Problem Set 8

Due 4pm, Wednesday, December 4th

Your name:

A worksheet for Problems 3 - 6 is provided at the end of this file. Extra copies and an excel version are available on the website.

You may discuss the problems on this assignment with other students in the class. You should run your own BLAST searches and base your solutions on the values you obtain from those searches. Note that the values returned by BLAST may change slightly from one day to the next as more sequences are added to the database.

List the names of the people you worked with:

Homework must be submitted by 4pm in MI650 or electronically to Canvas or mstolzer@andrew.cmu.edu.
1. *(6 points.) BLAST statistics:* For ungapped alignments, the expected number of high scoring pairs (HSP’s) with score at least $S$ found in the alignment of two random sequences is

$$E = K m' n' e^{-\lambda S},$$

where $m'$ and $n'$ are the effective lengths of the sequences and $K$ and $\lambda$ are constants that can be derived from the theory and depend on the substitution matrix. We can define a “normalized” bit score

$$S_b = \frac{\lambda S - \ln K}{\ln 2}.$$ 

Show that the number of HSP’s with score at least $S_b$ is

$$E = m' n' 2^{-S_b}$$

*Rearranging the expression for $S_b$ to solve for $\lambda S$, we get*

$$\lambda S = S_b \ln 2 + \ln K.$$ 

*Substituting $\lambda S$ in $E$, we get*

$$E = K m' n' e^{-(S_b \ln 2 + \ln K)}$$

$$= K m' n' e^{S_b \ln 2} e^{-\ln K}$$

$$= \frac{K m' n' e^{\ln 2 - S_b}}{e^{\ln K}}$$

$$= \frac{K m' n' 2^{-S_b}}{K}$$

$$= m' n' 2^{-S_b}.$$
2. (26 points.) Relative entropy quantifies the difference between two probability distributions. Relative entropy can be used in a number of ways to solve problems in information theory and machine learning. In this course, we use the interpretation that the relative entropy is a measure of the expected amount of discriminatory information obtained from a single observation of a random variable to help us decide whether the underlying probability distribution of the random variable represents the alternative hypothesis $H_A$ or the null hypothesis $H_0$.

In class, we discussed the relative entropy of a series of coin tosses. The sample space for this random variable has two possible events, $H$ and $T$. The relative entropy is the amount of discriminatory information we can derive from each toss to distinguish between a fair coin ($H_0 : p_H = 0.5$) and a biased coin ($H_A : p_H = \pi, \pi \neq 0.5$).

In the Kimura 2 Parameter (K2P) framework, we can think of the sites in a pairwise alignment of nucleotide sequences as observations of a random variable with a sample of space consisting of three possible events: matches ($M$), transitions ($S$), and transversions ($V$). In this problem, you will calculate the relative entropy of the frequency of matches, transitions, and transversions in sequences related according to the K2P model, relative to the null distribution of matches, transitions, and transversions in alignments of random sequences.

(a) (2 pts) Let $P_S = P(S|H_A), P_V = P(V|H_A)$, and $P_M = P(M|H_A)$ be the probabilities of observing transitions, transversions and matches, respectively in alignments of sequences evolving according to K2P with parameters $\alpha, \beta$, and $t$. Give expressions, in terms of $\alpha, \beta$, and $t$, for $P_S$, $P_V$, and $P_M$.

\[
P_S = P(S|H_A) = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t})
\]

\[
P_V = P(V|H_A) = \frac{1}{2}(1 - e^{-8\beta t})
\]

\[
P_M = P(M|H_A) = 1 - (P_S + P_V)
\]

\[
= \frac{1}{4}(1 + 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}).
\]
(b) (2 pts) What are the null probabilities of observing transitions, transversions and matches in alignments of random sequences in which each of the four nucleotides appears with probability 0.25?

There are 16 possible combinations of two nucleotides. Four of those combinations are matches (AA, GG, CC, TT) and four are transitions (AG, GA, CT, TC). The remaining eight combinations are transversions. Thus,

\[ P(S|H_0) = q_S = \frac{4}{16} = \frac{1}{4} \]

\[ P(V|H_0) = q_V = \frac{8}{16} = \frac{1}{2} \]

\[ P(M|H_0) = q_M = \frac{4}{16} = \frac{1}{4} \]

(c) (6 pts) Give an expression, in terms of \( P_S, P_V, \) and \( P_M, \) for the relative entropy of the distribution of transitions, transversions, and matches in related sequences evolving according to the K2P model.

Recall that

\[ H = \sum_i p_i \log_2 \frac{p_i}{q_i} \]

Therefore

\[ H(\alpha, \beta, t) = P_M \times \log_2 \frac{P_M}{\frac{1}{4}} + P_S \times \log_2 \frac{P_S}{\frac{1}{4}} + P_V \times \log_2 \frac{P_V}{\frac{2}{4}} \]

\[ H(\alpha, \beta, t) = P_M \times \log_2 4P_M + P_S \times \log_2 4P_S + P_V \times \log_2 2P_V \]
(d) (2 pts) Suppose that $\alpha = 0.002$ transitions per million years, $\beta = \frac{\alpha}{2}$ transversions per million years, and $t = 10$ million years. According to your expression for the relative entropy, how many bits per position are available to discriminate related sequences from chance similarity?

Given $\alpha = 0.002$, $\beta = 0.001$, and $t = 10$

\[
\begin{align*}
P_S &= P(S|H_A) = \frac{1}{4}(1 - 2e^{-4(\alpha + \beta)t} + e^{-8\beta t}) \\
&= \frac{1}{4}(1 - 2e^{-0.12} + e^{-0.08}) \\
&= \frac{1}{4}(1 - 2 \times 0.89 + 0.92) = \frac{1}{4}(0.15) = 0.04
\end{align*}
\]

\[
\begin{align*}
P_V &= P(V|H_A) = \frac{1}{2}(1 - e^{-8\beta t}) \\
&= \frac{1}{2}(1 - e^{-0.08}) \\
&= \frac{1}{2}(1 - 0.92) = \frac{1}{2}(0.08) = 0.04
\end{align*}
\]

\[
\begin{align*}
P_M &= P(M|H_A) = \frac{1}{4}(1 + 2e^{-4(\alpha + \beta)t} + e^{-8\beta t}) \\
&= \frac{1}{4}(1 + 2e^{-0.12} + e^{-0.08}) \\
&= \frac{1}{4}(1 + 2 \times 0.89 + 0.92) = \frac{1}{4}(3.70) = 0.92
\end{align*}
\]

\[
\mathcal{H}(\alpha, \beta, t) = P_M \times \log_2 4P_M + P_S \times \log_2 4P_S + P_V \times \log_2 2P_V
\]

\[
\mathcal{H}(0.002, 0.001, 10) = 0.92 \times \log_2(4 \times 0.92) + 0.04 \times \log_2(4 \times 0.04) + 0.04 \times \log_2(2 \times 0.04) \\
= 0.92 \times 3.70 + 0.04 \times 1.5 + 0.04 \times 0.08 \\
= 0.92 \times 1.9 - 0.04 \times 3.7 - 0.04 \times 2.7 \\
= 1.74 - 0.10 - 0.14 \\
= 1.5
\]
(e) (2 pts) Suppose that $\alpha = 0.002$ transitions per million years, $\beta = \frac{\alpha}{2}$ transversions per million years, and $t = 100$ million years. According to your expression for the relative entropy, how many bits per position are available to discriminate related sequences from chance similarity?

Given $\alpha = 0.002$, $\beta = 0.001$, and $t = 100$

$$P_S = P(S|H_A) = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t})$$

$$= \frac{1}{4}(1 - 2e^{-1.2} + e^{-0.8})$$

$$= \frac{1}{4}(1 - 2 \times 0.30 + 0.45) = \frac{1}{4}(0.85) = 0.21$$

$$P_V = P(V|H_A) = \frac{1}{2}(1 - e^{-8\beta t})$$

$$= \frac{1}{2}(1 - e^{-0.8})$$

$$= \frac{1}{2}(1 - 0.45) = \frac{1}{2}(0.55) = 0.28$$

$$P_M = P(M|H_A) = \frac{1}{4}(1 + 2e^{-4(\alpha+\beta)t} + e^{-8\beta t})$$

$$= \frac{1}{4}(1 + 2e^{-1.2} + e^{-0.8})$$

$$= \frac{1}{4}(1 + 2 \times 0.30 + 0.45) = \frac{1}{4}(2.05) = 0.51$$

$$\mathcal{H}(\alpha, \beta, t) = P_M \times \log_2 4P_M + P_S \times \log_2 4P_S + P_V \times \log_2 2P_V$$

$$\mathcal{H}(0.002, 0.001, 100) = 0.51 \times \log_2(4 \times 0.51) + 0.21 \times \log_2(4 \times 0.21) + 0.28 \times \log_2(2 \times 0.28) = 0.51 \times \log_2 2.05 + 0.21 \times \log_2 0.85 + 0.28 \times \log_2 0.55 = 0.51 \times 1.0 - 0.21 \times 0.2 - 0.28 \times 0.9 = 0.53 - 0.04 - 0.24 = 0.24$$
(f) (6 pts) Suppose you have query sequence of effective length $m' = 900$ and a database of size $n' = 10^{11}$. What is the minimum number of bits required to identify a related sequence with significance $E = 1$?

Recall that the bits need to achieve a match with E-value cut-off $E_T$ is

$$\frac{\log 2(m'n')}{E_T}$$

Therefore, when $E_T = 1$ and the given query and database sizes, the minimum number of bits required is

$$\log_2(900 \times 10^{11}) = 46.4$$

(g) (3 pts) What is the length of the shortest query sequence for which you could hope to find a related sequence with significance $E = 1$, when $\alpha = 0.002$ transitions per million years, $\beta = \frac{\alpha}{2}$ transversions per million years and $t = 10$ million years?

Recall that the shortest query sequence can be estimated as

$$\frac{\log 2(m'n')}{E_T} \mathcal{H}$$

In this case the shortest possible query is

$$\frac{46.4}{1.50} = 31 \text{ nucleotides}$$

(h) (3 pts) What is the length of the shortest query sequence for which you could hope to find a related sequence with significance $E = 1$, when $\alpha = 0.002$ transitions per million years, $\beta = \frac{\alpha}{2}$ transversions per million years and $t = 100$ million years?

In this case the shortest possible query is

$$\frac{46.4}{0.24} = 190 \text{ nucleotides}$$
For the following problems, you will carry out several BLAST searches with the same query sequence and different parameter settings to explore how these parameters influence the information that can be retrieved.

The query sequence for these searches is the 3-methylornithine–L-lysine ligase (PylC) sequence from archaeal species, *Methanosarcina barkeri* (accession ID: WP_048102615.1). *M. barkeri* uses an unusual genetic code that encodes 21 amino acids: the 20 canonical amino acids and the non-canonical amino acid, pyrrolysine. This expanded genetic code requires that the genome encode the biosynthesis machinery required to make the non-canonical amino acid and the associated tRNA. PylC is one of the enzymes in the pyrrolysine biosynthesis pathway.

To run your Blast searches, open a new tab in your browser, and go to [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov). Find “BLAST” in the list of “Popular Resources” on the right hand side of the page and follow that link. Select “Protein BLAST” on the right hand side of the page. You will be directed to a page entitled “Standard Protein BLAST”. This is the front end to the BLAST search engine for amino acid sequence databases. You will see a box labeled “Enter Query Sequence” in the upper left hand corner.

For each of the searches in Problems 3 and 5, you will run BLAST with a different set of parameter settings. In each case, when the BLAST search completes, a BLAST results page will be displayed. At the top of the page, below the blue NIH banner, you will see a gray ribbon with “BLAST >> blast suite >> results for RID-xxxx...”, where “xxxx...” is an alphanumeric string. “RID-xxxx...” is the request ID for this search. You can use this id to retrieve your search (for example, if you closed your browser) for up to 24 hours.

Below the gray ribbon, you will see a table with information about the search parameter settings (database used, molecule type, query length, etc), followed by the results of your search under a green banner. The results consist of the set of database sequences that are significantly similar to the query sequence, i.e., that have an E-value that is lower (i.e., more significant) than the E-value threshold for the search\(^1\). NCBI provides a fact sheet [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\_NewBLAST.pdf) with a detailed description of the output format. Briefly, there are different views of these results available via four different tabs:

**Descriptions** This view presents a one line description of each database sequence that is significantly similar to the query sequence. The individual fields in this table are described on p. 2 of the NCBI fact sheet.

**Graphic Summary**: A graphical representation of the search results, shown as colored lines. The color and length of each line indicates the strength and extent of the corresponding match.

**Alignments**: High-scoring local alignments (HSP’s) of the query sequence with each matching sequence.

**Taxonomy**: A summary of the species represented by the database sequences retrieved in the search.

For this assignment, we will focus on the information provided in the Descriptions tab.

\(^1\)On the Blast web site, this threshold is called the “Expect threshold”.
3. *(3 points.*) Find the blastp search engine as described on the previous page. You will run three searches in this question and additional searches in Questions 4 and 5. For all of these searches, you will follow the same basic procedure.

Copy the PylC sequence accession (WP_048102615.1) into the query box. Skip over “Choose Search Set” and “Program Selection” (you’ll be working with the default algorithm, blastp).

Next to the blue “BLAST” button, select “Show results in a new window”. This feature makes it easy to repeat your search with different parameter settings and compare the output.

Under the BLAST button, click the “+” next to “Algorithm parameters” to expand the parameter options. You will modify these in Searches 2 and 3.

A search is launched by clicking the blue BLAST button. This will launch a search against the nr database (the default). A new tab will open that shows the progress of the Blast search, which may take a while to complete.

When the search completes, **extract the following information from the Results page and enter it in Table 1 in the worksheet provided at the end of this file (extra copies and an excel version are available on the website):**

- At the top of the page, immediately below the gray banner with the text “results for RID-...”, click on “Search Summary”. A pop-up table will appear with parameter values used in this search. The “Hitlist size” entry in this table is the value of the “max target sequences” field from the blast query page. **Record the values of the following parameters:** “Hitlist size”, “Expect value”, and the “Number of letters” in the database. (Note that “Hitlist size” refers to the maximum number of sequences that BLAST will retrieve in a given search, not the number of sequences that were actually retrieved. The Hitlist size can be modified by changing the setting for “Max target sequences”. You will do this in Search 2.)

- Scroll down to the results and check that the “Descriptions” tab is selected. Immediately below “Sequence producing significant alignments”, the “select all” box should be checked by default. If it is not, check it. Immediately to the right, you is the text “xxx sequences selected”, where xxx is the total number of sequences that were actually retrieved in the search. **Record the number of sequences retrieved.**

- Scroll down to the bottom of the Descriptions. **Record the Bit score (max score) and the E-value of the least significant match.**
**SEARCH 1 (1 pt):** Find the “Choose Search Set” section (immediately below “Enter Query Sequence”). Use the “Database” pull down menu to select “Reference proteins (refseq_protein)”.

For this search, we will use the default settings for all other parameters.) Click the blue BLAST button to launch Search 1. When the search completes, enter the relevant information into the first row of Table 1 in the worksheet.

**SEARCH 2 (1 pt):** Go back to the Standard Protein BLAST search tab. The PylC accession should still be in the query box. Scroll down to the blue “BLAST” button and make sure that the “Show results in a new window” is still selected. Under the BLAST button, click the “+” next to “Algorithm parameters” to expand the parameter options.

Change “Max target sequences” to 500. Click the blue “BLAST” button to launch another search.

When this search completes, use the same steps as in Search 1 to record the appropriate information in row 2 of Table 1 in the worksheet.

**SEARCH 3 (1 pt):** Go back to the Standard Protein BLAST search tab. The PylC accession should still be in query box. Scroll down to the “Algorithm parameters” section. Make sure that “Show results in a new window” is still selected. Keep the “Max target sequences” set at 500 and change the “Expect threshold” parameter to 0.002. Click the blue “BLAST” button to launch another search.

When this search completes, use the same steps as in Search 1 to record the appropriate information in row 3 of Table 1 in the worksheet.

Keep the tab for Search 3 open. You will use the results again in Problem 6.
4. (16 points.) Compare the results from Searches 1, 2, and 3 to determine the impact of max target sequences and E-value threshold (“Expect”) on sequence retrieval.

(a) (8 points.) Compare Search 1 and Search 2:

i. (1 pt) What parameter values differed between these two searches?

*The maximum number of target (database) sequences to be retrieved was increased from 100 to 500.*

ii. (1 pt) Did the number of sequences retrieved increase or decrease in Search 2, compared with Search 1?

*Increase. In Search 2, 500 sequences were retrieved, instead of just 100.*

iii. (1 pt) Did the least significant match found in Search 2 have a higher or lower bit score than the least significant match found in Search 1?

*The least significant match in Search 2 had a bit score of 47.8, which is lower than the least significant match in Search 1 (187).*

iv. (1 pt) Did the least significant match found in Search 2 have a more significant or less significant E-value than the least significant match found Search 1?

*The least significant match in Search 2 had an E-value of 0.035, which is higher (less significant) than the least significant match in Search 1 (6e−52).*
v. (2 pts) Explain any changes you observe in terms of what you know about the behavior of the Blast program. Consider both the number of sequences retrieved and the score and significance of the least significant matching sequence.

Search 1 returned 99 (100) sequences, which is the number of sequences allowed by the “max target sequences” parameter. The least significant match has an E-value which is much lower (much more significant) than the “Expect” threshold, which is 10. This suggests that “max target sequences” is the factor that is limiting the number of sequences retrieved and that the database contains more than 100 sequences that are significantly similar (E < 10) to the query. In Search 2, when the value of “max target sequences” was increased to 500, additional significant sequences were retrieved and those sequences had E-values that were still significant (E < 10), but less significant than the least significant hit in Search 1.

vi. (2 pts) Do you think that Search 2 retrieved all sequences in refseq that are significantly similar to the query? If so, why? If you think it is possible that Search 2 did not retrieve all significant matches, what BLAST parameters would you change to find out if there are more?

Search 2 retrieved 500 sequences and the least significant sequence has an E-value of 0.035, which is lower (more significant) than the “Expect” threshold of 10. This suggests that “max target sequences” is also the limiting factor in Search 2.

To determine whether there are more database sequences that are significantly similar to the query, we can increase “max target sequences” to 1000.
(b) (8 points.) Compare Search 2 and Search 3:

i. (1 pt) What parameter values differed between these two searches?

In Search 3, the E-value (“Expect”) threshold was decreased from 10 to 0.002. Recall that lower E-values are more significant, so this is a more stringent threshold.

ii. (1 pt) Did the number of sequences retrieved increase or decrease in Search 3, compared with Search 2?

The number of sequences decreased from 500(ish) to 208(ish).

iii. (1 pt) Did the least significant match found in Search 3 have a higher or lower bit score than the least significant match found in Search 2?

The least significant match in Search 3 had a bit score of 52.0, which is higher than the least significant match in Search 2 (47.8).

iv. (1 pt) Did the least significant match found in Search 3 have a more significant or less significant E-value than the least significant match found Search 2?

The least significant match in Search 3 had an E-value of 0.002, which is lower (more significant) than the least significant match in Search 2 (0.035).
v. (2 pts) Explain any changes you observe in terms of what you know about the behavior of the Blast program. Consider both the number of sequences retrieved and the score and significance of the least significant matching sequence.

*In Search 3, the number of sequences retrieved (208) was much lower than the maximum number allowed (500). Moreover, the least significant match had an E-value of 0.002, which is the same as the E-value threshold. This suggests that in Search 3, the E-value threshold is the parameter that controls the number of sequences retrieved.*

vi. (2 pts) Do you think that Search 3 retrieved all sequences in refseq that are significantly similar to the query at the current E-value threshold? If so, why? If not, what Blast parameters would you change to find out if there are more?

*Search 3 retrieved 208 sequences, which is much lower than the “max target sequences” of 500. This suggests that for an “Expect” threshold of 0.002, BLAST returned all database sequences with significant E-values when compared to the query.*
5. (11 points.) SEARCH 4: Go back to the Standard Protein BLAST search tab. The PylC accession should still be in query box. Find the “Choose Search Set” section (immediately below “Enter Query Sequence”). **Use the “Database” pull down menu to select “UniProtKB/Swiss-Prot (swissprot)”.** Swissprot is a curated protein sequence database that is much smaller than refseq.

Leave “Max target sequences” set to 500 and “Expect threshold” at 0.002. The “Show results in a new window” should still be selected. Click the blue “BLAST” button to launch another search.

(a) (1 pt) When the search completes, click on “Search Summary” at the top of the results page. According to the “Search Summary” pop-up table, how many amino acids are there in the swissprot database?

*The database was changed from refseq, which contains roughly 155 million sequences and 59 billion amino acids, to Swissprot, which contains roughly 473 thousand sequences and 179 million letters.*

(b) (1 pt) Scroll down to the descriptions. Check that “select all” has been checked. Click on the “Download” menu to the right of “Sequences producing significant alignments” and select “Hit table (text)”. Save this file making sure that the file name contains YOUR NAME, and turn it in with your homework.

(c) (1 pt) How many sequences were retrieved in this search?

*The number of sequences (4) retrieved in Search 4 decreased dramatically, compared to Search 3.*

(d) (1 pt) What were the bit score and E-value of the least significant match in this search?

*The bit score is 45.1, and the E-value is 0.001.*

(e) (1 pt) Compare the results of this search using the swissprot database with the results of Search 3, in which you searched the refseq database with the same parameter settings. Did the least significant match found in Search 4 have a higher or lower bit score than the least significant match found in Search 3?

*The bit score of the least significant match in Search 4 (45.1) is lower than the bit score of the least significant match in Search 3 (52.0).*

(f) (1 pt) Did the least significant match found in Search 4 have a more significant or less significant E-value than the least significant match found Search 3?

*The E-value of the least significant match in Search 4 (0.001) is more significant (lower) than the E-value of the least significant match in Search 3 (0.002).*
(g) (2 pt) Explain any changes you observe in terms of what you know about the behavior of the Blast program. Consider both the number of sequences retrieved and the score and significance of the least significant matching sequence.

The least significant match in Search 4 has a lower bit score, but also a more significant E-value. This is because the size of the database is much smaller in Search 4. The E-value is related to the bit score by this equation:

\[ E = mn2^{-S_b} \]

In Search 4, \( m \) is unchanged compared to Search 3, but \( n \) has increased by a factor of \(~325\). Since \( n \) is much smaller in Search 4, fewer bits (i.e., a lower bit score) are required to obtain the same E-value.

(h) (2 pt) Suppose your goal is to find sequences that match a given query sequence with particular E-value. Given the current sizes of the refseq and swissprot databases, how many more bits are required to find such sequences in refseq, compared with swissprot? Answer this question using what you have learned about the information content of database searches. Show your work.

The refseq database contains roughly 325 times as many letters as the Swissprot database. The E-value is related to the bit score by

\[ E = mn2^{-S_b} \]

Since the same E-value is desired in both searches, we obtain

\[ mn_{\text{refseq}}2^{-S_{\text{refseq}}} = mn_{\text{sp}}2^{-S_{\text{sp}}} \]

where \( s_x \) is the minimum score required to find a match in a database \( x \) of size \( n_x \). Since the query used in both searches is the same and \( n_{\text{refseq}} \approx 325 n_{\text{sp}} \), this yields

\[
\begin{align*}
2^{-S_{\text{refseq}}} &= \frac{n_{\text{sp}}}{n_{\text{refseq}}} \\
2^{-S_{\text{sp}}} &= \frac{325 n_{\text{sp}}}{n_{\text{sp}}} \\
2^{S_{\text{sp}} - S_{\text{refseq}}} &= \frac{1}{325} \\
S_{\text{sp}} - S_{\text{refseq}} &= \log_2 \frac{1}{325} \\
S_{\text{refseq}} - S_{\text{sp}} &= -\log_2 \frac{1}{325} = 8.35 \text{ bits.}
\end{align*}
\]
(i) (1 pt) Compare the bit scores of the least significant matches obtained in Searches 3 and 4. How many more bits were required to find the least significant match in refseq (Search 3), compared with the least significant match in swissprot (Search 4)? How does this value compare with the number of bits you estimated in the previous question?

The difference in the number of bits for the least significant matches in Search 3 and Search 4 is 6.9 bits.

This is very similar to the number of bits in the previous question. The difference has to do with the fact that the least significant match in Search 4 is slightly more significant than the match in Search 3.
6. (38 points.) This question explores how the choice of matrix influences the results of a Blast search. Suppose our goal is to find all species that include pyrrolysine in their genetic code. Since PylC is an essential amino acid in the pyrrolysine synthesis pathway, possession of a PylC gene is a marker for species that use pyrrolysine.

Here, we’ll investigate the accuracy of a BLAST search, using the *M. barkeri* PylC sequence as a query, for finding species that encode PylC. Searches with this query will find orthologs of PylC. Orthologs of PylC will only be found in the genomes of species that use the expanded genetic code with pyrrolysine. These sequences are typically described as “PylC”, “3-methylornithine–L-lysine ligase” or “pyrrolysine synthetase”. For our question, these are the true positives.

Searches with this query will also find more distantly related sequences that harbor a similar enzymatic domain, but use that domain to catalyze a different reaction. These include

- ATP-grasp domain-containing proteins,
- D-alanine–D-alanine ligases,
- pyridoxal-phosphate dependent enzymes,
- carbamoyl-phosphate synthases.

These sequences are not evidence of pyrrolysine synthesis and so, for our purposes, these are false positives. Ideally, we will be able to find an E value threshold that separates these two sets.

We will examine how the choice of substitution matrix influences the ability of BLAST to retrieve PylC orthologs versus sequences from these more distantly related families. The parameter settings for the BLAST searches for this problem are:

**Database:** refseq
**Max target sequences:** 500
**Expect:** 0.002

For this problem, run four BLAST searches the following substitution matrices. For all parameters not specified, use the default settings.

**Search 5** PAM30
**Search 6** BLOSUM62
**Search 7** BLOSUM45
**Search 8** PAM250

Note that Search 6 is the same as Search 3 from Problems 4 and 5. If you still have that tab open, you can use the results without rerunning the search.
To track how bit scores and E-values change in response to changes in parameter settings, we will focus on four specific “sentinel” sequences:

**WP_048193606.1** This sequence is a PylC sequence in *Methanococccoides methylutens*. This is an orthologous enzyme in a closely related species; both *M. methylutens* and *M. barkeri* are archaeal species.

**WP_084573891.1** This sequence is a PylC sequence in the bacterial species, *Sporomusa malonica*. This is an orthologous enzyme in a distantly related species.

**WP_086636974.1** This sequence encodes an enzyme in the archael species, *Methanotheronarchaeum thermophilum*. This enzyme is not PylC, but like PylC harbors an ATP-grasp domain. This is a distantly related enzyme in a closely related species.

**WP_018970047.1** This sequence encodes a D-alanine-D-alanine-ligase sequence from a domain-containing protein in the bacterium, *Rubritalea marina*. This is a distantly related enzyme in a distantly related species.

For each search, enter the bit scores and E-values of the four sentinel sequences in Table 2 in the worksheet. If the sequence was not retrieved, note “Not found”.

For each of these searches, tabulate the following information in Table 3 of the worksheet:

- The total number of sequences retrieved
- The number of retrieved sequences annotated with “PylC”. (Make sure not to include the query sequence in this count.)
- The number of retrieved sequences annotated with “PylC”.
- The number of sequences with “ATP-Grasp” in the description
- The number of sequences with “carbamoyl” in the description
- The number of sequences with “D-alanine” in the description. (Be careful. With this search term it’s easy to count the same sequence twice.)
- The number of sequences with “Pyridoxal” in the description
- The number of sequences annotated as “hypothetical protein”.

Rank Searches 5 - 8 based on the level of divergence of the matrix used in each search. What is the relative entropy of each of the matrices used in these searches?

Search 1 is most conserved; searches 7 and 8 are the most diverged. Divergence increases with search number:

<table>
<thead>
<tr>
<th>Sequence Identity</th>
<th>Relative Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM30</td>
<td>75</td>
</tr>
<tr>
<td>BLOSUM62</td>
<td>30</td>
</tr>
<tr>
<td>BLOSUM45</td>
<td>20</td>
</tr>
<tr>
<td>PAM250</td>
<td>20</td>
</tr>
</tbody>
</table>
(d) (8 pts) Sentinel sequences. Use the values you tabulated in Table 2 to answer the following questions.

i. (3 pts) Were all four sentinel sequences found in all four searches? If not, which sequences were not retrieved in which searches? Explain why these sequences were not found in terms of what you know about the statistics of MSP scores and the Blast report threshold.

WP_018970047.1 in R. marina was not found in Search 8 (PAM 250). This suggests that the target frequencies for PAM250 are so poorly matched with the target frequencies of this family that the best HSP obtained with this sequence did not have a significant score. That is, the score of the best HSP was too low to obtain and E-value better than 0.002.

ii. (3 pts) Given the M. barkeri PyC sequence as a query, which matrices do you think is (are) best suited for finding

- other PyC sequences,
  
  BLOSUM62 gave the most significant E-values for the two orthologous PyC sequences WP_048193606.1 and WP_084573891.1.
  
  Looking at Table 3, PAM30, BLOSUM62, and BLOSUM45 returned a similar number (and the most) PyC sequences.
  
  Therefore, BLOSUM62 is probably the best.

- ATP-grasp domain containing proteins,
  
  All four matrices gave similar E-values for the sentinel sequence that contains ATP-grasp (WP_086636974.1).
  
  Looking at the numbers in Table 3, BLOSUM45 returned the most sequences with ATP-grasp domains.
  
  Therefore, BLOSUM45 is probably the best.

- D-alanine-D-alanine-ligase sequences?
  
  PAM250 could not find the sentinel D-alanine-D-alanine-ligase sequence (WP_018970047.1). The other three matrices gave similar E-values for that sentinel sequence.
  
  Looking at the numbers in Table 3, BLOSUM62 returned the most D-alanine-D-alanine-ligase sequences.
  
  Therefore, BLOSUM62 is probably the best.

Given the M. barkeri PyC sequence as a query, which matrices do you think is (are) best suited for not finding

- other PyC sequences,
  
  PAM250. The sentinel sequences were least significant and it returned the fewest PyC sequences.
• ATP-grasp domain containing proteins,

  PAM250. The sentinel sequence was least significant; PAM250 and PAM30 returned the fewest sequences with ATP-grasp domains.
• D-alanine-D-alanine-ligase sequences?

  PAM250. The sentinel sequence was not found; in fact, no D-alanine-D-alanine-ligase sequences were returned.

iii. (2 pts) Explain your reasoning.

  SEE ABOVE.
(e) **(19 pts)** Precision and Recall. Use the values you tabulated in Table 3 to answer the following questions.

i. **(2 pts)** For each search, calculate the total number of *false* positives retrieved. This is the sum of the number of ATP-grasp domain-containing proteins, D-alanine–D-alanine ligases, pyridoxal-phosphate dependent enzymes, and carbamoyl-phosphate synthases retrieved in the search. Hypothetical proteins cannot be classified as true or false positives.

<table>
<thead>
<tr>
<th></th>
<th>Search 5:</th>
<th>Search 6:</th>
<th>Search 7:</th>
<th>Search 8:</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives</td>
<td>PAM 30</td>
<td>BLOSUM62</td>
<td>BLOSUM45</td>
<td>PAM250</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>72</td>
<td>40</td>
<td>3</td>
</tr>
</tbody>
</table>

ii. **(2 pts)** For each search, calculate the total number of *true* positives retrieved. This is the number of sequences annotated “PylC”.

<table>
<thead>
<tr>
<th></th>
<th>Search 5:</th>
<th>Search 6:</th>
<th>Search 7:</th>
<th>Search 8:</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives</td>
<td>PAM 30</td>
<td>BLOSUM62</td>
<td>BLOSUM45</td>
<td>PAM250</td>
</tr>
<tr>
<td></td>
<td>131</td>
<td>132</td>
<td>132</td>
<td>126</td>
</tr>
</tbody>
</table>

iii. **(1 pt)** As a sanity check, verify that the numbers of true positives, false positives and hypothetical proteins sum to the total number of sequences retrieved.

iv. **(2 pts)** For each search, calculate the precision; that is, the number of true positives normalized by the total number of sequences retrieved.

<table>
<thead>
<tr>
<th></th>
<th>Search 5:</th>
<th>Search 6:</th>
<th>Search 7:</th>
<th>Search 8:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>PAM 30</td>
<td>BLOSUM62</td>
<td>BLOSUM45</td>
<td>PAM250</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.64</td>
<td>0.76</td>
<td>0.96</td>
</tr>
</tbody>
</table>
v. *(4 pts)* Estimate the total number of true positives in refseq, by taking the maximum, over all four searches, of the number of true positives found.

*By this estimate, there are 132 PylC sequences in reseq.*

Note, by estimating the number of true positive in this way, we are implicitly making the assumption that at least one search found all of the true positives in the refseq database. This is not necessarily a good assumption.

Do you think there is evidence to support the assumption in this case? Why or why not? (Hint: consider the ranking of true and false positives when the output from each of the four searches is sorted based on E value.)

*In the BLOSUM 62 search, the least significant PylC sequence has an E-value of 3e-29. The search looks at sequences with E-values as weak as 0.002. The search goes quite deep into the twilight zone making it unlikely that there is another PylC sequence with an E-value that is too weak to be reported.*

vi. *(2 pts)* For each search, calculate the recall; that is, what fraction of PylC genes in refseq, based on your estimate in the previous step, were retrieved by this search?

<table>
<thead>
<tr>
<th></th>
<th>Search 5:</th>
<th>Search 6:</th>
<th>Search 7:</th>
<th>Search 8:</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 30</td>
<td>Recall</td>
<td>BLOSUM62</td>
<td>BLOSUM45</td>
<td>PAM250</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.95</td>
</tr>
</tbody>
</table>

vii. *(7 pts)*

<table>
<thead>
<tr>
<th></th>
<th>Search 5:</th>
<th>Search 6:</th>
<th>Search 7:</th>
<th>Search 8:</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 30</td>
<td>True Positives</td>
<td>131</td>
<td>132</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>False Positives</td>
<td>9</td>
<td>72</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.92</td>
<td>0.64</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Which search(es) gave

- *(1 pt)* the best precision?

*PAM250 had the best precision.*

- *(1 pt)* the best recall?

*Both BLOSUM62 and BLOSUM45 have perfect Recall.*

- *(1 pts)* did the best job overall at finding PylC sequences and excluding more distantly related enzyme families?

*PAM250 and PAM30 have excellent Recall and excellent Precision — PAM30 has almost perfect Recall while PAM250 has better Precision. Considering the F1 score of \( \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \), both PAM250 and PAM30 perform similarly with a score of 0.96.*
(2 pts) Given your results in Part (d), does your answer surprise you?

The performance on the sentinel sequences suggests that the two BLOSUM matrices perform best with the PylC family, i.e., have target frequencies that match that family. In addition, PAM30 and PAM250 represent two very different levels of divergence. Therefore, yes, it is surprising that PAM30 and PAM250 have the best overall performance.

(2 pts) For the matrix that did the best overall performance, what are the properties of that matrix that resulted in good performance for this particular search?

PAM250 likely performs best overall due to its superior ability to exclude false positives. The bit scores obtained when aligning distantly related enzymes with PylC sequences using PAM250 are much lower than the bit scores obtained with any other matrix.

PAM30 likely did well because, although it returned more false positives, it wasn't many more false positives, and it was able to return more true positives.
<table>
<thead>
<tr>
<th>Search</th>
<th>Database size: Number of letters</th>
<th>Database used</th>
<th>E-value threshold</th>
<th>Max number target sequences</th>
<th>Number sequences</th>
<th>Least significant hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58,298,572,787</td>
<td>Ref Seq</td>
<td>10</td>
<td>100</td>
<td>99</td>
<td>Score 187.0 E-value 6E-52 Accession WP_073033370.1</td>
</tr>
<tr>
<td>2</td>
<td>58,298,572,787</td>
<td>Ref Seq</td>
<td>10</td>
<td>500</td>
<td>498</td>
<td>Score 47.8 E-value 0.035 Accession WP_010851342.1</td>
</tr>
<tr>
<td>3</td>
<td>57,373,772,419</td>
<td>Ref Seq</td>
<td>0.002</td>
<td>500</td>
<td>208</td>
<td>Score 52.0 E-value 0.002 Accession WP_143981508.1</td>
</tr>
<tr>
<td>4</td>
<td>178,963,502</td>
<td>Uniprot</td>
<td>0.002</td>
<td>500</td>
<td>4</td>
<td>Score 45.1 E-value 0.001 Accession P59448.1</td>
</tr>
</tbody>
</table>

Table 1

Least significant hit column:
- Score
- E-value
- Accession
<table>
<thead>
<tr>
<th>Sentinel sequence</th>
<th>Search 5: Pam 30</th>
<th>Search 6: Blosum62</th>
<th>Search7: Blosum45</th>
<th>Search 8: PAM250</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bit Score</td>
<td>E value</td>
<td>Bit Score</td>
<td>E value</td>
</tr>
<tr>
<td>WP_048193606.1</td>
<td>489</td>
<td>4E-143</td>
<td>488</td>
<td>1E-170</td>
</tr>
<tr>
<td>M. methylutens</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP_084573891.1</td>
<td>182</td>
<td>9E-43</td>
<td>170</td>
<td>2E-45</td>
</tr>
<tr>
<td>S. malonica</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP_086636974.1</td>
<td>153</td>
<td>4E-34</td>
<td>134</td>
<td>3E-32</td>
</tr>
<tr>
<td>M. thermophilum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP GRASP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP_018970047.1</td>
<td>55.8</td>
<td>2E-04</td>
<td>60.5</td>
<td>1E-06</td>
</tr>
<tr>
<td>R. marina D-ala-D-ala-ligase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1
<table>
<thead>
<tr>
<th></th>
<th>Search 5: Pam 30</th>
<th>Search 6: Blosum62</th>
<th>Search 7: Blosum45</th>
<th>Search 8: PAM250</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total sequences retrieved</strong></td>
<td>142</td>
<td>206</td>
<td>174</td>
<td>131</td>
</tr>
<tr>
<td><strong>PyLC sequences</strong></td>
<td>131</td>
<td>132</td>
<td>132</td>
<td>126</td>
</tr>
<tr>
<td><strong>ATP-grasp sequences</strong></td>
<td>2</td>
<td>18</td>
<td>28</td>
<td>3</td>
</tr>
<tr>
<td><strong>Carbamoyl</strong></td>
<td>0</td>
<td>38</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><strong>D-alanine</strong></td>
<td>7</td>
<td>16</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td><strong>pyridoxal</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Hypothetical sequences</strong></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>