**Lab 6: BLASTp**

**Due:** Wednesday, Oct 28, 2015 [beginning of class]

**Purpose:** This lab gives you practice using the BLASTp search tool, programming an experiment to better understand the probability distribution of string alignment scores, and programming the first part of the BLASTp search that finds high-scoring matches.

**Directions:** You should complete this lab in pairs, and you should take turns as the “driver” and “navigator” in pair programming. Choose a partner so that one person has more experience in biology and one person has more experience in computer science. You should swap roles at least as often as every ten minutes. All submitted work should be typed and include your name(s). Submit your code files, your output files, and write-up (as a single .zip file) *electronically* to Moodle.

**Background Information:** Review your notes and textbook about the BLAST search heuristic and read the lab preamble.

**Lab Exercises:**

1. Recall the two yeast genes that you and your partner chose for lab 2. Choose one of them. If you forgot, you can choose a gene from the list (see Lab 2 Gene List file). What are the two genes?

Gene name in yeast: \_\_\_\_\_\_\_\_**PFK1**\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Go to [yeastgenome.org](http://www.yeastgenome.org/). After searching for your gene, go to the protein tab. Download or copy the amino acid sequence for this gene. Now, go to:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Choose protein blast for the BLAST type. Enter the amino acid sequence or browse to upload the fasta file you downloaded. Use non-redundant protein sequences (nr) for the database. Use blastp for the algorithm.

Answer these questions for the write-up:

1. What are the top 10 organisms (other than yeast) that have the best alignments (actual or hypothetical) to your gene? For each, list the bits score and the E values for these ten alignments.

Table 1: Top 10 organisms that score high with blastp with yeast gene: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
| **Organism** | **Bits score** | **E value** |
| Saccharomyces arboricola H-6 | 2004 | 0 |
| Saccharomyces eubayanus | 2003 | 0 |
| Candida glabrata CBS 138 | 1712 | 0 |
| Naumovozyma dairenensis CBS 421 | 1695 | 0 |
| Kazachstania Africana CBS 2517 | 1680 | 0 |
| Naumovozyma castellii CBS 4309 | 1677 | 0 |
| Torulaspora delbrueckii | 1645 | 0 |
| Kazachstania naganishii CBS 8797 | 1643 | 0 |
| Tetrapisispora blattae CBS 6284 | 1615 | 0 |
| Zygossaccharomyces bailii ISA 1307 | 1606 | 0 |

1. Looking at your **entire** set of search results, what are the ranges (min and max) of the following values listed in Table 2.

Table 2: Ranges of values from the result of a BLASTp search with gene: \_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
| **Value** | **Min** | **Max** |
| Alignment score (header is titled max score in BLAST) | 905 | 2039 |
| Bits score | 905 | 2039 |
| E value | 0.0 | 0.0 |

2. Do a protein BLAST query with that same gene and modify the search parameters with the conditions shown below. Compare the top 20 results to see if there are any new results found (compared to what you found in part 1) and list those. To edit the parameters, click on the plus sign next to “Algorithm Parameters”.

Table 3: BLASTp results when parameters are altered with gene: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| **Search parameters** | **Organisms in top 20 in this search not found in part 1** |
| Max target sequences: 100  Expect threshold: 10  Word size: 6  Max matches in query range: 0  **Matrix: PAM 30**  Gap costs: existence: 11 extension: 1  Compositional adjustments: Conditional compositional score matrix adjustment | No differences in top 20 |
| Max target sequences: 100  **Expect threshold: 11**  Word size: 6  Max matches in query range: 0  Matrix: BLOSUM 62  Gap costs: existence: 11 extension: 1  Compositional adjustments: Conditional compositional score matrix adjustment | No differences in top 20 |
| Max target sequences: 100  Expect threshold: 10  **Word size: 3**  Max matches in query range: 0  Matrix: BLOSUM 62  Gap costs: existence: 11 extension: 1  Compositional adjustments: Conditional compositional score matrix adjustment | No differences in top 20 |
| Max target sequences: 100  Expect threshold: 10  Word size: 6  Max matches in query range: 0  Matrix: BLOSUM 62  **Gap costs: existence: 6 extension: 2**  Compositional adjustments: Conditional compositional score matrix adjustment | Saccharomyces eubayanus (replaced Saccharomyces arboricola H-6) |

3. Download the starter code for randomAlignments.py. Complete the functions in the program.

printHistogram – takes a histogram (list of values) and an output file as parameters. Prints the list in the following format to the output file:

1. 0
2. 2
3. 3
4. 5

… etc

where the histogram list consists of [0, 2, 3, 5, etc]. Per row: the first number to print is the list index, then a tab \t, then the value at list[index], then a newline.

createRandomSequence – takes a number n (number of nucleotides) and returns a random DNA sequence of that length. Each base, A or C or G or T, should be generated at a rate of 25%. A previous lab exercise will be helpful for this task.

alignManyRandomSequences – creates a histogram list with length 151 of all 0’s (scores for aligning 2 separate 30-long DNA sequences could range from 0 to 150)

hist = [0] \* 150 # python code for creating a list of 150 0’s

# algorithm for this function

For numberOfAlignments (parameter passed to function) times:

Generate a random string s1 (using function above) with 30 nucleotides

Generate a random string s2 (using function above) with 30 nucleotides

Calculate the best local alignment score for s1 and s2 (copy your solution from lab 5)

Add one to hist[score]

print the histogram (using function above)

localAlignmentScore – returns the best local alignment of parameters s1 and s2. You may copy much of the code you wrote for lab 5.

4. Download the starter code for HSWords.py. Also, download the BLOSUM62.txt file. This program should simulate the first step in the BLASTp algorithm, where high-scoring pairs are found between the amino acid words in the query sequence and the database. BLASTp uses length 6 words by default. In this lab, you will use words of length 3. Read through the first function in the program (reads the scoring matrix from a file and creates a data structure to store the 2D scoring matrix and the 1D legend (‘A’, ‘R’, ‘N’, etc.) of the amino acid letter order).

Complete the functions:

matchScore – takes word1, word2, and the scoring matrix data structure as parameters. Returns the match score of the two words, given the scoring matrix. Note that if either or both words does not have length 3, return 0.

highScoringWords -- takes a word, a threshold value T, a scoring matrix data structure, and output file as parameters. It should write to the output file all high-scoring words with the given word. High-scoring words have scores >= T. This function should call matchScore on word and every possible word of size 3 (note there are 24 distinct letters in the amino acid legend, so there are 24x24x24 distinct 3-letter words). See below for the first few lines of a properly formatted output file (word, tab, score, newline after the first line).

High scoring words for AHK, Threshold: 11

AHA 11

AHR 14

AHN 12

AHD 11

If you have time, you could modify the code so that it works for any length word (not just words of length 3). It should not take much code modification.

**Write-up (60 points):**

Create a word or pdf file to complete your writeup of this lab to show that you completed the above exercises and tested your code. Be sure to include your name(s) and lab number in your file.

1. (4 pts) Insert Table 1 from above here.
2. (4 pts) Insert Table 2 from above here.
3. (4 pts) Insert Table 3 from above here.
4. (4 pts) Import the text file your program produces from 1000 randomly generated tests into excel. Create a bar chart to show the distribution of local alignment scores, with scores along the x-axis and frequency along y-axis. Include the chart in the write-up here. (If there are several values of 0 for large alignment scores, you can delete them from the graph.)
5. (4 pts) Suppose someone gave you a pair of DNA sequences, each with length 30, and you calculated the optimal local alignment score for the pair. Based on the histogram you produced and graphed in exercise 4, what is the minimum optimal alignment score S that this pair could have such that 5% of randomly generated sequences have a score >= S? [This corresponds to a p-value of .05.]
6. (4 pts) Based on your histogram, what is the minimum optimal alignment score S that this pair could have such that 1% of randomly generated sequences have a score >= S? [This corresponds to a p-value of .01.]
7. (4 pts) Assume you determine that the best local alignment score for the pair of DNA sequences given to you is 40. Using your histogram, what is the **percentage** of randomly generated sequences with a score >= 40? [This is the p-value for a score of 40.]
8. (4 pts) Now run the experiment with 10000 randomly generated sequences and determine the local alignment scores. Import the results file into excel and create a bar chart. Provide the chart in the write-up here. Is this similar to the chart you produced for Q4?
9. (3 pts) When performing BLAST searches, an important step is assessing the significance of the results. In experimental biology and other scientific studies, a p-value of 0.05 or lower is often the threshold for significance. However, for BLAST searches, researchers sometimes use more stringent criteria for accepting results as non-random. For example, genome databases often use E-value cutoffs in the 10-3 to 10-6 range for sequence similarity. Why would a researcher use such stringent criteria for BLAST searches?
10. (2 pts) Determine all high-scoring words for “BHK” with threshold T = 11. Include this file in your zip file. Include this file in your zip file named “BHK\_T11\_matches.txt”.
11. (2 pts) Determine all high-scoring words for “AAA” with threshold T = 11. Include this file in your zip file named “AAA\_T11\_matches.txt”.

1. (2 pts) Determine all high-scoring words for “WKM” with threshold T = 11. Include this file in your zip file named “WKM\_T11\_matches.txt”.
2. (2 pts) Now find all high-scoring words for “BHK” with threshold T = 13. Include this file in your zip file named “BHK\_T13\_matches.txt”.
3. (3 pts) Which 3-letter amino acid word should have the most high-scoring words for threshold value T = 11? (look at the BLOSUM 62 matrix to be smart about this; or you could run lots of code)
4. (3 pts) Suppose a BLASTp is done with word size = 3 and a threshold value T of 11. Let’s call that search A. Now, suppose you do the same BLASTp search with word size = 3 and now a threshold value of T = 14. Let’s call that search B. Which search would you expect to be faster? Why?
5. (3 pts) In the BLOSUM 62 matrix, the size is more than 20 x 20. Earlier in the course, we learned that there are 20 amino acids. What do B, Z, and X mean as single characters encoding amino acids? (A reference is posted to moodle to help you answer this question)
   1. B =
   2. Z =
   3. X =
6. Read the lab 6 preamble.

A. (4 pts) Suppose you did a BLAST search with 100 target sequences. Now, suppose you increase the “max target sequences” parameter value in BLAST for the same search. Will the **sensitivity** increase, decrease, or stay the same?

B. Will the **specificity** increase, decrease, or stay the same?

1. A. (4 pts) Suppose you did a BLAST search with an expect threshold of 10. Now, suppose you increase the “expect threshold” parameter for the same search. Will the **sensitivity** increase, decrease, or stay the same?

B. Will the **specificity** increase, decrease, or stay the same?

**Appendix A: Authorship (please include statement in your write-up)**

The code and write-up submitted for this lab were authored by the named person(s) on this lab report. All external sources to BIO/CS423 are cited properly.

**Appendix B: Code (25 points, based on correctness and style):**

Copy and paste the code (both files) you wrote for lab 6 here (use Courier 8-pt font). Also, upload the code files to Moodle as part of you zip folder.