Biological Data Analysis (CSE 182) : Assignment 2

Logistics

Submit a hard copy containing the code and results. Create a compressed file containing the code and output as separate files, and email Julio Ng. Do not deviate significantly from the suggested file names. *Late submission will incur a penalty of 10pts for every day that the assignment is late.*

Sequence Alignment and Gap penalties

1. (40 pts.) Implement a space-efficient program *locAL* to align two long DNA sequences with linear gap-penalty. The program should take two DNA sequences as input, along with user defined values for *match-score*, *mismatch-score*, *indel*. Its should be invoked as follows:
   `locAL <seq_file1> <seq_file2> -m <match> -s <mismatch> -d <indel>`
   and should output the following:

   - Score of the best local-alignment.
   - Length of the best local-alignment
   - The alignment itself.

   Apply the program to aligning the two pairs of sequences (available on the course web-site) with the following parameters: match:1, mismatch -30, indel -20. Submit the code, and the output.

2. (30 pts.) Write a program to generate random DNA sequence ($pr[A] = Pr[C] = Pr[G] = Pr[T] = 0.25$) with a specified length. Generate 500 pairs of length 1000bp each, align them using *locAL* using two sets of parameters:

   - **parameter P1:** match 1, mismatch 0, indel 0
   - **parameters P2:** match 1, mismatch -30, indel -20

   Plot the lengths of the local alignment using paramaters P1, and P2. Are the lengths of the optimal local alignments different? If so, why? Try the same experiment with random pairs of different length.

   Define $l_p(n)$ as the expected length of the optimal local alignment for a pair of random sequences of length n. Your computations should give you estimates of $l_{P1}(n)$, and $l_{P2}(n)$ for different values of n. Can you guess the form of $l_{P1}(n)$, and $l_{P2}(n)$, as a function of n?

3. (20 pts.) **Phase Transition of Local Alignments:** Define $l_P(n)$ as in Problem 2. Clearly, the parameter set P can change $l_P(n)$.

   (a) Plot the values of $l_P(n)$ for a variety of parameter settings which go from mismatch= -30, to mismatch = 0. For example, you can choose mismatch=indel from $\{-30, -20, -10, -1, -0.5, -0.33, -0.25, 0\}$.

   (b) Is there an abrupt change in the value of $l_P(n)$? If so, can you give the parameters at which the change happens?

4. (Extra credit:10pts.) Can you give a theoretical justfication of your answers in Problems 2, and 3?

5. (8pts.) Go to the NCBI web-site, and BLAST the two sequences (available on the course web-page), after switching OFF all filters. What is the number of hits for each sequence? Why is the number different for the two sequences?

6. (2pts.) What language did you use? How much time did you take to do the assignment? Who did you discuss your homework with?