Biological Data Analysis (CSE 182) : Assignment 3

Logistics

Submit a zip archive of your code, written answers, output and graphs by email.

Sequence Alignment and Gap penalties

We have had quite a bit of discussion about the impact of different scoring functions on the alignment. In this assignment (Q1-3), you will look at the impact of scoring functions on the length of the optimal alignment.

Finally, we go back to the biology. In Q4. you find that running Blast on two sequences of similar lengths have very different outcomes in terms of running times, and what you see as the output. Why is that? This assignment requires a bit of biology, and computer science, so you should feel free to collaborate with colleagues from 'the other side'. Also feel free to use the class mailing list for general queries.

0.1 Problems

1. (30 pts.) (a) Modify *locAL* to output the *length* of the optimal local alignment, in addition to the score. (b) Next, write a program to generate random DNA sequence ($pr[A] = Pr[C] = Pr[G] = Pr[T] = 0.25$) with a specified length. Your program should be invoked as follows:
   `randomDNA.pl <number of seq> <size of seq>`
   It should output the random sequences, one per line. After the sequences are all output, the program should calculate and output a summary of the observed nucleotide frequencies in the set of sequences.

   (c) Generate 50-500 pairs of sequences (Larger number of pairs is better but you can choose a number depending on the speed of your aligner). Each pair has two sequences 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

   Note that you do not have affine gap penalties, and do not need to reconstruct aligments, so you can work with simpler and faster code that only computes the S matrix. Plot the lengths of the local alignment using parameters 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?

2. (30 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?

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

4. (20pts.) Go to the NCBI web-site, and BLAST the two sequences (available on the course web-page). You should use 'blastn', search human genomic database, and switch OFF all filters. What is the number of hits for each sequence? Why is the number different for the two sequences?

5. (18pts.) Google for "retrotransposons", "human repeats", and write a short paragraph (0.5 pages) describing the repeat structure of the human genome. Write 2-3 lines describing the impact of this repeat structure for a search tool like BLAST.

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