Algorithms in Bioinformatics
Fall 2014

Homework 2, week 6

1. A sequence assembly engine? (theory) (15 points)

   You're given two sequences x and y that represent overlapping sequencing reads from the same region of a chromosome. You know (from other data) that x and y overlap, that they're reads from the same strand, and that x is to the left of y: that is, the overlap is like this:

```
   1             i      M
x: ----------------------
y:               -----------------------
                 1      j               N
```

   so that the optimal alignment of the two reads involves x_i..M and y_1..j. This problem is a hybrid of local and global alignment. We want the overhangs to be free, as in local alignment (the gap penalties for x_1..x_i-1 and y_j+1..y_N are 0). Within the aligned section, though, we use normal gap penalties.

   Using a scoring system of +1 for a match, -3 for a mismatch and a linear gap penalty of -10 per gap symbol, give the dynamic programming algorithm that finds an optimal alignment of this type. Please show your algorithm complexity analysis for your algorithm as well.

   Bonus （5 points）：Prove that your solution is optimal.

2. Parsing Smith/Waterman output (Perl) (10 points)

   You are the new person in a bioinformatics group that's just decided that everything in their informatics pipeline will be converted from WU-BLAST searches to Smith/Waterman searches. Fortunately you notice that the program ssearch, part of Bill Pearson's FASTA package, a robust implementation of the Smith/Waterman local alignment algorithm, is installed on our server for you to use. (Do a man ssearch to see the man page.)

   You also have a legacy Perl script that takes a WU-BLAST output file and parses it to find the name of the query seuqence, and the name, socre and P-value of the top scoring hit. The source code for the script is here (/share/home/ccwei/pab/2014/hw2/blastparser.pl) . An example of WU-BLAST output is here (/share/home/ccwei/pab/2014/hw2/blast.out) . Make a copy of this script and the output in files called blastparser.pl and blast.out, and make the parser executable as a program (chmod +x blastparser.pl). When you run the script on the sample output file, it produces a single summary line of output as follows:

```
/share/home/ccwei/pab/2014/hw2/blastparser.pl blast.out
```

```
Best hit to RU1A_HUMAN is: K08D10.3, with score 378, P-value
3.2e-53.
```

Your task is to modify the script to do the same job on an ssearch output file. An example of ssearch output with the same query is here(/share/home/ccwei/pab/2014/hw2/ssearch.out). Parse the file to get the query name, and the name, Z-score and E-value of the top hit. Have the script print out a summary line similar to what the BLAST parser script did, showing this information.

Getting ahead…

p.s. Want to run the searches yourself? The query sequence is the human U1A RNA binding protein, and its sequence is here (/share/home/ccwei/pab/2014/hw2/u1_human.fa) . Get the sequence and save it to a file called u1_human.fa. The database is Wormbase C.elegans 215, the complete genome of C.elegans (/share/home/ccwei/pab/2014/hw2/C.elegans/Proteome/ws_215.protein) . The command line I used to run the ssearch program was:

```
/share/home/ccwei/tools/fasta-36.2.6/bin/ssearch36 -q u1_human.fa /share/home/ccwei/pab/2014/hw2/C.elegans/Proteome/ws_215.protein.fa > ssearch.out
```

the command line I used to run the BLAST search was

```
/share/home/ccwei/tools/wu-blast/blastp
/share/home/ccwei/pab/2014/hw2/C.elegans/Proteome/ws_215.protein u1_human.fa filter=seg+xnu > blast.out
```

All materials can be downloaded from the website of our course. However, the FASTA tool and BLAST tool are not included in the tar file. IN order to use these two tools, please contact Dr. Wei for your login information to the server.

Turning in your work

Email your algorithm for (1) and your Perl script for (2)(all in one file with name like your_name_hw2.doc, please) to ccwei@sjtu.edu.cn by 10:00AM on Oct. 31th, 2014.

--cut-----here---

独立作业承诺：（请选择一个， 并签名）

1. 本人，__________，保证本次作业由自己独立完成。

   签名

   时间　　年　月　日

或者

2. 本人，__________，保证本次作为和______________同学讨论后，由自己独立完成。

讨论内容包括____

签名　　　　　　　　，

时间　　　年　月　日