L519: Bioinformatics: Theory & Application (3CR)

HW3 (Due: **Oct. 17** Midnight)

http://darwin.informatics.indiana.edu/col/courses/L519

----------Section 1 ----------

For section 1, you are required to write a Perl script to do the following task.

1. **Write a program that performs Smith-Waterman Local sequence alignment.**
   A. You will implement a dynamic programming via Smith-Waterman Local Sequence Alignment Algorithm. You have to first understand the algorithm of Smith-Waterman alignment. Please refer to the course material for detailed algorithm as well as course materials on the course webpage.
   http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Similarity/simsrch8.html
   http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html
   B. Scoring Scheme
   You will use a simple scoring scheme instead of substitution matrix (like PAM or BLOSUM). Then you need three components **Match**, **Mismatch**, **Gap**. Let user specify these values. If not specified by users, you may use default values of your choice.
   C. Program name: **SMAlignment.pl**
   D. Suppose you have two FASTA format sequence files seq1.fas and seq2.fas, each containing a **single** sequence. And you want the script to save alignment result into 'alignment.out'. Then sample usage would be
   >SMAlignment.pl –s1 seq1.fas –s2 seq2.fas –o alignment.out –match 2 –mismatch -1 –gap -1
   (DO NOT CHANGE THESE OPTION NAMES)
   -s1: sequence 1
   -s2: sequence 1
   -o: output
   -match: match score
   -mismatch: mismatch score
   -gap: gap penalty

E. Evaluation will be based on the followings

i. Your script should work correctly.

ii. You should provide a proper README file including introduction to your script, error handling, simple explanation to your input and output files. Use a simple **text file** not MS-WORD.

iii. Your script should display proper error message.

iv. Your script should save the found alignments into a file.

v. You should follow the instruction given above regarding options.

vi. Alignments can be done on any pair of sequences not limited to DNA and protein. Since we are using FASTA format files, however, let's keep them either DNA or protein. So your script should display an error message when non-nucleotide character or non-amino acid character is encountered.

vii. Submit a single file containing all of your script, README, and sample sequences

F. Bonus Credit

i. If your script can find **every possible optimal alignment**, if more than one, you will get bonus credit.

ii. If your script can use a BLOSUM62 scoring matrix, you will get bonus credit. [-matrix BLOSUM62]

G. Sample output

i. Refer to the accompanying "alignment.out" file. Remember it's only a reference and you don't need to stick to the sample format. However, make sure that you show the pairwise alignment.

----------Section 2 ----------

FOR SECTION 2, SUBMIT A SINGLE MS-WORD DOCUMENT CONTAINING YOUR ANSWERS TO THE PROBLEM 2 THROUGH 4.

2. As a way of measuring closeness between two sequences, one can do alignments of them as seen in the class in a traditional ways like Smith-waterman and Needleman-Wunsch algorithms. There are many different ways of measuring such sequence similarity and one of them is using k-mer distribution by 'D2 distance'. D2 distance can be defined as the number of k-mer (tuple) they share. Suppose we are using 5-mers as basic units. Then we may have the following distribution from two input sequences.

	Sequence 1	Sequence 2
AAAAA	2	1
AAAAG	2	1
AAAAC	0	2
AAAAT	0	0
~		
TTTTT	2	2

Score = $\sqrt{(\sum_i (f1_i - f2_i)/4^k}$

(f1 and f2 : Occurrence of each unit in sequence 1 and 2,
k : number of unit mer / here k=5)

When compared to the traditional alignment methods,

Q1) What kinds of sequence similarities CAN be more efficiently revealed by D2 method than traditional sequence alignments?

Q2) What kinds of sequence similarities CAN be efficiently revealed by traditional sequence alignments than D2 method?

3. During the course of the determination of new DNA sequences, we have to deal with different sources of errors. Progress in sequencing techniques should contribute to reduce the frequency of random errors, but systematic errors will remain very difficult to resolve. In general systematic errors involve single base substitution and have little effect on the quality of the final sequence. However, a few of them concern base insertions or base deletions (indels) and produce '**frameshifts**' in coding regions that can remain undetected if another indel re-establishes the original frame and if the frameshifts do not introduce a stop codon in the main coding frame of the sequence. Such errors are the most annoying since they corrupt the amino acid sequence over several positions and can compromise future analyses.

Original :	AUG AGC ACG AAU CCU AAA CAA AGG GUA AAG
	M S T N P K Q R V K
Error :	AUG AGC ACG AAU CCU AAA CAA AGG **U**AG UAA
	M S T N P K Q R * *

Insertion of U created an abrupt stop codon which altered corresponding protein products significantly.

Suppose that you are given a codon usage table of coding sequences of your sequenced organism, please formulate or devise an algorithm that can employ this codon usage table and **detect frameshift sequencing errors**. Please refer to class material about '**Gene Finding'.** Many of the necessary concepts like ORF (Open Reading Frame), codon usage, and etc will be covered during the classes.

4. As genome sequencing projects of many different species are completed, we now have tremendous amount of valuable DNA sequence information. One of the most important tasks after genome sequencing is annotating all genes in the genome. Even with huge efforts to reveal the possible functions of genes, many genes are yet to be functionally undetermined, or sometimes referred as hypothetical. The following protein sequence is obtained from a newly completed genome sequencing project and seems to have multiple domain. Find out all possible domains and provide multiple alignments of each domain by using each domain's close homologues. You should submit alignments in **MSF** format.

   Reference sites: NCBI : http://www.ncbi.nlm.nih.gov
   Expasy : http://www.expasy.rog
   ClustalW : http://www.ebi.ac.uk/clustalw

```
>Hypothetical_Protein
MRRMFLVSRNGYKRWPEKQKLQTFRVLISTYIFSQPPILAQYHTIPCSNLKLYCRRPRSTMAKKNNKKSK
AKAKKAAPSVAVPATNVASSVAADPAAVETASSTSASLSVPESAAATETTASTSPTSPVTNVSPLAEAIA
GDEGSGSIPDDEEDIDEEEETGTEESAEAEAEPESPEGTTPVLTVQPNALAAEPTVETVDVVVPVPVEPE
EDEQESEVKAEIDPVVAEIEQPEPPMASDKDFFSTLSKEEDGEEDVKEVATSTTTDVATPVATATPSITQ
VEKHLESRFMGNPVVENEKKHKEDEIREVEDQLTERLMRDPVVEGIKRERAEESAQVEEKLEKRFMEDPV
VEQIKQEQADEYRKVRDQLETGSEEPNDDFFANLGSEEAKLQAKDTEPDTKHEDDFFSTLGGSSDETGKK
TVAETQSEPETKTIETAKAEEAKPIETKTEPVAEDDFFSGLGKSEQEQGFVPEHTTAASTTTKEEDDFMA
SLAQPQNPEQTDPLPQDDFFSQLAEENVQSVSAEPEHKSPVAAPAATKAAAAEDDFFSQLGKSPEKSVPA
AATSDDAPRRRHKVQPSGADFFDQLGVSETEAPPALHEPPQPKPITLSLDEDDDLLLTDEDDQAEVLAEA
GKPNPPVQSFAFLEDDDLLESDQDFLETDDEDEEPMAAQTGGDFRANQQHQNTYFPSVPVAPVSTSRYST
PTVPSVSQFGGYGAAQSTPNMYQPVSSQSTPSAAAPGAGKRVDKNKSDAFDFPTGMIPKVVKKARSQQQL
PQAHGAPGVTPGLMPAFGAAPGAPPTPGVPPMGPPVARPASNPYAPAPVQPPTTAPKKNFFEELPPIPVK
PISRQPSYALSPPRAPFAQEASQQPRRVSDGYGMPSHGGPHSGVHSAPVSHHNSVSGPPAAAPHNPYAPP
SGPSAVPPKTSSPYAHPPAAVPPKTSSPYAPPPPAVPPKSSSPYAPPPVSHSAPPAGGPPAGPPRGTSRG
VPVPQPQPPVAAAAAAAAAAAAAAAGTASGPPPAGPPRVSSRNAYAPPRGAPPRGTTPPVVAPAMAYSPNV
SPKRVIQQPQTQSPRRYSEFKDIGQKSTVSDEALRKRQFPIFRWSNSKNATCLIAPQIGYATAVSQTSVR
LLDVSKVVSTGEDITRFPGPLFTPNKGPVKSKKKELEKWVADHVNLLDSQNADADRVLLWKLVRVYLAND
GVINSAQVRAFLTPHFEQSATGDGSAFTSAMDLSSSSFVPQQGDNGPSFSGSDANHVLRHLQSGSKDAAI
RHCMDRRLWGHSMLIASTMGPEKWKDVVSEFIKDDVRPLYKPTLQFLYSSFGGVIPSSEAYGDWKETVSY
LLSNAKDDGSDLSSLVSLGDDLVQKGYVAAGHFCYVVSRAPLTDKITLLGSEGRDLDAILLSETYEFALG
LKTNVAIPQMQLYKLVHAEVLADLGNVAQAQRYAEYLNQALKSFTDKSSIIQPAYISRLIALSDRVSSTP
GATTGSWFSRPKLDKVLGHLDKSLSKFVAGEEANVAGASSASDTVFSQIAATPGISRTTSVVDLSQQSAV
TPGYQQHQPYGVPPTRASTGNILRPQGGSGPYDSRPSLPRSSSAMDAGPSGYGYERAPSVASVHSVQSDY
PRVMSPIDVGGVSGGNSAYAPVGGAGIYPLPAGGSSAANAYAPGAGGNANAPPSGATANAYTPGATSSGP
SPYALPSGNVYAPPSAPSASGASPYAPSASSFSPRQPAYGRSYASTPEHHIEEEEETEAVDQEQEPAADY
SHLENENERYEQKSEPEPEPMAHPPPAQKAPPAAPPAQQAPAQKAGQKPPARKVAPPPKPSVKSVYNPYD
PGSPAKEKKSSAAASNKYAPANSAYSPGNSSAPAPTANKEPEPATTSEAYGYGYGGGYGGYDPYSGYPAA
EEEGAEPTETEVENEKAGEAEAAAADYPDYGVPSYGYQSQAADDAGDYGDDEGAIYTPAAVTLPPISNLP
PVPLPGLEPATSAAPPASRYSAPTAAAEEEDDDDFGVGNKKPVAKKEEKKAEEKEEPKDGGKKGWFGGWF
KKGEQQPADDKKVYKAKLGEKSNFYYSEEHKRWINGDLPIEDQIKGAGGPPPPPKAKKPAGPPAGGTPPP
PSGGAPPVGASRGATPPVATPPAADTPPVPGAGGAPRPAATGAPRPKVVDPLEGFLTGGPPTGAPRKGAR
KSAKSRYVDIMNN
```

5. Mini Group Project

We now have the second mini class group project. The purpose of this project is to create a web site that has an integrated list of specific gene families. For this project, our target organism is Drosophila menanogaster (or so-called fruitfly). Each of your groups is to choose any one interesting gene family (not limited to the following families listed below) and find the genomic coordinates of the potential genes of this family and implement a website to report these results.

a. GPCR (G-protein Coupled Receptor)
b. Protein Kinase
c. Transporter
d. Hydrolase

- Drosophila melanogaster sequence data is available at Biokdd
  /tmp/L519FALL2005/Drosophila_menanogaster
  Copying these files to your directory is not recommended due to the quota limit.
- FlyBase : http://flybase.bio.indiana.edu/
- Genome Browsers: UCSC Genome Browser, Ensembl