CISC 841 Bioinformatics, Spring 2006
Homework 1
(100 points)

Handed out: February 28, 2006
Due date: March 9, 2006

Sequence pairwise alignments

1. [25 points] Let two sequences be s1 = CDAA and s2 = AEECA, and a scoring matrix

	A	C	D	E
A	2	-2	-2	-1
C		1	0	0
D			2	-2
E				2

   Find the highest score by aligning s1 and s2 globally when the gap penalty is 1.8 + 0.4 g, where g is the length of the gap. Show the best alignment.

2. [60 points] Implement Smith-Waterman algorithm with affine gap penalty for protein sequences. The following are the specifications regarding your implementation.

   a) Get sequences from input file (in FASTA format), and write to the standard output.
   b) Command line option **`-o 1`** to output the DP table. The default is to report only the best alignment and score.
   c) Command line **`-s <filename>`** to read substitution score matrix. (A sample matrix is available at http://www.cis.udel.edu/~lliao/cis841s06/blosum62, which should be hard-coded into your program as default score matrix.)
   d) Command line option **`-g a b`** to allow for choices gap penalty, where **a** and **b** are two integers serving respectively as penalty for opening a gap and extending a gap
   e) Name your script as "xxxx_align", where xxxx is your last name followed by the initial of your first name.

   `Synopsis:` **`xxxx_align [-o 1] <input>`**
   Instructions for submission:
   1) email me your code and *readme* file with subject: cisc841 hw1, your last name
   2) hand in a hardcopy of your finished assignment to me in class on the due day.

   N.B. You are strongly encouraged to use Perl.

3. [15 points] Run your program with default scoring matrix and gap penalty (-11, 1) on the two sets of sequences available at `http://www.cis.udel.edu/~lliao/cis841s06/hw1_dataset`. For each set, take one sequence and compare to the rest sequences in the set and record the best scores in a form of histogram. Repeat the same procedure for the other set. Compare the histograms and discuss.