Computer Science 303 – Algorithms
Programming Assignment #2: Global Sequence Alignment
Due: 4/2/2014

In this programming assignment, you will implement the sequence alignment algorithm.. More specifically, you will implement the pair-wise global sequence alignment based on fasta file format.

Objective

- Get a better understanding of dynamic programming
- Understand the usefulness of computer science algorithms in solving important problems from other sciences.

Task

You should write a code that accepts two input sequences based on fasta format. Fasta stands for "Fast All" which is the generalization of "Fast Protein" (aka fastp) and "Fast Nucleotide" (fastn). It is a commonly used format to store sequences either for protein or DNA/RNA.

A sequence in FASTA format begins with a single-line description, including the name of the gene, the starting position of the gene, the reference number of the gene in the gene bank, and a short description of the gene. Following the first lien is the lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. Usually the sequence data has 70 characters in each line. An example sequence in FASTA format is:

```
>gi|189181665|ref|NM_000520.4| Homo sapiens hexosaminidase A (alpha polypeptide) (HEXA), mRNA
AGTTGCCGACGCCCGGCACAATCCGCTGCACGTAGCAGGAGCCTCAGGTCCAGGCCGGAAGTGAAAGGGC
AGGGTGTGGGTCCTCCTGGGGTCGCAGGCGCAGAGCCGCCTCTGGTCACGTGATTCGCCGATAAGTCACG
GGGGCGCCGCTCACCTGACCAGGGTCTCACGTGGCCAGCCCCCTCCGAGAGGGGAGACCAGCGGGCCATG
ACAAGCTCCAGGCTTTGGTTTTCGCTGCTGCTGGCGGCAGCGTTCGCAGGACGGGCGACGGCCCTCTGGC
CCTGGCCTCAGAACTTCCAAACCTCCGACCAGCGCTACGTCCTTTACCCGAACAACTTTCAATTCCAGTA
CGATGTCAGCTCGGCCGCGCAGCCCGGCTGCTCAGTCCTCGACGAGGCCTTCCAGCGCTATCGTGACCTG
CTTTTCGGTTCCGGGTCTTGGCCCCGTCCTTACCTCACAGGGAAACGGCATACACTGGAGAAGAATGTGT
TGGTTGTCTCTGTAGTCACACCTGGATGTAACCAGCTTCCTACTTTGGAGTCAGTGGAGAATTATACCCT
GACCATAAATGATGACCAGTGTTTACTCCTCTCTGAGACTGTCTGGGGAGCTCTCCGAGGTCTGGAGACT
TTTAGCCAGCTTGTTTGGAAATCTGCTGAGGGCACATTCTTTATCAACAAGACTGAGATTGAGGACTTTC
CCCGCTTTCCTCACCGGGGCTTGCTGTTGGATACATCTCGCCATTACCTGCCACTCTCTAGCATCCTGGA
CACTCTGGATGTCATGGCGTACAATAAATTGAACGTGTTCCACTGGCATCTGGTAGATGATCCTTCCTTC
```

For more information about fasta format, you can refer to http://www.ncbi.nlm.nih.gov/blast/fasta.shtml for details.

Your code should read in two sequences with file names specified by the user and then calculate the optimal sequence alignment with the following parameters

- Gap penalty: -5
- Match: +2
- Mismatch: -1

Your code should not only report the score of the best alignment, but also report the optimal alignment itself. The alignment should be output in a format similar as below.

```
AGTTGCCGACGCCCGGCACAATCCGCTGCACGTAGCAGGAGCCTCAGGTCCAGGCCGGAA
|||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||
AGTTGCCGACGCCCGGCACAATCCGCTGCACGCAGCAGGAGCCTCAGGTCCAGGCCGGAA
```

If there are more than 1 optimal alignments, you only need to report one of them. You can test your program by using the three sequences I have uploaded on ANGEL.

These three sequences are from a gene called HEXA that exist in many mammals. Its mutation is believed to be responsible for tay-sachs disease.

Tay-Sachs disease, a heritable metabolic disorder commonly associated with Ashkenazi Jews, has also been found in the French Canadians of Southeastern Quebec, the Cajuns of Southwest Louisiana, and other populations throughout the world. The severity of expression and the age at onset of Tay-Sachs varies from infantile and juvenile forms that exhibit paralysis, dementia, blindness and early death to a chronic adult form that exhibits neuron dysfunction and psychosis. For more information about this disease, you can go to http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gnd&part=taysachsdisease

Note

- Your code should report the pair-wise sequence alignment on the three sample sequences (i.e. human~chimp, chip~mouse, human~mouse).
- Your code must work for any input file in the correct format.
- For your reference, the correct score for Mouse and Chimp alignment should be 28.
- Please put your honor pledge as comments in the code you turn in.