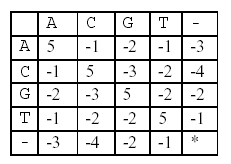
Assignment 4 Description:

The instructions for creating and running a living organism are encoded in DNA sequences. Most multicellular organisms are made of eukaryotic cells that contain mitochondria, an organelle that manages cellular metabolism. Interestingly, mitochondria contain their own unique DNA sequence which is passed down through the mother. Organelles are mini-factories within cells that perform essential functions. It is thought that early in the evolution of eukaryotic cells,  a  more primitive cell that was "captured" by the more complex eukaryotic cells. This simple cell became mitochondria. This explains why mitochondria have their own DNA. In this assignment, you are to implement a simple DNA alignment algorithm and apply it to analyze a set of mitochondrial DNA.

DNA is a large linear chain of nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases (cytosine [C], guanine [G], adenine [A] or thymine [T]). So a DNA sequence can be represented as a string of characters, where each character is one of C, G, A, T.

Modify your simple edit distance code to apply it to DNA alignment. Here we will use a slight variation in how we count an "edit step". During the evolution of living organisms, the probability of substitutions, deletions, and insertions is not identical but a function of specific nucleotide due to biochemical and selection pressure effects. This leads to an edit distance function represented as the following table. Using this maxtrix the problem becomes a maximization of alignment problem.



First test the algorithm on some short DNA sequences. Write the traceback routine and the code to print out the alignment vertically.

Do some simple experiments to estimate the speed of the algorithm as a function of the length of the DNA sequence. Here generate random sequences of ACGTs of increasing lengths. What size of the solution matrix does the algorithm become impractical?

Next review the mitochondrial DNA sequences from the following sources:

Prototypical human: [http://www.ncbi.nlm.nih.gov/nuccore/251831106 (Links to an external site.)](http://www.ncbi.nlm.nih.gov/nuccore/251831106)

Neandertal: [http://www.ncbi.nlm.nih.gov/nuccore/196123578 (Links to an external site.)](http://www.ncbi.nlm.nih.gov/nuccore/196123578)

Human diversity: [https://www.ncbi.nlm.nih.gov/nucleotide/NC\_012920.1 (Links to an external site.)](https://www.ncbi.nlm.nih.gov/nucleotide/NC_012920.1)

Great Apes: [https://www.ncbi.nlm.nih.gov/nuccore/NC\_001645.1 (Links to an external site.)](https://www.ncbi.nlm.nih.gov/nuccore/NC_001645.1)

In all these files, the DNA sequence is given at the bottom of the file following the keyword "ORIGIN". The text must be read in and cleaned up by removing the gaps, new lines, and line numbers.

In this assignment, you will apply your algorithms to measure the alignment distance between DNA sequences from these data sets.

First study the relationship between Human and Neandertal sequences (see [http://genome.ucsc.edu/Neandertal/ (Links to an external site.)](http://genome.ucsc.edu/Neandertal/) for some background). Download the sequence from [http://www.ncbi.nlm.nih.gov/nuccore/196123578 (Links to an external site.)](http://www.ncbi.nlm.nih.gov/nuccore/196123578) and [http://www.ncbi.nlm.nih.gov/nuccore/251831106 (Links to an external site.)](http://www.ncbi.nlm.nih.gov/nuccore/251831106), compute the distance between the two sequences and generate their alignment (as a text file).

Next study the relationship between the diversity of human mitochondrial sequences available from [https://www.ncbi.nlm.nih.gov/nucleotide/NC\_012920.1 (Links to an external site.)](https://www.ncbi.nlm.nih.gov/nucleotide/NC_012920.1)[(Links to an external site.)](http://www.mtdb.igp.uu.se/). Pick 10 sequences and run the alignment algorithm to compute all pairwise distances between the sequences. Produce a table of your results.

Finally, apply your algorithm to study the distance between a prototypical human sequence ([http://www.ncbi.nlm.nih.gov/nuccore/251831106 (Links to an external site.)](http://www.ncbi.nlm.nih.gov/nuccore/251831106)) and samples taken from the great apes ([https://www.ncbi.nlm.nih.gov/nuccore/NC\_001645.1 (Links to an external site.)](https://www.ncbi.nlm.nih.gov/nuccore/NC_001645.1)[(Links to an external site.)](http://www.mtdb.igp.uu.se/), bottom of page). Produce a table similar to the study of human diversity that includes all pairwise distances. Additionally, produce six text files each containing the alignments between the prototypical human and the 6 great ape sequences.

Write a brief summary discussing your observations on the algorithm performance and the scientific impact of this efficient alignment algorithm.