CS 466 Final Project Report

jlhuang2

kbatra3

Goal:

One of the main uses of the sequence alignment technique is to construct phylogenetic trees, which depicts the evolutionary relationships between different species. In our project, we will compare the result of an alignment between the genome sequences of a cat and a tiger and the alignment between the genome sequences of a dog and a tiger. We expect the cat and the tiger alignment to have a higher score and be better aligned, as a cat and tiger likely have a more recent common ancestor than a dog and a tiger and are also in a common family (Felidae).

Alignment Code:

* global\_align.py:
* subs.txt:

Data (not included in zip due to size):

* Data/GCF\_000002285.3\_CanFam3.1\_genomic.fna
* Data/GCF\_000181335.3\_Felis\_catus\_9.0\_genomic.fna
* Data/GCF\_000464555.1\_PanTig1.0\_genomic.fna

Implementation:

To implement this project, we used the dynamic programming algorithm for alignment we talked about in lecture. In order for it to be run more efficiently on large datasets, we implemented explicit memorization to fill in the matrix iteratively and get rid of unnecessary computation. We discovered that in the genome files, aside from the four types of bases (A, C, G, T) there was also the letter N, which represent unknown sequences in the DNA and are there to maintain proper distance between known sequences. To work around this, when reading in the sequences, we do not read a portion of a sequence if it is aligned with N or contains N, as we only want to align known sequences.

Results:

The greatest challenge