

CIS 4930 / CIS 6930 (Recent Advances in Bioinformatics) Spring 2009, Homework 1

Due date: 02 / 16 / 2009
Turn in hard copy in class

February 4, 2009

The purpose of this homework is to learn the edit operations in the frequency domain and gain familiarity with the DNA data type in FASTA format. You will implement a program that computationally analyzes data files that contain DNA sequences in FASTA format as described below.

Data Source. You will download data from GenBank (<ftp://ftp.ncbi.nih.gov/genomes/>) Pick three different organisms, one of them should be human (*H.sapiens*). *H.sapiens* chromosome will be your query sequence and the other two organisms will be your database sequence.

Download one chromosome from each of these three organisms in *fasta* format. The extension of the fasta files in GenBank are “.fa”. For example, you can get human chromosome 22 at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/CHR_22/hs_ref_chr22.fa.gz. Make sure that each of these three files have at least 5 million bases (i.e., nucleotides). If a chromosome is much longer than 5 million bases, you can keep the first 5 million bases and remove the rest.

Code. Your code will read the query and a database sequence and perform a query for each of the query lengths $L = \{50, 100, 200, 400\}$ as follows. For each L do the following for each database file.

1. Select 100 random query sequences from the query file by randomly selecting subsequences of length L . Compute the frequency vectors of these query sequences and store them.
2. Partition the database file into nonoverlapping subsequences of length L and store their frequency vectors. There will be roughly $N = \frac{5\text{million}}{L}$ such vectors.

3. Compute the frequency distance between each query frequency vector and database frequency vector. Totally, there will be around $100 \times N$ distance computations per query set. Store these distance values in a histogram.

Remarks. The following remarks can be helpful for this homework.

- You can learn about fasta format from many resources. Just google “fasta format”. An example location is <http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>
- Nucleotides are shown with letters A, C, G and T. Sometimes the fasta files contain the letter N. This usually means that that base is not known precisely. Ignore all Ns.
- A fasta file may contain more than one sequences. If each one is less than 5 million bases, you can merge them to make them longer.
- The discussion on the frequency vectors and the frequency distance, as discussed in the class, are available in <http://www.cise.ufl.edu/~tamer/papers/vldb2001.pdf>

Return. You will return the following in your report.

- The name of the organisms and the name of the chromosomes you have selected in your dataset.
- Plots of all the histograms. This corresponds to totally $2 \times 4 = 8$ plots.
- A brief discussions of all the plots (e.g., impact of L , distribution of distance values in different query sets, etc).

I do not need the code you implemented at this stage. Save your code though. I may request some of you to send me your implementation based on the results.