ECE-S690-503: Genomic Signal Processing: Introduction to Genomic Databases and the Matlab Bioinformatics Toolbox Homework Assignment 1

Due on:

April 21st

Introduction

This homework will familiarize you with using online databases to find DNA sequences and familiarize you with the Matlab Bioinformatics toolbox.

Reading for next week:

- Fitch JP, Sokhansanj B. Genomic engineering: moving beyond DNA sequence to function. Proc. IEEE, 88(12): 1949-1971, Dec. 2000.
- Dimitris Anastassiou. "Genomic Signal Processing," IEEE Signal Processing Magazine, 2001.
- Vaidyanathan and Yoon. "The role of signal-processing concepts in genomics and proteomics," Journal of the Franklin Institute, Special Issue on Genomics, 2004.
- E. R. Dougherty, A. Datta and C. Sima, "Research Issues in Genomic Signal Processing," IEEE Signal Processing Magazine, Vol. 22, No. 6, 46-68, November 2005.

1

a. Read the Entrez Tutorial (http://www.ncbi.nlm.nih.gov/entrez/query/static/help/entrez_tutorial_BIB.pdf) and the Entrez Manual (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp). Do the questions on MLH1 (colon cancer) (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/cover_coloncancer_exercises.html) and PER2 (circadian) (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/cover_circadian_exercises.html).

You do not need to hand in answers to these questions as they are provided; the point is to increase your familiarity with the Entrez query system. After doing this tutorial, find the location of the SNP which causes sickle-cell anemia. Print a screenshot of the webpage and include it with your HW.

b. Do the interactive BLAST Tutorial (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html). Then, using the skills you developed in the BLAST tutorial, find the proteins which are sequence similar to CC_3035, a cell cycle regulatory protein in Caulobacter crescentus. This will be a web page on NCBI with either BLAST results or BLINK results (pre- cached BLAST). Print a screenshot and include it with your HW.

2

Sometimes when you are searching for a gene or locus, you do not necessarily know the name of the gene you are searching for. The ability to search on keywords can be a useful tool.

- You will do a text search using the query: `heart disease`
- Go to NCBI MapViewer homepage and search the human Map Viewer for the term heart disease. Ignore the hits on the Celera genome,b ut record the number of locus of each hit on the reference genome.
- Hint: You can use the Assembly button to change which assemblies are searched and displayed.
- Create a table with the chromosome number, Map element, type and Map(s) for each match to the refrence genome.
- Go to the UCSC genome website, open the human genome browser and search using the term `heart disease`.

1. What was the difference between the type of hits found using MapViewer and those found using the UCSC genome browser?
2. Which search turned up more results? Why?
3. Starting with the cytogenetic regions (not genes) found in the MapViewer search, if you use those regions as a query in a text search of the UCSC genome browser do you get a match? Show an example of one of the regions and print out the Genbank report and the UCSC report for the region. Include in your Homework.

3

Using SEQSTATSDEMO in Matlab as a reference, calculate the:

a. Explore Genbank. (http://www.ncbi.nlm.nih.gov). Pick a sequence of interest and give two paragraphs of background information (accession number, locus, information from PubMed, how many articles are available on PubMed about it, etc.) about your sequence that you can find through Genbank.

b. NTDensity (nucleotide density) of your sequence. What is the window length that the function uses to calculate the NTDensity? Change the window length to half the size, and compare this NucleoTide Density plot to the old one. (Include both plots. Please label all your plots with the Accession Number of your sequence in the title and any additional information (in this case, the window size).

c. Do a basecount of the sequence. (Give results).

d. What is CG content of the sequence (percentage of Cs + Gs of the sequence NC+NG/NTotal)? . (Give percentage).

e. Even if you did not pick a protein-coding sequence, pretend that there is a 3-base window incrementing 3 nucleotides every iteration. Calculate the basecount for each base position. Code a Matlab script that will take a sequence and calculate the basecount for each subsequence.

Example, say you have a sequence: `CAGTGCATTATGGAT` (1)

Base position: `012012012012012` (2)

The basecount for position 0 would be: A: 2, C: 1, G:1, T: 1

Please provide your code and results.

f. Do a codon count of your sequence. (Give results).

g. Plot the codon bias for each amino acid via pie charts. (see codonbias) (Give results and explain how this differs from a plain codon count).

h. Generate a random nucleotide sequence that has the same length as the sequence you analyzed in step 3. Redo question 3 for the random sequence. Note any significant differences in statistics from an actual sequence to a real sequence.

i. Please report on any further findings that you have discovered in the toolbox.

4

(Original question from Alterovitz et al.)

1. Klenk H.-P. and colleagues published the complete genome sequence of the organism Archaeoglobus fulgidus in Nature. [Nature 390:364-370(1997)].

(a) Find the protein sequence of the hypothetical protein AF1226 precursor. What is the sequence of amino acids (in single letter representation) from positions 141-147?

(b) Write a Matlab function that calculates how many nucleotide sequences can give rise to an arbitrary amino acid sequence (hint: revgeneticcode). The input to the function should be a string (amino acid sequence) and the output should be the number of potential nucleotide sequences.

(c) Using the Matlab function from part 1b, calculate the number of sequences that could give rise to the 7 amino acid sequence found in part 1a.

(d) Create a Matlab function `nt2aa_cgbiased.m` that will take an amino acid sequence as the input and will choose the corresponding codons with the highest CG content. Note: If there is more than one sequence that satisfies this condition, pick one at random. What is the nucleotide sequence that could give rise to the 7 amino acid sequence found in part 1a.