L519: Bioinformatics: Theory & Application

HW4 (Due: **Nov. 4** Midnight)

http://darwin.informatics.indiana.edu/col/courses/L519

----------**Section 1** ----------

For section 1, you have two Perl script assignments..

1. **Periodicity of DNA sequence**

Periodicity is the quality of occurring at regular intervals (e.g. of time) and can occur in different contexts like clock, metronome. Imagine that we are dealing with DNA sequence and want to check whether there is periodicity in the sequence or not. For an interval ***j***, we computer the correlation score C(j) as following,

$$C(j) = \sum_i \frac{c(S_i, S_{i+j})}{N}$$

1. N = Number of total pairs of (S_i, and S_{i+j}). In fact, this is getting very close to the total number of characters in the given sequence if it's long enough.
2. c(S1, S2) = 1 (S1=S2)
   0 (S1≠S2)

If the input sequence has a strong periodicity *j*, this correlation score C(*j*) should be significantly higher than other scores (different *j*).

For example, suppose we are checking the interval of 2 for the following sequence,

A1C1A2C2G1C3G2T1

C(2) = {c(A1,A2) + c(C1,C2) + c(A2,G1)+c(C2,C3)+c(G1,G2)+c(C3,T1)}/6

= { 1 + 1 + 0 + 1 + 1 + 0}/6

= 4/6 = 0.67

which indicates relatively high level of periodicity.

A. Write a program that calculates the correlation score within a given range of intervals from a DNA sequence.
   i. Input : A FASTA file containing a DNA sequence (ATGC only)
   ii. Output : A File containing correlation score within a given interval ranges
   iii. RESTRICTION
      1. Submit a README file named "**README.PERIODICITY**"
      2. Submit a Perl Script named "**PERIODICITY.pl**"
      3. Options: **-i <InputFile> -o <OutputFile>**
         -min <MinPeriodicity> -max <MaxPeriodicity>
         A. Ex) PERIODICITY.pl –i test –o result –min 2 –max 20
      4. If you don't keep the restrictions above, you will lose some credit.

B. Plot a graph showing periodicity score vs periodicity interval from 2 to 20. DNA sequences will be given to you.
   i. Image file should be submitted.
   ii. Comment whether you found any notable periodicity from the given sequences in the README file
   iii. If you want to use gnuplot which is a scientific plotting tool, here are some references.
      1. http://www.gnuplot.info/
      2. http://www.cs.uni.edu/Help/gnuplot/
      3. And many many others from Google

2. Write a Perl script that finds the longest ORF (Open Reading Frame) from a given DNA sequence.
   A. Input: A FASTA format sequence file
   B. Output: A File containing the following information

      # Sequence File : Test.fas

      # Input Sequence Length : 2000

      # Longest ORF Frame : +1

      # Longest ORF Starts : 379

      # Longest ORF Ends : 679

      # Longest ORF Translation

      MSSHLVEQPPPPHNNNNNCEEGEQSLPPPAGLNSSWVELPMNSSNGNDNG

      NGKNGGLEHVPSSSSIHNGDMEKILLDAQHESGQSSSRGSSHCDSPSPQ

   C. RESTRICTION (KEEP ALL OF THE FOLLOWING)
      i. Submit a README file named "**README.ORF**"
      ii. Submit a Perl Script named "**ORF.pl**"
      iii. Options " **–i <InputFile> -o <OutputFile>**"
         1. <InputFile> and <OutputFile> name should be the full name of the file. If your file is "Test.fasta", then <InputFile> should be "Test.fasta" not "Test"
      iv. Your sample input sequence and output result file.
      v. If you don't keep the restrictions above, you will lose some credit.

----------Section 2 ----------

FOR SECTION 2, SUBMIT A SINGLE MS-WORD DOCUMENT CONTAINING YOUR ANSWERS TO THE PROBLEM 2 THROUGH 4.

3. As introduced during the class, many amino acids are encoded by several codons, and this is called 'codon degeneracy'. According to the general codon table, the third nucleotide change tends to be synonymous which means the corresponding amino acids usually don't change. In this view point, the third codon position will generally endure higher rate of mutation across species than the other two positions. Can you propose a method to make use of information to predict gene regions from multiple alignments of genomic sequences?

TABLE 4-1 The Genetic Code (RNA to Amino Acids)*

First Position (5′ end)	Second Position U	Second Position C	Second Position A	Second Position G	Third Position (3′ end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu (Met)*	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met (start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val (Met)*	Ala	Glu	Gly	G

*AUG is the most common initiator codon; GUG usually codes for valine, and CUG for leucine, but, rarely, these codons can also code for methionine to initiate a protein chain.

4. Microarray Practice.

A. Introduction to Microarray

For comprehensive introduction to Microarray and Data Analysis, please refer to the PDF document on the **Biokdd** server.

/tmp/L519FALL2005/Microarray/

B. Microarray Data Repository

There is a public repository for microarray experiments on the web. SMD (Stanford Microarray Database: http://genome-www5.stanford.edu/) contains numerous downloadable microarray data. Go to SMD and download any microarray experiment(s) of your interest. Once you obtained your data, apply clustering program and report the final clustering results. Don't forget to specify the parameters you used to discover the clusters.

C. Clustering program to use:

Use Cluster developed by Michael Eisen at UC Berkeley. The cluster program can perform various types of clustering algorithms like hierarchical clustering, SOM (Self-organizing map), k-means clustering, and PCA (Principal component analysis). In order to view the clustering result in a graphical way, you need to download 'TreeView' from the same website.

We will have some demonstration on Oct 28th (Friday) but you better try to figure out how to use the program beforehand.

5. Tiling array is a new type of microarray that interrogates genomes with high-density probes. In a typical tiling array, probes are distributed along chromosomes approximately evenly at a density of one probe per 10~100 bp. Owing to the high density of the probes, a whole genome can be surveyed in an unbiased manner at a high resolution. The whole genomic tiling arrays consisting of oligonucleotides or DNA fragments spanning all high complexity DNA sequences from large genomic regions are emerging as powerful tools for functional genomics.

Propose a method to predict genes in a Eukaryotic genome that makes use of the hybridization data of genome tiling with mRNA.

* References for basic idea of genome tiling array.

A. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays

B. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments

C. And many others from PubMed