L519: Bioinformatics: Theory & Application

HW5 (Due: **Nov. 16** Midnight)

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

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

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

1. **Conserved Profile Search**

We discussed the motif finding problem in the regulatory regions of a set of co-regulated genes. Discovery of TFBS (Transcription Factor Binding Site) can also be performed through the analysis of homologue DNA regulatory regions from multiple species (e.g., human, mouse, rat, and chicken). The motif of TFBS are presumably more conserved across closely related species than the background DNA sequences owing to the high evolutionary pressure on TFBS sequence. This method is often referred to as DNA fingerprinting.

You are to write a Perl script that can discover conserved profiles from a multiple alignment of many DNA sequences. A conserved profile refers to the profile of aligned sequence block with following three restrictions.

A. No gaps
B. Block length is longer than given length *l*
C. Entropy of the block is lower than given threshold *e*

- Program name: **CProfile.pl**
- Input
  A. Multiple alignment file: This file should be in Clustal format
  B. Minimum block length
  C. Maximum entropy threshold
- Sample usage
  D. >CProfile.pl –i MSA.aln –l 10 –e 1

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

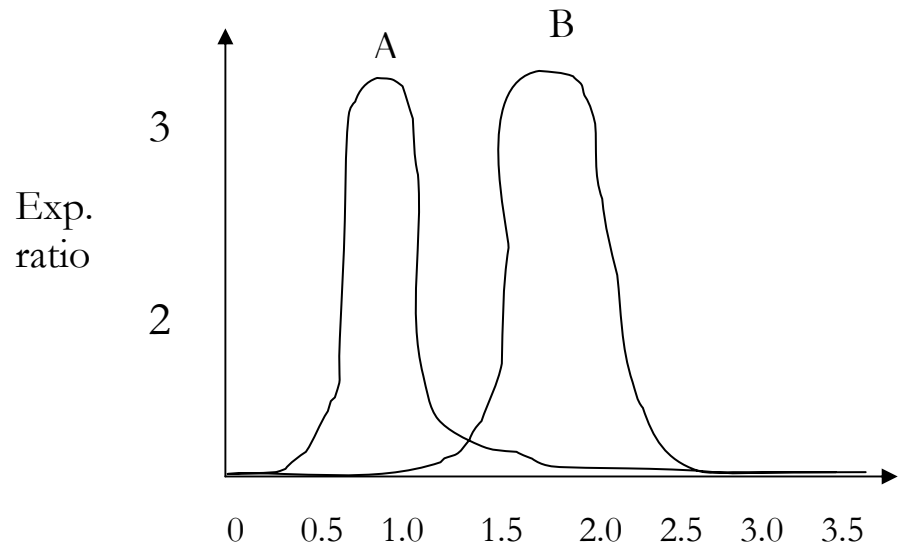
FOR SECTION 2, SUBMIT A SINGLE MS-WORD DOCUMENT CONTAINING YOUR ANSWERS TO THE PROBLEM 2 THROUGH 5.

1. Suppose you are given a set of microarray data, in which the expression levels of *N* genes are measured from many consecutive time points during a particular biological process such as cell cycle. If you want to compare the expression profiles of two genes, you can compare the distance of correlation between two vectors that represent the expression profiles over the time courses as we discusses in the class.

   There can be many genes that are involved in a certain biological process under different time scales. For example, gene A and B in the Figure 1, are both activated during, but the expressions level of the gene *A* changes much slowly than the gene *B*.

   Derive a method that can reveal the similarity of expression profiles between genes such as *A* and *B*.

   Hint. The starting responding time point of two genes should be synchronized and the interval should be calibrated.



2. Microarray experiments can reveal the expressional profiles of genes along the times. Suppose you can measure distance (or similarity) between two gene's expression profiles. Propose a method that can show the significance of such distance of a given pair of genes.

3. Suppose you have genotyping results of a family (father, mother, and a child) for three genetic markers.

	SNP1	SNP2	SNP3
Father	A/A	A/T	G/A
Mother	C/C	A/T	G/A
Child	A/C	A/T	G/G

What are the possible combinations of haplotypes from the parents to give a birth to this child?

SECTION 3. Minigroup project#3

As you experienced during the class and the lab session, there are many different motif discovery programs. For the mini-group project#3, you are supposed to compare currently available motif discovery programs. Your group should prepare your benchmark results on a web site and have to prepare a five minute presentation.

Programs: Gibbs sampler, MEME, and as many as other available programs. (The more, the better)

Features to check: Accuracy and running time (speed)

Dataset: http://bio.cs.washington.edu/assessment/

You can refer to Martin Tompa's paper about assessment of computational tools for motif discovery of transcription factor binding sites.
http://www.cs.washington.edu/homes/tompa/papers/assessment.pdf

Hint: Unix command 'time' can tell you the running time of a process.

```
>time perl CProfile.pl
real    0m2.133s
user    0m0.030s
sys     0m0.040s
```

But remember that the actual running time will depend on the current load of the machine you use.