L519: Bioinformatics: Theory & Application (3CR)

HW2 (Due: **Oct. 3** Midnight)

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

INTRODUCTION:

There are two sessions to be completed. The session 1 is for Perl programming and the session 2 is for algorithm. You should submit your assignments to Oncourse. Any document format like MS Word (doc), Acrobat (PDF), and PostScript (ps) are welcomed.

QUESTION:

Don't hesitate to contact me (Haixu Tang: hating@indiana.edu) or AI (Junguk Hur: juhur@indiana.edu).

INSTRUCTION:

1. Please start to work the homework as soon as possible. For some of you without enough computational background may need much more time than others.
2. Include README file for each programming assignment. This is not supposed to be lengthy but should contain concrete and enough information;
   A. Function of the script
   B. Input / Output
   C. Sample usage
3. You should submit a single compressed file for the session 1. On the biokdd server, do as following.
   A. Go to your 'L519FALL2005' directory.
   B. >tar –zcvf YourNetworkID_HW2.tgz ./HW2 (Suppose HW2 is your subdirectory)
4. Please enjoy learning and practicing new things.
5. There is no group project for HW2.

----------Section 1 ----------

For section 1, you are required to write Perl scripts to do the following tasks. (2 Scripts)

- Note: Sequence file should be in **FASTA** format. Please refer to the following site for further information on FASTA format; (Reference 1, Reference 2)

1. **Read the following introduction to Entropy and answer TWO questions, which should be included in README file. You also have to write a Perl script that calculates Entropy.**

   Entropy (information theoretic) is really a means trying to quantify information using some kind of "currency", usually bits. The rarer (or equivalently more interesting) a thing is, the more bits it's worth. The converse is true as well-the more common a thing, the fewer bits it's worth. Given a set of these things, we can give some ***average*** worth in terms of bits. This average is what ***entropy*** is. Another way to think about entropy is as a single number that describes a probability distribution, but this is less exciting. We will discuss this concept more at length during the semester, but we will begin to get a better feel for its wide use by applying it to sequences. Given an alphabet $\sum = \{x_1, \ldots, x_k\}$ and word $\omega \in \sum^*$, we write $|w|_{xi}$ to mean the number of times x_i occurs in w and $|w|$ to mean the number of symbols in w. The entropy of w, denoted $H(w)$, is then

   $$H(w) = \sum_{i=1}^{k} p_i (\log_2 1/p_i)$$

   where $p_i = |w|_{xi} / |w|$. For example, consider the string $w = \alpha\beta$ over $\sum = \{\alpha,\beta\}$. Then the entropy $H(w) = 1/2 + 1/2 = 1$. From continuity arguments 0log1=0 can be treated as 0.

   (a) The organism *Micrococcus phlei* has the following frequencies in its DNA: p_A = :164; p_C = :337; p_G = :337; p_T = :162. **Find H(DNA$_{Mp}$).**

   (b) Given a string, what is its maximum entropy? Intuitively, we can make every letter as important as every other-so they must all have the same frequency (f). Suppose there are k different symbols. Then, $\sum^{k} (f/kf) * \log_2 \frac{1}{(f/kf)}$. **Find the maximal entropy of DNA?**

(c) Write a program that produces the entropy when supplied a file as an argument containing a single string - presumably containing a DNA, RNA, or protein sequence using single letter encoding, but it doesn't really matter. Here are the particulars

* The program name is 'entropy.pl'

* As a preamble, the program displays the number of different characters. **# char = x**, the possible maximal entropy, **poss. max entropy = y**, and then the actual entropy to **three** decimal places, **entropy = z**. Here's a sample run (assume the file *sample* contains the string Mm

```
> entropy.pl sample
# char = 2
poss. max entropy = 1
entropy = 1
```

* The program should provide reasonable error messages if the appropriate inputs are not supplied or contain errors.

* The program should include a README file.

2. A teammate of yours has terrible habits in the maintenance of his laboratory notebook. After he left the laboratory, no one else can recognize the right DNA sequences from his notebook. Although he has written them in a typical FASTA format, nobody can recognize whether some of the letters he wrote are in fact A, T, C or G. After your careful examination, most of the handwriting can be reconstructed, but three distinct types of symbols remain mystic denoted as X, Y and Z. Each of these symbols apparently represents a nucleotide, however it remains unclear which one they really represent. Fortunately, some of the DNA sequences he obtained have similar (homologues) sequences in the GenBank. Can you write a script that takes use of the database searching tool BLAST to decode the letters X, Y and Z?

The input of your program will be a set of (> 10) sequences in the alphabet (A, C, G, T, X, Y, Z). You have to call BLAST program (either from Web or local server) to search against NCBI nr database, post-process the search results and report the most possible representation of X, Y, Z.

----------**Section 2** ----------

For section 2, you are NOT required to write scripts.

3. Michael Crichton's fantasy about cloning dinosaurs, Jurassic Park, contains a putative dinosaur DNA sequence. A) Identify the real source of the following sequence in the NCBI non redundant nucleotide database, nr.

```
>DinoDNA from JURASSIC PARK  p. 103 nt 1-1200
GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGC
GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCG
TGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGC
TGCTCACGCTGTACCTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTG
CCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA
AGTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAG
ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACT
CCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATGGCGGCCGACGCGCTGGGCT
GGCGTTCGCGACGCGAGGCTGGATGGCCTTCCCCATTATGATTCTTCTCGCTTCCGGCGG
CCCGCGTTGCAGGCCATGCTGTCCAGGCAGGTAGATGACGACCATCAGGGACAGCTTCAA
CGGCTCTTACCAGCCTAACTTCGATCACTGGACCGCTGATCGTCACGGCGATTTATGCCG
CACATGGACGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAA
CAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAA
GCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGG
CTTTCTCAATGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTG
ACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCA
ACACGACTTAACGGGTTGGCATGGATTGTAGGCGCCGCCCTATACCTTGTCTGCCTCCCC
GCGGTGCATGGAGCCGGGCCACCTCGACCTGAATGGAAGCCGGCGGCACCTCGCTAACGG
CCAAGAATTGGAGCCAATCAATTCTTGCGGAGAACTGTGAATGCGCAAACCAACCCTTGG
CCATCGCGTCCGCCATCTCCAGCAGCCGCACGCGGCGCATCTCGGGCAGCGTTGGGTCCT
```

NCBI scientist Mark Boguski noticed this obvious "contaminant" and supplied Crichton with a “better” sequence, shown below, for the sequel, The Lost World. B) Identify the source of this sequence. Did Mark do a good job to find a “better” sequence? Why? C) Mark imbedded his name in the sequence he provided. Identify the sequence MARK imbedded.

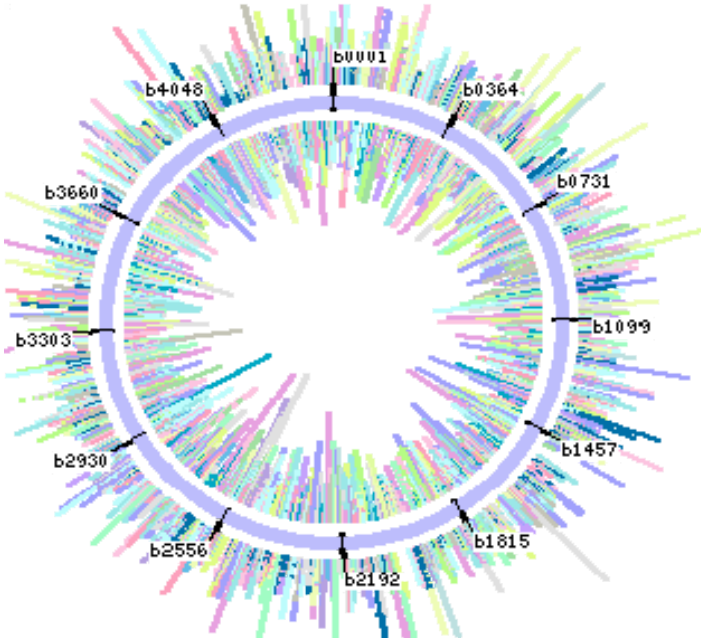
```
>DinoDNA from THE LOST WORLD  p. 135
GAATTCCGGAAGCGAGCAAGAGATAAGTCCTGGCATCAGATACAGTTGGAGATAAGGACG
```

GACGTGTGGCAGCTCCCGCAGAGGATTCACTGGAAGTGCATTACCTATCCCATGGGAGCC
ATGGAGTTCGTGGCGCTGGGGGGGCCGGATGCGGGCTCCCCCACTCCGTTCCCTGATGAA
GCCGGAGCCTTCCTGGGGCTGGGGGGGGGCGAGAGGACGGAGGCGGGGGGGCTGCTGGCC
TCCTACCCCCCCTCAGGCCGCGTGTCCCTGGTGCCGTGGGCAGACACGGGTACTTTGGGG
ACCCCCCAGTGGGTGCCGCCCGCCACCCAAATGGAGCCCCCCCACTACCTGGAGCTGCTG
CAACCCCCCCGGGGCAGCCCCCCCCATCCCTCCTCCGGGCCCCTACTGCCACTCAGCAGC
GGGCCCCCACCCTGCGAGGCCCGTGAGTGCGTCATGGCCAGGAAGAACTGCGGAGCGACG
GCAACGCCGCTGTGGCGCCGGGACGGCACCGGGCATTACCTGTGCAACTGGGCCTCAGCC
TGCGGGCTCTACCACCGCCTCAACGGCCAGAACCGCCCGCTCATCCGCCCCAAAAAGCGC
CTGCTGGTGAGTAAGCGCGCAGGCACAGTGTGCAGCCACGAGCGTGAAAACTGCCAGACA
TCCACCACCACTCTGTGGCGTCGCAGCCCCATGGGGGACCCCGTCTGCAACAACATTCAC
GCCTGCGGCCTCTACTACAAACTGCACCAAGTGAACCGCCCCCTCACGATGCGCAAAGAC
GGAATCCAAACCCGAAACCGCAAAGTTTCCTCCAAGGGTAAAAAGCGGCGCCCCCCGGGG
GGGGGAAACCCCTCCGCCACCGCGGGAGGGGGCGCTCCTATGGGGGGAGGGGGGGACCCC
TCTATGCCCCCCCCGCCGCCCCCCCCCGGCCGCCGCCCCCCCTCAAAGCGACGCTCTGTAC
GCTCTCGGCCCCGTGGTCCTTTCGGGCCATTTTCTGCCCTTTGGAAACTCCGGAGGGTTT
TTTGGGGGGGGGGCGGGGGGTTACACGGCCCCCCCGGGGCTGAGCCCGCAGATTTAAATA
ATAACTCTGACGTGGGCAAGTGGGCCTTGCTGAGAAGACAGTGTAACATAATAATTTGCA
CCTCGGCAATTGCAGAGGGTCGATCTCCACTTTGGACACAACAGGGCTACTCGGTAGGAC
CAGATAAGCACTTTGCTCCCTGGACTGAAAAAGAAAGGATTTATCTGTTTGCTTCTTGCT
GACAAATCCCTGTGAAAGGTAAAAGTCGGACACAGCAATCGATTATTTCTCGCCTGTGTG
AAATTACTGTGAATATTGTAAATATATATATATATATATATATATCTGTATAGAACAGCC
TCGGAGGCGGCATGGACCCAGCGTAGATCATGCTGGATTTGTACTGCCGGAATTC

4. In biology, the term **genome** of an organism contains the whole hereditary information, which was first coined, in 1920, by Hans Winkler, Professor of Botany at the University of Hamburg. More precisely, the genome of an organism is a complete DNA sequence of one set of **chromosomes**.

In contrast to the linear chromosome structures of the eukaryotic cells, many bacterial strains have single, covalently closed, **circular** chromosomes. The figure 1 illustrates an example of circular chromosome, Escherichia Coli K12. Most bacterial plasmid genomes were also shown to be circular. Some bacteria may have even multiple circular chromosomes, whereas some other bacteria have linear chromosomes and linear plasmids. (Plasmid: a usually circular, double-stranded unit of DNA that replicates within a cell independently of the chromosomal DNA. Plasmids are often found in bacteria and are very useful in recombinant DNA technology to be the vehicle of transferring genes between cells).

Describe an algorithm to align the two circular DNA sequences, e.g. bacterial circular genomes.



This image is obtained from

http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=115

5. It is often desirable to quickly estimate the size of a genome before performing a whole genome sequencing project. Since a significant portion of eukaryotic genomes is composed of repetitive sequences, it is also of great interests to estimate percentage of the repetitive sequences within a genome. One way to accurately estimate the genome size is to perform a mere (low coverage) shotgun sequencing of the target genome. *Suppose we have obtained shotgun reads from a genome with a low coverage. From these sequences, describe a computational method to a) estimate the size of the genome and b) estimate the percentage of repetitive sequence in the genome.*