I529: Bioinformatics in Molecular Biology and Genetics: Practical Applications **(3CR)**

HW1 (Due: **Jan. 29 BEFORE** Lab session)

http://darwin.informatics.indiana.edu/col/courses/I529-16

INTRODUCTION:

There are three sessions to be completed. The section 1 is for programming using Python or C/C++, the section 2 consists of problems related to computational methods and algorithms and the section 3 is for the group project to be completed by all members in each group. Please submit your completed homework (all sessions, for 1st and 3rd sections, source code should be included along with a report) on the Oncourse. Pdf files are encouraged for session 2; handwritten document scanned in pdf files are accepted, but not preferred for session 2. Each group may submit only one copy of the answer (source code along with a report) for section 3 by one of the group members, and **in the report the responsibility of each group member should be briefly described**.

QUESTION:

Don't hesitate to contact me (Haixu Tang: hatang@indiana.edu).

INSTRUCTION:

1. Please start to work on 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 program
   B. Input / Output
   C. Sample usage
3. You should submit a single compressed file for the session 1.

WARNINGS: **YOU ARE SUPPOSED TO WORK IN GROUP FOR THE MINI CLASS PROJECT. HOWEVER, YOU MUST DO HOMEWORK SESSION 1 AND 2 ON YOUR OWN.**

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

For section 1, you are required to write a Python or C/C++ program to do the following tasks.

- Note: Sequence file should be in **FASTA** format. Please refer to the following site for further information on FASTA format; (Reference 1, Reference 2), **40 points.**

Many applications in this course require generating many sequences with the same length and residue frequencies as a given input DNA or protein sequence. There are in general two ways of achieving this: (1) random sequence generation, in which one of the four residues (A, C, G or T) is selected randomly based on the pre-calculated frequencies of all residues as the input sequences for each position in the output sequence; (2) random sequence permutation, randomly permute the input sequence. Implement these two methods.

Results:

(1) two programs: RANSEQ1 and RANSEQ2 running in the same syntax
RANSEQ1 –i inputfiles –n N –o outputfiles
RANSEQ2 –i inputfiles –n N –o outputfiles

Inputfile stands for the name of input file, which should contain one DNA sequence in FASTA file format; the program should be able to report an error message if the input file is in the wrong format. N stands for the number of random sequences to be generated. Outputfile stands for the name of output file, which should contain N DNA sequences in FASTA file format, with the same residue frequencies.

(2) Benchmark and report the performance of your program. Submit your report in a word document.
   a. Generate three DNA sequences with length of 10, 100 and 1000, respectively.
   b. Run both of your programs on these three input sequences, and generate 10 output sequences in each case;
   c. Compute the residue frequencies of the input sequence and of each output sequences;
   d. Compute the mean and standard deviation of the residue frequencies for output sequences in each case;
   e. Report these results and discuss which method is better.

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

For section 2, you are NOT required to write scripts. **30 points**

1. A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good: it is 100% sensitive (it is always correct if you have the disease) and 99.99% specific (it gives a false positive result only 0.01% time). (1) Do you want to take the test? Why? (2) If you are forced to screen it, how many times of tests you should take to get a confident result, assuming each time of test provides an independent result?
2. Suppose that weather has three states: rain, sunny and cloudy. Tomorrow's weather depends on the weather in last two days. If it is sunny for the past two days, it will be sunny tomorrow with probability 0.7 and be cloudy with probability 0.2. If it rains for the past two days, it will rain tomorrow with probability 0.5 and be cloudy 0.3. In all other cases, the weather tomorrow will be same as today with probability 0.6 and will be the remaining two states with probability 0.2 each. Build a Markov chain for this weather forecast model. If it rains on Jan 1 and Jan 2, 2007, what is the probability of raining on Jan 4? And raining on Jan 8?

----------**Section 3** Mini Group Project # 1 ----------

Mini group project #1 is sequential to the HW Section 1. **30 points**

- GOAL
  - Build a simple codon-usage based gene finder for finding genes in E.coli (in Python or C/C++)
  - Extra points

- Procedure (hints)
  - Collect 1000 gene sequences from E. coli;
  - Compute the codon usage based on these genes (and the translated protein sequences from them);
  - Build a probabilistic model based on the codon usages;
  - For a given DNA sequence (and one selected reading frame), compare your model with a random sequence model;

- Result
  - Two FASTA files for the collected 100 genes and 100 translated protein sequences;
  - The printed codon usage table;
  - A program named ECgnfinder, running with the syntax as
    ECgnfinder –i inputfile
  - Inputfile stands for the name of input file, which should contain one DNA sequence in FASTA file format; the program should be able to report an error message if the input file is in the wrong format.
  - The output should be printed to the standard output as (xxx stands for the likelihood)
    ORF1: xxx
    ORF2: xxx
    ……
  - Each group needs to submit only one set of results in the report.