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

HW2 (Due: **Feb. 19 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 Python scripts 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), **25 points.**

In last mini group assignment, we built a probabilistic model of gene finding based on the codon usages. This time we want to build a gene finding model based on the 1st order Markov chain of codons. You can utilize parts of the codes from the last group project in this assignment.

- Procedure (hints)
  - Collect 1000 E. coli gene sequences as the training set;
  - Collect another 500 E. Coli gene sequences, and 500 non-coding sequences as test sets; (Note: the sequences of the gene and non-coding sequences should be in similar length).
  - Build the 1st order Markov chain of codons using the training set; the program (ECgnfinder_mc) should take the same kinds of format for input and output as ECgnfinder from the last assignment;
  - Evaluate the performance of ECgnfinder on the test, and compare it with the performance of ECgnfinder; if the performance is different, try to explain.

- Result
  - The program ECgnfinder_mc, including the source code and a short readme file.
  - Two FASTA files for the collected 1000 genes as training set and 500 genes as test set;
  - A report on the performance evaluation.

--**Section 2** --

For section 2, you are NOT required to write programs. **45 points**

1. Given 10 DNA Segments of the same length L =10 (DNA source: H-NS, Histone like, nucleoid-associated DNA-binding protein),

   CGCCTGAATA
   CGAGAAAGTT
   CGCCGGAATT
   GGCATGAATA
   TAAAGGAATC
   TAATTTAATT
   CAATTAAATT
   GACATGAATC
   TGGCTAATTT
   CAACTGAATT

   answer the following questions:

   a) Building a Position-Specific Scoring Matrix (PSSM), Θ_1;

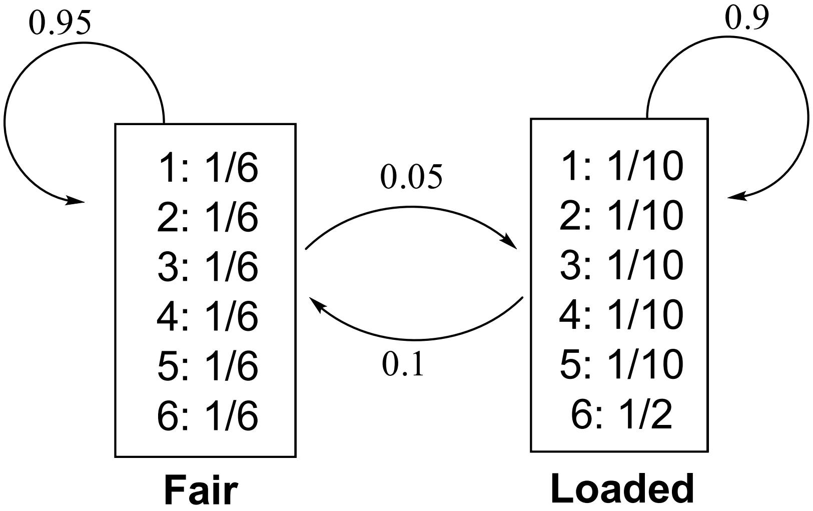
   b) Building a PSSM, Θ_2 , incorporating prior probability;

   c) Compute the relative entropy H for both models;

   d) Given another sequence S_0, CAAATTATTT, compare two models Θ_1 and Θ_2.

2. Devise a hidden Markov model for the prediction of protein secondary structure using Q3 representation. Explain the five elements of your HMM.

3. In a casino they use a fair die most of the time, but occasionally they switch to a loaded die. The switch between dice is a Markov process shown below:



Compute the most likely sequence of the dices that were used in 6 consecutive experiments, if 6 consecutive ‘6’ (“6, 6, 6, 6, 6, 6”) were observed.

---------------------------------- Mini Group Project # 2 ---------------------------------------

Mini group project # 2 should be completed by each group. **30 points**

Membrane proteins compromise a large fraction of eukaryotic proteins, and carry out many important protein functions as ion transporter, signal transduction and cell-cell recognition. Membrane proteins consist of *transmembrane domains* that can attach to the cellular membranes. The protein sequences for the transmembrane domains are enriched with hydrophobic amino acids, and shows a different amino acid patterns as the other kinds of inter-cellular globular proteins. In this project, we want to build a prediction model to identify transmembrane domains in a given protein sequence. Note that the input protein sequence may contain no transmembrane domain if it is not a membrane protein.

- GOAL
  - Download a number of membrane protein sequences and their annotations as the training set and testing set from the given database at /tmp/TM/TMseq.ffa (you can partition the whole dataset into training and testing data);
  - Build a GHMM for transmembrane domain prediction;
  - The results will be presented by each group at the lab section on 2/21.
  - Extra points: compare the performance of your program with TMHMM.

- Result
  - A program named ProdictMP_ghmm, running with the syntax as
    ProdictMP_ghmm –i inputfile –o outputfile
  - Inputfile stands for the name of input sequence file, in FASTA file format; the program should be able to report an error message if the input file is in the wrong format.
  - Each group needs to submit only one set of results, and present them on 2/19.