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

HW4 (Due: **April. 8, before lab session**)

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

INTRODUCTION:

There is only one session in this HW, which consists of problems related to computational methods and algorithms. Please submit your completed homework on the Oncourse. Pdf files are encouraged; handwritten document scanned in pdf files are accepted, but not preferred.

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. **Please ENJOY learning and practicing new things.**

WARNINGS: **YOU MUST DO OTHER PARTS OF HOMEWORK ON YOUR OWN.**

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

You are NOT required to write programs. **100 points**

1. Hidden Markov models can be trained in a progressive fashion. An initial dataset A can be used to build a model θ_A. When a second training set B is available, B can be combined with A to build a new model θ_{AB}. However, this requires to keep all training data during the process. Another approach, which assumes dataset A is not available any more when B is ready, is to improve θ_A to a new model θ'_{AB} by using the data set B alone. This process will continue when more training data becomes available. 1) Devise a Bayesian approach to achieve this goal; 2) Argue if this approach will result in the same model θ_{AB}.

2. Consider the following multiple alignment of five DNA sequences:

```
1          C A - - - G
2          C A C C - G
3          C A C A A T
4          - A C A A G
5          G A C - - G
```

   1) Derive a profile HMM (by depicting its model architecture and the model parameters), assuming the columns 1, 2, 3 and 6 as matching states;
   2) Compute the most likely hidden state sequence generating the observation sequence CACT, given the profile HMM derived in 1).

3. Profile HMMs can model protein sequences from the same family. Some multi-domain protein families, however, may consist of several protein domains with shuffled orders across different proteins (see [1] for a comprehensive introduction). How can you generalize the regular profile HMM to incorporate these domain shuffling events?

[1] Copley RR, Ponting CP, Schultz J, Bork P. Sequence analysis of multidomain proteins: past perspectives and future directions. Adv Protein Chem. 2002;61:75-98.

4. We used a shotgun strategy to sequence an unknown DNA molecule. Assume we obtained reads (fragments) with coverage 10, i.e. in average, each nucleotide in the target DNA is covered by 10 different reads (as the following figure). Suppose the sequencing errors distribution is dependent on the real nucleotides, i.e. the distribution P(X|Y) for different nucleotide Y (=A, C, G, or T) in the unknown target DNA may be different. Explain how to find the most likely (unknown) target DNA sequence as well as the error distribution using an EM algorithm (Note: you can assume the error rate, i.e. P(X|Y) is very low, <<1/10)

```
…ACCCCTGCCCGTCCCCTGG… target DNA (unknown)
```

```
…ACTCCTGCCCGTCCCCTGG… reads
…ACCCCTGCCCGTCCCCTGG…
…ACCCCTGCCCGACCACTGG…
…ACCGCTGCCCGTCCCCTGG…
…ACCCCTGCCCGTCCCCTGG…
…ACCCCTGCCCGTCGCCTGC…
…ACCCCTGCGCGTCCTCTGG…
…ACCCCTGCCCGTCCCCTGG…
…ACCCCTGCCCGTCCCCTGA…
…ACCCCTGCCCGTCCCCTAG…
```

5. Protein-protein interaction network can be experimentally determined by different techniques, e.g. the two-hybrid systems, tandem affinity purification (TAP) coupled to mass spectrometry, etc. However, these techniques may have different false positive P(-|+), and false negative rate P(+|-). Assume we have applied *m* distinct techniques to elucidate the interactions between *N* proteins in an organism. Each technique may give a different result for the protein-protein interactions. Describe an EM algorithm that can predict the consensus protein-protein interactions as well as false positive and false negative rate for each technique.