EECS 458 – Homework 3

1. (1pts) Assume a sequence X=*a*1*a*2*a*3…*a*n (where *a*i is from { A,T, G, C } for 1<=i<=n) is generated independently at each position, using maximum likelihood method to estimate the frequencies of A, T, G and C of the underlying model.
2. (1pts) Given the Markov model below and assume that the transition from any state (A,T, G, C) to the end state (e) has the same probability τ. Show that the sum of the probabilities



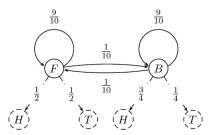
over all sequences of length *m* is τ(1-τ)*m*-1. (Durbin et al. Page 50, exercise 3.2.)

1. (1pts) Show that argmaxπP(π|x) and argmaxπP(x, π) are equivalent. (Notice the difference of P(π|x) and P(πi|x).) And why we are not interested in argmaxπP(x|π), i.e. why the path defined by argmaxπP(x|π) is not a good choice for state decoding? (Durbin et al. Page 57, exercise 3.4.)
2. (1pts) Derive the expected account of letter *b* emitted by state *s* from *k* sequences, i.e.,



(Durbin et al. Page 66, exercise 3.6.)

1. (1pts) Baum-Welch algorithm is an iterative algorithm. Show the time complexity of the algorithm in one iteration. Assume we have *n* sequences with same length *L*, the number of states is *s*.
2. (4pts)



The Hidden Markov Model (HMM) above is used to describe the dishonest casino mentioned in class. The dealer has two coins “**F**air” and “**B**iased”. Each coin has a “**T**ail” and a “**H**ead”, but with different distributions. The dealer can switch the two coins using the specified probabilities without being noticed by anyone. The first time the dealer chooses one of the coins with probability 0.5 each. Now 1) decode the following sequence of coin tosses **HHHHHTTTTT** using Viterbi algorithm. 2) What is the posteriori probability that the "T" at the seventh position is generated by the biased coin?

For both questions, your answer should including the necessary dynamic programming tables along with your answer to the questions themselves. For computer science background students, you should write a small but general program (you program can decode any input sequence and give posteriori probability for any position) to obtain the answers. For students without CS backgrounds, you can either write a program if you feel comfortable to do so, or you can actually use excel sheet to do the calculation for this specific example.

1. (2pts) For a pair of siblings i and j in a family, at each SNP position, they may share 0, 1 or both alleles (sharing means inherited the same allele from the same parents, termed as identical by decent, or IBD). Given genotype data of a chromosome for both individuals, how can you use a HMM to infer the IBD status at each position? What are the states? How to model transition probabilities and emission probabilities? For this question, it is not necessary to define a precise model, but you should clearly states what is your model structure (hidden states, observations), and what factors you should consider when you define your parameters (transition probabilities and emission probabilities).