Algorithms in Computational Biology 236522, Spring 2017

Home Assignment #2 – HMM and EM

General instructions

- Last submission date 31/05/2017 23:55
- Submission in pairs

1. (30pts) Recently you won some cash in lotto and now considering its investment in CS Inc. You know, that once (4 months ago) this company was good and since then its stock is increasing each month. After some research you learn that probability of stock increase (on monthly scale) is 60% when company is in good condition and 20% in bad condition. Also you learn that company usually sticks to its previous condition with 90%.
   a. (10pts) Build and explain the model describing the data.
   b. (10pts) What is the probability of the stock increase in the following (#5) month?
   c. (10pts) You’ve found a report indicating that 2 months ago (in month #2) the condition of company was bad. How does it affect your calculations in (1a)? What is the new probability of stock increase?

2. (30pts) Yeast organism can be either in haploid or diploid state, while the probability of state change in an hour interval is 10%. Once in haploid state, the gene haplo is expressed 80% of time, while in diploid state it is expressed only 30% of time.
   We assume that our yeast is initially in a diploid state.
   a. (10pts) Provide an HMM model that describes the haploid-diploid state change based on the expression of haplo gene.
   b. (10pts) What is the probability that the yeast is in a diploid state after 3 hours, given that that haplo gene is continuously expressed for the entire 3-hour period?
   c. (10pts) What is the sequence of most probable haploid-diploid states, given that that haplo gene is continuously expressed for the entire 3-hour period?
3. (40pts) Evolution of the following 3 species A,B,C is described by:

In this question we will relate to 2 independent sites. Each site has two values: The value 0 is A or T, and the value 1 is C or G.

We will assume that P(X=0)=0.5 in the two sites. The unknown parameter $\theta$ is the probability to have a switch (mutation) between father and son, meaning for X, and its derived Y:

$\theta = P(Y=1|X=0) = P(Y=0|X=1)$

Performing measurements on the two independent sites, the following values were found:

First site: A=1, B=0, C=1
Second site: A=0, B=1, C=0

a. (8pts) Write a formula as a function of $\theta$ for the following probability: $P(A=1,B=0,C=1|\theta)$
b. (8pts) Write a formula as a function of $\theta$ for the following probability: $P(A=0,B=1,C=0|\theta)$
c. (8pts) Write a formula as a function of $\theta$ for the following probability:
   $P(X=0,Y=1|A=0,B=1,C=0,\theta)$
d. (8pts) Write the EM functions for finding $\theta$
e. (8pts) Find the value of $\theta$ that maximizes the probability of all the data. Hint: can be done also without EM algorithm