Computational Genomics
© Benny Chor and Amos Tanay
School of Computer Science
Tel Aviv University
January 5, 2005

Computational Genomics: Assignment No. 4 (last, but not least)
due on Jan 20th, 2005

Warm up - maximum likelihood:

1. (10 points) We are given a k-sided die, and we toss it n times. Outcomes of different tosses are independent. We are told that out of n tosses, n_1 have result 1, n_2 have result 2, ..., n_k have result k. What are the maximal likelihood estimates of the die probabilities $p_1, p_2, \ldots, p_k$? (p_i is the probability of getting result i, and clearly $p_1 + p_2 + \ldots + p_k = 1$)? Prove your answer.

Hidden Markov models:

2. (15 points) Generalize the EM algorithm from HMMs to the pair-HMM for global alignment defined in class (Durbin chapter 4 if you missed the recitation). Assume that the forward and backward probabilities are given - how should we update the parameters $\epsilon, \delta, \tau, P_{xy}, P_x, P_y$. Would your answer be different if we require $P_x = P_y$?

3. (25 points) You are interested in modeling genes of a certain family and thought of using an HMM to do it. Your genes have two exons and an intermediate intron. Exons have a typical codon distribution and introns are expected to behave like the background, consisting of random nucleotides along some genomic distribution.

   a. (10 points) Suggest an HMM based model that fits the biological scenario

   b. (15 points) Can you do better if you know that introns lengths are distributed according to a known normal distribution? In other words, can you approximate better the biological scenario using a refined model?

4. (10 points) Given an HMM with parameters θ and an L-long observation $\mathbf{O} = O_1 \ldots O_L$, let $\xi_t(i,j) = Pr(q_t = S_i, q_{t+1} = S_j \mid \mathbf{O}, \theta)$. In words, $\xi_t(i,j)$ is the probability that the sequence of states was at state i in time t and then switched to state j in time $t+1$. Prove that

$$\xi_t(i,j) = \frac{\alpha_t(i) \cdot a_{i,j} \cdot b_i(O_t) \cdot \beta_{t+1}(j)}{Pr(\mathbf{O} \mid \theta)} \ ,$$

where $a_{i,j}$ is the transition probability from state i to state j, $b_i(O_t)$ is the emission probability of the letter O_t at state i, and $\alpha_t(i), \beta_{t+1}(j)$ are the forward and backward probabilities (as defined in class), correspondingly.

Approximation algorithm for TSP:

5. (25 points) We are given a complete graph G with edge weights that reflect distances. The goal is to find the shortest tour visiting all nodes. Recall that in an *Eulerian* graph all node degrees are even and that in such a graph it is easy to construct an Eulerian path (which is also a tour).

   a. (10 points) Assuming the triangle equality holds, show that the minimum spanning tree (MST) for G can be used to construct a 2-approximation to the TSP. Note that computing MSTs is an extensively studied polynomial problem.

   b. (15 points) Another well explored polynomial problem on graphs is minimum cost maximum matching (a matching is a set of disjoint edges). Use the minimum cost maximal matching on all odd degree nodes of the MST to construct a 3/2 approximation to the TSP problem. (Hint: what is the relation between the cost of the matching and the optimal TSP solution?)

   In this question you are not asked to specify the details of the graph algorithms you are using as black boxes. Nevertheless, if you don't know how to solve MSTs or minimum cost maximal matching - you should definitely consider reading about it (a comprehensive a very practical account is "Network flows" by Ahuja, Magnanti and Orlin, but there are many others).

Radiation Hybrid Mapping:

6. (30 points) The first step in ordering markers using data gathered from radiation hybrids experiments is to estimate the breakage probability between every pair of markers. Let a and b be two markers, and let θ denote their breakage probability. That is, θ is the probability that at least one radiation induced breakpoint occurs between a and b. In this question you will compute two different estimations for θ.

   Let n^{--} denote the number of hybrids that contain both markers, and define n^{-+}, n^{+-}, n^{++} similarly. In class we showed that the probability of getting the outcome $\langle n^{--}, n^{-+}, n^{+-}, n^{++} \rangle$ given that the breakage probability is θ is,

   $$[q(1-p\theta)]^{n^{--}}[pq\theta]^{n^{-+}+n^{+-}}[p(1-q\theta]^{n^{++}} \tag{1}$$

   where p is the retention probability, and $q = 1 - p$.

a. (5 points) Use the maximum likelihood method to find $\hat{\theta}$, the value of θ that maximizes (1).

b. (10 points) Assume that the values of $n^{--}, n^{-+}, n^{+-}, n^{++}$ equal exactly their expected values (e.g., $n^{-+} = mpq\theta$).

Compute the value of your estimator, $\hat{\theta}$, under this assumption.

Another (simpler) approach for estimating the value of θ is the following. Denote by $n^{=}$ the number of hybrids in which either both markers appear or both are absent (i.e., $n^{=} = n^{--} + n^{++}$). Similarly, denote by $n^{\neq}$ the number of hybrids in which exactly one of the markers appears (i.e., $n^{\neq} = n^{-+} + n^{+-}$).

c. (5 points) Compute the probability for the $\langle n^{=}, n^{\neq} \rangle$ given that the breakage probability is θ.

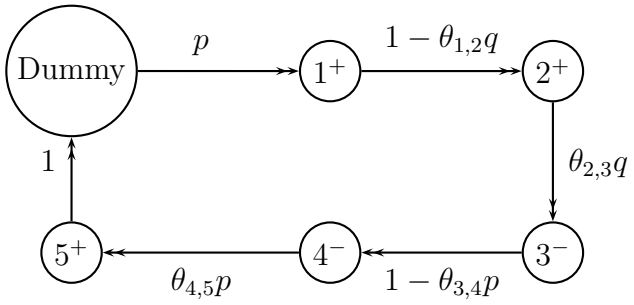
d. (5 points) Use the maximum likelihood method to find $\bar{\theta}$, the value of θ that maximizes the expression you found in **c**. Notice that $\hat{\theta} \neq \bar{\theta}$.

e. (5 points) As in **b**, assume now that the values of $\langle n^{=}, n^{\neq} \rangle$ equal exactly their expected values.

Compute the value of $\bar{\theta}$ under this assumption.

7. (15 points) In this problem we want to explore how edge weights in the "likelihood TSP" can be made symmetric, so that the "standard" TSP heuristics can be successfully applied to the graph.

a. (5 points) Suppose we have five markers, 1 to 5, and an hybrid where the measurements are $[+ + - - +]$. Suppose the estimate for the separation probability between i and j is $\theta_{i,j}$. The following graph (a subgraph of the complete graph) can be used to find the likelihood of the data, given the arrangement [12345] (by multiplying the probabilities along the directed circuit).



Draw the corresponding graph for the likelihood of the data given the *reverse* arrangement [54321]. Write down the expressions for the likelihood of traversing the graph in these two directions.

b. (10 points) Observe that some of the edges are assigned different values when their direction is reversed, while others do not. Use this observation to derive a method for changing the weight of the edges in the graph in an "undirected way" (weight(i,j)=weight(i,j)) , such that the weight of any traveling salesman tour remains invariant (same as its two directed counterparts).

8. **And the golden star question is:** At what country do members of the Green Party use wild animals' fur in parliament, and why?