

CSE 590BI: Algorithms in Molecular Biology

Assignment #2

January 16, 1996

due: Tuesday, March 5

Choose as many of the following problems as you care to work on, and take each as far as you can. Some are routine; there are some that 1–3 of the instructors don't know how to solve. We don't expect you to tackle all the problems, nor even all the problems that we know how to solve. We will get much more excited about promising partial progress on an open problem than about long solutions to all the routine problems. Keep your answers clear and concise. If you use references, please include citations.

1. A coin with probability of heads p is tossed n times, giving k heads and $n - k$ tails. Assuming that the coin tosses are independent, prove that the maximum likelihood estimator of p is $\frac{k}{n}$.
2. The following approach to Radiation Hybrid Mapping is called *minimization of obligate breaks*. With each pair of STSs one associates a weight equal to the number of observed retention patterns in which exactly one member of the pair is retained. An ordering of the STSs is then sought in which the sum of the weights of adjacent pairs of STSs is minimized. Show that this minimization problem can be expressed as a Traveling-Salesman Problem.
3. In our discussion of Radiation Hybrid Mapping we considered the following problem: given a retention probability p , an ordering π of a set of n STSs and parameters $\theta_1, \theta_2, \dots, \theta_{n-1}$, where θ_i denotes the probability of at least one break between the i th and $i + 1$ st STSs in the ordering, determine the probability of a given retention pattern, presented as an array of $n +$ and $-$ signs. We gave both a left-to-right and a right-to-left method of expressing this probability as a product of n factors. In these expressions the factor involving a consecutive pair of STSs depended on the order of their occurrence. We then gave a “symmetrized” expression in which the factor associated with two consecutive STSs is independent of their order. Prove that the symmetrized expression correctly represents the probability of the retention pattern.
4. This question concerns the lecture by Elizabeth Thompson. Give a precise definition of optimal peeling sequence and prove that the problem of finding an optimal peeling sequence is NP-hard.
5. A Hidden Markov Model is specified as follows:

- A set of states $\{S_1, S_2, \dots, S_N\}$;
- A set of observation symbols $\{V_1, V_2, \dots, V_M\}$;
- A $N \times N$ matrix (a_{ij}) of transition probabilities, where a_{ij} gives the probability that the next state is S_j , given that the present state is S_i ;
- A $N \times M$ array (b_{jk}) of output probabilities, where b_{jk} is the probability of generating output V_k in state S_j ;
- A length- N vector (π_i) of initial probabilities, where π_i is the probability of state S_i at time 1.

For a given output sequence O_1, O_2, \dots, O_T we showed how to compute the function $\alpha_t(i)$, where t ranges from 1 to T , i ranges from 1 to N , and $\alpha_t(i)$ is the probability of generating the output symbols O_1, O_2, \dots, O_t in the first t time steps and being in state S_i at time t .

Let O_1, O_2, \dots, O_T be a sequence of outputs.

- (a) Let $\beta_t(i)$ be the probability of generating outputs O_t, O_{t+1}, \dots, O_T at times $t, T+1, \dots, T$, given that the state at time t is S_i . Give an efficient algorithm for computing $\beta_t(i)$ for $t = 1, 2, \dots, T$ and $i = 1, 2, \dots, N$. Hint: Work backwards.
 - (b) Assuming that the functions $\alpha_t(i)$ and $\beta_t(i)$ have been tabulated, give an efficient method of computing the expected number of transitions from state S_i to state S_j during the first T time steps, given that the output sequence is O_1, O_2, \dots, O_T .
6. The following diagram gives phenotype data at the ABO blood type locus for a pedigree. A labeling is also shown, specifying the flow of alleles. You are given that allele A has frequency .28 in the population, allele B has frequency .06 and allele O has frequency .66. Determine the probability that the given phenotypes would occur, given the labeling.

7. In class we stated that, given the phenotype data for a pedigree at an ordered sequence of loci, the recombination probabilities between successive loci could be estimated by applying the E-M algorithm to a Hidden Markov Model.
 - (a) Give a precise definition of this Hidden Markov Model.
 - (b) How does the number of states in this Hidden Markov Model depend on the size of the pedigree?
 - (c) How could this approach to estimating recombination probabilities be extended to the case where phenotype data is given at the same loci for many small pedigrees rather than one?
8. Construct a PQ-tree representing all the linear orderings of the set $\{A, B, C, D, E, F, G\}$ in which the elements of each of the following sets occur consecutively: $\{A, F, G\}$, $\{C, D, F\}$, $\{B, G\}$, $\{A, C, D, E, F\}$.
9. Given noisy data about the occurrences of STSs on clones, we want to assign a cost to a proposed ordering of the STSs. The data for each clone consists of an array of zeros and ones. The successive positions in the array correspond to the successive STSs in the proposed ordering. A 1 in a given position indicates that the corresponding STS is measured as occurring on the clone, and a 0 indicates that the STS is measured as not occurring on the clone. If the data were error-free then, unless the clone is chimeric, the ones in the vector would occur in a consecutive block. In the presence of error the ones may not be consecutive. In that case, we look for the least costly way to change some of the bits so that the ones in the resulting array are consecutive. Assuming that the cost of changing a 1 to a 0 (i.e., designating the 1 as a false positive) is a positive constant A, and the cost of changing a 0 to a 1 (i.e., designating the 0 as a false negative) is a positive constant B, give an efficient algorithm for finding the least costly set of changes.