

BCB568 Spring 2010 Final examination

May 5, 2010, 9:45AM–11:45AM

Sign your name ____Super Solver ____

No books, notes, or other media.

For full credit, you must succinctly explain your answers.

Problem 1 [10 points]

Problem 2 [4 points]

Problem 3 [8 points]

Problem 4 [8 points]

Problem 5 [10 points]

TOTAL [40 points]

1. The basic structure of a transmembrane-spanning protein can be represented by assignment of the location of each amino acid as located external to the cell, within the membrane, or internal to the cell. From experimental determination it is known that there are distinct compositional biases in the amino acid sequences for each of these locations. Your task is to provide a method that would predict the basic structure of a potential transmembrane-spanning proteins by inspection of its amino acid sequence.
 - (i) [2 pts] Set up a 3-state Hidden Markov Model for the problem where the hidden states correspond to the location possibilities and the outputs correspond to observable amino acids.
 - (ii) [4 pts] Describe in mathematical detail an efficient algorithm to calculate the most probable state sequence for an observed protein sequence regarded as being generated by the Hidden Markov Model, assuming that all parameters of the model you proposed in (i) are known.
 - (iii) [2 pts] Without mathematical detail, give a strategy how you would estimate parameters for your model from experimental data.
 - (iv) [2 pts] How would you change your model to explicitly model a restriction on the basic structure that only allows complete traversals of the membrane? In other words, the model should prohibit any structures in which the location assignments go from external to membrane back to external, or from internal to membrane back to internal; only external to membrane to internal and internal to membrane to external should be allowed.

Solution:

(i) We provide the solution for a general Hidden Markov Model with s states S_1, S_2, \dots, S_s , initial state probabilities π_{S_j} , state transition probabilities $\tau_{S_k S_l}$, and output probabilities $P(O|S)$. Here, $s = 3$ and S_1, S_2 , and S_3 represent the states “external”, “membrane”, and “internal”, respectively. The output in each state is one of the twenty amino acids.

(ii) Define $F_{ij} = \text{Prob}\{O_1, O_2, \dots, O_i, Q_i = S_j\}$. Then

$$\text{Prob}\{O_1 O_2 \dots O_N\} = \sum_{j=1}^s F_{Nj}. \quad (1)$$

We can calculate the F_{ij} recursively as follows:

$$F_{1j} = \text{Prob}\{O_1, Q_1 = S_j\} = P(O_1|S_j)\pi_{S_j}, \quad j = 1, 2, \dots, s \quad (2)$$

$$F_{ij} = P(O_i|S_j) \sum_{k=1}^s F_{i-1,k} \tau_{S_k, S_j}, \quad i = 2, 3, \dots, N, \quad j = 1, 2, \dots, s \quad (3)$$

Define $V_{ij} = \max_{Q_1 \dots Q_{i-1}} \text{Prob}\{O_1, O_2, \dots, O_i, Q_1, Q_2, \dots, Q_{i-1}, Q_i = S_j\}$. Then $V_{1j} = F_{1j}$ for $j = 1, 2, \dots, s$, and

$$V_{ij} = P(O_i|S_j) \max_{k=1, \dots, s} \{V_{i-1,k} \tau_{S_k, S_j}\}, \quad i = 2, 3, \dots, N, \quad j = 1, 2, \dots, s. \quad (4)$$

Further, $\max_{\{Q\}} P(Q|O) = \max_{k=1, \dots, s} \{V_{N,k}\} / \text{Prob}\{O_1 O_2 \dots O_N\}$, where the denominator is calculated efficiently as per equation (1).

(iii) For a collection of proteins with known transmembrane structure, output probabilities can be estimated as observed residue frequencies in the respective locations. Almost all transmembrane-spanning proteins have the N-terminus cell-external, and thus the initiation probabilities should be adjusted to reflect that fact. Transition probabilities could be estimated from the relative lengths of the observed segments, or more thoroughly by expectation maximization (Baum-Welch algorithm).

(iv) We could introduce a 4-state model with states EM, MI, IM, and ME, representing locations "external pre membrane", "membrane pre internal", "internal post membrane", and "membrane pre external", and allow only transitions EM to MI, MI to IM, IM to ME, and ME to EM.

2. Describe concisely the principle of each of following tree-making methods: (1) Distance method; (2) Maximum parsimony method; (3) Maximum likelihood method; and (4) Bayesian method.

Solution:

(1) In distance methods, evolutionary distances are computed for all pairs of taxa, and a phylogenetic tree is constructed by considering the relationships among these distance values. The theoretical basis of this method for the correct inference is the principle of minimum evolution.

(2) Maximum parsimony (MP) methods were originally developed for morphological characters but now widely used in analyzing molecular data. The smallest number of nucleotide (or amino acid) substitutions that explain the entire evolutionary process for the topology is computed. This computation is done for all potential topologies, and the topology that requires the smallest number of substitutions, also called the shortest tree length, is chosen to be the best tree.

(3) Maximum likelihood (ML) methods to infer the phylogenetic tree have the following basic steps. First, given a phylogeny, calculate the likelihood function of each site (a column in the multiple alignment), based on the Markov chain property. Second, the likelihood for the entire sequence alignment is the product of likelihood for all sites, or the log-likelihood of the entire tree is the sum of log-likelihood of all sites. And third, the likelihood of observing an entire sequence alignment under a specific substitution model is maximized for each topology, and the topology that gives the highest maximum likelihood is chosen as the final tree.

(4) Bayesian methods provide a computationally efficient approach to infer the phylogenetic tree, by calculating the posterior distribution of phylogenetic trees, given the multiple alignment of sequence data and the prior probability of phylogenies. Moreover, for each of these possible trees, the calculation has to be integrated over all possible values of the branch lengths of the tree and over the parameters of the model of sequence evolution. This method becomes practically feasible because of the implementation of MCMC algorithm.

3. (i)[2 pts] Let q_t be the probability of identical nucleotide at a site between two homologous sequences with the divergence time t . Let r be the substitution rate from one nucleotide to another. Under the JC (Jukes-Cantor) model, we have

$$q_{t+1} \approx (1 - 2r)q_t + \frac{2}{3}r(1 - q_t)$$

Show how the above equation can be approximated to the following differential equation

$$\frac{dq}{dt} = \frac{2r}{3} - \frac{8r}{3}q_t$$

(hints: consider $q_{t+1} - q_t$)

- (ii) [4 pts] With the initial condition $q = 1$ at $t = 0$, show the solution of above differential equation as follows

$$q = 1/4 + (3/4)e^{-8rt/3}$$

and the JC distance ($d = 2rt$)

$$d = -(3/4) \ln(1 - 4p/3)$$

where $p = 1 - q$ is the proportion of different sites between two DNA sequences.

- (iii) [1 pt] Calculate the JC distance between two DNA sequences when $p = 0$, $p = 0.1$ and $p = 0.75$, respectively.

- (iv) [1 pt] Let d be the evolutionary distance defined by Question 3(ii) and 3(iii). Determine which one of following claims is correct: (a) d is the average number of nucleotide substitutions per site. (b) d is the average number of nucleotide substitutions per gene.

Solution:

- (1) We first note that

$$q_{t+1} - q_t = -\frac{8}{3}r \left(q_t - \frac{1}{4} \right)$$

and then approximate $q_{t+1} - q_t \approx \Delta q / \Delta t$ after assuming one time unit (year or generation) is very small with respect to the between-species divergence time. This leads to be

$$\Delta q / \Delta t \approx -\frac{8}{3}r \left(q_t - \frac{1}{4} \right)$$

The differential equation is obtained by letting $\Delta t \rightarrow 0$ so that $\Delta q / \Delta t \rightarrow dq / dt$.

(2) Note that

$$\frac{dq}{dt} = -\frac{8}{3}r \left(q - \frac{1}{4} \right)$$

Then we have

$$\int \frac{dq}{q - 1/4} = -\frac{8}{3}r \int dt$$

resulting in

$$\ln(q - 1/4) = -\frac{8}{3}rt + c$$

The initial condition allows to determine the constant $c = \ln 3/4$. We therefore have

$$q = 1/4 + (3/4)e^{-8rt/3}$$

Let $p = 1 - q$. One can show

$$p = \frac{3}{4} (1 - e^{-8rt/3})$$

It follows that $d = 2rt$ is given by

$$d = -(3/4) \ln(1 - 4p/3)$$

(3) $d = 0$ when $p = 0$; $d = -0.75 \ln(1 - 0.1/0.75) \approx 0.107$ when $p = 0.1$; and $d = \infty$ when $p = 0.75$.

(4) (a) is correct and (b) is not.

4. (i) [2 pts] For the unrooted phylogeny shown in Fig. Question-4, write down the four-point conditions.

(ii) [5 pts] Using the non-math style, describe the major steps in the algorithm of NJ method to infer the phylogenetic topology. You are only required to describe how to infer the topology, and the detail of branch length estimation is NOT required. You are also allowed to use any hypothetical tree with at least 6 species for illustration.

(iii) [1 pt] What is the relationship between the distance method and the four-point condition?

Solution:

(1) The four-point conditions associated with the given phylogeny are

$$d_{13} + d_{24} < d_{12} + d_{34}$$

and

$$d_{13} + d_{24} < d_{14} + d_{23}$$

where d_{ij} is the evolutionary distance between taxa i and j .

(2) Construction of a tree by the NJ method begins with a star tree, which is produced under the assumption that there is no clustering of taxa (star-tree). The key step is to find a pair of taxa as a neighbor. Since we do not know which pair of taxa are true neighbors, we consider all pairs of taxa as a potential pair of neighbors and compute the sum of branch lengths (S_{ij}) for the i -th and j -th taxa as a neighbor. We then choose taxa i and j that show the smallest S_{ij} value.

Once a pair of neighbors are identified, they are combined into one composite taxon. Suppose that the smallest S_{ij} is determined. We can create a new node (A) that connects taxa i and j . The next step is to compute the distance between the new node (A) and the remaining taxa ($k \neq i, j$), denoted by d_{Ak} . Apparently, the number of taxa for the operation is reduced by one. The procedure of finding a new neighbor is then repeated until the final tree is produced.

(3) If the evolutionary distance is additive strictly, the tree that all four point conditions hold must be the correct one. Since the estimated evolutionary distance may not be strictly additive, many distance methods approximate to find the tree topology with the largest number of valid four-point conditions.

5. (i) [3 pts] List all possible bifurcating trees that depict the possible phylogenetic relationships between the following species: soybean (S), maize (M), grape (G), and rice (R).

(ii) [2 pts] How many different trees (topologies) are there when you add another species (e.g., barley)?

(iii) [3 pts] Given the following multiple sequence alignment, find the most parsimonious phylogenetic tree for the four species (state the total number of required substitutions for the tree):

soybean	G	T	C	T	C	T
maize	A	T	C	C	C	G
grape	G	T	G	T	T	T
rice	A	T	G	C	C	G

(iv) [2 pts] Consider the number of substitutions for each pair of sequence as a distance measure between the species. What phylogeny is supported by a distance-matrix approach to phylogenetic reconstruction? Are the data consistent with an additive tree?

Solution:

(i) For four species there are three topologies, characterized by distinct nearest-neighbors: I = {SM}{GR}, II = {SG}{MR}, and III = {SR}{MG}.

(ii) A fifth species could be added at any of the five branches in a given topology for four species. Because there are three topologies for four species, the total number of topologies for five species is $3 \times 5 = 15$.

(iii) Only positions 1, 3, 4, and 6 are informative (giving different numbers of required substitutions for the different topologies). Looking only at those positions, topology I requires $2 + 1 + 2 + 2 = 7$ substitutions, topology II requires $1 + 2 + 1 + 1 = 5$ substitutions, and topology III requires $2 + 2 + 2 + 2 = 8$ substitutions. Thus, the most parsimonious phylogeny corresponds to topology II which pairs soybean with grape and maize with rice.

(iv) The substitution distances are as follows: $d(SM) = 3$, $d(SG) = 2$, $d(SR) = 4$, $d(MG) = 5$, $d(MR) = 1$, and $d(GR) = 4$. Hence,

$$d(SG) + d(MR) = 3 < d(SM) + d(GR) = 7 < d(SR) + d(MG) = 9,$$

indicating that the distance-based tree is not additive but supports topology II which pairs soybean with grape and maize with rice.