

BCB568 Spring 2010 Midterm Examination II

March 11, 2010, 9:30AM–10:50AM

Sign your name ____Super Solver ____

**You may not use books, notes, or other media,
except (unnecessary) calculators.**

Explain your answers.

None of the answers should take more than a few lines!

Problem 1 [9 points]

Problem 2 [14 points]

Problem 3 [7 points]

TOTAL [30 points]

(b) $S_{00} = 0; S_{0j} = w(j); S_{i0} = w(i)$

(c) The number of additions is seen to be $\sum_{i=1}^M \sum_{j=1}^N [1 + 2(j-1) + 2(i-1)] = MN + 2M \frac{N(N+1)}{2} - 2MN + 2N \frac{M(M+1)}{2} - 2MN = MN^2 + NM^2 - MN$. Thus, for $M=N$, the algorithm is of $O(N^3)$.

2. For a 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 $P(O|S)$:
- (a) [6 points] Show how to efficiently calculate $\text{Prob}\{O_1 O_2 \dots O_N\}$.
- (b) [4 points] Determine the exact number of addition and multiplication operations required.
- (c) [4 points] Explain how you could equally efficiently determine $\max_{\{Q\}} P(Q|O)$.

Solution:

(a) 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:

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

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

(b) Collecting terms from equations (1), (2), and (3), the number of required operations are seen to be

$$\text{Additions:} \quad (s-1) + 0 + (N-1)s(s-1) \quad (4)$$

$$\text{Multiplications:} \quad 0 + s + (N-1)s(s+1). \quad (5)$$

Adding up, we see that, apart from constant terms, $2s^2N$ operations are required.

(c) 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

$$V_{ij} = P(O_i|S_j) \max_{k=1, \dots, s} \{V_{i-1,k} \tau_{S_k, S_j}\}, \quad i = 2, 3, \dots, N, \quad j = 1, 2, \dots, s. \quad (6)$$

Further, $\max_{\{Q\}} P(Q|O) = \frac{\max_{k=1, \dots, s} \{V_{N,k}\}}{\text{Prob}\{O_1 O_2 \dots O_N\}}$.

3. Let $X = X_1 X_2 \dots X_N$ be a sequence of scores derived from independently identically distributed random variables X_i for which $\Pr\{X_i = s_j\} = p_j$, $j = 1, 2, \dots, r$ with the restrictions $\text{Prob}\{X_i > 0\} > 0$ and $E[X] = \sum_{k=1}^r p_k s_k < 0$. The maximal segmental score S in X approximately follows an extreme value distribution such that

$$\Pr\{S > \frac{\ln N}{\lambda} + x\} = 1 - e^{-Ke^{-\lambda x}},$$

where λ is the unique positive root of $E[e^{\lambda X_i}] = 1$ and K is a function of λs_i .

(i) [2 points] Describe how to graphically determine S and the corresponding segment.

(ii) [1 point] Determine x_c such that $\Pr\{S > \frac{\ln N}{\lambda} + x_c\} = p$.

(iii) [2 points] For a scoring scheme $t_i = \rho s_i$, determine the equivalent offset x_c^* giving the same probability p , i.e. find x_c^* such that $\Pr\{T > \frac{\ln N}{\lambda'} + x_c^*\} = p$ where T is the maximal segmental score and λ' is the parameter for the t_i scoring scheme. Use the result to explain why λ can be interpreted as a scale factor.

(iv) [2 points] The threshold value S_p for the maximal segmental score to be significant at the p -level is $S_p = \frac{\ln N}{\lambda} + x_c$ with x_c determined as in (ii). A common choice for λ is $\lambda = \frac{\ln 2}{2}$. In this case, determine the p -level threshold S'_p when considering a sequence of length $N' = 2N$. Briefly discuss the consequences for statistical evaluations of matches against growing databases of sequences.

Solution:

(i) The maximal segmental score corresponds to the highest peak in the excursion plot of E_k versus k , where $E_0 = 0$ and $E_k = \max\{E_{k-1} + X_k, 0\}$, and the coordinates of the maximal scoring segment are from the beginning of the excursion (first positive scoring position of the excursion) to the position where the peak is achieved.

(ii) We look for the solution of $1 - e^{-Ke^{-\lambda x_c}} = p$, which after a little bit of manipulation is seen to be

$$x_c = \frac{\ln K - \ln \left[\ln \frac{1}{1-p} \right]}{\lambda}.$$

(iii) By definition, λ is the unique positive root of $E[e^{\lambda X_i}] = 1$. In the t_i scoring scheme, all the X_i are multiplied by ρ , and thus λ' is seen to be $\frac{\lambda}{\rho}$. As $T = \rho S$, it is clear that the solution is $x_c^* = \rho x_c$. Thus, multiplying all scores by ρ does not change the probability assignments. Equivalently, we could select a particular λ' value and scale given scores s_i by the appropriate ρ factor.

(iv) The centering value $\frac{\ln N}{\lambda}$ becomes $\frac{\ln N'}{\lambda} = \frac{2 \ln 2N}{\ln 2} = \frac{\ln N}{\lambda} + 2$. Thus, $S'_p = S_p + 2$. In the context of a database search, as the database size doubles, the significant score threshold increases by 2. Note that this merely refers to statistical significance, as the underlying biological relatedness of sequences does not change.