

# BCB/GDCB/STAT/COM S 568 Spring 2010

## Solutions to Homework 5

February, 2010

### Problem 1:

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  $\Pr\{X_i > 0\} > 0$  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

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

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

(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) [1 point] 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  $\lambda'$  is the parameter for the  $t_i$  scoring scheme..

(iv) [2 points] Explain why  $\lambda$  can be interpreted as a scale factor.

(v) [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). For  $\lambda = \frac{\ln 2}{2}$ , determine the  $p$ -level threshold  $S'_p$  when considering a sequence of length  $N' = 2N$ .

(vi) [2 points] Altschul (1998; Proteins 32:88) defines a normalized score as

$$S' = \frac{\lambda S - \ln K}{\ln 2}.$$

Making use of the result that the number of separate high-scoring segments, i.e. segments with scores exceeding  $\frac{\ln N}{\lambda} + x$ , is closely approximated by a Poisson distribution with parameter  $K \exp\{-\lambda x\}$ , prove his assertion that the expected number of distinct segment pairs with normalized score greater than or equal to  $y$  is well approximated by the formula

$$E(S' \geq y) \sim \frac{N}{2^y}.$$

**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,  $\lambda$  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  $\rho$ , and thus  $\lambda'$  is seen to be  $\frac{\lambda}{\rho}$ . As  $T = \rho S$ , it is clear that the solution is  $x_c^* = \rho x_c$ .

(iv) As  $K$  is a function of the  $\lambda s_i$  and by result (iii), we can multiply the  $s_i$  scores by a factor  $\rho$ , and all we would need to change in the formulae is to replace  $\lambda$  by  $\lambda' = \frac{\lambda}{\rho}$ . Equivalently, we could select a particular  $\lambda'$  value and scale given scores  $s_i$  by the appropriate  $\rho$  factor.

(v) The centering value  $\frac{\ln N}{\lambda}$  becomes  $\frac{\ln N'}{\lambda} = \frac{2 \ln 2N}{\ln 2} = \frac{\ln N}{\lambda} + 2$ . Thus,  $S'_p = S_p + 2$ .

(vi) Subtract  $\frac{\ln K}{\lambda}$  from both sides of the inequality  $S > \frac{\ln N}{\lambda} + x$  and multiply by  $\frac{\lambda}{\ln 2}$  to get

$$\frac{\lambda S - \ln K}{\ln 2} > \frac{\lambda}{\ln 2} \left[ \frac{\ln N}{\lambda} + x \right] - \frac{\ln K}{\ln 2},$$

or  $S' > y$  where  $y = \frac{\ln N - \ln K}{\ln 2} + \frac{\lambda}{\ln 2} x$ . Solving for  $x$  gives  $x = \frac{1}{\lambda} [y \ln 2 - \ln N + \ln K]$ . Inserting into  $\exp\{-\lambda x + \ln K\}$  gives  $\exp\{-y \ln 2 + \ln N\} = \frac{N}{2^y}$ . Thus,  $\text{Prob}\{S' > y\} = 1 - \exp\{-\frac{N}{2^y}\}$ , and the assertion holds by the cited Poisson approximation.

**Problem 2:**

Dayhoff's approach to generating amino acid substitution scores appropriate for various levels of protein divergence involves the following steps:

- (1) Observe counts  $c_{ij} = c_{ji}$  of amino acid  $j$  to  $i$  substitutions (and *vice versa*) in a set of proteins with overall residue frequencies  $f_j$ .
- (2) Define relative mutabilities as

$$m_j = \frac{c_{.j}}{f_j} k$$

where  $c_{.j} = \sum_{i \neq j} c_{ij}$  and  $k$  is some positive constant.

- (3) Set

$$M_{ij} = \rho m_j \frac{c_{ij}}{c_j} \quad \text{for } i \neq j \quad \text{and} \quad M_{jj} = 1 - \rho m_j.$$

(i) Derive a substitution matrix with transition probabilities to be used in a Markov model for which each time step represents an expected number of 5 accepted point mutations per 100 sites.

(ii) Describe how, using the result of (i), you could derive a symmetric amino acid substitution scoring matrix appropriate for comparing protein sequences at a divergence level of 25 accepted point mutations per 100 sites.

**Solution:**

(i) The required condition for 5 accepted point mutations per 100 sites (PAM5) can be written as  $\sum_j f_j M_{jj} = 0.95$ . Thus,  $\sum_j f_j (1 - \rho m_j) = 1 - \rho \sum_j f_j m_j = 0.95$ , with solution  $\rho = \frac{0.05}{\sum_j f_j m_j}$ .

(ii)  $(M_{ij})$  with  $\rho$  as above is the 1-step transition matrix corresponding to 5 PAM. The 5-th power of this matrix,  $(M_{ij}^{(5)})$ , is the 5-step transition matrix corresponding to 25 PAM.

Symmetrical substitution scores can be defined as  $s_{ij} = \frac{1}{\lambda} \ln \frac{M_{ij}^{(5)}}{f_i}$ , where  $\lambda$  is a scale factor. [Note:  $\sum_i \sum_j f_i f_j e^{\lambda s_{ij}} = \sum_i \sum_j f_j M_{ij}^{(5)} = \sum_j f_j (\sum_i M_{ij}^{(5)}) = \sum_j f_j = 1$ , and thus  $\lambda$  is the same scale factor appearing in the Karlin-Altschul theory of sequence analysis with scores.]

To show that the scores are symmetrical, note

$$M_{ij} = \rho m_j \frac{c_{ij}}{c_j} = \rho \left( \frac{c_j k}{f_j} \right) \frac{c_{ij}}{c_j} = \rho \left( \frac{c_{ij} k}{f_j} \right) = \rho \left( \frac{c_i k}{f_i} \right) \left( \frac{c_{ji}}{c_i} \right) \frac{f_i}{f_j} = \left( \rho m_i \frac{c_{ji}}{c_i} \right) \frac{f_i}{f_j} = M_{ji} \frac{f_i}{f_j},$$

i.e.,  $\frac{M_{ij}}{f_i} = \frac{M_{ji}}{f_j}$ . By induction, also

$$M_{ij}^{(5)} = \sum_k M_{ik}^{(4)} M_{kj} = \sum_k M_{ki}^{(4)} \frac{f_i}{f_k} M_{jk} \frac{f_k}{f_j} = M_{ji}^{(5)} \frac{f_i}{f_j}.$$