

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

## Homework 5

February 23, 2010

Due one week later. Answers to selected problems will be posted.

### 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  $\text{Prob}\{X_i > 0\} > 0$  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

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

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

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

**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.