

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

## Solutions to Homework 4

February, 2010

### Problem 1:

For a Hidden Markov Model with  $s$  states  $S_1, S_2, \dots, S_s$ , initial state probabilities  $\pi_S$ , state transition probabilities  $\tau_{ST}$ , and output probabilities  $P(O|S)$  (where  $S$  and  $T$  are any of the states, and  $O$  is a particular output symbol), define for  $j = 1 \dots s$

$$F_{ij} = \text{Prob}\{O_1 O_2 \dots O_i, Q_i = S_j\} \quad i = 1 \dots N,$$

and

$$B_{ij} = \text{Prob}\{O_{i+1} O_{i+2} \dots O_N | Q_i = S_j\} \quad i = 1 \dots N-1; \quad B_{Nj} = 1.$$

Use these terms to derive a recursion for efficiently calculating  $P_{iST} = \text{Prob}\{Q_i = S, Q_{i+1} = T | O_1 \dots O_N\}$ .

### Solution:

We can write

$$\text{Prob}\{Q_i = S, Q_{i+1} = T | O_1 \dots O_N\} = \frac{F_{is} \tau_{ST} \text{Prob}\{O_{i+1} | Q_{i+1} = T\} B_{i+1,t}}{\text{Prob}\{O_1 O_2 \dots O_N\}},$$

where  $s$  is the index corresponding to state  $S$ , and  $t$  is the index corresponding to state  $T$ . All terms can be calculated recursively as shown previously.

### Problem 2:

For a  $k$ -th order Markov Model for DNA sequences, the transition probabilities  $\hat{P}(i | h_k)$  are typically obtained as the maximum likelihood estimates from a set of training sequences such that the probability of observing the letter  $i$  following the  $k$ -mer  $h_k$  is proportional to the observed count of  $h_k i$   $k+1$ -mers. For an Interpolated Markov Model, transition probabilities are determined by taking well-formed linear weighted sums of relevant fixed-order transition probabilities, for example

$$P_{\text{imm}}(i | h_k) = \sum_{l=-1}^k \mu(h_l) \hat{P}(i | h_l),$$

where  $h_l$  is  $l$ -mer preceding letter  $i$  ("history of length  $l$ ") and  $\hat{P}(i | h_{-1})$  is taken to be one over the cardinality  $\alpha$  of the alphabet (i.e., 0.25 for the common 4-letter nucleotide alphabet).

Show that the probabilities  $P_{\text{imm}}(i | h_k)$  are well defined (i.e., add up to 1 when summed up over all letters  $i$ ) provided  $\sum_{l=-1}^k \mu(h_l) = 1$ .

Alternatively, define interpolated transition probabilities recursively for  $l = 0, 1, \dots, k$  as follows:

$$P_{\text{imm}}(i | h_l) = \lambda(h_l)\widehat{P}(i | h_l) + [1 - \lambda(h_l)]P_{\text{imm}}(i | h_{l-1}) , \quad (1)$$

with boundary condition  $P_{\text{imm}}(i | h_{-1}) = \widehat{P}(i | h_{-1})$  and weights  $0 \leq \lambda(h_l) \leq 1$ .

Demonstrate the equivalence of the two alternative formulations by first showing that  $P_{\text{imm}}(i | h_k) = \sum_{l=-1}^k \mu(h_l)\widehat{P}(i | h_l)$  where

$$\mu(h_l) = \left\{ \begin{array}{ll} \lambda(h_k) & l = k \\ \prod_{j=l+1}^k [1 - \lambda(h_j)]\lambda(h_l) & l = 0, 1, \dots, k-1 \\ \prod_{j=0}^k [1 - \lambda(h_j)] & l = -1 \end{array} \right\}$$

and then proving that  $\sum_{l=-1}^k \mu(h_l) = 1$ .

**Solution:**

For the first part, we have

$$\sum_{i=1}^{\alpha} P_{\text{imm}}(i | h_k) = \sum_{i=1}^{\alpha} \sum_{l=-1}^k \mu(h_l)\widehat{P}(i | h_l) = \sum_{l=-1}^k \mu(h_l) \left[ \sum_{i=1}^{\alpha} \widehat{P}(i | h_l) \right] = \sum_{l=-1}^k \mu(h_l) = 1.$$

The alternative formulation follows from successive insertion into equation (1) starting from  $l = k$  and collecting terms.

To prove that the  $\mu(h_l)$  add up to 1, note that

$$\begin{aligned} \prod_{j=l}^k [1 - \lambda(h_j)] + \prod_{j=l+1}^k [1 - \lambda(h_j)]\lambda(h_l) &= \\ \prod_{j=l+1}^k [1 - \lambda(h_j)][1 - \lambda(h_l) + \lambda(h_l)] &= \prod_{j=l+1}^k [1 - \lambda(h_j)] \end{aligned}$$

for  $l = 0, 1, \dots, k-1$ .