

## Math 275: Homework # 2

Recall that an HMM has  $n$  hidden random variables  $X_1, \dots, X_n$  and  $n$  observed random variables  $Y_1, \dots, Y_n$ . There is a  $k \times k$  transition matrix  $S = (s_{ij})$  for each horizontal transition  $X_r \rightarrow X_{r+1}$  and a  $k \times l$  transition matrix  $T = (t_{ij})$  for each vertical transition  $X_r \rightarrow Y_r$ .

### Problem 1

Consider the hidden Markov model with  $k = 2, l = 4$  (corresponding to  $A, C, G, T$ ) and  $n = 4$ . Suppose that

$$S = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

and

$$T = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.125 & 0.125 \end{pmatrix}$$

Compute  $p_{ACGT}$ .

### Problem 2

Suppose that an HMM with  $k = l = 2$  and  $n = 3$  has  $p_{011} = p_{110}$  and  $p_{100} = p_{001}$ . Show that  $p_{000}$  and  $p_{111}$  cannot both be 0.

### Problem 3

Compute the prime ideal of the HMM with  $k = 2, l = 3$  and  $n = 2$ .

### Problem 4

How many possible explanations are there for the sequence  $\sigma = \{1, 1, 0\}$  in an HMM with  $k = l = 2$  and  $n = 3$ .

### Problem 5

GENSCAN is a freely available program for finding genes in DNA sequences. It is based on hidden Markov models.

a) Find the accessions U73304 and AF276990 at NCBI. In order to do this, simply paste the accession into the search box on the main page, and then click on "Nucleotide" to find the DNA sequence.

b) Submit the sequence U73304 to GENSCAN. The organism you will use is vertebrate. The easiest way to submit the sequence is to select the FASTA format for the sequence on the NCBI website, and then to copy and paste it into the GENSCAN window. You will see that GENSCAN finds the single exon in the DNA exactly. GENSCAN also annotates the polyA signal. What is this signal? Does GENSCAN get it correct?

c) Submit the sequence AF276990 To GENSCAN. This is a much longer (213343 bp) very recently sequenced part of the canis familiaris (dog) genome. Copy and paste is again the best way to put the sequence into GENSCAN.

d) BLAST each of the GENSCAN predicted peptides (these are the proteins that the predicted genes would code for) against the nr database using blastp. Which of the predictions do you believe? For each gene, either cite evidence for it being a true prediction, or explain why you think the prediction is false. You may also want to use tblastn, which translates the DNA sequences in the database and compares them to your protein query.

e) You will see that the 10th prediction is in fact the dog version of the RAD50 human gene. Do you think all the predicted exons are exactly right? If yes explain why, and if not describe the false exons and explain how the prediction could be corrected.

### Problem 6

Let  $S \in \{(,)\}^n$ . For example  $S = (()())()$ . Define a *gene parse* to be a sequence  $1 \leq i_1, \dots, i_{2k} \leq n$  where  $S(i_{2r-1}) = ($  and  $S(i_{2r}) = )$  for  $1 \leq r \leq k$ . In the example above  $(2, 6, 8, 10)$  is a gene parse. If  $S$  contains  $n$  open parentheses  $($  and  $m$  close parentheses  $)$ , what is the maximum number of gene parses  $S$  could contain?

**Problem 7**

Show that the explanation for an observation from an HMM of length  $n$  with  $k$  hidden states can be obtained in time  $O(kn \log n)$  and space  $O(k)$ .