Universität Bielefeld
Technische Fakultät, AG Genominformatik
WinterSemester 2007/2008

Advanced Algorithmic Techniques for Bioinformatics

Prof. Dr. Ferdinando Cicalese, Dr. Zsuzsanna Lipták
Dipl.-Inform. Roland Wittler

Exercise sheet no. 5 (out: 21 Dec. 2007, in: 9 Jan. 2008)

1. **HMM-Viterbi.**

   The following exercise asks you to complete the example seen during the lecture.

   Consider an HMM for DNA that includes two hidden states H and L, for higher and lower $C + G$ content, respectively. Initial probabilities for both H and L are equal to 0.5, while transition probabilities are as follows:

   $$a_{H\,H} = 1/2, a_{H\,L} = 1/2, a_{L\,H} = 3/5, a_{L\,L} = 2/5.$$

   Nucleotides T, C, A, G are emitted from states H and L with probabilities 0.2, 0.3, 0.2, 0.3, and 0.3, 0.2, 0.3, 0.2, respectively. Use the Viterbi's algorithm (computing the log of the probabilities) to define a most likely sequence of hidden states for the sequence $\mathbf{x} = GGCACTGAA$.

2. **Back and Forward.**

   For the hidden Markov model defined in the previous exercise and the DNA sequence fragment $\mathbf{x} = GGCA$

   (a) find $Pr(\mathbf{x})$ by both the forward and the backward algorithm.

   (b) find the probability that the state at time 4 (i.e., when A is emitted) is H and the probability that it is L.

3. **Model set up.** A bar uses two different coffee brands, one is German, the other American. On a day when the German coffee is used, the probability that the coffee (the drink) is good is 0.5, while on a day when the Americal brand is used, the probability that the coffee comes out bad is 0.7. The bar switches from one brand to the other with probability 0.3. We can assume that the first brand that the bar used was chosen completely randomly, i.e., American or German with probability 0.5.

   Describe an HMM for the above situation. Suppose that you observed that in the first 4 days of its life the bar's coffee was *good*, *bad*, *good*, *good*. What is the probability that on the 5th day the brand used will be German?

4. **Information source.** A source of information emits symbols from the alphabet $\Sigma = \{0, 1, 2, 3\}$.

   Let s_t be the symbol emitted at time t. For $t = 1$, let $Pr(s_1 = i) = 1/4$, for each $i \in \Sigma$.

   Moreover, for each $t \geq 2$, and each $i, j \in \Sigma$, let $Pr(s_t = j | s_{t-1} = i) = 2^{-d(i,j)}$ where

$$d(i,j) = \begin{cases} j - i \bmod 4 & \text{if } i \neq j \\ 3 & \text{if } i = j, \end{cases}$$

   for each $i, j \in \Sigma$.

   For example, if the symbol emitted at time t is 0, then we have that the emission probabilities at time $t + 1$ are given by

$$Pr(s_{t+1} = 0 | s_t = 0) = 1/8, \quad Pr(s_{t+1} = 1 | s_t = 0) = 1/2,$$

$$Pr(s_{t+1} = 2 | s_t = 0) = 1/4, \quad Pr(s_{t+1} = 3 | s_t = 0) = 1/8.$$

   (a)Which are the most probable sequences of the first 4 symbols emitted by the above information source?

   (b) What is the probability distribution of the 3rd emission?

   (c) Assume now to change the initial distribution so that the first symbol emitted is always 0. What is probability distribution of the second emission? And of the 3th emission?