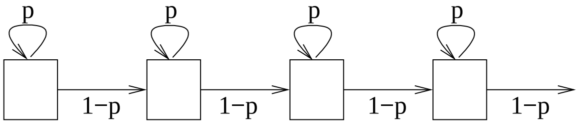
582653 Computational methods of systems biology, Autumn 2009
Homework 1
Thursday Nov. 12th 14-16 B119

General instructions

Problems for each exercise session will be distributed approximately one week before the session. You are expected to be prepared to present your solutions in the exercise session.

Assignments

1. [Durbin, Exercise 3.8] Show that the number of paths through an array of n states (see Figure below) is $\binom{l-1}{n-1}$ for length l.



2. Hidden Markov models can be used in algorithms of protein secondary structure prediction. One rather straightforward approach uses the secondary structure conformations α-*helix*, β-*strand*, and *turn* as the hidden states emitting observable amino acids. It is assumed that the frequencies of appearance of each of twenty amino acids in either conformation have been determined from analysis of the proteins with the three-dimensional structures known from experiment. Draw the state diagram of the HMM and describe the Viterbi and posterior decoding algorithms that could be used for predicting the protein secondary structure.

3. Real DNA sequences are inhomogeneous and can be described by a hidden Markov model with hidden states representing different types of nucleotide composition. Consider an HMM that includes two hidden states H and L for higher and lower $C + G$ contents, respectively. Initial probabilities for both H and L are equal to 0.5, while transition probabilities are as follows: $a_{HH} = 0.5$, $a_{HL} = 0.5$, $a_{LL} = 0.6$, $a_{LH} = 0.4$. 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 algorithm to define the most likely sequence of hidden states for the 'toy' sequence $x = GGCACTGAA$.

4. For the hidden Markov model defined in Exercise 3 and the DNA sequence fragment $x = GGCA$ find $P(x)$ by both the forward algorithm and the backward algorithm. Repeat the computation by the forward algorithm with use of scaling variables.

5. Find the posterior probability of states H and L at position 4 of the DNA sequence $x = GGCA$. Consider the hidden Markov model defined in Exercise 3.

6. [Durbin, Exercises 3.10 and 3.11] A prokaryotic gene is a continuous sequence of nucleotide triplets, codons. A gene starts with a start codon ATG and ends with one of three stop codons: TAA, TAG, TGA. Calculate the number of parameters in such a codon model. The data set contains on the order of $300\,000$ codons. Would it be feasible to estimate a second order Markov chain from this dataset? How can the above gene model be improved?