

1. Hidden Markov Models

Given is a hidden Markov model (HMM) with four states: $\mathcal{H} = \{A, B, C, D\}$. The observed sequence is composed of a three-letter alphabet: $\mathcal{O} = \{0, 1, 2\}$. The HMM has the following transition probabilities: $P(A|A) = 0.9, P(B|A) = 0.1, P(C|B) = 1.0, P(C|C) = 0.9, P(D|C) = 0.1, P(A|D) = 1.0$, zero otherwise. The emission probabilities are: $P(0|A) = P(0|C) = 1, P(1|B) = 1, P(2|D) = 1$, zero otherwise.

1. Draw a state diagram of the HMM.
2. Write down a typical sequence generated from this model.
3. Does the observed sequence satisfy any (possibly higher order) Markov property?
4. Sketch a general HMM as a probabilistic independence graph and explain why, in general, a sequence of observed characters generated from an HMM is not Markov of any order.

Comment: A sequence $\{y_1, \dots, y_n, \dots, y_N\}$ is called m th order Markov if $P(y_n|y_{n-1}, y_{n-2}, \dots, y_1) = P(y_n|y_{n-1}, y_{n-2}, \dots, y_{n-m}) \forall n > m$

2. Profile Hidden Markov Model

Consider the following hypothetical DNA sequence alignment:

	1	2	.	.	.	3
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C

Main states are indicated by a number in the first row, insert states are marked with a dot.

1. Draw a diagram of the corresponding profile HMM.

2. Determine the maximum likelihood estimate of the transition and emission probabilities.

3. Information Theory and DNA Sequence Alignments

Let $X \in \{A, C, G, T\}$ denote a random variable that represents a nucleotide at a given position in a DNA sequence alignment. The entropy of a discrete probability distribution $P(X)$ is defined as $H(P) = -\sum_X P(X) \ln P(X)$. The relative entropy of a probability distribution $P(X)$ with respect to another distribution $Q(X)$ is defined as $KL[P, Q] = \sum_X P(X) \ln \left(\frac{P(X)}{Q(X)} \right)$.

1. Show that $KL[P, Q] \geq 0$. (Hint: Use the relation $\ln x \leq x - 1$).
2. Show that $0 \leq H(P) \leq \ln(4)$.
3. Consider the following hypothetical DNA sequence alignment:

bat	A	A	G	G
rat	C	A	G	C
cat	A	A	G	T
gnat	C	A	G	A

Determine the maximum likelihood estimate of the nucleotide emission probabilities and the entropy for each position.

4. Phylogenetics

1. Determine (1) the total number of nodes and edges and (2) the number of different topologies for (a) a rooted tree and (b) an unrooted tree with n leaves.
2. An evolution model is called reversible if

$$P(Y|X, w)P(X) = P(X|Y, w)P(Y)$$

where X, Y denote nucleotides at given nodes of the tree and w is the length of the branch connecting these nodes. Show that in this case the likelihood is independent of the position of the root.