Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
September 19, 2018

Hidden Markov Model: Project

General information

The project should be handed in before noon Monday October 8, 2018, in Noors mailbox in the BiRC building.

Writing and organization of the report

Report your analysis in a clear, complete, and concise language. Organise the report in a logical order with descriptive headings and subheadings. This means that the headings should not be e.g. '1(a)', but rather 'Getting the evolutionary distance from the probability of genomic difference'.

The report should be written in danish or english.

A main part of the project is to write R code. In the report you should include in an Appendix the R code that you have written. The R code should be documented. This means that you should not just include your R commands, but also briefly describe what the R commands do and also describe the structure of the R code.

Project description

The project consists of the two exercises below.
The data and R commands for the HMM exercise can be downloaded from the homepage (PhyloFoot.dat and HMMproject.R).

1. **Jukes-Cantor model** (EG page 487)
   In the Jukes-Cantor model, the probability p that two nucleotides are different in a genomic position is given by

   $$p = 1 - I(t) = \frac{3}{4}\Big(1 - \exp(-4t/3)\Big),$$

   where t is the evolutionary distance (or, more precisely, the expected number of substitutions) between the two genomes.

   (a) Show in detail that we can invert the equation to obtain

   $$t = -\frac{3}{4}\log\Big(1 - \frac{4}{3}p\Big), \quad 0 \leq p < 3/4.$$

   (b) Show that $t \approx p$ for small p (use a Taylor expansion), and $t \to \infty$ for $p \to 3/4$. Argue why these two limiting cases make sense.

   (c) The probability of a base pair (nucleotide) position in the human genome being different from the homologous position in the chimpanzee genome is 1.37%, and the substitution rate is $0.5 \cdot 10^{-9}$ substitutions per base pair per year. What is the average divergence time (in million years) between human and chimpanzee?
   The difference between human and gorilla is 1.75% and human and orangutan is 3.40%. What is the divergence time between human and gorilla, and human and orangutan?

2. **Identifying neutral and functional regions**
(or: Phylogenetic footprinting)
We investigate two aligned sequences `S1` and `S2`. The sequences are of length $L = 293$ after removal of gaps. The first and the last part of the sequences look as follows:

```
S1: GCCTTGG...TGAGGGT
S2: GCCTTCG...AGAGGGT
```

We first translate the nucleotide sequences to a numerical vector. We translate `A,G,C,T` to 1,2,3,4. The first and the last part of the numerical sequences look as follows:

```
S1.num: 2 3 3 4 4 2 2 ... 4 2 1 2 2 2 4
S2.num: 2 3 3 4 4 3 2 ... 1 2 1 2 2 2 4
```

There is a total of 16 possible outcomes in each alignment column (4 possible outcomes in `S1` times 4 possible outcomes in `S2`).

We can translate the 16 possible outcomes to the numbers $1, \ldots, 16$ using the command

```
obs.seq <- 4*(S1.num-1)+S2.num
```

The first and last part of `obs.seq` look as follows:

```
> obs.seq
[1] 6 11 11 16 16  7  6 ... 13  6  1  6  6  6 16
```

a. Explain that a column with `A` in `S1` and `A` in `S2` is translated to the number 1, while a column with `G` in `S1` and `C` in `S2` is translated to the number 7.

Table 1 provides the translation of the 16 possible columns to the numbers $1, \ldots, 16$.

S1	A	A	A	A	G	G	G	G	C	C	C	C	T	T	T	T
S2	A	G	C	T	A	G	C	T	A	G	C	T	A	G	C	T
Translation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Table 1: Translation from possible columns in the alignment to the numbers $1, \ldots, 16$.

b. Explain that the following vector of length 16 indicates if an alignment column contain similar or different nucleotides:

```
diff.indx <- c(0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,0)
```

From `diff.indx` and `obs.seq` we make a new vector, `diff.seq`, of the same length as `S1` and `S2`, which indicates if the two nucleotides are the same or not. The R command and the first and last entries in `diff.seq` are

```
> diff.seq <- diff.indx[obs.seq]
> diff.seq
[1] 0 0 0 0 0 1 0 ... 1 0 0 0 0 0 0
```

We find the total number of different columns in the alignment with the command

```
sum(diff.seq)
```

and get the result

```
[1] 69
```

c. Use the previous exercise to estimate the total number of expected substitutions between the two sequences under the Jukes-Cantor model.

We now want to inspect the two sequences in more detail. In particular we want to divide the sequences into two types of regions: Neutrally evolving regions and functional regions. If a region is functional, the number of substitutions is expected to be smaller than if the region is evolving neutrally. In order to divide the sequences into these two types of regions, we formulate a hidden Markov chain with two hidden states. The hidden states correspond to functional regions (state 1) and neutral regions (state 2).

Let δ_1 be the probability for observing a column with different nucleotides in state 1, and let δ_2 be the probability for different nucleotides in state 2. Furthermore, let $\epsilon_i = 1 - \delta_i$ $(i = 1, 2)$ be the probability for observing a column in state i where the two nucleotides are the same. The emission probabilities are determined from the matrix $b_i(a)$ in Table 2, where $i = 1, 2$ and $a = 1, \ldots, 16$.

possible columns	S1	A	A	A	A	G	G	G	G	C	C	C	C	T	T	T	T
	S2	A	G	C	T	A	G	C	T	A	G	C	T	A	G	C	T
hidden states	1	$\frac{\epsilon_1}{4}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\epsilon_1}{4}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\epsilon_1}{4}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\delta_1}{12}$	$\frac{\epsilon_1}{4}$
	2	$\frac{\epsilon_2}{4}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\epsilon_2}{4}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\epsilon_2}{4}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\delta_2}{12}$	$\frac{\epsilon_2}{4}$

Table 2: Emission probabilities from the two hidden states. We have $\epsilon_1 = 1 - \delta_1$ and $\epsilon_2 = 1 - \delta_2$.

Neutrally evolving regions have more substitutions and therefore $\delta_2 > \delta_1$.

d. Explain why the emission probabilities sum to 1 within each row.

The transition matrix is determined by the 2×2 matrix

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix},$$

and the initial state is given by $\pi = (1/2, 1/2)$.

We want to estimate the parameters $\theta = (\delta_1, \delta_2, p_{11}, p_{22})$ using the EM-algorithm. Let $\mathcal{O} = (\mathcal{O}_1, \ldots, \mathcal{O}_L)$ be the observed sequence and suppose we know the hidden state sequence $Q = (q_1, \ldots, q_L)$.

e. Show that the full likelihood

$$L(\theta; Q, \mathcal{O}) \quad = \quad \pi_{q_1} b_{q_1}(\mathcal{O}_1) \prod_{l=2}^{L} \Big(p_{q_{l-1}, q_l} b_{q_l}(\mathcal{O}_l) \Big)$$

is proportional to

$$L(\theta; Q, \mathcal{O}) \propto p_{11}^{N_{11}} p_{12}^{N_{12}} p_{21}^{N_{21}} p_{22}^{N_{22}} \delta_1^{N_1^\delta} \epsilon_1^{N_1^\epsilon} \delta_2^{N_2^\delta} \epsilon_2^{N_2^\epsilon}, \tag{1}$$

and describe in words the meaning of N_{ij}, N_i^δ and N_i^ϵ.

f. Describe the M-step in the EM-algorithm (explain the maximization of (1)).

g. Describe what type of calculations are needed for the E-step in the EM-algorithm.

We now want to implement the EM-algorithm. We first implement a function, `EmissionFct`, which returns the matrix in Table 2 as a function of δ_1 and δ_2. Second we activate the function `HMMexpectationsFct` (previously used in the course). Third we apply the EM algorithm:

```
01 ## EM algorithm
02 IntPrb <- c(1/2,1/2)
03 TrnsPrb <- matrix(c(0.8,0.2,
04                     0.1,0.9),byrow=TRUE,nrow=2,ncol=2)
05 delta.1 <- 0.2
06 delta.2 <- 0.4
07 EmsPrb <- EmissionFct(delta.1,delta.2)
08 ## Number of iterations
09 nIter <- 20
10 for (iter in 1:nIter){
11  HMMexpct <- HMMexpectationsFct(IntPrb,TrnsPrb,EmsPrb,observed.seq)
12  TrnsPrb <- HMMexpct$TransCnt/rowSums(HMMexpct$TransCnt)
13  N.delta.1 <- sum(HMMexpct$PostProb[,1]*diff.seq)
14  N.epsil.1 <- sum(HMMexpct$PostProb[,1]*(1-diff.seq))
15  N.delta.2 <- sum(HMMexpct$PostProb[,2]*diff.seq)
16  N.epsil.2 <- sum(HMMexpct$PostProb[,2]*(1-diff.seq))
17  delta.1 <- N.delta.1/(N.delta.1+N.epsil.1)
18  delta.2 <- N.delta.2/(N.delta.2+N.epsil.2)
19  EmsPrb <- EmissionFct(delta.1,delta.2)
20 }
```

h. Explain the R program above.

After 20 iterations the parameter estimates are

$$\hat{\theta} = (\hat{\delta}_1, \hat{\delta}_2, \hat{p}_{11}, \hat{p}_{22}) = (0.12, 0.34, 0.84, 0.86)$$

i. What is the expected number of substitutions between the two sequences under the hidden Markov model?

j. Decode the sequence alignment using (a) the Viterbi algorithm and (b) posterior decoding. Make a plot that summarizes the data and corresponding decoding. Do the regions with few differences correspond to predicted functional regions?