

# Theory and Algorithms: Exercise sheet 2

Magnus Rattray

## 1 Casino shenanigans

The dishonest casino uses a loaded die with which sequential throws are not necessarily independent. In particular, every time a six is thrown there is a 50% chance that the next throw is also a six, while all other numbers are equally likely. If any of the other five numbers are thrown then the next throw is independent and all numbers are equally likely.

- (a) Design a Markov chain model of the casino.
- (b) Design a hidden Markov model of the casino.
- (c) If the parameters of the two models were not known in advance, how could you learn them from data?
- (d) Which representation do you think is most natural for this problem?

The casino introduces a new game that involves creating long sequences of coin tosses. Sometimes they use a coin with heads on both sides. When using the fake coin they switch to a fair coin with a 30% probability after every coin toss and when they are using the fair coin they switch back to the fake coin with 5% probability after every coin toss.

- (e) Design a hidden Markov model of the coin flipping process.
- (f) Which algorithm could you use to detect when the fake coin was in use? What does this algorithm compute and what is its computational complexity on this problem?
- (g) Is it possible to model this process with a 1st order Markov chain model?

## 2 Markov chains and CpG islands

We define the parameters of a Markov chain model to be the transition probabilities from nucleotide  $X$  to nucleotide  $Y$ ,

$$t_{XY} = P(x_{i+1} = Y | x_i = X) .$$

Let  $T_{XY}$  be the number of times  $X$  is followed by  $Y$  in our observed sequence. The maximum likelihood parameters are then given by,

$$t_{XY} = \frac{T_{XY}}{\sum_Z T_{XZ}}$$

and the probability of a sequence  $\{x_1, x_2, \dots, x_N\}$  under the model is,

$$P(x|t) = \prod_{i=1}^{N-1} t_{x_i x_{i+1}} .$$

- (a) Show that  $\sum_Y t_{XY} = 1$ . Why is this condition necessary for a Markov chain model to be well-defined? What constraint does this impose on a Markov chain diagram? What constraint does this impose on a transition probability matrix? (eg. a PAM matrix)
- (b) Given the following table of transition counts,

$T_{XY}^+$	A	C	G	T
A	560	855	1324	385
C	1144	2462	2526	607
G	1241	2582	2905	982
T	179	840	955	452

determine the missing values in the corresponding table of transition probabilities below using maximum likelihood estimates.

$t_{XY}^+$	A	C	G	T
A	0.179	0.274	0.424	0.123
C	0.170	0.365	0.375	0.090
G				
T				

- (c) The table in part (b) was determined using sequences thought to be CpG islands. The following table gives transition probabilities calculated from non-CpG islands,

$t_{XY}^-$	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

To discriminate between models we use the log-odds ratio for a particular sequence  $x$ ,

$$\begin{aligned}
 S(x) &= \log \frac{P(x|t^+)}{P(x|t^-)} \\
 &= \sum_{i=1}^{N-1} \beta_{x_i x_{i+1}} ,
 \end{aligned}$$

where we have defined,

$$\beta_{XY} = \log \frac{t_{XY}^+}{t_{XY}^-} .$$

Determine the table of  $\beta$  values given the above results for  $t_{XY}^+$  and  $t_{XY}^-$ . Which transition is most informative when scoring a sequence?

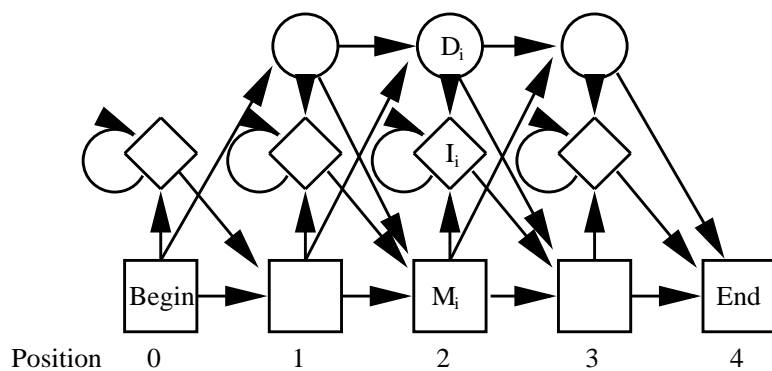
- (d) Calculate the score for the following sequence,

GCCCCTTCACCGCGAGGCTG

Do you think it is likely to come from a CpG island?

- (e) How do you think the score defined in part (c) could be adapted to include information about the relative abundance of CpG islands?

### 3 Profile HMM example



Match ?	x	x	-	-	-	x
seq. 1	A	T	-	-	-	T
seq. 2	A	G	-	-	-	-
seq. 3	A	G	A	T	-	T
seq. 4	A	G	C	G	T	T
seq. 5	A	-	-	-	-	T
Position	1	2	.	.	.	3

- (a) For each sequence in the alignment above work out the corresponding path through the profile HMM model.
- (b) Fill in the missing entries in the following tables of emission and transition counts.

Position		0	1	2	3
Match Emissions	A	-	5	0	0
	C	-			
	G	-			
	T	-			
Insert Emissions	A	0	0	1	0
	C				
	G				
	T				

Position		0	1	2	3
State Transitions	M-M	5	4	1	4
	M-D				-
	M-I				
	I-M				
	I-D	-	-	-	-
	I-I				
	D-M	-			
	D-D	-			-
	D-I	-	-	-	-

- (c) How can the counts in these tables be converted into maximum likelihood parameter estimates ?
- (d) Outline one problem associated with using maximum likelihood estimates for the transition and emission parameters in this example. How might one overcome this problem ?
- (e) Name an algorithm that can be used for aligning a new sequence to the model. What is the computational complexity of this algorithm?