

HMM I: Markov chains

Introduction

In assessing whether patterns of bases are under- or over-represented, we need a model for the occurrence of the bases. In earlier lectures we used models in which the bases were assumed to be laid down independently (for example, when we were studying the significance of sequence alignments). In this lecture, we introduce *Markov chains*, the natural generalization of a sequence of independent trials. A useful reference is the text by Kemeny and Snell (1976). Many applications of Markov chains in computational biology are described in Durbin et al. (1998).

Markov chains

We study a sequence of random variables $\{X_n, n = 0, 1, 2, \dots\}$ taking values in the state space \mathcal{S} (think of \mathcal{S} as the set $\{A, C, G, T\}$ as the obvious example when describing a DNA sequence). The sequence $\{X_n, n \geq 0\}$ is called a *Markov chain* if it satisfies the *Markov property*:

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i) \quad (1)$$

for $n = 1, 2, \dots$ and for all $i, j, i_{n-1}, \dots, i_0 \in \mathcal{S}$. This property says that the future evolution of the sequence is determined by its present position, and not by its earlier history.

If $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ is the same for every value of n , we call the chain *homogeneous*; the common value is written p_{ij} :

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i), \quad i, j \in \mathcal{S}. \quad (2)$$

The matrix $P = (p_{ij})$ is called the (one step) transition matrix of the chain. We study only homogeneous chains in this lecture. When the rows of P are identical, the next move has the same distribution no matter what the current position is. In this case the chain corresponds to a sequence of independent trials.

Properties of the chain are determined by P , and the *initial distribution* π of the chain,

$$\pi_i = \mathbb{P}(X_0 = i), \quad i \in \mathcal{S}.$$

To see this, we calculate

$$\begin{aligned}
\mathbb{P}(X_2 = j \mid X_0 = i) &= \sum_k \mathbb{P}(X_2 = j, X_1 = k \mid X_0 = i) \\
&= \sum_k \mathbb{P}(X_2 = j \mid X_1 = k, X_0 = i) \mathbb{P}(X_1 = k \mid X_0 = i) \\
&= \sum_k \mathbb{P}(X_2 = j \mid X_1 = k) \mathbb{P}(X_1 = k \mid X_0 = i) \\
&= \sum_k p_{ik} p_{kj} \\
&= (P^2)_{ij}.
\end{aligned}$$

The first equality comes from the law of total probability (summing over all the possibilities for X_1). The second equality comes from the conditional probability formula $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid B \cap C) \mathbb{P}(B \mid C)$. The third equality comes from the Markov property (1). The final equality comes from the definition of matrix multiplication. In order to find the distribution of X_2 , we condition on the value of X_0 :

$$\begin{aligned}
\mathbb{P}(X_2 = j) &= \sum_i \mathbb{P}(X_2 = j \cap X_0 = i) \\
&= \sum_i \mathbb{P}(X_2 = j \mid X_0 = i) \mathbb{P}(X_0 = i) \\
&= \sum_i \pi_i (P^2)_{ij} \\
&= (\pi P^2)_j.
\end{aligned}$$

The final line above is the j th element of the product of the row vector π with the square matrix P^2 .

The arguments above can be generalized to any time point. Writing

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i)$$

we have

$$p_{ij}^{(n)} = (P^n)_{ij}, \quad n = 0, 1, \dots, \quad (3)$$

where we define $P^0 = I$, the identity matrix. The distribution of X_n is given by

$$\mathbb{P}(X_n = j) = (\pi P^n)_j, \quad n = 0, 1, \dots, \quad (4)$$

A distribution π with the property that

$$\pi_j = \sum_i \pi_i p_{ij} \text{ for all } j \quad (5)$$

is called a *stationary distribution* of the chain. In matrix notation, the condition is

$$\pi = \pi P.$$

If this holds, and if X_0 has this π as its distribution, then

$$\begin{aligned} \mathbb{P}(X_n = j) &= (\pi P^n)_j \\ &= ([\pi P] P^{n-1})_j \\ &= (\pi P^{n-1})_j \\ &= \dots \\ &= \pi_j, \end{aligned}$$

showing that X_n has the same distribution for every n (but remember that the X_n are not in general independent).

Irreducible and aperiodic chains

We are studying chains with finite state space, where the theory breaks into two main parts: chains that wander all over the state space, and chains that do not. Here we study the first sort. To set the scene, consider the example with

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Some simple matrix multiplication shows that

$$P^{2n} = I, \quad P^{2n+1} = P, \quad n = 0, 1, \dots$$

If the chain starts in state 1, then it hits state 1 at time points 2,4,6,... It can never be in state 1 at times 1,3,... We call such a chain *periodic*. They make the theory uglier, so we assume our chains are aperiodic. For the next example, assume

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This chain is boring: if it starts in state 1, it stays there for ever, so state 2 is never visited. We want to rule out this sort of behavior too. We achieve this by assuming that the chain is *ergodic*:

$$\text{For some } n, \quad p_{ij}^{(n)} > 0 \text{ for all } i, j. \quad (6)$$

Such chains have a number of important properties, listed below.

1. They have a unique stationary distribution π
2. The limit distribution of the chain is also π : $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = j) = \pi_j$, no matter what the initial state of the chain was. This says that the effects of the initial position wears off as time goes on.
3. Think of a chain starting in state i . Because of the ergodic property, the chain must eventually hit state i again. Let T be the number of steps this takes, and let μ_i be the mean of T , given $X_0 = i$. Then we have

$$\pi_i = 1/\mu_i, \quad i \in \mathcal{S}. \quad (7)$$

The simplest example of an ergodic chain is

$$P = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

for $0 < \alpha, \beta < 1$. To find the stationary distribution of the chain, we solve the matrix equation $\pi = \pi P$, getting

$$\begin{aligned} \pi_1 &= \pi_1(1 - \alpha) + \pi_2\beta \\ \pi_2 &= \pi_1\alpha + \pi_2(1 - \beta) \end{aligned}$$

the first equation shows that

$$\pi_1\alpha = \pi_2\beta,$$

and using the fact that $\pi_1 + \pi_2 = 1$, we can solve the system to get

$$\pi_1 = \frac{\beta}{\alpha + \beta}, \quad \pi_2 = \frac{\alpha}{\alpha + \beta}.$$

It follows that the average number of steps taken to return to state 1 starting from state 1 is $(\alpha + \beta)/\beta = 1 + \alpha/\beta$.

Some examples

We use the theory in the last section to study the occurrences of patterns in a sequence of DNA bases. Lets start with the simplest model for the letters: independent and identically distributed trials, with probabilities p_A, p_G, p_C, p_T of the next letter being A, G, C or T . Now think of a (mathematical) restriction enzyme reading along the DNA and cutting every time it sees AA . In the sequence $TACTAATCGGATAACCAAACA \dots$, we would get fragments $TATCAA, TCGGATAA, CCAA$ and $ACA \dots$. First question: what is the average size of the fragments? Now suppose another cutter cuts at AB , where B denotes the set $\{C, G, T\}$. How long are those fragments? What if it cuts at AC ? Second question: what is the chance that one sort of cut appears before another?

To answer the question about the average length of an AA fragment we use a bit of a trick. We construct a Markov chain with state space determined by the pattern of interest as follows: add a state for the first letter of the pattern, one for the first two letters of the pattern, and another to denote all other states. In our example, we have state space $\mathcal{S} = \{A, AA, B\}$. X_n then records the letters in the original chain ending at position n . Thus in the sequence $TATCAA$ we have $X_1 = B, X_2 = A, X_3 = B, X_4 = B, X_5 = A, X_6 = AA$. What is the transition matrix P of our chain? Write $p = p_A, q = 1 - p = p_C + p_G + p_T$. Then P is determined by

$$P = \begin{array}{c|ccc} & A & B & AA \\ \hline A & 0 & q & p \\ B & p & q & 0 \\ AA & p & q & 0 \end{array}$$

The initial state of the chain is AA (we want the mean length of a fragment, so we can assume we have just cut the DNA sequence at AA , and we are looking for the next AA .) Now we can find the stationary distribution π . First, $\pi_1 = \pi_1 0 + \pi_2 p + \pi_3 p = p(1 - \pi_1)$, so that $\pi_1 = p/(1 + p)$. From the second element we get $\pi_2 = q(\pi_1 + \pi_2 + \pi_3) = q$, and finally $p\pi_1 = \pi_3$, so that $\pi_3 = p^2/(1 + p)$. The answer to our first question is now given by (7):

$$\text{The mean length of an } AA \text{ fragment is } \frac{1}{p_A} + \frac{1}{p_A}. \quad (8)$$

How do we find the average length of AC fragments? We can do the same sort of thing, but the state space is now $\{A, C, G \cup T, AC\}$ and the initial

state is AC . The transition matrix is

$$P = \begin{array}{c|cccc} & A & C & G \cup T & AC \\ \hline A & p_A & 0 & p_G + p_T & p_C \\ C & p_A & p_C & p_G + p_T & 0 \\ G \cup T & p_A & p_C & p_G + p_T & 0 \\ AC & p_A & p_C & p_G + p_T & 0 \end{array}$$

Deriving the stationary distribution gives $\pi_1 = p_A$, $\pi_2 = p_C(1 - \pi_1) = p_C(1 - p_A)$, $\pi_3 = p_G + p_T$, and $\pi_4 = p_C\pi_1 = p_Cp_A$. Hence

$$\text{The mean length of an } AC \text{ fragment is } \frac{1}{p_{APC}}. \quad (9)$$

Notice that your ‘first guess’ for the average length of the AA fragments was $1/p_A^2$. This turns out to be wrong because of the appearance of overlaps in the string AA . When there are no overlaps (as in AC) the original guess is correct, as shown by (9). Questions like this are simple examples of problems involving pattern statistics, a very well-developed part of the computational biologist’s arsenal. The package R’MES cited below can guide you through some of the complexities.

Estimation

Suppose that we have a realization x_0, \dots, x_n of a Markov chain with unknown transition matrix P . An obvious question: how would we estimate P ?

To do this, let $n(i, j)$ be the number of times the chain has state j following state i . The likelihood of the observations is

$$L = \prod_{t=1}^n p_{x_{t-1}, x_t}.$$

By counting how many times the chain has j following i , we can write this as

$$L = \prod_{i, j \in \mathcal{S}} p_{ij}^{n(i, j)}.$$

We now want to maximize this subject to the constraint that each row of the transition matrix sums to 1, i.e. $\sum_{j \in \mathcal{S}} p_{ij} = 1$ for each $i \in \mathcal{S}$.

Calculus gives the result that the MLEs of the parameters P are given by

$$\hat{p}_{ij} = \frac{n(i, j)}{\sum_{k \in \mathcal{S}} n(i, k)}, \quad i, j \in \mathcal{S}.$$

We will use this result as motivation during our discussion of hidden Markov chains.

References

Durbin J, Eddy S, Krog A, Mitchison G (1998) *Biological Sequence Analysis*. Cambridge University Press.

Kemeny JG, Snell JL (1976) *Finite Markov Chains*. Second Edition. Springer Undergraduate Texts in Mathematics.

R'MES is available at <http://www-bia.inra.fr/J/AB/genome>