4.3 Markov Chains

In real DNA molecules, the nucleotide bases are highly dependent on their neighbors. A probability model that allows for such neighborhood dependence is called a Markov chain. The Markov chain is a classical stochastic model that has been widely used in modeling sequential phenomena where dependence among neighboring observations is apparent. It is very natural that it is popularly adopted as a model for DNA sequences. In this section, we discuss Markov chain in the context of nucleotide sequences. For a good introduction to the theory of Markov chains, consult the book *Introduction to Probability Models* (6th edition) by Sheldon M. Ross.

Mathematically, a Markov chain can be defined as a sequence of discrete random variables X_0, X_1, X_2, ...,each of which takes on a countable number of possible values. The value taken by X_n is referred to as the state of the Markov chain at time n. At any n = 0, 1, 2,..., the conditional probability distribution of the random variable X_{n+1} given the values of its predecessors X_0, X_1, ..., X_n, depends only on the value of its immediate predecessor X_n but not on those values of X_0, X_1, ..., $X_{n\text{-}1}$. In mathematical notation, we write

(4.3.1) $P\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0\} = P\{X_{n+1} = j \mid X_n = i\}$.

The probability above, denoted by p_{ij} is known as the transition probability from state i to state j. That is to say, if the Markov chain is in state i at the present time, p_{ij} will represent the probability that the chain is in state j one unit of time later.

In the context of DNA sequence analysis, each nucleotide base can be considered a random variable X_n which takes on four possible values *A, C, G, T*. The time index n is actually the position of the base in the DNA sequence. The transition probability p_{ij} is the probability that the base at the next position is j given that the base at the present position is i. It is most convenient to present the transition probabilities in the form of a 4 by 4 matrix

$$\mathbf{P} = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$

Note that all entries of this matrix must be nonnegative numbers and the sum of each row must equal 1. Matrices with these properties are called stochastic matrices.

Using the transition probability matrix, it will be easy to write down the probability distribution of X_{n+1} when we know the probability distribution of X_n. It is a simple application of the Law of Total Probability. If we let f_n stand for the probability distribution (also referred to as the probability mass function) of X_n, we have

$$\begin{aligned} f_{n+1}(b') &= P(X_{n+1} = b') \\ &= \sum_b P(X_{n+1} = b' \mid X_n = b) P(X_n = b) \\ &= \sum_b P_{bb'} f_n(b) \end{aligned}$$

For those who knows matrix multiplication, a convenient way of representing the probability distribution of X_n is writing it as a row vector $\mathbf{ð}_n = (f_n(A), f_n(C), f_n(C), f_n(T))$. In this form, one can easily verify that

(4.3.2) $$\mathbf{ð}_{n+1} = \mathbf{ð}_n \mathbf{P}.$$

The right hand side of the above equality refers to a matrix multiplication of $\mathbf{ð}_n$ to $\mathbf{P}$.

Other Markov chain models

There are many other alternative Markov models for studying DNA. For example, if we are interested in the distribution of strong/weak (resp. purine/pyrimidine) bases, we can construct the sequence X_0, X_1, X_2, ... to be binary random variables taking values S/W (resp. R/Y). In such cases, the transition probability matrix will be 2 by 2.

It is also possible that the nucleotide bases of DNA are not only dependent on its immediately preceding base, but rather a few, say *m*, preceding bases. This is referred to as an *m*th order Markov nucleotide sequence. Strictly speaking, an *m*th order Markov chain is not Markov, but can still be formulated as a suitable Markov chain model using a slightly modified set up.

Let us look at a simple example where $m = 2$. Suppose that each base depends on the two preceding bases. Denote the probability of observing base *k*, given that the two immediately preceding bases are (*i, j*), by $P_{(i,j),k.}$ We can construct a sequence of random variables Y_0, Y_1, Y_2, ... where $Y_n = (X_{n,} X_{n+1})$ actually represents the pair of bases at positions *n* and *n*+1. There are 16 possible values for the *Y*'s, namely all the dinucleotide pairs (*A*, *A*), (*A*, *C*), ..., (*T*, *T*), and it is not hard to see that the sequence Y_0, Y_1, Y_2, ... satisfy equation (4.4.1) and hence is itself a Markov chain. It has a 16 by 16 transition probability matrix. The transition probability between the dinucleotide pair (b_1, b_2) and (b_1' , b_2 ') must be 0 unless $b_2 = b_1'$.

The original Markov chain model that allows each base to depend on the preceding base only can be called the first order Markov model. The model with independently generated nucleotide base model can be called a 0^{th} order Markov chain.

Question: What can you say about the four rows of the transition probability matrix in a 0^{th} order Markov nucleotide sequence model?

Answer: ________________________

Multi-step transition probability matrix

By iterating equation (4.4.2), we obtain for any integers $n \geq 0$ and $m > 0$,

$$\boldsymbol{\eth}_{n+m} = \boldsymbol{\eth}_{n+m-1}\mathbf{P} = \boldsymbol{\eth}_{n+m-2}\mathbf{P}^2 = \cdots = \boldsymbol{\eth}_n\mathbf{P}^m$$

So, suppose we are currently at base n of the DNA sequence and the probability distribution is $\boldsymbol{\eth}_n$, then m bases later, the probability distribution $\eth_{n+m}$ is equal to $\eth_n$ multiplied to $\mathbf{P}^m$, the mth power of the transition probability matrix $\mathbf{P}$. In other words, the matrix $\mathbf{P}^m$ tells us with what probabilities the Markov chain transition from the current base to m bases later. Hence, $\mathbf{P}^m$ is called the m-step transition probability matrix. Of course, $\mathbf{P}$ would be the one-step transition probability matrix.

The stationary distribution

Under suitable mathematical conditions on the matrix $\mathbf{P}$, it can be proved that independent of the initial probability distribution $\boldsymbol{\eth}_0$, $\boldsymbol{\eth}_n$ approaches a unique limiting probability distribution $\boldsymbol{\eth} = (f(A), f(C), f(G), f(T))$ which satisfies the equation

(4.3.3) $$\boldsymbol{\eth} = \boldsymbol{\eth}\mathbf{P}$$

as n approaches infinity. This unique distribution $\boldsymbol{\eth}$ is called the stationary distribution of the Markov chain. In a very long Markov nucleotide sequence, as we move away from the beginning of the sequence, the base distribution will get very close to $\boldsymbol{\eth}$. It can also be interpreted as the long run proportions of observing the four bases. This distribution $\boldsymbol{\eth}$ can be obtained quite easily using the mathematical result in Markov chain theory which states that the stationary distribution is the unique probability distribution which is a left eigenvector with eigenvalue 1 for the matrix $\mathbf{P}$. This probability distribution $\boldsymbol{\eth} = (f(A), f(C), f(G), f(T))$ must be the one and only solution to the linear system of equations:

$$\begin{aligned}
f(A)(P_{AA}-1) + f(C)P_{CA} + f(G)P_{GA} + f(T)P_{TA} &= 0 \\
f(A)P_{AC} + f(C)(P_{CC}-1) + f(G)P_{GC} + f(T)P_{TC} &= 0 \\
f(A)P_{AG} + f(C)P_{CG} + f(G)(P_{GG}-1) + f(T)P_{TG} &= 0 \\
f(A) + f(C) + f(G) + f(T) &= 1
\end{aligned}$$

Solving for the stationary distribution $\boldsymbol{\eth}$ for a Markov nucleotide sequence involves solving a system of four equations in four unknowns, a task which generally takes a while to accomplish working by hand. If a higher order Markov model is used, the size of the linear system of equations grows exponentially. We will be working with 4^m equations in 4^m unknowns. This is much better done with the help of mathematical or statistical software. The statistical package S-Plus comes with a built-in functions such as "solve" to solve linear systems, and "eigen" that finds eigenvectors. In a later exercise, you will use these to obtain stationary distributions.