Stochastic Processes

Stochastic processes are simply descriptions of probability events that happen over time. Random events happen with time and we may track the events themselves or some cumulative aspect of the events. We will call the possible set of outcomes at any given time, the ***state space*** of the stochastic process. For example, an extremely simple stochastic event is the outcome of repeated throws of a coin with the state space {H, T}. One might have an outcome of the following sort…

Example 1:

"#HHHTHTTH.."

(Here I used the character "#" to denote the initial state where nothing has happened.)

We will use the notation *S(t)* to refer to the random variable for the events at the *t*-th time. In the above, we have the outcome S(1) = H, S(2) = H, S(3) = H, and so on.

As another example, one might start with the number zero and then throw a coin and add +1 if the coin is heads and add −1 if the coin is tails. So, we might have the outcome:

Example 2:

"0:1:2:1:2:3:2:1:0:-1" (corresponding to the coin tosses HHTHHTTTT; I used ":" colons to separate the tosses.)

In the first example the event space for each time is {H,T} while for the second example the event space is {…-2, -1, 0, 1, 2, …}, i.e., the integers.

In both cases we notice that *S(t)* has a distribution indexed by time. In the case of Example 1, we have the following probability distributions, assuming a fair coin:

Prob{S(1) = H} = ½, Prob{S(1) = T} = ½
Prob{S(2) = H} = ½, Prob{S(2) = T} = ½
Prob{S(3) = H} = ½, Prob{S(3) = T} = ½
…

That is, Prob{*S(t)* = H} = Prob{*S(t)* = T} = ½ for all time.

In the case of Example 2, it is a bit more complicated. Let's call the initial starting point at number zero, time zero. Then we have:

Prob{S(0) = 0} = 1, and Prob{S(0) = i} = 0 for all i not equal to zero.
Prob(S(1) = 0} = 0, Prob{S(1) = -1} = Prob{S(1) = 1} = ½, and Prob{S(1) = i} = 0 for all other values of i.
Prob{S(2) = 0} = ½, Prob{S(2) = 1} = Prob{S(2) = -1} = 0, Prob{S(2) = 2} = Prob{S(2) = -2} = ¼, and Prob{S(2) = i} = 0 for all other values of i.
…and so on...

So, in Example 2, the distribution changes with respect to time.

Example 3: Suppose we toss a fair coin and keep track of number of heads like this:

"0:0:1:1:1:2:3..." corresponding to the outcome "TTHTTHH..."

Then the state space is the positive integers and we have the following probabilities.
Prob{S(1) = 0} = ½, Prob{S(1) = 1} = ½, and Prob{S(1) = i} = 0 for all other values of i.
Prob{S(2) = 0} = ¼, Prob{S(2) = 1} = ½, Prob{S(2) = 2} = ¼, and Prob{S(2) = i} = 0 for all other values of i.
Prob{S(3) = 0} = 1/8, Prob{S(3) = 1} = 3/8, Prob{S(3) = 2} = 3/8, Prob{S(3) = 3} = 1/8, and Prob{S(3) = i} = 0 for all other values of i.
...and so on...

Again, the distribution changes with time.

Stochastic process specification

A stochastic process is defined by giving rules and parameter values for computing *S(t)* for all time. In the above examples, the rules were given in words and the parameter for the process was the probability of heads for the coin toss. (Note, the time index is also a parameter, but we will assume that this is a given and not discuss it specially unless we need to.) In all of our computations, we assumed that we had a fair coin and used p = ½ for the probability of heads—but obviously this could be some other value.

Path of a stochastic process.

Sometimes we are interested in computing the probability of particular sequence of outcomes. We will call such sequence of outcomes a "*path*" of the stochastic process. For example, in the process described in Example 3, we might want to know the probability of the particular sequence "0:0:0"—that is, no heads in three throws. A little bit of thinking will show that this probability is simply ½ x ½ x ½ = 1/8.

Quantities of interest in a stochastic process

There are certain quantities that are useful for modeling or characterizing a stochastic process. The first is the ***marginal distribution***. The marginal distribution is simply the probability distribution of the states at some fixed time. For example, the marginal distribution of the Example 1 process at time 2 is Prob{S(2) = H} = ½ and Prob{S(2) = T} = ½. In fact, for this process Prob{*S(t)* = H} = Prob{*S(t)* = T} = ½ for all time t. A particular kind of marginal distribution of interest is that at time infinity. That is, we want the probability distribution at S(∞). We call this the ***equilibrium distribution***. For a given stochastic process, the equilibrium distribution may or may not exist. In the Example 1 stochastic process, the equilibrium distribution is Prob{S(∞) = H} = Prob{S(∞) = T} = ½. Consider a stochastic process that goes like this: "If the previous state was H then switch to T with probability 1; if it was T then switch to H with probability 1" That is, flip-flop between heads and tails (such deterministic process can also be seen as a special case of a stochastic process). This process has no equilibrium distribution since it is either H or T dependent on whether we have taken even number of time steps or odd number of time steps.

A more interesting quantity is the ***waiting time***. Here, we are interested in the time until some event of interest happens. First thing to note is that the time until some event occurrence is a random variable—that is, the waiting time is a random quantity with a distribution. In the Example 1 stochastic process, we could ask "what is the distribution of the waiting time until the first head appears?" Call the random variable for the waiting time W. Then,

Prob{W = 1} = ½ (Throw heads on the first trial.)
Prob{W = 2} = ¼ (Throw a tail then throw a head.)
Prob{W = 3} = 1/8 (Throw a tail, another tail, then throw a head.)
...and so on...

Types of stochastic processes

There is a very complex zoo of stochastic processes. The study of stochastic processes is one of the most complex areas in probability theory and we could teach a whole year course without making much of a dent. Here, I will just give a simple categorization and then concentrate on a particular class of processes that is commonly used in computational biology. The main classification we will use is (1) whether the state space is discrete or not; (2) whether the time index is counted discretely or continuously; and (3) how the probabilities at time t is computed. Items (1) and (2) are not fundamentally important from a model point of view and only categorizes the processes into what kinds of techniques are required for studying it. Item (3) is a modeling issue and some of the possible types are (a) probability of *S(t)* can be computed without explicit knowledge of any prior events (as in Example 1); (b) probability of *S(t)* depends on the state of the system at prior time steps; and (c) the probability of *S(t)* depends on the state of the system at prior time steps and the absolute value of the time itself. As an example of (c), one might imagine a stochastic process describing a poker table that changes betting rule after midnight.

Here, we will concentrate on a very simple but widely used stochastic process, the so-called ***Markov process***. In the Markov process, we are given rules by which we can compute the probability of *S(t)* given knowledge of the state of the system at some prior time *s*—but we do not need to know any other prior time states. Or equivalently, to compute the probability of *S(t)* at some future time, we only need to know the state of the system at the current time, but we do not need to know the system at any prior times. More precisely,

Let $F(S(t) \mid S(s))\ (s<t)$ be the conditional distribution of the state of the system at time *t* given the distribution at time *s*. Consider *n* time points prior to *s*, say $r_1..r_n$.
We say the stochastic process has the **Markov property** if

$$F(S(t) \mid S(s), S(r_1),\ldots,S(r_n)) = F(S(t) \mid S(s))$$

Within the class of Markov processes, we will mostly study the Markov process where the state space is discrete such as {A, C, G, T} since most of the genomics data have a discrete characterization. Such Markov processes are sometime called ***Markov chains.***

The most important part of specifying a Markov chain is the ***transition probabilities***. Suppose we denote the possible outcomes from the state space as 1, 2, 3...k, and so on. For example, we might say "A" is the 1st state, "C" is the 2nd state, "G" is the 3rd state, and "T" is the 4th state. Then the transition probabilities are given by filling in values for quantities of the form *P{ith state at time t+s | kth state at time t}*. That is, specifying the conditional probability that the system will be in the *i*th state at time *t+s* given that the system was in the *k*th state at time *t*. Note, that to fully specify the process we need to know these conditional probability values for all combination of states at the two time points. For example, if the state space is {A, C, G, T}, then we need to know the conditional probabilities for all 16 possible combination of the states at time *t+s* and *t*. If the number of state is finite, then we can write this conveniently in matrix form as:

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} & \cdots \\ t_{21} & t_{22} & \cdots & \cdots \\ t_{31} & t_{32} & & \\ \vdots & \vdots & & \end{bmatrix} \qquad (1)$$

where we read the ith row, jth column element as giving Prob{S = jth state at time *t+s* | S = ith state at time *t*}. Note the ith row denotes the all the possible transitions from ith state, thus these probabilities have to sum to one; i.e., the rows have to all sum to one. A non-negative matrix whose rows sum to one is called a ***stochastic matrix***, because they can be used to represent transition probabilities for a stochastic process.

If time is counted discretely then it is natural to only consider the transition probabilities in increments of one time step—all the other time steps can be easily computed from this "one time step" transition probability by the Markov property.

Our goal is to compute *S(t+1)* given what we know about *S(t)*. Suppose we would like to know *P(S(t+1) = i).*

(From here on we will just use "i" to denote "ith state" and *P()* to denote $Prob\{\}$.)

Then, one possibility is that at time t, the system was in state 1 and then subsequently made a transition to the ith state. The probability of this kind of scenario is P(S(t+1) = i | *S(t)* = 1)P(*S(t)* = 1). That is, the probability that the system was in state 1 at time t multiplied by the conditional probability that the system jumps to state i at time t+1. We know how to compute this if we know the transition probabilities and the marginal probability at time t. [Thus, for a Markov process we can compute probabilities at time t, if we know the transition probabilities and the marginal probability at some prior time s (s<t).]

Another possible scenario is that the system was in state 2 at time t and then made a transition to state i at time t+1. By the same reasoning this probability is P(S(t+1) = i | *S(t)* = 2) P(*S(t)* = 2). Now we can see that the probability that the system is in ith state at time t+1 requires us to consider the possibility that the system was in kth state at time t, and made the k → i transition for all possible k. That is, we need to consider all the possible ways in which we arrive at ith state from the system at the previous time. Thus,

$$P(S(t+1)=i)=\sum_{k=1}^{n} P(S(t+1)=i \mid P(S(t)=k)P(S(t)=k) \qquad (2)$$

Formulas of the form (2) can be compactly represented by matrix algebra. Suppose we use the row vector $\vec{s}(t)=(s_1^t, s_2^t, \cdots, s_n^t)$ to denote marginal probabilities of ith states at time t. (Note, we used the superscript t, just to keep track of the fact that we are discussing probabilities at time t.) Also, suppose that we use matrices of the form (1) to denote transition probabilities. Then,

$$\vec{s}(t+1)=\vec{s}(t)\cdot T$$

$$\Rightarrow$$

$$(s_1^{t+1}, s_2^{t+1}, \cdots, s_n^{t+1}) = (s_1^t, s_2^t, \cdots, s_n^t)\begin{bmatrix} t_{11} & t_{12} & t_{13} & \cdots \\ t_{21} & t_{22} & \cdots & \cdots \\ t_{31} & t_{32} & & \\ \vdots & \vdots & & \end{bmatrix}$$

That is, the usual matrix-vector multiplication is in fact equivalent to doing (2) type of computations.

Now we can use the Markov property to compute such probabilities for time t+s given the marginal distribution at time t. Namely we see the following sequence:

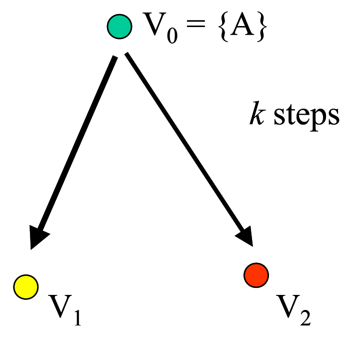
$$\begin{aligned}
&\vec{s}(t+s) = \vec{s}(t+s-1)\cdot T\\
&\vec{s}(t+s-1) = \vec{s}(t+s-2)\cdot T\\
&\vdots\\
&\vec{s}(t+1) = \vec{s}(t)\cdot T\\
&\Rightarrow\\
&\vec{s}(t+s) = \vec{s}(t)\cdot T^{s}
\end{aligned}$$

That is, we take the product of the T matrix s times.

Side Note: Suppose we are faced with taking the product of a matrix A, say 100 times. It is not computationally efficient to actually do the matrix multiplication 100 times. What is useful to note is that we can do such multiplications like this:

$$\begin{aligned}
&B = A\cdot A\\
&C = B\cdot B = A^{4}\\
&D = C\cdot C = A^{8}\\
&E = D\cdot D = A^{16}\\
&F = E\cdot E = A^{32}\\
&G = F\cdot F = A^{64}\\
&A^{100} = G\cdot F\cdot C
\end{aligned}$$

We can now apply the computational rules to a problem in molecular evolution. Suppose we are looking at a sequence position. Somehow, we know that there is an ancestral sequence with the letter "A" at this position. This ancestor leaves two descendents k generations in the future as shown in the figure below. We wish to compute the probability of identity at this sequence position for the two descendents. We will also assume that the transition matrix has the form given in the picture.



	A	C	G	T
A	1-3p	p	p	p
C	p	1-3p	p	p
G	p	p	1-3p	p
T	p	p	p	1-3p

The first thing to note is that the sequence position at the descendents V1 and V2 will be identical if the following four cases occur: {V1=A and V2 = A}, {V1=C and V2 = C}, {V1 = G and V2 = G}, and {V1 = T and V2 = T}. If we know the probabilities of these four cases then the probability of identity is given as the sum since these are mutually exclusive events. Let's consider one of the cases and compute P(V1 = A and V2 = A| V0 = A), that is the probability that both of the descendents have "A" in this position given that the ancestor had an "A". After a little bit of thinking we realize that this is given by:

$$P(V1 = A \text{ and } V2 = A \mid V0 = A\} = P(V1 = A \mid V0 = A)\, P(V2 = A \mid V0 = A) \qquad (3)$$

because once we condition on the ancestor being letter "A", what happens in the lineage to V1 is independent of what happens in lineage V2. (For the more advanced students, you should think about whether this statement is still true if we didn't condition on the ancestral state.)

We also note that P(V1= A| V0 = A) = P(V2 = A| V0 = A), since in our simple model the process along the V1 lineage is the same process as that along V2 lineage. So now we only need to compute P(V1=A| V0 = A). We can compute this by taking the matrix shown in the figure and taking its kth power and looking at the row 1, col 1 element, because this will correspond to the conditional probability of starting with A, and after k steps, ending with A. Let's denote this by $t_{AA}(k)$, meaning the probability of starting with A and ending with A after k steps. Similarly, we will denote $t_{AC}(k)$ to mean the probability of starting with A and ending with C after k steps, and so on. Then (3) is computed as $P(V_1 = A \text{ and } V_2 = A \mid V_0 = A) = t_{AA}(k) \cdot t_{AA}(k)$. We have similar reasoning for the cases where we end up with C and C at the two descendents, and G and G, etc. So the desired probability of identity is:

$$P\{Identity\} = t_{AA}(k) \cdot t_{AA}(k) + t_{AC}(k) \cdot t_{AC}(k) + t_{AG}(k) \cdot t_{AG}(k) + t_{AT}(k) \cdot t_{AT}(k)$$

Continuous time

In the computations above, we dealt with cases when time was indexed in a discrete manner so we used simple matrix algebra to compute "*s* time steps" and so on. We run into technical difficulties if time is given as a continuous number. For example, we don't know how to take matrices to the 1.2 power. We are also unsure that if we take a stochastic matrix to, say 0.4, power, whether we will end up with another matrix that is a stochastic matrix (rows summing to one). It turns out that for continuous time Markov chain we have to use the special formula:

$$\mathbf{P(t+s)} = e^{\mathbf{R}s}\mathbf{P(t)} \qquad (4)$$

where **R** is a matrix that we call the rate matrix. Elements of the rate matrix are real valued numbers (both positive and negative numbers) and the rows of a rate matrix sum to zero. We interpret the i,j th element to mean the instantaneous rate of transition from ith state to the jth state. The quantity *exp(**R**s)* means "Take the matrix exponential of the rate matrix time the scalar number s". We discussed matrix exponentially in the "Matrix" handout. Continuous time models are commonly used in molecular evolution models and we will discuss them again later.