Lecture 2: Absorbing states in Markov chains. Mean time to absorption. Wright-Fisher Model. Moran Model.

Antonina Mitrofanova, NYU, department of Computer Science

December 18, 2007

1 Higher Order Transition Probabilities

Very often we are interested in a probability of going from state i to state j in n steps, which we denote as $p_{ij}^{(n)}$. For example, the probability of going from the state i to state j in two steps is:

$$p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}$$

where k is the set of all possible states. In other words it consists of probabilities of going from state i to any other possible state (in one step) and then going from that step to j. Interestingly, the probability $p_{ij}^{(2)}$ corresponds to (i, j)'th entry in the matrix

$$P^2 = P \times P$$

Similarly, going from i to j in n steps is defined as

$$p_{ij}^{(n)} = \sum_k p_{ik} p_{kj}^{(n-1)} = \sum_k p_{ik}^{(n-1)} p_{kj}$$

and correspond to the (i, j)'th entry in P^n matrix (therefore we can compute transition probabilities by taking matrix powers).

We define the n-step transition probabilities for the Markov Chain by

$$P[X_n = j | X_0 = i]$$

Since we know that

$$P[X_{n+1} = k_1, \ldots X_{n+m} = k_m | X_n = i] = P[X_1 = k_1, \ldots, X_m = k_m | X_0 = i]$$

then

$$P[X_n = j | X_0 = i] = p[X_{n+m} = j | X_m = i]$$

In other words, the probability that a path started at i and ended at j does not depend on the time at which it is initiated. Therefore

$$p_{ij}^{(n)} = P[X_n = j | X_0 = i]$$

We can also write

$$p_{ij}^{(n+m)} = \sum_k p_{ik}^{(n)} p_{kj}^{(m)}$$

which is also called *Chapman-Kolmogorov* equation.

It follows that we can compute a *uniconditional* probability (of X_n taking a value of j)as

$$P[X_n = j] = p_j^{(n)} = \sum_i p_i p_{ij}^{(n)}$$

Frequently we are interested in the time the system goes from some initial state to some terminal critical state, called absorbing state.

2 Absorption Probabilities

The state i is called absorbing if $p_{ii} = 1$. In other words, once the system hits state i, it stays there forever not being able to escape.

The most interesting absorbing states which arise in population genetics are at $i = 0$ and at $i = M$. Let us assume the population of 2N haploids , each having either allele A_1 or allele A_2. In this case $M = 2N$. And let X be a random variable which describes the frequency of allele A_1 in a given population. In population genetics, it is of interest to find out how fast the allele will go to either absorption state ($i = 0$ or $i = 2N$) given that the population started with i alleles A_1.

Many functionals (including absorption probabilities) on Markov Chain are evaluated by a technique called *first step analysis*. This method proceeds by the analyzing the possibilities that can arise at the end of the first transition.

Let us now fix k as absorbing state. The probability of absorption in this state depends on the initial state $X_0 = i$. Let us define U_{ik} as the probability of eventually reaching state k given that we started in state i. First possibility would be to go from state $X_0 = i$ to $X_1 = k$ immediately, which is described by p_{ik}. However, if the state k is not entered at X_1, then we must go to some other state $j \neq k$. Once we enter state j, the probability of ultimate absorption in k is U_{jk} by definition. Weighting all the possibilities gives

$$U_{ik} = p_{ik} + \sum_{j=0 \& j \neq k}^{M} p_{ij} U_{jk} = \sum_{j=0}^{M} p_{ij} U_{jk} \tag{1.1}$$

since $U_{kk} = 1$.
If we write u_i as the probability of absorption (say, in 2N), then we write

$$u_i = \sum_{j=0}^{M} p_{ij} u_j \tag{1.2}$$

3 Mean time until absorption

It is more difficult to assess the properties of the (random) time until absorption. However, it is common to evaluate the ***mean time*** until X reaches 0 or $2N$ starting from i (it is called mean absorption time).

We wish to make a calculation of a mean time until absorption for a general starting point i. We now introduce a concept, which is central in calculating the mean absorption time: Let us observe that starting from i the system will visit state j some number of times before absorption. This fact it true for all j (except 0 and 2N). Therefore, if we know the number of times the system visits state j (for all j) before absorption, then we can obtain an average time until absorption by summing up over average times the system is in a specific state, for each state.

Let us now formally define mean number of times that X takes the value j before absorbtion in 0 or $2N$ (given that it started in i) as $\{\bar{t}_{ij}\}$. Then the mean time to absorption given that we started at state i is the sum:

$$\bar{t}_i = \sum_{j=1}^{M-1} \bar{t}_{ij}$$

We will proceed by a first step analysis: if the system starts at i and then proceeds to k at its next step, then the mean number of visits to state j prior to absorption starting from state k now is $\{\bar{t}_{kj}\}$ by definition. However, observe that in this case, we miss the case when $i = j$. Therefore, there is a need to define an additional variable $\delta_{ij} = 1$ if $i = j$, and 0 otherwise. As a result, weighting by the probability of going to a state k at the first step, we obtain

$$\bar{t}_{ij} = \sum_{k=0}^{M} p_{ik} \bar{t}_{kj} + \delta_{ij}, \qquad \bar{t}_{0j} = \bar{t}_{Mj} = 0$$

By summing up the meant times the system is in state j for all j, we obtain

$$\sum_{j=1}^{M-1} \bar{t}_{ij} = \sum_{j=1}^{M-1} \sum_{k=0}^{M} p_{ik} \bar{t}_{kj} + \sum_{j=1}^{M-1} \delta_{ij}$$

By definition $\sum_{j=1}^{M-1} \bar{t}_{ij} = \bar{t}_i$ so that

$$\bar{t}_i = \sum_{k=0}^{M} p_{ik} \bar{t}_k + 1, \qquad \bar{t}_0 = \bar{t}_M = 0$$

Observe that since $\sum_{j=1}^{M-1} \delta_{ij}$ will equal to 1 only once (when $i = j$), then $\sum_{j=1}^{M-1} \delta_{ij} = 1$ in the above equation.

Now, we will discuss two most important models in population genetics, which describe the evolution of allele frequencies: Wright-Fisher model and Moran model.

4 Wright-Fisher Model

Let us assume a simple haploid model of the population of 2N genes (or alternatively – N diploid organisms) of random reproduction, with each haploid possessing either allele A_1 or allele A_2. Let us, for start, disregard mutation pressures and selective forces. At every time-step, each gene (allele) gives birth to some number of offsprings (which are the exact copies of himself) and dies immediately after that, thus living only one generation. This process describes how the genes get transmitted from one generation to the next. However, the processed of birth and death will have to remain unseen behind the curtain for a moment. Instead, we will only observe how the frequency of alleles will change from generation to generation. [1]

We will fix our attention at frequency of allele A_1 in the population of $2N$ haploids. Let us think of this process of going from one generation to the other as a Markov Chain, where the state X of the chain corresponds to the number of haploids (genes) of type A_1. Clearly, in any generation X takes one of the values $0, 1, \ldots, 2N$, which constitutes a state space. We will denote the value taken by X in generation t as X_t.

The model assumes that genes for the generation $t + 1$ are derived by sampling with replacement from the genes of generation t. Thus, the make up of the next generation is determined by $2N$ independent Bernoulli trials so that X_t is a binomial random variable. Let the initial generation consist of i genes of type A_1 and $2N - i$ genes of type A_2. Then we define a probability of success (resulting in allele A_1) p_i and a probability of failure q_i (resulting in allele A_2) for each Bernoulli trial as

$$p_i = \frac{i}{2N} \qquad q_i = 1 - \frac{i}{2N}$$

We generate a Markov Chain $\{X_n\}$, where X_n is the number of A_1 genes in the n'th generation, among a constant population size of 2N individuals. Basically, X_{t+1} is a binomial random variable with index $2N$ and parameter (probability of success) $X_t/2N$. Observe that the transition probabilities from $X_t = i$ to $X_{t+1} = j$ for this Markov Chain are computed according to the binomial distribution as

$$P(X_{t+1} = j | X_t = i) = p_{ij} = \binom{2N}{j} p_i^j q_i^{2N-j} = \binom{2N}{j} (i/2N)^j \{1 - (i/2N)\}^{2N-j}$$

Observe now that states 0 and 2N are completely absorbing. In other words, no matter what the value of X_0 is, eventually X_t will take the value 0 or $2N$. And once this happens, X will stay in that state forever. In the case of $X_t = 0$, the population will consist only of A_2 genes, while in the case of $X_t = 2N$ the population will be purely A_1-gene population.

[1] In this model, the allele frequency of the next generation is manipulated mainly by a genetic drift (a Genetic drift can be defined as a force that reduces heterozygosity by the random loss of alleles).

4.1 Absorption probability in Wright-Fisher model

Let us discuss absorption probabilities in Wright-Fisher model. In fact, the population can attain fixation and be composed of only A_1-genes ($X_t = 2N$) or A_2-genes ($X_t = 0$). It can be shown that with probability one, either of the absorbing states (either 0 or 2N) is eventually entered (and this is true for both Wright-Fisher and Moran models). And therefore $lim_{t\to\infty} P(X_t = j) = 0$.

We will discuss two cases of absorption: at 0 and at 2N.

4.1.1 Absorption at zero:

We write the probability of extinction (absorption at 0) of a gene given that it started with i copies as $lim_{n\to\infty} P(X_n = 0|X_0 = i)$

Let us find the probability of absorbing in state 0 by using $E(X_n)$. We express $E(X_n)$ by using the expectation of a conditional expectation

$$E(X_n) = E[E(X_n|X_{n-1})] = E(X_{n-1}) = E(X_{n-2}) = \cdots = E(X_0) = i$$

This property is called the *constancy of expectation* (and is also true for Moran model). Further we can write $E(X_n)$ as

$$E(X_n) = 0 \times u_{i,0} + 2N \times (1 - u_{i,0})$$

Now, since $lim_{n\to\infty} E(X_n) = i$,

$$i = 0 \times u_{i,0} + 2N \times (1 - u_{i,0})$$

and therefore

$$u_{i,0} = \frac{2N - i}{2N}$$

Observe that we ignore $P(X_t = j)$ since they are equal to zero as n goes to infinity.

4.1.2 Absorption at 2N:

We want to calculate probability that A_1 eventually becomes fixed in the population (absorption at 2N)and follow a similar argument:

$$i = 0 \times (1 - u_{i,2N}) + 2N \times u_{i,2N}$$

so that

$$u_{i,2N} = \frac{i}{2N}$$

A different argument (which is more relevant to a genetical point of view) is that eventually every gene in the population is descended from the unique gene in generation zero. The probability that such a gene (allele) is A_1 is simply the initial fraction of A_1 alleles, namely $i/2N$, and this also must be a fixation probability of allele A_1.

4.1.3 Absorption starting with one A_1 allele

Suppose that in a population of pure A_2 alleles a single new mutant A_1 allele (gene) arises. There are no more new mutations and therefore we can assume that we start with a population with one A_1 allele and $2N - 1$ A_2 alleles. According to the previous result, the probability of fixation for this allele is

$$u_{1,2N} = \frac{i}{2N} = \frac{1}{2N}$$

. On the other hand, the probability that the allele is lost is $1 - 1/2N$.

4.2 Mean time until absorption in Wright-Fisher model

The calculation of the mean time until absorption for the Wright-Fisher model is very computationally expensive. Therefore, it is useful to approximate this quantity, which will be described in the next section. However, it is relatively simple to derive the mean time until absorption starting with one allele of type A_1 (before the mutant is lost or before the mutant is fixed).

For this calculation we use the same analysis as before. We will use the expected number of visits to a state j along the path to absorption, starting from state $X_0 = 1$. We denote the mean number of generations to absorption in 0 or 2N , given that we started with one allele A_1, as $\bar{t}_1$. We need to sum up the expected number of such visits for all j, avoiding states 0 and $2N$:

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j}$$

where $\bar{t}_{1,j}$ is the mean number of times when the number of A_1 alleles takes the value of j (system is in state j) before reaching either 0 or $2N$. Both Fisher and Wright found that

$$\bar{t}_{1,j} \approx \frac{2}{j}$$

starting at $i = 1$. Since $\sum_{i=1}^{N} \frac{1}{i} = log(N) + \gamma$ where γ is a Euler's constant (0.5772 ...), we derive

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j} = \sum_{j=1}^{2N-1} \frac{2}{j} = 2 \sum_{j=1}^{2N-1} \frac{1}{j}$$

$$= 2(log(2N-1) + \gamma)$$

4.3 Approximating Mean Time until Absorption

Even though in principle we can find solutions of the mean time to absorption for a general i, in practice these solutions seem extremely difficult, and simple expressions for these mean times have not yet been found. Let us now present a simple approximation for $\bar{t}_i$. We again apply the first step analysis , where we start from state i and in the first step attend some intermediate step k. We define $M = 2N$, $i/M = x$, $k/M = x + \delta x$, and $\bar{t}_i = \bar{t}(x)$.

Then we can rewrite the equation

$$\bar{t}_i = \sum_{k=0}^{M} p_{ik} \bar{t}_k + 1$$

as

$$\bar{t}(x) = \sum P\{x \to x + \delta x\} \bar{t}(x + \delta x) + 1 \tag{1}$$

$$= E\{\bar{t}(x + \delta x)\} + 1 \tag{2}$$

$$\tag{3}$$

Now assuming that $\bar{t}(x)$ is a twice differentiable function of a continuous variable x, we can use Taylor series to approximate the above quantity. The Taylor series states that

$$f(y) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (y-a)^n \tag{4}$$

$$= \frac{f(a)}{0!}(y-a)^0 + \frac{\{f(a)\}'}{1!}(y-a)^1 + \frac{\{f(a)\}''}{2!}(y-a)^2 + \dots \tag{5}$$

$$= f(a) + \{f(a)\}'(y-a) + \frac{1}{2}\{f(a)\}''(y-a)^2 + \dots, \tag{6}$$

We re-write $\bar{t}(x)$ by applying Taylor's series (showing three leading terms):

$$\bar{t}(x) = E\{\bar{t}(x+\delta x)\} + 1 \tag{7}$$

$$\approx \bar{t}(x) + E(\delta x)\{\bar{t}(x)\}' + \frac{1}{2}E(\delta x)^2\{\bar{t}(x)\}'' + 1, \tag{8}$$

$$\tag{9}$$

All expectations are conditional on x. Using the fact that the expectation of the binomial random variable is $E(Y) = np$, we can rewrite then

$$E(x+\delta x) = E(j/M) = \frac{E(j)}{M} = \frac{M \times \frac{i}{M}}{M} = \frac{i}{M}$$

, where $p = \frac{i}{M}$. We can as well use the fact established before that $E(X_n|X_{n-1} = i) = i$. At the same time $x = \frac{i}{M}$ and $E(x) = x$; therefore $E(\delta x) = 0$. As a result the term $E(\delta x)\{\bar{t}(x)\}' = 0$.

Let us calculate $E(\delta x)^2$. In our case, the $E(\delta x)^2 = Var(\delta x)$ since $Var(\delta x) = E(\delta x)^2 - [E(\delta x)]^2$ and $[E(\delta x)]^2 = 0$, as shown from the previous result.

The variance of the binomial random variable

$$Var(x+\delta x) = Var(j/2N) = \frac{Var(j)}{4N^2} = \frac{2N\frac{i}{2N}(1-\frac{i}{2N})}{4N^2} = \frac{x(1-x)}{2N}$$

.

The above gives

$$\bar{t}(x) \approx \bar{t}(x) + \frac{1}{2}\frac{x(1-x)}{2N}\{\bar{t}(x)\}'' + 1$$

$$-1 \approx \frac{1}{2}\frac{x(1-x)}{2N}\{\bar{t}(x)\}''$$

$$-4N \approx x(1-x)\{\bar{t}(x)\}''$$

The solution to this equation, subject to the boundary conditions $\bar{t}(0) = \bar{t}(1) = 0$ is

$$\bar{t}(x) = \int\int -4N\frac{1}{x(1-x)} \tag{10}$$

$$= -4N\int\int \frac{1}{x(1-x)} \tag{11}$$

$$= -4N\int\int \left(\frac{1}{x} + \frac{1}{1-x}\right) \tag{12}$$

$$= -4N\int \left(\int \frac{1}{x} + \int \frac{1}{1-x}\right) \tag{13}$$

$$= -4N\int (ln(x) + ln(1-x)) \tag{14}$$

$$= -4N\left\{\int ln(x) + \int ln(1-x)\right\} \tag{15}$$

$$= -4N\left\{(xln(x) - x) + ((1-x)ln(1-x) - (1-x))\right\} \tag{16}$$

$$\approx -4N\{xlogx + (1-x)log(1-x)\} \tag{17}$$

where $x = \frac{i}{2N}$, the initial frequency of allele A_1. This is also is called *diffusion approximation to the mean absorption time*.

When we initially start with one A_1 allele $x = \frac{1}{2N}$, then the mean time to absorption is

$$\bar{t}\left(\frac{1}{2N}\right) = -4N\{\frac{1}{2N}log\left(\frac{1}{2N}\right) + (1-\frac{1}{2N})log\left(1-\frac{1}{2N}\right)\} \tag{18}$$

$$\approx 2 + 2log2N \tag{19}$$

$$\tag{20}$$

At the same time, when $x = \frac{1}{2}$, then

$$\bar{t}(\frac{1}{2}) \approx 2.8N$$

Observe that for equal initial frequencies ($x = \frac{1}{2}$), the mean time is relatively long.

5 Moran model

In each generation of the Moran model, one gene is chosen at random to give 2 offsprings and one gene is chosen to die (all other genes survive to the next generation). As opposed to Wright-Fisher model, Moran model has overlapping generations. This model is also known as a birth-and-death model. We still consider a constant population size of $2N$ haploids, each of which has either allele A_1 or allele A_2. Let us (for now) ignore mutation or selection pressures.

Again, we define X to be a random variable, which represents the number of alleles of type A_1 in the population. It is of interest to calculate transition probabilities for the implied Markov chain. Suppose that in population t (which corresponds to state X_t in underlying Markov chain) the number of alleles A_1 is i. Then in population $t+1$, the number of alleles A_1 can be either ($j = i-1$), ($j = i+1$), or $j = i$.

The system can go from i to $i-1$ if A_2 individual is chosen to give 2 offsprings and A_1 individual is chosen to die:

$$p_{i,i-1} = \left(\frac{2N-i}{2N}\right)\left(\frac{i}{2N}\right)$$

To go from i to $i+1$, the opposite should be true: A_2 is chosen to die and A_1 is chosen to reproduce:

$$p_{i,i+1} = \left(\frac{2N-i}{2N}\right)\left(\frac{i}{2N}\right)$$

and for going from i to i, it takes either A_1 to reproduce and die or A_2 to reproduce and die:

$$p_{i,i} = \left(\frac{2N-i}{2N} \times \frac{2N-i}{2N}\right) + \left(\frac{i}{2N} \times \frac{i}{2N}\right) = \frac{i^2 + (2N-i)^2}{(2N)^2}$$

Observe that $p_{ij} = 0$ for all other values of j since it is impossible to make other transitions.

5.1 Properties of a Continuant matrix in Moran model.

In the case of Moran model, the transition probability matrix is Continuant, which means that $p_{ij} = 0$ iff $|i-j| > 1$. Now we can apply the standard Continuant matrix theory to our model so that the probability of fixation and mean time to absorption can be found explicitly.

In particular, we can use a "birth-and-death" process concepts to calculate the desired quantities. The birth-death process is a special case of Continuous-time Markov process where the states represent the current size of a population and where the transitions are limited to births and deaths. When a birth occurs, the state i goes to state $i+1$, defined by the birth rate $p_{i,i+1} = \lambda_i$. When the death occurs, the process goes from state i to state $i-1$, defined by the death rate $p_{i,i-1} = \mu_i$.

For now, we will just use facts from the theory of birth-death processes, without proving them; however, we would like to return to formal definitions and proofs in one of the future lectures.

We define

$$\rho_i = \frac{\mu_1\mu_2\ldots\mu_i}{\lambda_1\lambda_2\ldots\lambda_i}$$

and $\rho_0 = 1$. If 0 and $M = 2N$ are both absorbing states, then the probability of absorption in either of them becomes

$$u_i = \sum_{k=0}^{i-1} \rho_k / \sum_{k=0}^{M-1} \rho_k$$

Proceeding further with the above argument, we can calculate the mean number of times the system is in state j given that it started in state i as

$$\bar{t}_{ij} = \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j}, (j = 1, \ldots, i)$$

$$\bar{t}_{ij} = \frac{u_i \sum_{k=j}^{M-1}\rho_k}{\rho_j \lambda_j}, (j = i+1, \ldots, M-1)$$

from this we can derive $\bar{t}_i$

$$\bar{t}_i = \sum_{j=1}^{M-1} \bar{t}_{ij} = \sum_{j=1}^{i} \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j} + \sum_{j=i+1}^{M-1} \frac{u_i \sum_{k=j}^{M-1}\rho_k}{\rho_j \lambda_j}$$

5.2 Probability of fixation in Moran model

Let us now observe that in the Moran model

$$\lambda_i = \mu_i = i(2N-i)/(2N)^2$$

so that

$$\rho_i = \frac{\mu_1}{\lambda_1}\frac{\mu_2}{\lambda_2}\cdots\frac{\mu_i}{\lambda_i} = 1$$

for $i = 0, 1, \ldots, 2N$. It can be shown that, similarly to Wright-Fisher model, $E(X_t) = i$.

Following all of the above, the probability of fixation (given that we started with i copied of A_1) is

$$u_i = \frac{i}{2N}$$

given that $M = 2N$.

5.3 Expected absorption time

Using the fact that $\rho_i = 1$, we can derive the following

for $j = 1, \ldots, i$

$$\bar{t}_{ij} = \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j} \tag{21}$$

$$= \frac{(1-\frac{i}{2N})\sum_{k=0}^{j-1}1}{1\frac{2N-j}{2N}\frac{j}{2N}} \tag{22}$$

$$= \frac{(\frac{2N-i}{2N})j \times 2N \times 2N}{(2N-j)j} \tag{23}$$

$$= \frac{(2N-i)\times 2N}{2N-j} \tag{24}$$

and for $j = i+1, \ldots, M-1$

$$\bar{t}_{ij} = \frac{u_i \sum_{k=j}^{M-1} \rho_k}{\rho_j \lambda_j} \tag{25}$$

$$= \frac{\frac{i}{2N} \sum_{k=j}^{M-1} 1}{1 \frac{2N-j}{2N} \frac{j}{2N}} \tag{26}$$

$$= \frac{\frac{i}{2N}(2N-j) \times 2N \times 2N}{(2N-j)j} \tag{27}$$

$$= \frac{i \times 2N}{j} \tag{28}$$

$$\tag{29}$$

Now we can calculate the expected time to absorption as

$$\bar{t}_i = \sum_{j=1}^{M-1} \bar{t}_{ij} \tag{30}$$

$$= \sum_{j=1}^{i} \frac{(1-u_i) \sum_{k=0}^{j-1} \rho_k}{\rho_{j-1} \mu_j} + \sum_{j=i+1}^{M-1} \frac{u_i \sum_{k=j}^{M-1} \rho_k}{\rho_j \lambda_j} \tag{31}$$

$$= \sum_{j=1}^{i} \frac{(2N-i) \times 2N}{2N-j} + \sum_{j=i+1}^{M-1} \frac{i \times 2N}{j} \tag{32}$$

$$= (2N-i)2N \sum_{j=1}^{i} \frac{1}{2N-j} + 2Ni \sum_{j=i+1}^{M-1} \frac{1}{j} \tag{33}$$

It is possible to condition on the fact that, say A_1 eventually fixes. In this case, it is not difficult to derive simpler expressions for mean absorption times. Such derivations will be discussed in the future lectures.

References

[1] Warren Ewens, "mathematical Population Genetics", Second edition, 2004, pp 20-23, 86-91, 92-99, 104-109.

[2] Howard Taylor and Samuel Karlin, "An Introduction to Stochastic Modeling", Third edition, 1998, pp 95-138.

[3] Sidney Resnick, "Adventures in Stochastic Processes", 1992, pp72-116.