Applications of Markov Chains in Genetics (MA597 Assignment)

Mehta Apurva R.(06212318)

November 25, 2007

1 Introduction:

Markov chain models have been the most widely used ones in the study of random fluctuations in the genetics compositions of populations over generations. Besides being a convenient theoretical tool, markov chains have provided rather satisfactory theoretical explanations to some observed long run phenomena related to the genetic structure of populations.

Here we shall discuss some of the fundamental and classical research done in this area with emphasis on the most pioneering classical work done by **S.Wright** and **R.A.Fisher**, presently known as the **"Wright-Fisher Model".** We shall first discuss two relatively simple models, namely **selfing** and **sibmating.**

2 Selfing:

For an autosomal gene with two alleles A and a we have three genotypes AA, Aa and aa. Let us name the states as 1, 2, 3. Let us consider selfing and follow a lone of descent. Thus, if an individual is in state 1 or 3 then all his descendants will be in the same state. If on the other hand, an individual is in state 2 then his descendant in the next generation will be in the states 1 or 3 with the probability 1/4 each or in state 2 with the probability 1/2. Thus we have a Markov chain (X_n) with three states 1, 2 3. Its transition matrix is given as follows,

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 0 & 1 \end{pmatrix}$$

In this entire analysis, we assume that there are no mutations or fitness constraints. The states 1 and 3 are absorbing while state 2 is transient. This matrix is simple enough to allow a full analysis as follows. The matrix has eigenvalues 1, 1 and 1/2. The corresponding right eigenvectors are $(1,1,1)$,$(1,2,3)$ and $(0,1,0)$ while the left eigenvectors are $(3/2,0,-1/2)$,$(-1/2,0,1/2)$ and $(-1/2,1,-1/2)$. Thus, P can be diagonalized as follows,

$$P = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 3/2 & 0 & -1/2 \\ -1/2 & 0 & 1/2 \\ -1/2 & 1 & 1/2 \end{pmatrix} = \Lambda D \Lambda^{-1}$$

This makes it possible to explicitly calculate the n-step transition matrix $P^n = \Lambda D^n \Lambda^{-1}$ from which it easily follows that,

$$p_{21}^{(n)} = \frac{1}{2}[1 - 2^{-n}],\ p_{22}^{(n)} = 2^{-n}\ and\ p_{23}^{(n)} = \frac{1}{2}[1 - 2^{-n}]$$

$$p_{11}^{(n)} = p_{33}^{(n)} = 1;$$

$$b_{12} = P(X_n is\ eventually\ 1 | X_0 = 2) = \frac{1}{2},$$

$$b_{32} = P(X_n is\ eventually\ 3 | X_0 = 2) = \frac{1}{2}.$$

One can of course get all these by direct probabilistic calculations without bringing the matrix P or P^n. Thus starting from state 2, the absorption probabilities to the two states 1 and 3 are $1/2$ each as is expected from symmetry. Let T be the absorption time that is, T=n iff $X_n = AA$ or aa but $X_{n-1} = Aa$.Then $P(T = n | X_0 = 2) = (1/2)^{n-1} - (1/2)^n = (1/2)^n$. One easily has $E(T|X_0 = 2) = 2$.That is starting from state 2, the system takes two generations on an average to get absorbed in one of the two states 1 or 3.

The above Markov Chain (X_n)models the genotype sequence of a line of descent under selfing.

3 Sibmating:

In case of selfing an individual has only one parent and the transition from the genotype of the father to the genotype of the offspring is modeled by a Markov chain. But in sibmating each of the offspring has two parents and hence the genotype of an individual depends on those of both its parents. It is therefore evident that simply the individual genotype changes from generation to generation can not form a Markov chain. To build a markovian model, we consider the evolution of genotypic pairs of sibs. In other words, we look at the line of descent of sibs as follows. Consider two sibs of a generation; form their offsprings select two sibs at random; form their offsprings again select two sibs at random and so on. For instance if the present sibs are (Aa, Aa) then their offsprings consists of $\frac{1}{4}AA + \frac{1}{2}Aa + \frac{1}{4}aa$ in both males and females, so that if two independent choices are made one from males and one from females then the sibs so formed will be of type (AA,AA) with chance 1/16, of type (AA,Aa) with chance 1/4 etc. While considering genotypic pairs of sibs, we do not attach any importance to which on e of the pairs is a male member and which one is female.

Thus the state space of the Markov chain is {(AA, AA), (aa, aa), (AA, aA), (aa, aA), (aA, aA), (aa, AA)} numbered as 1,2,3,4,5 and 6. The transition matrix is

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/2 & 0 & 0 & 1/4 \\ 0 & 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1/16 & 1/16 & 1/4 & 1/4 & 1/8 & 1/4 \end{pmatrix}$$

A direct calculation shows that P has eigenvalues $\lambda_1 = \lambda_2 = 1$,$\lambda_3 = \frac{1+\sqrt{5}}{4}$,$\lambda_4 = \frac{1}{2}$,$\lambda_5 = \frac{1}{4}$ and $\lambda_6 = \frac{1-\sqrt{5}}{4}$. Thus we conclude that the rate of absorption is given by $\frac{1+\sqrt{5}}{4}$.

4 The Wright-Fisher Model:

All the previous analysis before the Wright-Fisher Model failed to capture the phenomena of genetic evolution in finite populations, where the sampling fluctuations play a central role. This suggests adapting models that capture this component of randomness in evolution. This was already realized by Pearson, Yule and Fisher himself earlier. Since then this model, now known as the Wright-Fisher Model, has occupied the centre stage in the mathematical models of genetics. We proceed to describe this model.

Let us consider, as usual an autosomal gene with two alleles A and a, so that there are three genotypes AA, Aa and aa. We wish to study the evolution of genotypic frequencies in a given population. Ideally what we wish to do is the following. Suppose that initially there are N_1males with composition $N_{11}AA + N_{12}Aa + N_{13}aa$ and N_2females with composition $N_{21}AA + N_{22}Aa + N_{23}aa$. Let us assume random mating. For the k-th generation we want to affect some simplifications.

Let us assume that for all k, $N_{1i}^k = N_{2i}^k$ for i=1,2,3, that is in all generations the genotypic frequencies are the same for both the sexes. Of course, this will imply that $N_1^k = N_2^k = N^k$,say. To put it differently, we consider unisex population, as for example, plants. Then the problem simplifies to describing the 3-tuple (N_1^k, N_2^k, N_3^k). As a further simplification we assume that $N^k = N$for all k, that is variation in the total population size is also ruled out. This may look like a gross over simplification, far removed from reality. However, it can be given the following interpretation. Imagine a "real" population evolving in time with possibly changing size. But to facilitate calculations we concentrate on N individuals randomly chosen from each generation. The cautious reader would of course realise that this is not completely truthful interpretation. Truely speaking, the population is constrained to have N individuals neither more or less, in each generation where N is fixed in advance. Under this simplification it suffice to know how many AA and how many aa are there. Thus the problem is reduced to describing the evolution of the pair (N_1^k, N_2^K) only.

Even after all these simplifications, the problem still remains quite intractable. Therefore, we are going to simplify it further. However, in the subsequent sections, we shall return to the problems described above. For the time being we decide to concentrate only on the variations in the gene frequencies rather than the genotypic frequencies. In an generation, the N individuals carry a total of 2N genes, some of which A and the rest are a. Let X_k be the number of A genes in the k-th generation so that $2N - X_k$ is the number of a genes. We are going to study the evolution of X_k.Of course, this would have been perfectly alright if, to start with we had a haploid population of 2N individuals in which case there are only two genotypes A and a.

We now come to the specific hypothesis concerning how a generation gives rise to the next generation. We assume that the 2N genes of a generation are obtained by simply taking a random sample of size 2N with replacement from the 2N genes of the parent generation. This is the classical Wright-Fisher Model. It is clear, that for each n, X_n the number of A genes in the n-th generation is a random variable taking values 0,1,...,2N. The above assumption really means firstly, that the conditional distribution of X_{n+1} given $X_0, X_1, ..., X_n$ depends only on $X_n$and secondly given $X_n = j$, $X_{n+1}$is distributed as the binomial variable $B(2N, \frac{j}{2N})$.In other words, the process (X_n) $n \geq 0$forms a Markov Chain with the state space {0,1,...,2N} and transition probabilities

$$p_{kj} = \begin{pmatrix} 2N \\ k \end{pmatrix} \theta_j^k (1-\theta_j)^{2N-k} \qquad for \qquad 0 \leq j,k \leq 2N, \quad where \quad \theta_j = \frac{j}{2N}.$$

For this chain it is clear that the states 0 and 2N are absorbing while others are transient. Thus, no matter where we start, the chain eventually gets absorbed in one of the two absorbing states. Thus $X_\infty = lim_{n\to\infty} X_n$exists and takes the two values 0 and 2N with probability one.

The first important question that we would like to address is the following. Given that the number of A genes is i to start with (that is $X_0 = i$),what are the probabilities $b_0(i)$ and $b_{2N}(i)$of the chain to be absorbed in the states) and 2N respectively. Note that $b_0(i)$ = and $b_{2N}(i)$=$P(X_\infty = 2N|X_0 = i)$. Here is a beautiful alternative due to Feller.

Observe that the process, (X_n)has the property that,

$$E(X_{n+1}|X_n) = X_n \quad for \quad every \quad n.$$

Indeed, since the conditional distribution of X_{n+1} given $X_n = j$, is binomial $(2N, \frac{j}{2N})$. We have E$(X_{N+1}|X_N = j) = 2N.\frac{j}{2N} = j$.Because of the Markov property the above equation is the same as,

$$E(X_{n+1}|X_0, X_1, ..., X_n) = X_n \qquad for \quad every \quad n.$$

In particular for all n, $E(X_n|X_0 = i) = i$. Since the random variables are uniformly bounded it follows that $E(X_n|X_0 = i) = i$.But of course $E(X_\infty = 0|X_0 = i) = 0.b_0(i) + 2N.b_{2N}(i)$. This yields,

$b_{2n}(i) = \frac{i}{2N}$ and $b_0(i) = 1 - \frac{i}{2N}$.

Thus initially there are i many A genes then eventually the number of A genes would be 0 with probability $1 - \frac{i}{2N}$.

Having thus obtained the absorption probabilities, we now turn to the rate at which the absorption takes place. We recall that it suffice to know the largest eigenvalue of the transition matrix which is smaller than one in modulous.

We define,

$\lambda_0 = 1$and for $a \leq r \leq 2N, \quad \lambda_r = 1(1-\frac{1}{2N})...(1-\frac{r-1}{2N})$ or equivalently, $\lambda_r = \begin{pmatrix} 2N \\ 2 \end{pmatrix} \frac{r!}{(2n)^r}$ for $0 \leq r \leq 2N$. Note that $\lambda_0 = \lambda_1 = 1 > \lambda_2 > \lambda_3 > ... > \lambda_{2N}$. It can be shown that these are precisely the eigenvalues of P from which it would follow that convergence takes place geometrically at the rate $\lambda_2 = (1 - \frac{1}{2N})$.

We have discussed the simplest case of the Wright-Fisher Model here. A realistic model would also take into account the mutations that happen in nature all the time.

References

[1] Lecture notes of Eric Anderson, University of California, Berkely.

[2] Lecture notes of A. Goswami, ISI Calcutta.