

A Simple Markov Chain

Qian Zhai *

University of Illinois at Urbana-Champaign, United States

Abstract. In this paper, Markov chain is used to model the reproduction of the fixed finite population, and use binomial distribution to discuss the probability of gen inheritance between generations population genes and establish transfer matrix ,using conditional mathematics expect to obtain the probability of absorpation at 0 and at $2N$ only depends on the initial fraction of A gene, and this also must be a fixation probability of gene A , and this conclusion is the same as the results of MATLAB simulation.

1 Introduction

The population size is assumed to be fixed throughout every step in our model. Our population shall contain $2N$ individuals either having a type-a genetic characteristic or a type-A characteristic. One individual of the parent generation is randomly chosen and the successor is from the same genetic type. In a population, some genes are passed on to the next generation, while others are not. If there is no selective mutation, no selective pressure, what is the effect on the inheritance of genes in the population, whether there will be a deletion of a gene, whether there will be gene fixation. This has important implications for the evolution and extinction of species, hence, We want to solve the following questions:

Question 1: What is the probability that a randomly chosen individual of the parents' generation is of type-a or of type-A respectively?

Question 2: What's the duplication process?

Question 3: we want to know whether a certain type of gene will disappear, and if so, how likely is that to happen?

The general idea of Markov process is that the future is only about the present, not the past. , In our model, The gene frequency at next generation only depends on present generation, not depend on the previous generations. So we use Markov chains to solve the above three problems.

2 Assumptions

Our model follows the following rules:

(1) The population size is assumed to be fixed throughout every generation in our model, our population shall contain $2N$ individuals.

(2) Each individual either has a type-a gene or a type-A gene, there are no other types of genes;

(3) Generations do not overlap and die simultaneously;

(4) Whether or not each ancestor has descendants is completely random every time;

(5) There is no mutation[1].

3 Model analysis

In our model. Let A and a denote two genetic characteristics in a given population. Our model is a discrete, non-overlapping and no mutation. Discrete means that the offspring appear together at the same time, but not continuously; non-overlapping means that the individuals of the t-generation will not be left over to the $t + 1$ -generation. It can be understood as: the parents in the population give birth to offspring at the same time. When the offspring appear, all the parents disappear, ensuring that the generations do not interfere with each other; no mutation means that all offspring of A individuals continue to contain only type-A.This is an ideal state, and the reality is more complicated than this.

As an example, there is a small population of 4 individuals. This is unrealistic, but we want to illustrate the duplication process of the model through this simple example.

3.1 Duplication process

since we assume that the population is a constant, there will be 4 individuals in the next generation again. In the first generation, the number of type-A in our model can be 0,1,2,3, 4.but if there are 4 type-A gene in ancestor, there will be no type-a gene in the next generation. Similarly, if there are 0 type-A gene in ancestor, there will be no type-A gene in the next generation.

The number of type- A and the corresponding probabilities are presented in the below table1 and figure 1(n represent the number of type-A ,p represent the corresponding probabilities)

Table 1. The number of type-A and the corresponding probabilities in first generation(edited by the author)

n	0	1	2	3	4
p	0.0625	0.25	0.375	0.25	0.0625

*qzhai4@illinois.edu

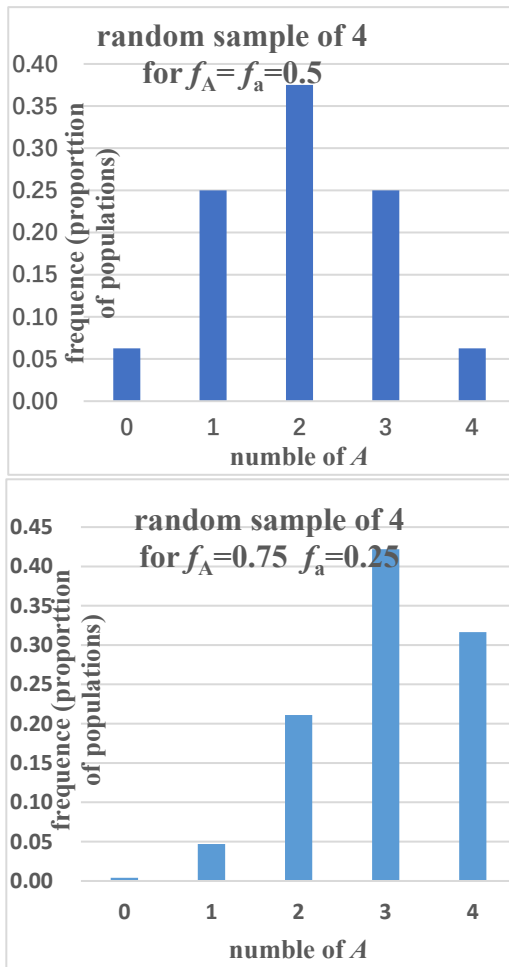


Fig. 1. Random samples of 4 for different possibilities(edited by the author)

To illustrate how the duplication process is, we represent each individual with a small circle. The blue circles represent the individuals having a type-A gene, and the yellow circles represent the individuals having a type-a gene (see Figure 2). Without loss of generality, we assume that there are 3 individuals having type-A. Each individual from the offspring generation picks a parent at random from the previous generation, so the number of individuals with type-A in the next generation may be 0,1,2,3, and 4. and parent and child are linked by a line. Each offspring inherits the genetic type of the parent. One of the possible results for the second generation is shown in figure 2.

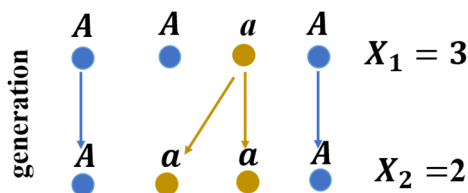


Fig. 2 The possible result for the second generation[2]

After a few generations, it will look something like Figure 3. There are tangled and untangled version. This diagram shows the same process, except that those individuals have been shuffled a little to avoid the confusion of many lines crossing. The genealogical relationship is still the same, except that the children of one parent are now placed together, close to the parents[3].

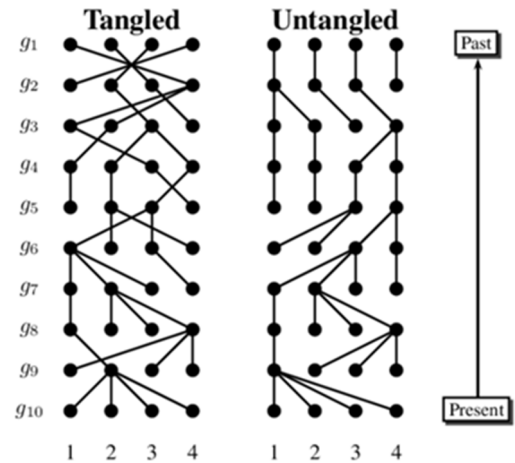


Fig. 3 Evolution of the model(edited by the author)

3.2 Distribution

This is a small population of size $2N = 4$. we will focus on how the frequency of genes will change from generation to generation, Now, if in generation 1, There are three individuals with gene- A, the frequency of A gene is 0.75, and what is the probability of having 0, 1 A's in the second generation?

In a given population with 4. One individual of the parent generation is randomly chosen and the offspring is of the same genetic type, each individual has the same chances to be the ancestor of the next generation. So the second generation is determined by 4 independent binomial trials. If the number of A gene is i ($0 \leq i \leq 4$) and the number of a gene is $(4 - i)$, then the probability that the offspring has gene A is $i/4$, the probability that the offspring has gene a is $(1 - \frac{i}{4})$, So in the second generation, denote p_j the probability of the number of gene A is j .

$$p_j = C_4^j \left(\frac{i}{4}\right)^j \left(1 - \frac{i}{4}\right)^{4-j}, \quad 0 \leq i, j \leq 4$$

Let us calculate the probabilities from formula of the binomial distribution ($i = 2$). Table 2 shows the genetic situation of each generation if $i = 2$

Table 2. Probabilities in second generation(edited by the author)

p	0	1	2	3	4
0	1	0	0	0	0
1	0.316	0.422	0.211	0.047	0.004
2	0.0625	0.25	0.375	0.25	0.0625
3	0.004	0.469	0.211	0.422	0.0316
4	0	0	0	0	1

3.3 Markov chain

The state space

We will fix our attention at frequency of gene A in the population of $2N$. The general idea of Markov process is that the future is only about the present, not the past. The discrete time process is an example of Markov chain[4]. The gene frequency at generation $t + 1$ only depends on generation t , not depend on the states before t . So the process of inheritance between generations can be regarded as a Markov chain.

The state X of the chain corresponds to the number of genes of type A. In any generation, X takes one of the values $0, 1, \dots, 2N$, which constitutes a state space. Let X_t be the number of the gene A at generation t . The state space of this Markov chain is the set of possible numbers of the type A, obviously $X_t \in \{0, 1, \dots, 2N\}$.

transition matrix

In our model, one individual of the parent generation is randomly chosen and the offspring is of the same genetic type, each individual has the same chances to be the ancestor of the next generation. Thus, the next generation is determined by $2N$ independent Bernoulli trials. So X_t

is a binomial random variable. If the initial generation consist of i genes of type A and $2N - i$ genes of type a, the probability of resulting in gene A is $p_i = i/2N$ and the probability of resulting in gene a is $q_i = 1 - i/2N$ for each Bernoulli trial, as we generate a Markov Chain $\{X_t\}$, where X_t is the number of A genes in the t 'th generation, since population size is a constant $2N$. X_{t+1} is a binomial random variable with index $2N$ and the probability of resulting in gene A is $X_t/2N$. The transition probabilities from $X_t = i$ to $X_{t+1} = j$ for this Markov Chain are computed according to the binomial distribution. The probability is given by the binomial formula (in which $2N$ is the population size and p the frequency of gene A and therefore, the probability that an individual picks a parent with gene A). So, we can write the probability transition matrix for the Markov chain as $P = (P_{ij})$, since $X_t \in \{0, 1, \dots, 2N\}$, then the transition probability matrix has a size of $(2N + 1) \times (2N + 1)$. If $X_t = i$ in generation t , then the probability of $X_{t+1} = j$ in the next generation with gene A is given by the binomial formula.

$$p_{ij} = C_{2N}^j \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}, \quad 0 \leq i, j \leq 2N$$

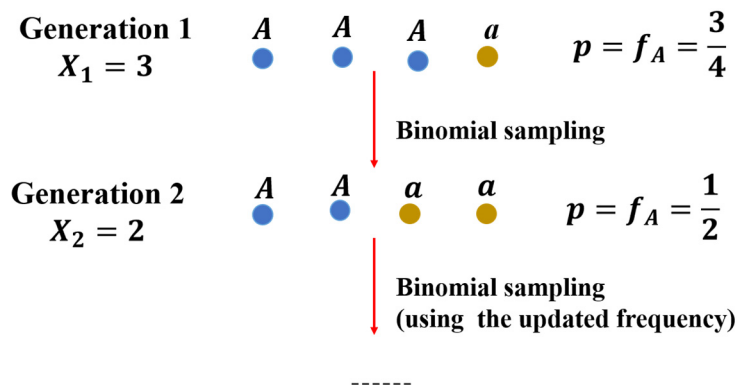


Fig. 4 The binomial distribution[5]

Figure 4 shows what the second generation might look like if $p=0.75$. We can multiply the transition matrix to predict the distribution of gene frequencies at any number of generations in the future using the updated frequency.

3.4 simulate evolution of population

When simulating allele frequencies, we don't need to pick a random parent for each individual one by one. We can just pick a random number from the binomial distribution (with the appropriate $2N$ and p) and use this as the frequency of the allele in the next generation

A simple MATLAB function to simulation
function zq()

```
f=WF(100, 0.5, 2000);
plot(f);
```

```
function t=WF(2*N, p, G)
t=zeros(1, G);
t(1)=p;
for i=2:G
    t(i)=binornd(2*N, t(i-1), 1)/(2*N);
end
```

Fig.5 MATLAB program(edited by the author)

Run it in MATLAB when $p = 0.5$, $p = 0.75$, $p = 0.25$, The code is shown in Figure 5, we can get the follow figures:

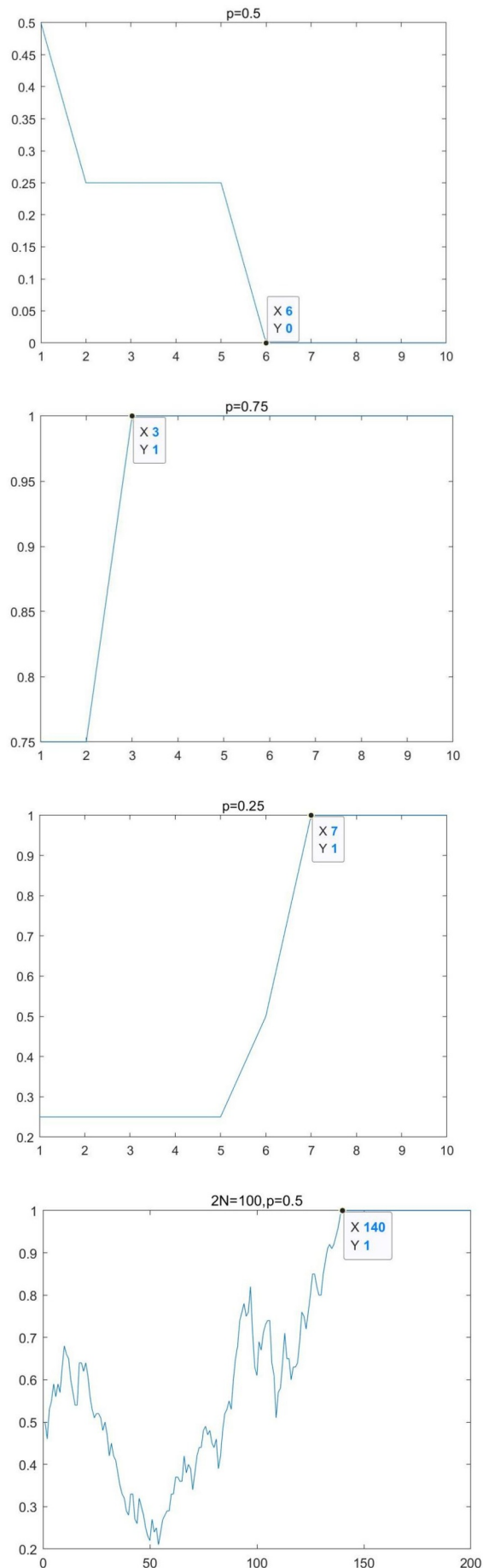


Fig. 6 The model simulation(edited by the author)

As shown in Figure 6, we can conclude that the probability of losing the gene is much higher if p is smaller.

3.5 the absorption state

The absorption state of a Markov chain is a state that once entered can never be exited. Since there are no mutations in the model, eventually one of the two genes will be lost (and the other fixed). In other words, States 0 and $2N$ are completely absorbing. no matter what the value of X_1 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 genes, while in the case of $X_t = 2N$, the population will be purely A-gene population.

The population can attain fixation and be composed of only A-genes or a-genes. Either of the absorbing states (either 0 or $2N$) will be entered. Therefore $\lim_{t \rightarrow \infty} P(X_t = i) = 0$ ($0 < i < 2N$). And we will discuss absorption probabilities in two cases[6].

Assume the number of A-gene is i in the first generation, then it can be obtained by using the mathematical expectation formula of binomial distribution.

$$E(X_1) = 2N \frac{i}{2N} = i \quad (1)$$

absorption at 0

Using the expectation of a conditional expectation

$$E(X_t) = E[E(X_t|X_{t-1})] = E(X_{t-1}) = \dots = E(X_1) = i \quad (2)$$

$$\text{Let } \lim_{t \rightarrow \infty} P(X_t = 0|X_1 = i) = p_{0,i},$$

On the other hand,

$$E(X_t) = 0 \cdot p_{0,i} + 2N \cdot (1 - p_{0,i}) \quad (3)$$

$$\text{Since } \lim_{t \rightarrow \infty} E(X_t) = i$$

$$\text{So } i = 0 \cdot p_{0,i} + 2N \cdot (1 - p_{0,i}) \quad (4)$$

$$p_{0,i} = \frac{2N-i}{2N} \quad (5)$$

absorption at $2N$

Similarly, we can get the probability of absorption at $2N$.

$$\text{Let } \lim_{t \rightarrow \infty} P(X_t = 2N|X_1 = i) = p_{2N,i}$$

$$\text{Then } E(X_t) = 0 \cdot (1 - p_{2N,i}) + 2N \cdot p_{2N,i} \quad (6)$$

so

$$p_{2N,i} = \frac{i}{2N} \quad (7)$$

For example, we can calculate the probability of losing A gene $P[X = 0] = 0.5$, with $2N = 10$ for $p = 0.5$;

$P[X = 0] = 0.8$ for $p = 0.2$;

$P[X = 0] = 0.99$ for $p = 0.01$.

the probability of losing A gene $P[X = 0] = 0.5$ with $p = 0.5$ whether $2N = 10$ or $2N = 100$.

the probability of fixing A gene $P[X = 2N] = 0.5$ with $2N = 10$ for $p = 0.5$;

$P[X = 2N] = 0$ for $p = 0.2$;
 $P[X = 0] = 0.01$ for $p = 0.01$.
the probability of fixing A gene $P[X = 2N] = 0.5$ with $p = 0.5$ whether $2N = 10$ or $2N = 100$.

4 conclusion

According to the above analysis, Under the assumption that the total quantity is fixed, the probability of absorption at 0 and at $2N$ only depends on the initial fraction of A gene which equal $i/2N$, and this also must be a fixation probability of gene A. If the proportion of gene A at the first generation, the probability of absorption at $2N$ is greater. This is the same as the conclusion of MATLAB simulation.

References

1. Jost Jürgen, and Tran, Tat Dat (2017) information geometry and population genetics Springer, 17-45
2. C.A. Muirhead (2016) Genetic drift, models of Random. Journal of Encyclopedia of Evolutionary, volume(4), 136-143
3. Fox, Charles W.(EDT) ,Wolf. Jason B.(EDT)(2006) Evolutionary Genetics, Oxford Univ Pr 21-23
4. Howard Taylor and Samuel Karlin (1998) An Introduction to Stochastic Modeling, Third Edition Academic Press, 95-138
5. PaulanTataru,Maria Simonsen, Thomas Bataillon,Asger-Hobolth (2017) Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data Journal of Systematic Biology volume(66), e30–e46
6. Warren Ewens, (2004),mathematical Population Genetics, Springer Second edition, , 20-23, 86-91, 92-99, 104- 109.