ISBN 978-608-4904-04-5(електронско издание)
ISBN 978-608-4904-03-8(печатено издание)
UDC: 519.217.2:57

APPLICATION OF MARKOV CHAINS IN BIOLOGY

Natasha Stojkovikj, Limonka Koceva Lazarova, Aleksandra Stojanova Ilievska

Abstract. The Markov chain is a random process with Markov characteristics, which exists in the discrete index set and state space in probability theory and mathematical statistics. Markov chains are powerful tools for stochastic modelling that can be useful in any science discipline. In this paper, we give an overview of some basic applications of the Markov chains in biology. We will describe the application in crossbreeding of the animals in close relation and carcinogenesis.

1. INTRODUCTION

Markov Chain is a powerful and effective technique to model stochastic processes with discrete time space and states space. Markov chains can be used to model many real-life processes. They have very wide applications in various fields such as: physics, chemistry, biology, medicine, music, game theory, sports, economics etc. They can be used for animal life populations mapping, to search engine algorithms, for music composition, speech recognition etc., [1].

In chemistry, Markov chains are used when physical systems closely approximate the Markov property. More chemical reactions can be considered as Markov chains. The reaction networks can be modelled with Markov chains, also the model of enzyme activity, Michaelis–Menten kinetics, can be viewed as a Markov chain. There are many advantages of using the Markov-chain model in chemistry. Some advantages are: physical models can be presented by state vector and a one-step transition probability matrix, it is easy to obtain all distributions of the state vector from the Markov-chain solution. Also, with Markov chain can be modelled various processes in chemical engineering by combination of flows, recycle streams, plug-flow and perfectly mixed reactors [2,3].

Claude Shannon, the father of Information theory, used Markov chains to model the English language. Through this model, he introduced the concept of entropy. In this language model, he assumed that letters from some text have a certain degree of randomness and are dependent on each of others. Also, this Markov model allow to produced text similar to text written to English language. Hence Markov models are widely used in Natural Language Processing and Computational Linguistics. Markov model can be used for effective data compression through entropy encoding techniques. Even without describing the full structure of the system perfectly, such signal models can make possible very effective data compression through entropy encoding techniques such as arithmetic coding [4].

Also, Markov chains are a base for hidden Markov models (HMM). These models are used in telephone networks, speech recognition and bioinformatics [5].

Process of birth and death that are basis of queueing theory are homogeneous Markov process. More of the queue systems (M/M/n, M/m/,∞) can be modelled

by using Markov process or Markov chain. For example, for the queue M/M/m, the time spent by a client in the queue is a Markov process and the number of clients in the queue is a Markov chain [6,7].
The Markov chain has applications in Internet applications. For example, Google's PageRank algorithm of a webpage is defined by a Markov chain [8].
In economics and finance, Markov chains are used to predict macroeconomic situations like market crashes and cycles between recession and expansion. Other areas of application include predicting asset and option prices and calculating credit risks [9].
In this paper, we will consider the application of Markov chains in biology. Concretely, we will regard two applications of Markov chains: the genetic problem of interbreeding animals in close relatives and application in carcinogenesis.

2. RANDOM (STOCHASTIC) PROCESS

The neediness for introducing the concept of stochastic (random) process follows from the work of different systems and variables that are random by their nature. Also, those variables depend on one or more parameters such as time, length, elevation, and others. Stochastic processes are widely used as mathematical models of systems and phenomena that appear to vary in a random manner. The variables that are considered by meteorological research like: temperature, humidity, pressure, concentration of smoke, sulfur dioxide, winds speed on the certain place are functions of the time, latitude, altitude of the place, are random.

Brownian motion of particles, voltage and power of electric current, number of car accidents, number of earthquakes, the speed of the vehicles etc., are the real processes that are random functions of the time [10,11].

Let (Ω, F, P) is a probability space and nonempty parameter set T. The stochastic or random process $\{X_t, t \in T\}$ is usually defined as a family of random variables.

Depending on the parameter set T, random process can be:

1. Random process with discrete parameter set (discrete random sequence).
2. Random process with continuous parameter set.

If the distribution of random variables $\{X_t, t \in T\}$ is considered, then random process can be:

1. Discrete random process.
2. Continuous random process.

Usually, T is one-dimensional set and the parameter can be interpreted as **time**.

For each fixed $t \in T, X_t$ is a random variable that is called intersection of process at the moment t. For fixed $\omega \in \Omega, X_t(\omega)$ is a function of t, that is called realization (trajectory) of the random process $\{X_t, t \in T\}$.

Definition 2.1 A random process $\{X_t, t \in T\}$ is said to be Markov, if for each $n \in \mathbb{N}, t_1 < t_2 < \ldots t_n, t_i \in T, i = 1,2,\ldots n$ and for each $x_1, x_2, \ldots x_n$:

$$P\{X_{t_n} < x_n \mid X_{t_{n-1}} = x_{n-1}, X_{t_{n-2}} = x_{n-2}, \ldots, X_{t_1} = x_1\} =$$
$$= P\{X_{t_n} < x_n \mid X_{t_{n-1}} = x_{n-1}\}$$

Because of that, it is said that a random process $\{X_t, t \in T\}$ is a Markov if the "future" state X_{t_n} is independent of the "past state" $X_{t_{n-2}}$, if the "present state" $X_{t_{n-1}}$ is given.

For discrete Markov process the following holds:

$$P\{X_{t_n} = x_n \mid X_{t_{n-1}} = x_{n-1}, X_{t_{n-2}} = x_{n-2}, \ldots, X_{t_1} = x_1\} =$$
$$= P\{X_{t_n} = x_n \mid X_{t_{n-1}} = x_{n-1}\}.$$

A discrete Markov process with a discrete set of states and a discrete parameter set is called a Markov chain, i.e.

Defeniton 2.2 A discrete random process $\{X_i\}_{i=0}^{\infty}$ is said to be Markov chain, if for each $n \in \mathbb{N}$, and for each $x_1, x_2, \ldots x_n$, the following holds:

$$P\{X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\} =$$
$$= P\{X_{n+1} < x_{n+1} \mid X_n = x_n\}.$$

The random process is a chain when the state space is discrete. The name Markov refers to Andrei. A. Markov, a Russian mathematician, who works described the Markov chains.

Definition 2.3 The one step transition probability, denoted as $p_{ij}^{(n)}$ is defined by the condition probability:

$$p_{ij}^{(n)} = P\{X_n = j \mid X_{n-1} = i\}.$$

This is the probability that process is in state j at the moment n, given that the process was in state i at the moment $n-1$. The probability $p_{ij}^{(n)}$ are called transition probability from state i to state j at the moment n. These probabilities form matrix of transition probabilities $\mathbf{P}^{(n)} = \left[p_{ij}^{(n)} \right]$.

3. APPLICATION OF DISCRETE MARKOV CHAINS IN BIOLOGY

The discrete Markov chains have wide application in biology.
With Markov chain the general process of birth and death in discrete time can be modelled. In this model is assumed that the size of the population is maximal. The theory developed from the random walk can be used for analysing the birth and death process. This theory is used to analyse, the probability of population extinction, the expected time of population extinction, and the distribution conditioned on nonextinction, known as the quasistationary distribution [12].
The other application of Markov chains in discrete time are the epidemiology models. Discrete Markov chains can be used for modelling of some chain based as SI, SIS and SIR models in epidemiology. The Susceptible-Infected-Susceptible (SIS) model describes the transmission of disease when recovered individual from the population do not have permanent immunity. Recovered individual can immediately become infectious again. The results for SIS model in [13] show normal distribution nature of the quasi-stationary distribution in the case when the population is large, and the reproduction number is greater than 1. With the SIR model population is divided into three subgroups: susceptible, infected, and recovered individuals. In this model a susceptible individual gets infected with disease and recovers from it and have a permanent immunity. The main aim of this model is to predict the trajectory of epidemic transmission. The transitions are made from one to another population [14-16].
The other application of discrete Markov chains in epidemiology is known as the binomial chain model.
Epidemiological models of binomials chains were first developed in 1920 and 1930 by Reed, Frost and Greenwood, so according to them is the model named. For these models, the duration and extent of the epidemic are calculated [17].
Also, discrete Markov chain is used to proliferating epithelial cells.
In this paper, two classical biological applications of Markov chains in discrete time will be considered [17].

3.1 The genetic problem of interbreeding animals in close relatives

Genetics is an important area in the biology which studies genes, genetic variation, and heredity. It studies how living organisms receive common characteristic from previous generation. Heredity depends on the information that are contained in chromosomes. Every cell of a human being contains 46 chromosomes (23 from the mother and another 23 from the father), or 23 pairs named as ‘diploid’: 22 pairs of autosomes, and one pair of sex chromosomes, called X and Y.

Genes are arranged, one after another, at specific locations on chromosomes in the nucleus of cells. The genes are made of sequence of DNA. A location on the chromosome where the gene is found are called locus. The gene is located within a determined region on the chromosome and is composed of the different

base pairs (GATC). Alleles are variants of the same gene that occur on the same place on a chromosome [18,19].

Assume that there are only two types of alleles for a given gene, denoted as a and A. A diploid individual can have one of three different allele combinations: AA, Aa or aa, known as locus genotypes. As aa and AA have the same homogenous composition, they are called homozygotes, while Aa is called heterozygotes. [20]:

In [21] the problem of genetic pairing of animals in close relatives is regarded. This process form Markov chain. Suppose that two individuals are randomly paired. Process of pairing between siblings and close relatives continues every year. This process can be formulated as a finite Markov chain in discrete time, whose states consist of 6 types of mating:

1. $AA \times AA$
2. $AA \times Aa$
3. $Aa \times Aa$
4. $Aa \times aa$
5. $AA \times aa$
6. $aa \times aa$

It is assumed that the parents are type 1, $AA \times AA$. Then the next generation descendants of these parents will be AA individual, and then the pairing of siblings will be only type 1, $p_{11} = 1$. Analogously for parents type 6, $aa \times aa$, where $p_{66} = 1$.

Now, it is assumed that the parents are type 2, $AA \times Aa$. Let X represents their randomly chosen descendant. Let $Y_1 \in \{A, A\}$ is allele that will be transmitted on to descendant from parents with genotype AA, and $Y_2 \in \{A, a\}$ represents the allele that will be transmitted to descendants from parents with genotype Aa. Then the following holds:

$$P\{X = AA\} = P\{Y_1 = A, Y_2 = A\} = P\{Y_1 = A\} P\{Y_2 = A\} = 1\frac{1}{2} = \frac{1}{2},$$

$$P\{X = Aa\} = P\{Y_1 = A, Y_2 = a\} = P\{Y_1 = a\} P\{Y_2 = A\} =$$

$$= P\{Y_1 = 1\} P\{Y_2 = a\} = P\{Y_1 = a\} P\{Y_2 = A\} = 1\frac{1}{2} + 0 = \frac{1}{2},$$

$$P\{X = aa\} = P\{Y_1 = a, Y_2 = a\} = P\{Y_1 = a\} P\{Y_2 = a\} = 0$$

From this generations there are two genotypes. If X_1 and X_2 are two randomly chosen generations. Then:

$$P\{X_1 \times X_2 = AA \times AA\} = P\{X_1 = AA\}\, P\{X_2 = AA\} = \frac{1}{4}$$

$$P\{X_1 \times X_2 = AA \times Aa\} = P\{X_1 = AA\}\, P\{X_2 = Aa\} + P\{X_1 = aA\}\, P\{X_2 = AA\} = \frac{1}{2},$$

$$P\{X_1 \times X_2 = Aa \times Aa\} = P\{X_1 = Aa\}\, P\{X_2 = Aa\} = \frac{1}{4}.$$

The probabilities are:

$$p_{12} = \frac{1}{4}, p_{22} = \frac{1}{2}, p_{32} = \frac{1}{4}.$$

If the parents are type 3, $Aa \times Aa$ the descendant is in proportions $\frac{1}{4}AA, \frac{1}{2}Aa$ and $\frac{1}{4}aa$, then the pairing between siblings give $\frac{1}{16}$ type 1, $\frac{1}{4}$ type 2, $\frac{1}{4}$ type 3, $\frac{1}{4}$ type 4, $\frac{1}{8}$ type 5 and $\frac{1}{16}$ type 6.

The transition matrix P is:

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

$$= \left(\begin{array}{c|cccc|c} 1 & 1/4 & 1/16 & 0 & 0 & 0 \\ \hline 0 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1 & 0 \\ 0 & 0 & 1/4 & 1/2 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 0 & 0 \\ \hline 0 & 0 & 1/16 & 1/4 & 0 & 1 \end{array}\right) = \begin{pmatrix} 1 & A & 0 \\ 0 & T & 0 \\ 0 & B & 1 \end{pmatrix}.$$

The Markov chain is reducible and there are three classes of communication, $\{1\},\{6\}$ and $\{2,3,4,5\}$. The first two classes are positive recurrent, and the third class is transient. The states 1 and 6 are absorbing states, $p_{ii}=1, i=1$ and $i=6$. Let notice

$$p^n=\begin{pmatrix}1 & A_n & 0\\ 0 & T^n & 0\\ 0 & B_n & 1\end{pmatrix},$$

Where A_n and B_n are functions of T, A and B, $A_n=A\sum_{i=0}^{n-1}T^i$ and $B_n=B\sum_{i=0}^{n-1}T^i$. First T^n need to be found. Because of T belongs to the transient class, it holds that $\lim_{n\to\infty}T^n=0$.

Also,

$$\lim_{n\to\infty}B_n=B\left(I-T\right)^{-1},\lim_{n\to\infty}A_n=A\left(I-T\right)^{-1}.$$

3.2 Restricted random walk and its application in carcinogenesis

The random walk is one of the most fundamental models in probability theory, demonstrating power of mathematical properties.

A one-dimensional random walk is a Markov chain with finite or infinite state space. In the simple one-dimension random walk, two movement are allowed. The movement to the right for one position, i.e. from the position x to the position $x+1$ and movement from the position x to the position $x-1$ i.e. movement to left for one position. Let p is a probability of moving to the right and q is a probability of moving to the left, [22].

A restricted random walk corresponds to a random walk in the presence of a boundary. If the state space is finite, $\{0,1,2,\ldots,N\}$ then 0 and N are boundary. If the state space is, $\{0,1,2,\ldots\}$ then boundary is in 0. There are three types of boundary behaviour: absorbed, reflected and elastic. Absorbed boundary in the $x=0$ assumes that transition probabilities for one step $p_{00}=1$. Reflected boundary in the $x=0$ assumes that transition probabilities for one step are: $p_{00}=1-p,\ p_{10}=p,\ 0<p<1$. And the elastic boundary in the $x=0$ assumes that one-step transition probabilities are: $p_{21}=p,\ p_{11}=sq,\ p_{01}=\left(1-s\right)q,\ p+q=1, p_{00}=1$ for $0<p,s<1$.

Cancer is a human genetic disease. It is caused by mutations that occurred in a more number of genes that controlling growth. Cancer is multi-stage process. The genes that caused cancer can exist from birth, increasing a chance of getting cancer. The transition of a normal cell into a cancerous cell are happened in more

steps (stages). The number of stages is a number of mutations that are required to creating a cancerous cell.
The number of stages can be regarded as the number the state in the random walk. For this reason, we will study a simple random walk model. Let $S = \{0,1,2,\ldots,N\}$ is a number of states. One movement in the random walk (when the process transits from one to other state) corresponds on the transition from one to other stage of the one cancerous cell.
The state 0 represents the stage (state) of total recovery. This model requires several successive mutations, each of which produces a clone of mutated cells. State N indicates completion of the mutation process in which malignant cells are created. [17]
Let $\{X_n\}, n \geq 0$ represents random walk, that corresponds to the mutation process. A step forward implies transition in the next stage(state). This transition is occurred with following probability: $P\{x \to x+1\} = p_x = \frac{x}{N}$. A step back implies transition in the previous stage (this is a move toward recovery) and this transition is occurred with probability: $P\{x \to x-1\} = q_x = \frac{N-x}{N}$.
The state 0 and state N are absorbing states. If the process come in this state stay here:

$$P\{N \to N\} = P\{0 \to 0\} = 1.$$

The other states different from state 0 and state N are reflecting states. Probability the process to stay in these states is equal to zero: $P\{x \to x\} = 0$, for $x = 0,1,2,\ldots,N-1$.

Let π_0 is a stationary probability of complete recovery (the process is in state 0), π_N is a stationary probability in cancerous state and π_x is a stationary probability in state x, $1 \leq x \leq N-1$. For these probabilities following differential equation is obtained:

$$\pi_x = \frac{x}{N}\pi_{x+1} + \left(1 - \frac{x}{N}\right)\pi_{x-1}, 1 \leq x \leq N-1$$

with initial conditions

$$\pi_0 = 0,\ \pi_N = 1.$$

Let $A(t) = \sum_x \pi_x t^x$ is a probability generating function for $\{\pi_x\}$. By using of simple mathematical operations, we write the differential equation in the following way:

$$\pi_x = \frac{x+1}{N}\pi_{x+1} - \frac{1}{N}\pi_{x+1} + \left(1 - \frac{x-1}{N}\right)\pi_{x-1} - \frac{1}{N}\pi_{x-1}, 1 \le x \le N-1$$

If is taken $\pi_{N+k} = 1,\ k \ge 0$,

$$A(t) = \sum_{x=0}^{\infty}\left\{\frac{x}{N}\pi_x t^{x-1} - \frac{1}{N}\pi_x t^{x-1} + \left(1 - \frac{1}{N}\right)\pi_x t^{x+1} - \frac{x}{N}\pi_x t^{x+1}\right\}$$

are obtained that

$$\left(A(t)\right)^{-1}\frac{dA}{dt} = t^{-1} + (1-t)^{-1} + (N-1)(1-t)^{-1}.$$

With solving of the differential equation, the following equation is obtained:

$$\left(A(t)\right)^{-1} = Ct(1+t)^{N+1}(1-t)^{-1} = Ct\left[\sum_{x=0}^{N-1} C\binom{N-1}{x}t^x\right]\sum_{y>0} t^y,$$

where C is a constant.

With using the limit conditions,

$$1 = \sum_{x=0}^{N-1} C\binom{N-1}{y} C2^{N-1},$$

is obtained that

$$\pi_x = \sum_{x=0}^{x-1} C\binom{N-1}{y}2^{1-N}, 0 \le x \le N.$$

After the initial process of the carcinogenesis, it is assumed the state is 1. From there:

$$\pi_N = \pi_1\binom{N-1}{0}2^{N-1} = 2^{N-1}.$$

Because of the random walk is simple, we obtain that:

$$\pi_0 = 1 - \pi_N = 1 - 2^{N-1}.$$

4.CONCLUSION

Markov chains are useful tools in statistics modeling in all fields of applied mathematics. They have great application in the modeling of natural phenomena and sciences. In this paper, we consider application of Markov chains in biology. By help, of two applications of Markov chain in biology, we can conclude that they are powerful tool for modeling of many problems in real life. In this paper, we have considered application of Markov chain in the genetic problem of interbreeding animals in close relatives and application in carcinogenesis which is very important for their analysis.

References

[1] P. Von Hilgers, A.N. Langville, *The five greatest applications of Markov chains*, Proceedings of the Markov Anniversary Meeting, Charleston SC, (2006), 155–168.

[2] A. Tamir, *Applications of Markov chains in chemical engineering*, Elsevier, Amsterdam, 1998.

[3] S. Chao Du, C. Kou, *Correlation analysis of enzymatic reaction of a single protein molecule*, The Annals of Applied Statistics, *6* (3) (2012).

[4] C. Ames, *The Markov Process as a Compositional Model: A Survey and Tutorial*: *Leonardo*, The MIT Press, 22 (2) (1989), 175-187.

[5] D. Pratas, M.R. Silva, J.A. Armando J. Pinho, J.S.G.P Ferreira, *An alignment-free method to find and visualise rearrangements between pairs of DNA sequences*: Sci Rep. 2015;5:10203.

[6] F.M. Neuts, *Markov Chains with Applications in Queueing Theory, Which Have a Matrix-Geometric Invariant Probability Vector*, Advances in Applied Probability, 10 (1) (1978), 185-212.

[7] G. Giambene, *Markov Chains and Queuing Theory: Queuing Theory and Telecommunications*, Springer, Boston, MA , (2005), 305-383.

[8] F. Mahfuz, *Markov Chains and their applications*: University of Texas at Tyler, Master Theses, (2021).

[9] S.D. Myers, L. Wallin, P. Wikström, *An introduction to Markov chains and their applications within financeM*: VE220 Financ. Risk Read. Proj, (2017).

[10] H.M. Taylor, S. Karlin, *An Introduction to Stochastic Modeling*, Academic Press, (1998).

[11] M.S. Ross, *Stochastic Process*, Second Edition: John Wiley and Sons, Inc., (1996).

[12] A. Tolver, *An Introduction to Markov Chains: Lecture Notes for Stochastic process*, Department of Mathematical Sciences University of Copenhagen, (2016).

[13] I.Nassel, *On the quasi- stationary distribution of stochastic logistic epidemic*, Mathematical Bioscience, (1999), 21-40.

[14] W.O.Kermack, *A contribution to the mathematical theory of epidemics*, Proceedings of the Royal Society of London, (1927), 700-721.

[15] *L. K. Lazarova, N. Stojkovikj, A. Stojanova, M. Miteva, Application of differential equations in epidemiological model*, Balkan Journal of Applied Mathematics and Informatics, 4 (2) (2021), 91-102.

[16] *L. K. Lazarova, N. Stojkovik, A. Stojanova, M. Miteva, M. Ljubenovska, Mathematical model for predictions of COVID-19 dynamics*, International Journal of Applied Mathematics, 34 (1), (2021), 119-133.

[17] J.S. L .Allen, *An Introduction to Stochastic Process with Application in Biology*, Pearson Education, (2010).

[18] M. Kimura, *Some Problems of Stochastic Process in Genetics: The Annals of Mathematical Statistics*, 28 (4) (1957), 882-901.

[19] B. Armstrong, R. Wilkinson, *Genetics*, ISBT Science Series, 15 (S1), (2020), 112-122.

[20] L.R.Nussbaum, R.R. Mcinnes, F.W. Huntington, *Genetic Variation in Population*, Thompson & Thompson Genetics in Medicine, Chapter 9, (2016) 155-170.

[21] J.T.N. Bailey, *The Elements of Stochastic Processes with Applications to the Natural Sciences*, Wiley-Interscience, (1991).

[22] J. Nicolau, Stationary Processes That Look like Random Walks: The Bounded Random Walk Process in Discrete and Continuous Time: Econometric Theory, 18 (1) (2002), 99-118.

Faculty of computer sciences, Goce Delcev University, Stip, Macedonia
E-mail address: natasa.stojkovik@ugd.edu.mk.
Faculty of computer sciences, Goce Delcev University, Stip, Macedonia
E-mail address: limonka.lazarova@ugd.edu.mk.
Faculty of computer sciences, Goce Delcev University, Stip, Macedonia
E-mail address: aleksandra.stojanova@ugd.edu.mk.