"Science Stays True Here"
Journal of Mathematics and Statistical Science, Volume 2016, 336-342 | Science Signpost Publishing

Markov Chain for Estimating Human Mitochondrial DNA Mutation Pattern

Sandy Vantika and Udjianna S. Pasaribu*

Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jl. Ganesha 10, Bandung 40132, Indonesia

** Email: sandy.vantika@math.itb.ac.id*

Abstract

The Markov chain was proposed to estimate the human mitochondrial DNA mutation pattern. One DNA sequence was taken randomly from 100 sequences in Genbank. The nucleotide transition matrix and mutation transition matrix were estimated from this sequence. We determined whether the states (mutation/normal) are recurrent or transient. The results showed that both of them are recurrent.

Keywords: Estimate, Markov chain, recurrent, transient, transition matrix.

1. Introduction

Markov chain stochastic model is one that currently widely applied to study the distribution of the changes that occur in nature both for inanimate objects and living beings. As a tool in operations research to managerial decision, Markov chains have been widely used to analyze the displacement of the brand in marketing, accounts counters, rental services, sales planning, inventory issues, rejuvenation cash flow management, cross-setting the capacity of water and so on [3,5].

Markov chain has been employed as a mathematical technique used to perform modeling of various systems and business processes [Dony ICMNS 3]. Markov chain is a new application of dynamic programming to solve a stochastic process that can be described by a limited number of circumstances [5]. This technique can be used to predict the changes that will occur in the future based on some informations in the past. Here, the Markov chain would be applied to study the pattern of human mitochodrial DNA (deoxyribosenucleic acid) mutation. In fact, this study is a further analysis from Vantika and Pasaribu (2012). Mitochondria are the most important organelles in cell. They have the respiratory function of

living organisms, in addition have other cellular functions, such as fatty acid metabolism, pyrimidine biosynthesis, calcium homeostasis, signal transduction and generating cellular energy in the form of adenosine triphosphate on track catabolism [6]. In contrast to other cell organelles, mitochondria have their own genetic material whose characteristics are different from the genetic material in the cell nucleus called mitochondria DNA [6]. Mitochondria DNA has an event called mutations. Mutations can occur in mitochondria DNA (mtDNA) and cell nucleus DNA [4]. To determine the presence or absence of mutations in an individual, is done by comparing the base sequences with standard base (rCRS). If there is only one base different from the base of rCRS on certain position then it is said mutation occured. Until now, it is not known whether the base mutation is dependent or independent. Previously, Nielsen mentions that the base sequence follows a Markov chain stochastic processes [4]. Therefore, mutations problems will be assessed using Markov chain (in this case a mutation in mitochondrial DNA). The reason for the selection was based on things: (1) the rate of mutation in mitochondrial DNA is 5-10 times faster than mutations in DNA cell nucleus, (2) mitochondrial DNA strand consists of 16,569 bases, and (3) the variation of base composition in the HVR (the area that most commonly affected by mutations) in the D-loop of mitochondrial DNA [6].

2. Mitochondrial DNA Mutations

In contrast to other cell organelles, mitochondria have their own genetic material whose characteristics are different from the genetic material in the cell nucleus called mitochondrial DNA. Unlike the cell nucleus DNA which is linear, mitochondrial DNA is circular. Mitochondrial DNA (mtDNA) consist of 16,569 base pairs in the mitochondrial matrix, circular and has a double strand consisting of heavy strand (H) and light strand(L). Named such because H strand has a molecular weight greater than L strand, caused by a large content of purine bases [4, 6].

Mitochondrial DNA consist of coding region and non-coding region. Non-coding region contains areas that have high variation called the displacement loop (D-loop). D-loop has an area with a high rate of polymorphism so that the sequence varies among individuals, that is the Hypervariable Region (HVR). Each unit on the base pair of mitochondrial DNA is a combination of the bases: adenine (A), thymine (T), cytosine (C) and guanine (G) [4].

Types of mutations found in mitochondrial DNA is point mutations due to mitochondrial size (16,569 units of base pairs) is much smaller compared to the cell nucleus DNA (3,109 units of base pairs). The mechanism of point mutations are divided into two kinds which are base substitutions and frameshift

mutation due to the addition of bases or loss of bases (deletion). Point mutations caused by base substitution is called base substitution mutations [6].

In this paper, the type of mutation analyzed is base substitution mutations and to determine whether the occurrence of base substitution mutations at certain position (eg position *t*). That position influences the base at position *t* +1. To measure the dependency, the sign test is applied.

3. The Sign Test

Sign test (??) is used to examine the existence of a trend. Let the process

All pairs (B_t, B_{t+1}), B_t states base at position *t*, will be encoded with a positive sign '+', negative '-' or '0'. This test is a binomial test with *p* (probability of success) = ½ (as probability of '+' and '-' are considered to be equal), where the binomial random variable is the number of positive sign states '+'. If sign test indicates a trend, we will find the base in the position *t*+1 through Markov chain.

Nucleotides A, C, G, and T are coded with numbers 1, 2, 3, and 4 respectively, in other words $B_t \in$ {1, 2, 3, 4}. Positive sign '+' is given if $B_t < B_{t+1}$, '-' given if $B_t > B_{t+1}$ and '0' is given if $B_t = B_{t+1}$

Let the process

$$\{B_t, t = 1, 2, \ldots\} \tag{1}$$

be a Markov chain with state space B = {1, 2, 3, 4} corresponding to {A, C, G, T}, respectively. Then the transition matrix (first order) that describes the transition from one state to another state is as follows.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}, 0 \le a_{ij} \tag{2}$$

$$\sum_{j=1}^{4} a_{ij} = 1, i = 1, 2, 3, 4 \tag{3}$$

We have a Markov chain $\{B_t\}$ with 4 states and transition matrix A is unknown. Then, we have the observation at position [1, T] where $T \ge 1$ and let also $x_1, x_2, \ldots, x_T$ be the state of the observation. We can estimate the transition matrix by Maximum Likelihood Estimation as follows.

$$\widehat{a_{ij}} = \frac{N_{ij}}{N_i}, N_i = \sum_j N_{ij} \tag{4}$$

4. Results

As already explained before, variables used in the sign test is bivariate random variables. Therefore, we performed the analysis on the position (t, t +1). After that, we assign each pair (B_t, B_{t+1}) in the sign (+), (-) or (0), by first doing coding. We gives the code as follows: A = 1, C = 2, G = 3 and T = 4. Each base (A, C, G and T) has 3 probabilities to mutate into another base.

Table 1 shows the base appearing at position t+1 if mutated base at position t is known. In the case of mutations that yield cytosine as mutated base, the position t+1 will be filled by cytosine as well which is about 51.75% of 114 cases. Meanwhile, in the case of mutations that produce thymine as mutated base, the position t+1 will be filled by thymine as well which is about 44.90% of 66 cases.

According to **Table 1**, we present it in a transition probability matrix form as follows.

$$A = \begin{pmatrix} 0.2400 & 0.2800 & 0.4000 & 0.0800 \\ 0.3070 & 0.5175 & 0.1053 & 0.0702 \\ 0.4118 & 0.2941 & 0.1176 & 0.1765 \\ 0.1361 & 0.4082 & 0.0068 & 0.4490 \end{pmatrix} \tag{5}$$

Let a_{ij}^n be n-step transition probability of a mutation pattern when mitochondrial DNA base at position t mutates into base i and base at position t+n is the base j.

$$a_{ij}^n = P\left(B_{t+n} = j | B_t = i\right), 1 \le n, 1 \le i, j \tag{6}$$

We make a model for n = 2, the base that appears in position t+2 and obtain the two-steps transition probability matrix as follows.

$$A^2 = \begin{pmatrix} 0.3192 & 0.3624 & 0.1731 & 0.1454 \\ 0.2855 & 0.4134 & 0.1902 & 0.1110 \\ 0.2616 & 0.3741 & 0.2107 & 0.1536 \\ 0.2219 & 0.4346 & 0.1013 & 0.2423 \end{pmatrix} \tag{7}$$

However, based on the observed data, we obtain a two-steps transition probability matrix B as follows.

$$\widehat{A^2} = B = \begin{pmatrix} 0.2400 & 0.3200 & 0.0000 & 0.4400 \\ 0.2456 & 0.7105 & 0.0175 & 0.0263 \\ 0.3529 & 0.3529 & 0.0588 & 0.2353 \\ 0.2585 & 0.5850 & 0.0136 & 0.1429 \end{pmatrix} \tag{8}$$

Based on the two-steps transition probability matrix A^2, we could see that for every case of mutation (the mutated bases are A, C, G and T), base at position t+2 tends to be occupied by cytosine (C) with an inclination of 36% - 43%. Similarly, when we look at a two-steps transition probability matrix derived from the observed data, for every case of mutation (except for mutations that produce adenine (A), base that appear in position t+2 is also cytosine (C).

We want to know the (unconditional) probability that at a certain position (position t+1 and / or t+2) is occupied by a particular base (A, C, G and / or T). We can calculate it by using an unconditional transition probability as follows. Suppose

$$\alpha_i = P(B_t = i), i \in \{1,2,3,4\}, \sum_{i=1}^{4} \alpha_i = 1 \tag{9}$$

Unconditional probability can be calculated by requiring the initial state of the base at position t,

$$P(B_{t+n} = j) = \sum_{i=1}^{4} P(B_{t+n} = j \mid B_t = i) P(B_t = i) = \sum_{i=1}^{4} P_{ij}^{n} \alpha_i \tag{10}$$

There are no mutations of guanine (G) to cytosine (C) and thymine (T) to guanine (G). Trend is not found in mutations cases of adenine (A) to cytosine (C), adenine (A) to thymine (T), and cytosine (C) to guanine (G).

$$\begin{pmatrix} 6\ (0.24) & 7(0.28) & 10(0.4) & 2(0.08) \\ 35(0.307) & 59(0.5175) & 12(0.1053) & 8(0.0702) \\ 7(0.4118) & 5(0.2941) & 2(0.1176) & 3(0.1765) \\ 20(0.1361) & 60(0.4082) & 1(0.0068) & 66(0.449) \end{pmatrix}$$

Table 1. Overall Base Appearing In Position t+1

Mutat. Base	A	C	G	T
A	6;24.00%	7 (0.2800)	10 (0.4000)	2 (0.0800)
C	35; 30.70%	59 (0.5175)	12 (0.1053)	8 (0.0702)
G	7;	5 (0.2941)	2 (0.1176)	3 (0.1765)
T	20 (0.1361)	60 (0.4082)	1 (0.0068)	66 (0.4490)

If the mutated nucleotide base is cytosine, the nucleotide base afterward will be cytosine with probability 51.75%. In addition, if the mutated nucleotide base is thymine, the nucleotide base afterward will be thymine with probability 44.90%.

We calculated the unconditional probability that at position t+1 is occupied by a certain base (A, C, G, or T). Thus, we will compute $P(B_{t+1}= j), j \in \{1, 2, 3, 4\}$. In other words, there would be calculated the probability of a certain base appearing at position t+1 which is the position right after the mutated base position.

After that, the proportion of each base that appears in position t+1 is also calculated by using the data presented in **Table 1**. We did the same procedure to calculate the unconditional probability of a certain base appearing at position t+2 that is $P(B_{t+2}= j), j \in \{1,2,3,4\}$, and also calculated the proportion of each base that appears at position t+2 by utilizing existing data. We presents all the results of the calculations in **Table 2**.

From **Table 2**, we could see that the position t+1 would be occupied by cytosine (C) with similar unconditional probability and proportion which are equal to 43.23%. Similarly, the t+2 position will be occupied by cytosine (C) with unconditional probability and proportion respectively 41.72% and 59.74%.

Table 2. Unconditional Probability And Proportion Of Bases At Position t+1 And t+2

Num.	Types Of Base	Position t+1		Position t+2	
		Unconditional Probability	Proportion	Unconditional Probability	Proportion
1.	Adenine (A)	22.44%	22.44%	25.61%	25.74%
2.	Cytosine (C)	43.23%	43.23%	41.72%	59.74%
3.	Guanine (G)	8.25%	8.25%	14.68%	1.65%
4.	Thymine (T)	26.07%	26.07%	17.99%	12.87%

5. Conclusion

In this article, the authors assess the problems: (1) the relationship between the mutated nucleotide base and nucleotide base at one position afterward and (2) the pattern of mutations in mitochondrial DNA. The authors investigated the relationship between the mutated nucleotide base and base at one position afterward using the sign test. To find the pattern of mitochondrial DNA mutations, the authors used a Markov chain. The results obtained are: (1) the sign test indicated that the mutated nucleotide bases have a relationship with nucleotide bases at one position afterward, except mutation of adenine (A) to cytosine (C), adenine (A) to thymine (T), and cytosine (C) to guanine (G), (2) adenine (A), cytosine (C), guanine (G), and thymine (T) will mutate if bases at one position and two positions afterward are cytosine (C) with probability 43.23% and 41.72% respectively.

References

[1]. R. V. Hogg, J. McKean, and A. T. Craig, Introduction to Mathematical Statistics, New Jersey: Prentice Hall, 2005, pp. 10-15.

[2]. K. P. Donnelly, Theoritical Population Biology 23, 34-63 (1983).

[3]. S. Karlin and M. T. Howard, An Introduction to Stochastic Modeling, California: Academic Press, 1994, pp. 15-20.

[4]. R. Nielsen, Genetics 159, 401-411 (2001).

[5]. S. M. Ross, Stochastic Processes, New York: John Wiley and Sons, 1996, pp. 4-9.

[6]. K. S. Dimmer, S. Fritz, F. Fuchs, M. Messerschmitt, N. Weinbach, W. Neupert, and B. Westermann, Molecular Biology of the Cell 13, 847-853 (2002).

[7]. R. E. Walpole, Probability and Statistics for Engineers and Scientists, New Jersey: Prentice Hall, 2002, pp. 87-92.