Statistika, Vol. 8 No. 2, 93 – 96
Nopember 2008

Analyzing Pattern of Mutation in mtDNA Using Markov Chain

LIRA ADIYANI, SUTAWANIR DARWIS, AND ACHMAD SAIFUDDIN NOER

Institut Teknologi Bandung, Indonesia
Email: lyera_math@yahoo.com, sdarwis@math.itb.ac.id, noer@chem.itb.ac.id

ABSTRACT

Mutation in mtDNA becomes an interesting topic that needed to discuss. If someone has a mutation in his mtDNA, then it might be affect his health. Those effects could be some diseases or another variation that gives different characteristics. In study of mutation, there are two such things become the main problems: (1) does mutation occur dependently? and (2) what is the pattern? From the research (by 9 degrees of freedom χ^2), DNA sequence shows a positional dependence. In addition, we can also see a positional dependence in mtDNA sequence clearly (position i-1, i, i+1 are dependent with i define as mutation) by sign test, which means, it is possibly that there is a pattern of mutation. This paper uses Markov chain to quantify the pattern and as results all bases will mutate if position i+1 is C or cytosine (±40%). Moreover, A, C, and G will mutate (become T) if position i-1 is A or adenine (54.5%).

Keywords: *Markov chain, mtDNA, mutation, pattern, positional dependence.*

1. Introduction

Mutation could happen in mtDNA or nucleus DNA. To identify whether there is a mutation in human body, we compare the rCRS with their DNA. If there is one different base that makes the sequence's changed, we call it as a mutation. From research using chi square, there is a positional dependence in DNA sequence, but we still don't know yet what kind of dependencies that we have. That is why this paper will analyze it (quantify the pattern of mutation) and to answer it, Markov chain will be used.

mtDNA has many characteristics, such as mutate 5-10 times faster than nucleus DNA; has a Hyper variable Region 1/HVR-1 (a region that most commonly mutate); and consist of 16,569 bases/positions. From those positions, there are 360 positions that most able to change called HVR-1. So that, the analyzing process only done in HVR-1 or interval [16024, 16383]. To make counting and analyzing processes become easier, than HVR-1 is categorized to three positions, there are i-1, i, and i+1; which mutation is defined as position i (note: if i-1 or i+1 also mutate then i-1, i, i+1 will not be used). This paper uses data from gene bank and the total data is 9188 individual but only 8951 data is used because this paper only analyze base substitution mutation.

2. Methods and Empirical Results

Markov chain is one of stochastic process with a discrete parameter also discrete and finite state space (S = {0, 1,..., n}) that is: if a state is given then probability for one next state only influenced by current state. When by the time t the process is in state i, than this event can be written by $X_t = i$ and $\{X_t, t = 0,1,...\}$ is called Markov chain if:

$$P\{X_{t+1} = j \mid X_0 = i_0, ..., X_{t-1} = i_{t-1}, X_t = i\} = P\{X_{t+1} = j \mid X_t = i\}$$

Let X is a random variable with $p[X=k]=a_k, \quad k \geq 0, \quad \sum_{i=0}^{n} a_i = 1$. Assume there is a Markov chain $\{X_t\}$ with m states and observed during times $[0,t]_{t\geq 0}$. Let $i_0,...,i_t$ is the succession of state observed, then transition probability matrix, P, can be estimated by Maximum Likelihood Estimation with a likelihood function $L_n = L_n(P)$ is

$$\begin{aligned}
L_n(P) &= P[X_0 = i_0,..., X_t = i_t] \\
&= P[X_0 = i_0,..., X_{t-1} = i_{t-1}]\, P[X_t = i_t \mid X_0 = i_0,..., X_{t-1} = i_{t-1}] \\
&= P[X_0 = i_0,..., X_{t-1} = i_{t-1}]\, P[X_t = i_t \mid X_{t-1} = i_{t-1}] \\
&= P[X_0 = i_0,..., X_{t-1} = i_{t-1}]\, P_{i_{t-1}i_t} \\
&\;\;\vdots \\
&= P[X_0 = i_0, X_1 = i_1]... P_{i_{t-2}i_{t-1}}\, P_{i_{t-1}i_t} \\
&= P[X_0 = i_0]\, P[X_1 = i_1 \mid X_0 = i_0]... P_{i_{t-2}i_{t-1}}\, P_{i_{t-1}i_t} \\
&= a_{i_0}\, P_{i_0 i_1}... P_{i_{t-2}i_{t-1}}\, P_{i_{t-1}i_t}
\end{aligned}$$

if $l_n(P) = \log L_n(P)$ then:

$$\begin{aligned}
l_n(P) &= \log\,(a_{i_0}\, P_{i_0 i_1}... P_{i_{t-2}i_{t-1}}\, P_{i_{t-1}i_t}) \\
&= \log\,(a_{i_0} \prod_{(i,j)\in S\,x\,S} P_{ij}^{N_{ij}}) \qquad \text{with } N_{ij} = \sum_{k=0}^{n-1} 1_{[X_k = i,\, X_{k+1} = j]} \\
&= \log a_{i_0} + \sum_{(i,j)\in S\,x\,S} \log P_{ij}^{N_{ij}} \\
&= \log a_{i_0} + \sum_{(i,j)\in S\,x\,S} N_{ij} \log P_{ij}
\end{aligned}$$

Using Lagrange and induction, results:

$$\hat{P}_{ij} = \frac{N_{ij}}{N_i} \qquad , N_i = \sum_{j\in\{0,1,...,n\}} N_{ij}$$

So, to find out a dependency level of nucleotides in position *i*-1 and *i*+1 in this base substitution mutation cases, it can be obtained from the optimal empirical transition probabilities or $\max\{\hat{P}_{ij} : i, j \in \{\text{A,T,C,G}\}\}$. For instance, Table 1 shows the case for mutation from C to T where position *i*-1 doesn't contains any significant information but cytosine will mutate become timine if position *i*+1 is filled with also cytosine with dependency level about 48.4 %.

From Table 2, it seems that for all cases, dependency level of each base in position i-1 cannot be calculated. It is completely match with Markov chain principle that the probability of a state is only influenced by one state before. As a consequence, if state i is given then state i-1 is unknown. Also from Table 2, we can know most of all bases (A, T, C, and G) will mutate if i+1 is filled with cytosine or C. The conclusion is supported by information contained on Table 3.

Table 1. Transition Probability Matrix (Mutation from C to T)

i-1	*i*				Total
	A	T	C	G	
A	169 (0.001)	4929 (0.017)	286842 (0.982)	88 (0.000)	292028
T	39 (0.000)	1093 (0.008)	137243 (0.992)	22 (0.000)	138397
C	67 (0.000)	7633 (0.025)	298743 (0.975)	5 (0.000)	306448
G	31 (0.000)	1318 (0.017)	77930 (0.983)	7 (0.000)	79286

i	*i*+1				total
	A	T	C	G	
A	86 (0.281)	15 (0.049)	198 (0.647)	7 (0.023)	306
T	2058 (0.137)	5432 (0.363)	7250 (0.484)	233 (0.016)	14973
C	306399 (0.383)	133043 (0.166)	326284 (0.407)	35032 (0.044)	800758
G	46 (0.377)	11 (0.090)	64 (0.525)	1 (0.008)	122

Table 2. Transition Probability Matrix (Twelve Mutation Cases)

No	**Case**	**n**	***i*-1**	***i*+1**	**No**	**Case**	**n**	***i*-1**	***i*+1**
1	T to A	36	-	A / C (30.6%)	**7**	A to C	562	-	C (92.5%)
2	C to A	306	-	C (64.7%)	**8**	T to C	12653	-	C (47.3%)
3	G to A	2418	-	G (41.1%)	**9**	G to C	46	-	G (45.7%)
4	A to T	141	-	C (58.2%)	**10**	A to G	2482	-	C (46.9%)
5	C to T	14973	-	C (48%)	**11**	T to G	20	-	C (45%)
6	G to T	18	-	C (55.6%)	**12**	C to G	127	-	C (50.4%)

Table 3. All Mutation Cases

i	*i*+1				
	A	T	C	G	
A	465 (0.168)	95 (0.034)	1194 (0.433)	1006 (0.365)	2760
T	2094 (0.138)	5445 (0.360)	7342 (0.485)	251 (0.017)	15132
C	5232 (0.395)	929 (0.070)	6514 (0.491)	586 (0.044)	13261
G	647 (0.246)	163 (0.062)	1237 (0.471)	582 (0.221)	2629

To investigate a dependency level of base in position i-1, we categorized all data set into 16 transitions shown on Table 4.

Table 4. Transitions

No	***i*-1**	***i*+1**	**No**	***i*-1**	***i*+1**	**No**	***i*-1**	***i*+1**	**No**	***i*-1**	***i*+1**
1	A	A	**5**	T	A	**9**	C	A	**13**	G	A
2	A	T	**6**	T	T	**10**	C	T	**14**	G	T
3	A	C	**7**	T	C	**11**	C	C	**15**	G	C
4	A	G	**8**	T	G	**12**	C	G	**16**	G	G

The aim of this categorization is to know what kind of bases that most able to fill position i by choosing the optimal probabilities. From these sixteen transitions, there is only one transition that fit with gene bank data set shown on Table 5; that is ATC.

Table 5. Transition ATC

i-1	i				i+1
	A	T	C	G	
A	1139 0.164	3786 0.545	1711 0.247	305 0.044	C

It means, for all mutation to T (A to T, C to T, and G to T), base A, C, and G will mutate become T if position i-1 is filled with adenine and cytosine in position i+1, but for another nine cases, it still unknown the contain of i-1. A graphical presentation of diagram of the optimal transitions is shown on Figure 1.

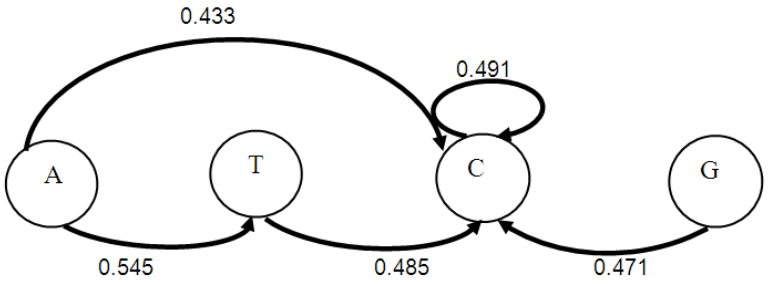

Figure 1. Diagram of Transition

References

[1]. Hadianti, R. (2006). *Queuing Theory*. Bandung: Bandung Institute of Technology.
[2]. Karlin, S and Howard M. T. (1994). *An Introduction to Stochastic Modeling*. California: Academic Press.
[3]. Resnick, S. (1992). *Adventures in Stochastic Processes*. Boston: Birkhäuser.