Journal of Data Mining in Genomics & Proteomics

Palit and Ngili, J Data Mining Genomics Proteomics
2018, 9:1
DOI: 10.4172/2153-0602.1000213

Research Article Open Access

The Highest Mutation in *mtDNA* Hypervariable Region and Application of Biostatistics with Nucleotide Base $X_{t\text{-}n}$ in Determining the Identity of the Mutation through a Transition Intensity Matrix

Palit EIY[1] **and Ngili Y**[2*]

[1]*Biostatistics Research Group, Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Cenderawasih, Jl. Kamp Wolker Waena, Kampus Baru, Jayapura-Indonesia*

[2]*Biochemistry Research Group, Department of Chemistry, Faculty of Mathematics and Natural Sciences, University of Cenderawasih, Jayapura, Indonesia*

[*]**Corresponding author:** Yohanis Ngili, Biochemistry Research Group, Department of Chemistry, Faculty of Mathematics and Natural Sciences, University of Cenderawasih, Jayapura, Indonesia, E-mail: yohanisngili@gmail.com

Received date: August 13, 2018; **Accepted date:** September 20, 2018; **Published date:** September 27, 2018

Abstract

Human mitochondrial DNA (mtDNA) have been used intensively in the field of forensic identification of victims or suspects of crime through biological evidence. The number of mtDNA molecules in a single cell are in the tens of thousands which enable analysis of samples very little or damaged. Till now there is no standard method for identification using mtDNA in a mass disaster victims such as natural disasters, wars and accidents so that the identification process can not run fast. This study found C16.223t variants in mtDNA sequences that can be used to divide the database into two groups so as to accelerate the process of identification through a mathematical algorithm. This variant has the highest frequency (29.7%) of the 91 polymorphic human mtDNA HVS1 along the 300 nucleotide (16,024-16,324) derived from the NCBI database as much as 142 sequences. MtDNA sequences obtained from data collection Papuan human mtDNA groups that have been published in the NCBI. The next variant that can be used as a classifier in a row in the sequence is 16,311; 16,304; 16,189; and 16,270 with the identity (T→c). For a matrix Q is reversible so the matrix and could have the opposite diagonal. Thus the above equation can be solved by using the diagonal method that can be written: $Q = S \wedge S^{-1}$. This equation could count the number of transitions and transversion substitution mutations that occur in a nucleotide sequence of mtDNA. With this grouping, the database can be reduced so as to accelerate the process of identification of samples. Expected method of grouping by the variant with the highest frequency can be developed in the codification database for forensic interest such as the police or the mtDNA database purposes of study anthropology and evolutionary biology.

Keywords: Mutations of *mtDNA*; Hypervariable region; Biostatistics; $X_{t\text{-}n}$; Transition intensity matrix

Introduction

Human mitochondrial DNA or mitochondrial DNA (*mtDNA*) has been used intensively in the field of forensic identification of victims or suspects of crimes through biological evidence. The number of *mtDNA* molecules in a single cell in the tens of thousands that make the analysis of samples very little or even damaged, for example bombing victims in Legian-Kuta, Bali (Bali Bombing), Jimbaran (Bali II), plane crash victims in Padang Bulan Medan, and other cases. Currently there is no standard method for identification using *mtDNA* in a mass disaster victims such as victims of bombings or natural disasters, wars and accidents so that the identification process is not running fast. Based on these problems, the analysis of *mtDNA* mutations in *mtDNA* hypervariable regions with the approach of biostatistics with application $X_{t\text{-}n}$ basis in determining the identity of the mutation through a transition intensity matrix becomes important to be used as a benchmark in the development of future biomolecular research [1-4].

MtDNA mutation rate ten times faster than nuclear DNA. This is because in mitochondria there are many free radicals are formed as a reaction byproduct of respiration and DNA polymerase enzyme χ *mtDNA* does not have a system repair during replication. Control region or D-loop mutations are relatively more tolerant because they do not encode proteins. This tolerance led to the D-loop mutations could accumulate, so the difference or D-loop variation between individuals is relatively high and therefore also called hypervariable regions. There are two hypervariable regions on *mtDNA* D-loop that is *hypervariable segment* 1 (HVS1) at position 15971-16569 and hypervariable region 2 (HVS2) at 1-589 positions. HVS1 area and HVS2 role in determining the identity or the identification process. D-loop diversity among human individuals in the world are used as a means of making up the phylogenetic tree of the human race [5].

MtDNA has been used in forensic even become evidence in court Europe and America [6-8]. The use of *mtDNA* especially for identification based on shared D-loop region individuals inline maternal lineage. Till now there is no standard method of identification based on the *mtDNA* sequences to sequence databases with a large number, for example in mass casualty disaster Bali or *mtDNA* fingerprint database in the police [9,10].

Materials and Methods

Amplification of *mtDNA in vitro*

MtDNA PCR process to obtain primary data is done by machine *Automatic thermal cycler* (Perkin Elmer) by 30 cycles. The initial stage

of the PCR process is the stage of initial denaturation at 94°C for 1 min, then go to programs PCR cycles with each cycle comprising three phases: denaturation performed at a temperature of 94°C for 1 min, annealing stage is performed at a temperature of 50°C for 1 min and extension or polymerization stage at a temperature of 72°C for 1 min. The end of all cycles carried out additional polymerization process at a temperature of 72°C for 4 min. DNA PCR results are stored at -20°C [11-15].

Analysis of the results of PCR and *mtDNA* sequencing method

PCR amplification results of the process are then analyzed by agarose gel electrophoresis 1.2% (w/v) using a Mini subTM DNA electrophoresis cell (Biorad). Agarose gel compositions can be made by dissolving agarose (Boehringer-Manheim) in 1X TAE buffer (Merck: 0.04 M Tris-acetate, 0.001 M EDTA pH 8.0). The solution is heated to its agarose late at all, then cooled to a solution temperature reaches 50°C-60°C. Electrophoresis process is carried out in 1X TAE buffer as a medium conductor current at a voltage of 75 volts for 45 min. Once this process is continued with *mtDNA* sequencing method using Sanger dideoxy method using Automatic DNA Sequencer is based on the dye terminator labeling method with materials and reactants from the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer) [15-17].

Sequence analysis and correction of *mtDNA*

The nucleotide sequences obtained from sequencing process has several nucleotide bases that can not be read so that coded N or shifting the reading so as to resemble a deletion or insertion and reading errors peaks. N code should be changed to correspond alkaline highest peak in the electropherogram legible. Deletions and insertions reassured by observing the distance between peaks readable and when an error occurs reading it must be repaired.

Correction with the help of computer applications is done using a program of ABI prism sequence navigator version 2.1. HVS1 sequences inserted in the table navigator sequence, number 16 024 is determined by looking for typical order TTCTTT and cut so Thymine 16 024 is equal to 1. The length of the sequence likened to cut the other end in order to obtain 300 sequences amount 16,324 nucleotides. Position on revised Cambridge Reference Sequence/rCRS (16024-16324) is included in the table navigator and conducted comparative sequence with sequences were corrected. Different nucleotide bases with rCRS or N of sequences that code is corrected will be marked on the line comparison. Simultaneously electropherogram images can be displayed and can be set zoomed in. Correction is done directly in the table navigator sequence. When data HVS1 has a pair complement HVS2 sequence data, the correction can be performed simultaneously by first doing the reverse complement sequences HVS2. HVS2 sequences that have been reverse complement sequence inserted in the table navigator and undergoing procedures as HVS1 sequence above.

Analysis of homology and polymorphism

Homology analysis manually can be done by comparing indirectly all sequences. Each sequence compared to rCRS and recorded variants possessed within Microsoft Excel table with the head of the column and the number of bases *mtDNA* sequence rCRS first row. The next line contains sequences with only variants only, which is no different bases filled rCRS point. Overall sequences set in the next lines. Homology analysis manually can be done with the command sequence sorting by selected columns, ie the most variants. With some time sorting and selecting the correct column homologous sequences could then be collected in adjacent rows.

Grouping of *mtDNA*

Based on the number and types of variants each number bases, selected variants with high frequency. Used variant with the highest frequency to divide the database into two groups, which have bases like the variant and which has a base such as rCRS. In each group is determined again variance with the next highest frequency and is used as a divider as the above. Grouping of stopped when the database can be reduced up to 10%-20% of the original large.

Results and Discussion

Electropherogram of the sequencing results and correction of electropherogram

Electropherogram of the identification number sequence by direct sequencing using a primer on HVS1 area of 256 sequences. Twenty-one of them electropherogram strecth poly C, which has a row of C (cytosine) at bases 16184-16194 and experience the chaos of peaks that can only be read up to 16194 positions only.

The use of DNA sequences in the study biostatistics and analysis of phylogenetic is that there is a change of bases according to the time that will be reconstructed an evolutionary relationship between one species and another, which changes the nucleotide bases in an organism/gene depends on various factors molecular, words another phylogenetic studies follow a stochastic process. Markov chain is a sequence of random variables $X=(X_0, X_1, X_2...)$ the nature, the opportunity of a state X_t at time t just rely on a fixed value n of the circumstances preceding X_{t-1}, X_{t-2}, ... X_{t-n}. Where n is the order of X. In other words, Markov chain is a stochastic process of the past that do not have an influence on the future when the present is known.

Correction of electropherogram done using applications on Sequence program navigator. Electropherogram corrected each peak to give more attention peaks of different variants with rCRS. Each variant is reassured by observing the peak height. Peak characteristics such as peak G is often weak and a row of peaks C is widened and disrupt the summit by his side to be considered. Possible heteroplasmi ignored for purposes of this study, bases selected from the highest peak in the event of symptoms heteroplasmy. For the record of heteroplasmy itself is considered in the process of identifying a missing person because it can be used as valid identification between two electropherogram people who have a maternal relationship [16-21].

Confirmation of the MITOMAP data

Variants with different nucleotide bases rCRS confirmed by the data on MITOMAP variant. Confirmation aims to convince variants were found. The confirmation of known, there are several variants that have not been reported in MITOMAP, variants are with identity 16,060 (G→t); 16,109 (A→g); 16,110 (G→c); 16,170 (A→g); 16,175 (A→g); dan 16,258 (A→t).

Transition intensity matrix (*Q*)

The process of substitution of the nucleotide sequence of DNA as described in the Poisson process, can be generalized by using a matrix *Q*, which is a matrix that determines the relative rate of change in each of the nucleotide bases along the DNA sequences. *Q* matrix can be formed by using the following theorem:

Theorem (matrix P (t) is the matrix of transition opportunities P(t)=pij(t) is given by:

$$Q = S \wedge S^{-1}$$

Proof:

$$p_{ij}(t) = P(X(t) = j|X(0) = i)$$

$$= \sum_{k=0}^{\infty} P(X(t) = j|X(0) = i,\ N(t) = k)P(N(t) = k|X(0) = i)$$

$$= \sum_{k=0}^{\infty} (P^k)_{ij} e^{-\mu t} \frac{(\mu t)^k}{k!}$$

(based theorem *k*-step matrix opportunities)

From the equation above, if the form $(P^k)_{ij}$ substituted in the form of a matrix will be obtained:

$$P_{ij}(t) = e^{-\mu t} \sum_{k=0}^{\infty} \frac{(P\mu t)^k}{k!}$$

If there are non-negative matrix *Q* and size *k* x *k*, then apply:

$$e^{Q} = \sum_{k=0}^{\infty} \frac{Q^k}{k!}$$

So the above equation becomes:

$$P(t) = p_{ij}(t) = e^{-\mu t} e^{\mu t}$$

$$= e^{Qt}$$

Thus obtained matrix A as follows:

Q = (P-I)μ

The rows of the matrix *Q* follows the order of the nucleotide bases A, C, G and T. The components of $p_{ij}(t)$ satisfies the following conditions:

1. $\sum_{j=1}^{n} p_{ij}(t) = 1$
2. $p_{ij}(t) > 0$ for $t > 0$
3. $P(t + s) = P(t) + P(s)$

This equation is known as the Chapman-Kolmogorov equation.

So that the above equation is fulfilled, then the matrix *Q* must satisfy:

$$\sum_{i=1}^{n} Q_{ij} = 0$$

So the equation above, can be obtained by defining:

$$Q_{ij} = -\sum_{i=1}^{n} Q_{ij}$$

Note that $Q_{ij}>0$ because it can be interpreted as a change of base *i* flow into base *j*, $Q_{ij}< 0$ is the total current leaving nucleotide changes in *mtDNA* $_i$, therefore Q_{ij} value is less than zero. The number of nucleotide substitutions per unit time is the total rate rate μ are:

$$\mu = -\sum_{i=1}^{n} \pi_i Q_{ij}$$

Chapman-Kolmogorov equation is then obtained a differential equation forward and backward:

$$\frac{d}{dt}P(t) = P(t)Q = QP(t)$$

This equation can be solved by $P(0)=I$ is the initial condition on an equation that will result in $P(t)=\exp((t)Q)$.

For a matrix Q is reversible, the matrix is to have the opposite/inverse and could shaped diagonal. Thus the above equation can be solved by using the method diagonal form which can be written: $Q = S \wedge S^{-1}$.

From the equation above, it can be substituted *mtDNA* mutation data as follows: 500 existing variants, a variation of 95% substitution of transition and transversion variant is only 5%. In terms of the point or base numbers that experience, variations in transition achieve 83% and 17% transversions mutation. Variations transition itself can be grouped into two: T>C and A>G with C>T and G>A. These groupings are based on complement; T>C complements A>G and C>T complements G>A. Which means when encountered variant T>C in the light chain with a primer to sequence HVS1 then the heavy chains that were sequenced in the region will find variants HVS2, A>G. By grouping these variants can be known tendency is formed, which is evident from the number of bases or of total *mtDNA* variants both have a value that is not much different.

Character variant in question is the type and amount of 91 polymorphic variants in approximately 300 bases. Variations transition is a variation of purine bases with purine or pyrimidine to pyrimidine whereas transversion variation is a variation of a purine with a pyrimidine or vice versa. Variations T>C in the complement light chain with a variation of A>G in the heavy chain and vice versa, as well as C>T complements G>A and C>A to G>T [19-21].

Grouping based on the highest frequency

Variants with the highest frequency can be seen from the discussion above, but with attention to the linkages between variants in the preceding table, the variant could used as a divisor is: C16,223t, T16,311c, T16,304c, T16,189c and C16,270t. Using these variants do grouping that split the data into two groups of the same base with rCRS and which are not. The group with the same base with rCRS further subdivided with the next highest frequency variants.

Conclusion

The results of this research has been to simplify the system of determining the mutation and found the system clustering algorithm uses a variant with the highest frequency to reduce the number of databases that need to be compared. Of the hundreds of sequences that were analyzed contained 91 polymorph with divergences from 0.3 to

5.6 and found six variants that have not been reported in MITOMAP, namely G16,060t, A16,109g, G16,110c, A16,170g, A16,175g and A16,258t. Varian C16,223t which is a variant with the highest frequency (29.6%) was found to divide the database into two groups. 16,223t group has 27 morph can be divided further by variant G16,129a, while group consisting 16,223C are 64 morph can be grouped further variants: T16,311c, T16,304c, T16,189c, C16,270t. Use of *mtDNA* sequences in the study biostatistics and phylogenetic analysis follows a stochastic process, of the random variable $X=(X_0, X_1, X_2...)$ the nature, the opportunity of a state X_t at time t just rely on a fixed value n of the circumstances before ie X_{t-1}, X_{t-2}, ... X_{t-n}. In this case the Markov chain is a stochastic process bases nucleotide changes with the change of time, with the odds of various nucleotide changes between states depends only on the previous state. Calculation of nucleotide substitution mutation by observing the intensity of the transition matrix then this equation can be solved by $P(0)=I$ is the initial condition of the equation that produces the form: $P(t)=\exp((t)Q)$. For a matrix Q is reversible, the matrix is to have the opposite/inverse and could shaped diagonal. Thus the above equation can be solved by using the method diagonal form which can be written: $Q=S\wedge S^{-1}$. By using this classification system, the database needs to be compared can be reduced and the identification process can be accelerated.

Acknowledgement

The author gratefully acknowlegde the support of this work by Directorate of research and community service (Ditlitabmas)-Ministry of Research, Technology and Higher Education-Republic of Indonesia on funding research through research grants scheme to EIYP. Thanks to the Chairperson of the Biochemistry Laboratory for the isolation and amplification of *mtDNA* and facilities in the Computing Laboratory, Faculty of Mathematics and Natural Sciences, University of Cenderawasih for simulation and calculation and determination of the identity of *mtDNA* mutations.

References

1. Bendall KE, dan Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57: 248.
2. Cavalier L, Cazin E, Jalonen P, Gyliensten U (2000) MtDNA substitution rate and segregation of heteroplasmy in coding and non coding region. Hum Genet 107: 45.
3. Yang Z, Rannala B (2000) Bayesian phylogenetic inference using dna sequences: A Markov Chain Monte Carlo Method. Mol Biol Evol 14: 717-724.
4. Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J Mol Evol 43: 304-311.
5. Ngili Y, Noer AS, Ahmad AS, Syukriani YF, Natalia D, et al. (2012) Variants analysis of human mitochondrial genome mutation: Study on Indonesian human tissues. Int J ChemTech Res 4: 720-728.
6. Ngili Y, Ubyaan R, Palit EIY, Bolly HMB, Noer AS (2012) Nucleotide mutation variants on D-Loop HVS1/HVS2 mitochondrial DNA region: Studies on Papuan Population, Indonesian. Euro J Sci Res 72: 64-73.
7. Ngili Y, Palit EIY, Bolly HMB, Ubyaan R (2012) Cloning and analysis of heteroplasmy in hypervariable segment 1 (HVS1) D-loop in mitochondrial DNA of human isolates of Timika and Wamena in Highlands of Papuan Province, Indonesia. J Appl Sci Res 8: 2232-2240.
8. Ngili Y, Bolly HMB, Ubyaan R, Jukwati, Palit EIY (2012) Studies on specific nucleotide mutations in the coding region of the ATP6 gene of human mitochondrial genome in populations of Papuan Province-Indonesia. Aust J Bas Appl Sci 6: 111-118.
9. Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708-713.
10. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, et al. (2005) M. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308: 1034-1036.
11. Attimonelli M, Altamura N, Benne R, Brennicke A, Cooper JM, et al. (2000) MitBASE : A comprehensive and integrated mitochondrial DNA database. The present status. Nucl Ac Res 28: 148-152.
12. Coskun PE, Pesini ER, Wallace DC (2003) Control region mtDNA Variant: Longevity, climatic, adaptation, and forensic conundrum. PNAS 100 (5): 2174-2176.
13. Fox TD (1987) Natural variation in the genetic-code. Annual Review of Genetics 21: 67-91.
14. Hill C, Soares P, Mormina M, Macaulay V, Meehan W, et al. (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 23: 2480-2491.
15. Sudoyo H, Marzuki S, Mastaglia F, Carroll W (1992) Molecular genetics of Leber's hereditary optic neuropathy: study of a six generation family from Western Australia. J. Neurol Sci 108: 7-17.
16. Wei YH, Lee HC (2003) Mitochondrial DNA mutation and oxidative stress in mitochondrial disease. Advances in Clinical Chemistry 37: 84-128.
17. Zeviani M, Tiranti V, dan Piantadosi C (1998) Reviews in molecular medicine. Mitochondrial disorders. Medicine 77: 59-72.
18. Li R, Greinwald JH Jr, Yang L, Choo DI, Wenstrup R.J, et al. (2004) Molecular analysis of the mitochondrial 12S rRNA and tRNASer(UCN) genes in paediatric subjects with non-syndromic hearing loss. Journal of Medical Genetics 41: 615-620.
19. Ghezzi D, Marelli C, Achilli A, Goldwurm S, Pezzoli G, et al. (2005) Mitochondrial DNA haplogroup K is associated with a lower risk of Parkinson's disease in Italians. European Journal of Human Genetics 13: 748-752.
20. Maksum IP, Farhani A, Rachman SD, Ngili Y (2013) Making of the A3243g mutant template through site directed mutagenesis as positive control in PASA-Mismatch three bases. Int J ChemTech Res 5: 441-450.
21. Siallagan J, Maryuni A, Jukwati, Ngili Y (2015) Single nucleotide mutations of intergenic and intragenic region in mitochondrial genome from different individuals from Papua-Indonesia.Der Pharma Chemica 7: 334-339.