March - April 2020
ISSN: 0193-4120 Page No. 11496 - 11502

Markov Model for Full Genome Sequence Generation

Foo Weng Lim[1], Yong Kheng Goh[2]

[1,2]Universiti Tunku Abdul Rahman, Malaysia
limfw@utar.edu.my

Article Info
Volume 83
Page Number: 11496 - 11502
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 16 April 2020*

Abstract
This work is devoted to introducing a Markov Chain method to generate a long sequence written in this four-letter alphabet namely; Adenine (*A*), Cytosine (*C*), Guanine (*G*) and Thymine (*T*). The algorithm can be used to generate a new genomic DNA sequence that captures the statistical properties of the original sequence as well as preserve its statistical properties of the sequence for any case of *N*-grams. An *N*-grams is a subsequence of length *N* in the genomic DNA. Later, by counting the occurrence of different *N*-grams, and a signature vector of a genetic text, called contrast value is constructed. With the contrast value vector and correlation as distance measures, a phylogenetic tree is constructed. The phylogenetic trees manage to group the organisms according to its kingdom which does not against the commonly accepted phylogenetic tree.

Keywords: *N*-gram, contrast value, Markov Chain, Metropolis-Hasting

1. Introduction

The Deoxyribonucleic acid or DNA is a long chain sequence made from four basic repeating units called nucleotides, the four basic nucleotides founded in the DNA sequence are adenine (*A*), cytosine (*C*), thymine (*T*) and guanine (G). The long-chain sequence carrying the biology instructions used in the growth, development, functioning, and reproduction of all living organisms. Every living organism possesses a genomic signature that does not depend on knowledge of individual genes or the alignment of homologous sequences. A genomic signature profile is invariant across the genome of an organism and is similar for closely related species and shows a dissimilarity pattern between non-related species. By analyzing the similarity of genomic signature among organisms, we could construct a phylogenetic tree.

In general, commonly used techniques in building phylogenetic trees are based on sequence alignment which compares the similarity of the fragment from different organisms, and the observed similarity measures were used to constructing a phylogenetic tree that shows the probable evolution of various organisms. Although the sequence alignment method was successful in building various phylogenetic trees, most are only use a small portion of the organisms. There are various limitations associated with sequence alignment method such as long sequence is used to be analyzed as mention on the complexity of multiple sequence alignment by (Susana Vinga, and Jonas Almeida, 2003) and review on multiple sequence alignment by (Biswanath and Gautam, 2017).

To address the limitation of the sequence alignment method, various alignment-free methods have been developed recently. Although the sequence alignment always provides reliable results compare to the alignment-free method, it fast and easy as reviewed by (Bonham, 2014), and the benefits of non-sequence alignment-based by (Susana Vinga, and Jonas Almeida, 2003), so alignment-free always chosen by the researcher. Generally, the alignment-free method can separate into two different categories. The first is based on the frequency of oligomer and the method that does not depend on the word frequency. In a non-sequence alignment base, the first method working on the statistic of word frequency in a sequence, and the distances are defined over a Cartesian space and then working on the information content of the frequency distribution. And the next techniques include the use of an algorithm of information theory called - Kolmogorov complexity which explains the sequence in the context of the computer program.

In alignment-free approach, the common method is based on word (oligomer) frequency, where the sequence will slice into a small segment with fixed word length-

March - April 2020
ISSN: 0193-4120 Page No. 11496 - 11502

namely as *N*-grams, later the frequency of each word will be used to constructing a frequency vector- namely as composition vector. The composition vector will be used to construct the phylogenetic tree with correlation coefficient measures as distance measures for each composition vector. We know that the average length of the human chromosome is about 3 million-based pairs, but the average length of bacteria chromosome is just about 130 thousands-based pair. So, the length of each chromosome organisms is an important feature used in classification.

In this paper, we use the Markov chain Monte Carlo (MCMC) techniques to generate a genomic sequence based on its composition vectors constructed from the segment of the original sequence. The simulated sequences later will apply the alignment-free approach to build the phylogenetic tree.

2. Methods

Markov chain Monte Carlo (MCMC) techniques first appeared in statistical physics. The idea of MCMC is simple, to sample randomly from a specific probability distribution then design a Markov chain whose long-time equilibrium is the desire distribution. In this study, we focus on the Metropolis-Hasting method to generate the new sequence with it preserve the statistical properties of original sequences.

2.1 Metropolis-Hasting algorithm

Consider a genomic sequence that combined from the four basic nucleotides namely; adenine (*A*), cytosine (*C*), thymine (*T*) and guanine (*G*), then one way of performing a feature extraction is to describe it in term of its subsequence. An *N*-grams is a subsequence of length *N*. Here we applied a Metropolis-Hasting scheme to generate a sequence while preserving it statistical distribution $\prod\left(\boldsymbol{X}_{i=4^N}^{(N)}\right)$, with *N* refer to it *N*-gram. For instance, the probability mass function $\prod\left(\boldsymbol{X}_{i=4^N}^{(N)}\right)$ describes as follows:

$N = 1$, $\prod\left(\boldsymbol{X}_{i=4^N}^{(1)}\right) =$

$$\left[\prod(\boldsymbol{X}_{A}^{(1)}),\prod(\boldsymbol{X}_{\boldsymbol{T}}^{(1)}),\prod(\boldsymbol{X}_{G}^{(1)}),\prod(\boldsymbol{X}_{\boldsymbol{C}}^{(1)})\right]_{1\times4} \quad (1)$$

$N = 2$, $\prod\left(\boldsymbol{X}_{i=4^2}^{(2)}\right) =$

$$\left[\prod(\boldsymbol{X}_{AA}^{(2)}),\prod(\boldsymbol{X}_{\boldsymbol{AT}}^{(2)}),\prod(\boldsymbol{X}_{AG}^{(2)}),\cdots,\prod(\boldsymbol{X}_{\boldsymbol{CC}}^{(2)})\right]_{1\times16} \quad (2)$$

$N = 3$, $\prod\left(\boldsymbol{X}_{i=4^3}^{(3)}\right) =$

$$\left[\prod(\boldsymbol{X}_{AA}^{(3)}),\prod(\boldsymbol{X}_{\boldsymbol{AAT}}^{(3)}),\prod(\boldsymbol{X}_{AAG}^{(3)}),\cdots,\prod(\boldsymbol{X}_{\boldsymbol{CCC}}^{(3)})\right]_{1\times64} \quad (3)$$

In general, the element of the probability mass function is described as following

For *N* ,

$$\prod(\mathbf{X}_{i=4^N}^{(N)}) = \frac{\text{frequency of } x_i \text{ in sequence x}}{L-N+1} \quad (4)$$

The *N*-gram can be obtained from its predecessor by dropping its first character and adding the next character at its end. In general, the number of components for $\prod\left(\boldsymbol{X}_{i=4^N}^{(N)}\right)$, is 4^N different subsequence with the length of *N* and the frequency of each element in an *N*-gram is overlapping occurrence in a sequence. The Metropolis-Hasting algorithms provide an approach to generate a sequence of a random variable which converges to $\prod\left(\boldsymbol{X}_{i=4^N}^{(N)}\right)$.

The Metropolis-Hastings (MH) algorithm is the prototype for a class of Markov chain Monte Carlo methods that propose transitions between states and then accept or reject the proposal. These methods generate a correlated sequence of random samples that convey information about the desired probability distribution.

Consider a sequence $X_1X_2X_3\ldots X_n$. Interpret X_n as the state of the sequence at time *n*. If there exists a set of numbers $P_{i,j}$, $i = j = 1,2,\ldots, n$, such that whenever the process is in state *i* then independent of the past state, the probability that the next state is *j*, then we say that the collection $\{Xn \ , \ n \geq 0\}$ constitutes a Markov chain having transition probabilities. Since the process must be in some state after it leaves states *i*, these transition probabilities satisfy

$$\sum_{j=1}^{n} P_{i,j} = 1, \ \ i = 1,2,\ldots,n \quad (5)$$

A Markov chain is said to be irreducible if for each pair of state *i* and *j* is a positive probability with initial state *i* that the process will ever go to the state *j*

According to the irreducible Markov chain. The $\prod_j$ is the long-run proportion of time that the process is in state *j*. The quantity $\prod_j$, $j = 1,2, \ldots, n$ can show to be the solution of the following set of problem:

$$\prod_j = \sum_{j=1}^{N} \pi_i P_{ij} \quad j = 1,2,\ldots, N \quad (6)$$

$$\sum_{j=1}^{N} \pi_i = 1 \quad (7)$$

2.2 Illustrations

Consider a sequence ATCGAC…. with length *L* and we split it into individual nucleotide as *A, T, C, G, A, C* … and it respective vectors as X_1, X_2, X_3, …, X_L and it transition matrix, *Q* for 1-grams as follows:

March - April 2020
ISSN: 0193-4120 Page No. 11496 - 11502

$$Q = \begin{array}{c} \begin{matrix} A & T & G & C \end{matrix} \\ \begin{bmatrix} P_{AA} & P_{AT} & P_{AG} & P_{Ac} \\ P_{TA} & P_{TT} & P_{TG} & P_{Tc} \\ P_{GA} & P_{GT} & P_{GG} & P_{Gc} \\ P_{CA} & P_{CT} & P_{CG} & P_{Cc} \end{bmatrix} \end{array} \begin{matrix} A \\ T \\ G \\ C \end{matrix} \quad (8)$$

where $P_{ij} = \frac{f_{ij}}{L-N+1}$ is the frequency of word *ij* in sequence, and $\prod_j = \sum_{j=1}^{N} \pi_i P_{ij}$, $j = 1,2,...,N$, with *N* = 1,2,3,… and $\sum_{j=1}^{N} \pi_i = 1$

For *k* = 1, $\sum_{j=1}^{4} \pi_j = \pi_A + \pi_T + \pi_G + \pi_c$ (9)

For *k* = 2, $\sum_{j=1}^{16} \pi_j = \pi_{AA} + \pi_{AT} + \pi_{AG} + ... + \pi_{cc}$ (10)
and so on.

2.3 N-grams

Consider a transition probability function $q(x_i , x_{i+1})$ that takes an *N*-gram in state x_i to state x_{i+1} which provides possible change in $\mathbf{X}_{(i)}$ for any value of *i*. Thus, if $\mathbf{X}_{(i)} = a$ then with probability $q(a,b)$ the value of *b* is for $\mathbf{X}_{(i+1)}$ and indeed with subsequent probability $\alpha(a,b)$ for some specified *acceptance function* $\alpha(a,b)$, we accept *b*, so that $\mathbf{X}_{(i+1)} = b = x_{i+1}$. However, we *reject b*, which occur with probability $1-\alpha(a,b)$, then $\mathbf{X}_{(i+1)} = a = x_{i+1}$, that is the value of $\boldsymbol{X}$ remain unchanged. Thus.

$$P\left(\boldsymbol{X}_{(i+1)=b} \middle| \boldsymbol{X}_{(i)=a}\right) = q(a,b)\alpha(a,b) \quad \text{for } a \neq b \quad (11)$$

As long as the acceptance probability function α is suitably chosen, then the resulting sequence $x_{(i+1)}$ converges as $i \to \infty$ to a series of values from π (x). This is dependent upon the transition probability function *q* satisfying certain standard conditions, given form of π (:) and *q*(:), the standard choice for alpha is

$$\alpha(a,b) = \min\left\{1, \frac{\pi(b)q(b,a)}{\pi(a)q(a,b)}\right\} \quad (12)$$

In order to generate a sequence, *S*, we start with constructing the transition matrix *Q*, when *Q* is irreducible and aperiodic, we seek the unique equilibrium distribution π that satisfy the following:

$$\pi = \pi Q \quad (13)$$

3. Result

Our aim is to used the Metropolis-Hasting to regenerate the new genomic sequence while preserving its original statistical properties. So, one of the commonly used technique in comparing two samples are drawn from the same distribution or not is Kolmogorov-Smirnov goodness of fit test (KS-test). Kolmogorov-Smirnov test goodness of fit test is used to compare two samples that are significantly different from each other by using the empirical distribution function.

Here, we name the two distribution as a targeted distribution which referring to original sequences, while the sample distribution referring to the simulated sequence which construct based on the statistical properties of original sequences. The frequency data set will be normalized for both target and sample data. For illustration, here we use the Arabidopsis thaliana mitochondrion, complete genome sequences (NC_001284.2), which downloaded from National Center for Biotechnology Information (NCBI) website.

3.1 Hypothesis Testing

H_0: There is no significant difference between the two distribution
H_1: There is a significant difference between the two distribution.

Here, the following formula using to calculate the value of Kolmogorov-Smirnov goodness of fit test:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i)$$

$$D_n = \sup_x \left|F_{n(x)} - F(x)\right|$$

In Table 1, it has shown the normalized frequency for Arabidopsis thaliana mitochondrion (NC_001284.2)

Table 1: Normalized frequency for Arabidopsis thaliana mitochondrion

N 1-gram	Targeted distribution Original sequence	Sample distribution Simulated sequence
A	0.279251	0.270546
T	0.273054	0.273220
G	0.222414	0.221049
C	0.225281	0.226186

where the value of the targeted distribution is normalized frequency count (*f*) based on original sequence and the value of sample distribution is the normalized frequency count based on generated sequence, both refer to *N* = 1. Here, both sample sequence and original sequence with length *L* = 366,924 base-pairs respectively. From Table 1. The KS-test value is 0.25 and *p*-value is 0.9969, so we can conclude that the

March - April 2020
ISSN: 0193-4120 Page No. 11496 - 11502

result is not significant at $p < 0.05$. Therefore, we do not reject H_0, and conclude that there is no significant difference between the original sequences and simulated.

Table 2: The KS-test statistics value for various 2-gram to 5-gram with significant level (0.05)

N	KS-test statistics	*p*-value
2	0.1250	0.9991
3	0.0625	0.9994
4	0.0664	0.6105
5	0.0402	0.3805

From Table 2, we can notice that, although increasing the size of *N*, the KS-test still do not reject the null hypothesis and conclude that there is no significant difference between the original sequences and simulated sequence at $\alpha = 0.05$.

3.2 Contrast value

Let the *S* be a genome sequence over the four-letter alphabet {*A, T, G, C*} with length *L*. We define an *N*-gram (ψ) of length *N* as a string of *N* characters over the given alphabet *A*. A genome sequence can be viewed as a stream of overlapping *N*-grams one after another. In the sequence *S*, there is total *L* - *N* + 1, *N*-grams of length *N*. *L* is the length of sequence *S*.

To map a long genome or sequence to a point in high dimension space for comparison, we need a way to describe the characteristic signature (genome signature) of a sequence and a similarity measure to rank the relatedness of different sequences. A simple genome signature and comparison of these sequences are calculating the correlation for the frequency vectors.

For any *N*-gram, let denote $f(\psi, S)$ by the observed frequency of *N*-gram ψ in the text *S*. The frequency ψ is the number of occurrences of ψ in the sequence *S* divided by (*L* – *N* + 1).

Given the observed frequencies of the *N*-grams of size (*N* – 2) and of size (*N* – 1), we can calculate the expected frequency of the *N*-grams in genome sequence *S* as

$$E(a_1..a_N, S) = \frac{f(a_1..a_{N-1}, S) \times f(a_2..a_N, S)}{f(a_2..a_{N-1}, S)} \quad (14)$$

In the case of *N* = 2, the formula (1) is reduced to

$$E(a_1a_2, S) = f(a_1, S) \times f(a_2, S) \quad (15)$$

There exist 4^N different *N*-grams of length *N*. For example, if *N* = 2, there exist 16 *N*-grams: *AA*, *AT*, ... , *TT*. Thus, the *N*-gram bias in *X* could be defined by the contrast value

$$q(\psi, S) = f(\psi, S) - E(\psi, S) \quad (16)$$

The *N*-vocabulary vector of the genome is the set of the characteristic values for all the *N*-grams of length *N*. We denote $\mathbf{q}(S) = \{q(\psi_1, S), ..., q(\psi_n, S)\}$ be the contrast *N*-vocabulary. Since the size of the genome alphabet is four, then the size of the *N*-vocabulary is $n = 4^N$. Thus, a genome with any length is mapped into a point in the 4^N dimensional space. To compare the similarity between two genomes S_1 and S_2, we use the correlation as the similarity measures:

$$C(S_1, S_2) = \frac{\mathbf{q}^T(S_1)\mathbf{q}(S_2)}{\|\mathbf{q}(S_1)\|\|\mathbf{q}(S_2)\|} \quad (17)$$

3.3 Phylogenetic tree construction

As mention early, our aim is applied the Metropolis-Hasting algorithms to regenerate a new sequence, in which the generate sequences possess an identical distribution as the original sequence. For illustration, we use 18sRNA sequences from various family, - birds, mammals, reptiles and amphibians to constructing the phylogenetic tree. 18sRNA is a component of the small eukaryotic ribosomal subunit and it is one of the basic components of all eukaryotic cells. First, for each 18sRNA sequence, we randomly select a segment of sequence *S*, with length *L* = 1000 from the complete 18sRNA sequences, then each sequence *S* will use to construct the transition matrix *Q* as mention in section 2.2. Then we solve the equation $\pi = \pi Q$ to obtain the π. With each π, *N*-gram distribution then genomic signature was created by applying the contrast value (in section 3.2) for each sequence *S*. With the genomic signature, applied the Pearson-correlation coefficient as distance measures, the phylogenetic tree shown as Figure 1. From the phylogenetic tree, we manage to group the species according to its family.

March - April 2020
ISSN: 0193-4120 Page No. 11496 - 11502

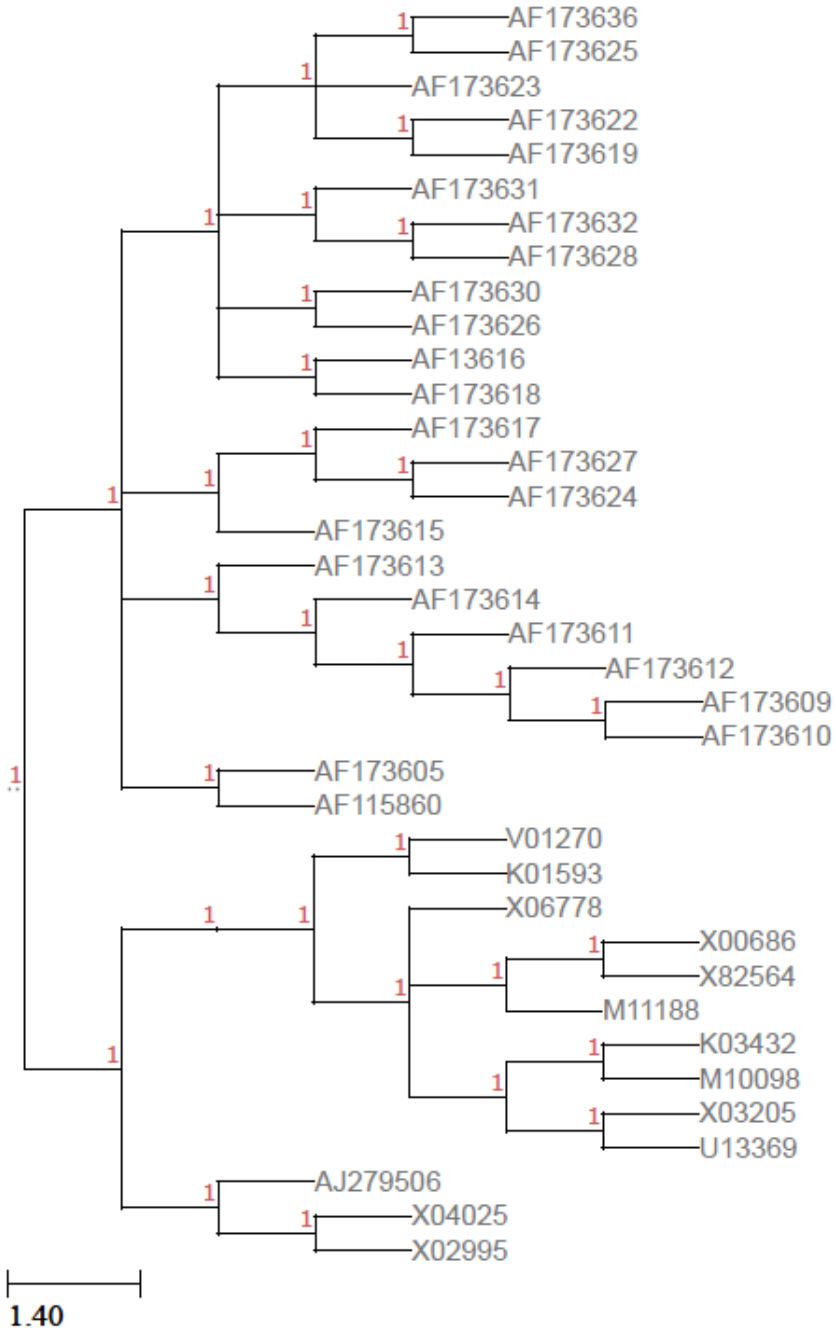

Figure 1: The phylogenetic tree for 18s rRNA sequence with estimation formula (14) with $N = 8$

4. Conclusion

For comparison of whole-genome sequences, multiple sequence alignment of a few selected genes is not appropriate, also sequence alignment is time-consuming. One alternative approach is to use an alignment-free method based on the frequency profiles of whole genomes. Every living organism possesses a genomic signature that does not depend on knowledge of individual genes. The genomic signature profile is invariant across the genome of an organism and is similar for closely related species and shows a dissimilarity pattern between nonrelated species. By analyzing the similarity of genomic signature. However, use frequency to create the genomic signature for every living organism will create an additional issue, as frequency relay on the length of the genome sequence. Here we proposed the

Metropolis-Hasting synthetic sequence generation approach which can be used to create the genomic signature which does not depend on the genome sequence length, also it preserves the statistical properties as the original sequence. Also, we use the generated sequence, we able to group the 18sRNA into the correct family. This approach particularly useful when we have only partial, incomplete genome sequences, we can apply the Metropolis-Hasting method to regenerate a new sequence which has statistical distribution identical to the original sequence.

References

[1] Ahmi, A., & Mohamad, R. (2019). Bibliometric Analysis of Global Scientific Literature on Web Accessibility. International Journal of Recent Technology and Engineering, 7(6), 250–258.

[2] A.E. Shannon, (1984). A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, 379-423.

[3] Almeida,J.S., Carrico,J.A., Maretzek,A., Noble,P.A. and Fletcher,M (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17, 429-437.

[4] Andreas Schulz, Daniela Zoller, Stefan Nickels, Manfred E. Beutel, Maria Blettner, Philipp S. Wild, and Harald Binder (2017), Simulation of complex data structures for planning of studies with focus on biomarker comparison, *BMC Medical Research Methodology*, 17. 90-102

[5] B.L. Hao, J. Qi, B. Wang, (2003). Prokaryotic Phylogeny Based on Complete Genomes without Sequence Alignment, *Modern Physics Letters B*, 2,1-4.

[6] Biswanath Chowdhury, Gautam Garai (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics* 109, 419-431.

[7] Bonham-Carter (2014). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in Bioinformatics*, 15,890-905.

[8] Buhlmann P, Wyner (1999). A Variable length Markov chains. *Annals of Statistics*, 27, 480-513.

[9] PK Burma, A Raj, JK Deb, SK Brahmachari (1992). Genome analysis: a new approach for visualization of sequence organization in genomes. *Journal. Bioscience*,17, 395-411.

[10] Cedric Notredame, Desmond Higgins, Jaap Heringa (). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302, 205-217.

[11] Cenac P, Chauvin B, Ginouillac S, Pouyanne N (2009). Digital Search Trees and Chaos Game Representation. *ESAIM-Probability and Statistics*, 13,15-37.

[12] Raymond H. Chan, Tony H. Chan, Hau Man *Y*eung, Roger Wei Wang (2012). Composition vector method based on maximum entropy principle for sequence comparison. *Journal IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9, 79-87

[13] Darren J.Wilkinson (2007), Bayesian methods in bioinformatics and computational systems biology. *Briefing in Bioinformatics*,8, 109-116.

[14] Deschavanne P, Tuffery P (2008), Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie*, 90, 615-625.

[15] Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology Evolution*, 16, 1391-1400.

[16] G.J. Phillips, J. Arold, R. Ivarie (1987). Mono-Through Hexanucleotide Composition vector of the Escherichia Coli Genome: A Markov Chain Analysis. *Nucleic Acids Research*, 15, 2611-2626.

[17] G.W. Stuart, K. Moffett, J.J. Leader (2002). A Comprehensive Vertebrate Phylogeny Using Vector Representations of Protein Sequences from Whole Genomes. *Molecular Biology and Evolution*, 19, 554-562.

[18] G.W. Stuart, M.W. Berry (2004). An SVD-Based Comparison of Nine Whole Eukaryotic Genomes Supports a Coelomate Rather than Ecdysozoan Lineage. *BMC Bioinformatics*, 5, 204-217.

[19] Gentleman J, Mullin R (1989). The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability. *Biometrics*; 45, 35-52.

[20] J. Gibbs, M. B. Dale, H. R. Kinns and H. G. MacKenzie (1971). The transition matrix method for comparing sequences; its use in describing and classifying proteins by their amino acids sequence. *Systematic Zoology*, 20, 417-425.

[21] Hao B, Qi J (2004). Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *Journal of Bioinformatic Computational Biology*, 2, 1-19

[22] Hussain, A., Mkpojiogu, E. O. C., Jamaisse, A., & Mohammed, R. (2018). Grab mobile app: A UX assessment on mobile devices. Journal of Advanced Research in Dynamical and Control Systems, 10(10), 1233–1238.

[23] W. K. Hastings (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, 97-109.

[24] Helena Skutkova, MartinVitek, Karel Sedlar, Ivo Provaznik (2015). Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of Theoretical Biology*, 385, 20-30.

[25] Miller,R.T., Christoffels, A.G., Gopalakrishnan,C., Burke,J., Ptitsyn,A.A., Broveak,T.R. and Hide,W.A (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Research.*, 9,143-1155.

[26] Orion Penner , Peter Grassberger , Maya Paczuski (2011). Sequence Alignment, Mutual Information, and Dissimilarity Measures for Constructing Phylogenies. *PLoS ONE*. 6. e14373

[27] P.A.S. Nuin, Z. Wang, E.R.M. Tillier (2006). The Accuracy of Several Multiple Sequence Alignment Programs for Proteins, *BMC Bioinformatics*,7,1-18.

[28] Pevzner, P.A (1992). Statistical distance between texts and filtration methods in sequence comparison. Computational. Applied. Bioscience, 8, 121-127.

[29] Qi J, Wang B, Hao B-I.(2004). Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *Journal of Molecular Evolution* 58, 1-11.

[30] Reinert,G., Schbath,S. and Waterman,M.S. (2002). Probabilistic and statistical properties of words: an overview. *Journal of Computational. Biology*, 7, 1-46.

[31] Roy A, Raychaudhury C, Nandy (1998). An Novel techniques of graphical representation and analysis of DNA sequences - A review. *Journal of Biosciences*, 23, 55-71.

[32] S.B. Hedges, K.D. Moberg, L.R. Maxson (1990). Tetrapod Phylogeny Inferred from 18S and 28S Ribosomal RNA Sequences and a Review of the Evidence for Amniote Relationships. *Molecular Biology and Evolution*,7, 607-633.

[33] S.B. Needleman, C.D. Wunsch (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal Molecular Biology*, 48, 443-453.

[34] Sims GE, Jun SR, Kim SH (2008). Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of National Academy of Science*, 106, 2677-2682

[35] Soares I, Goios A, Amorim (2012). A. Sequence comparison alignment-free approach based on suffix tree and L-words frequency. *Scientific World Journal* , 2012, 450124.

[36] Stuart, G.W., Moffett ,K. and Leader, (2002). A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Molecular Biology Evolutional*, 19, 554-562.

[37] Susana Vinga, Jonas Almeida (2003). Alignment-free sequence comparison review. *Bioinformatics*, 9, 513-523.

[38] Paul Viola and William M. Wells III (1995). Alignment by maximization of mutual information. *International journal of computer vision*, 24, 137-154.

[39] Wang L, Jiang T (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1, 337-48

[40] Wu X, Wan X, Wu G (2006). Phylogenetic analysis using complete signature information of whole genomes and clustered neighbor-joining method. *International Journal Bioinformatics Research Application*, 2, 219-48.

[41] Smith, Temple F., Waterman, Michael S (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology*. 147, 195-197.

[42] Wu, T.J., Hsieh,Y.C. and Li, L.A (2001). Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, 57, 441-443.

[43] X. Xia, Z. Xie, K.M. Kjer (2003). 18S Ribosomal RNA and Tetrapod Phylogeny. *Systematic Biology*, 52, 283-295.

[44] Xin Chen, Sam Kwong, Ming Li. (2014). A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison. *Genome Informatics*, 10, 51-61.

[45] Yang and Rannala (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology Evolution*, 14, 717-724.

[46] Z.G. Yu, X.W. Zhan, G.S. Han, R.W. Wang (2010). Proper Distance Metrics for Phylogenetic Analysis Using Complete Genomes without Sequence Alignment. *International Journal Molecular Sciences*,11, 1141-1154.