AMERICAN JOURNAL OF PHYSIOLOGY, BIOCHEMISTRY AND PHARMACOLOGY, 2021
VOL 11, NO. 5, PAGE 01-03

Short Review

Open Access

Short Review on Mathematical Model of Molecular Evolution through a Stochastic Analysis

Joel Valdivia Ortega
Faculty of Science from the National Autonomous University of Mexico, Mexico

ABSTRACT

On Mathematical model of molecular evolution through a stochastic analysis, I present a mathematical treatment for the molecular evolution on DNA chains from some species which allowed me to understand its modifications as random variables connected by a Markov chain. As a result of this, I was able to determine the probability for some given nucleotide on a given position in the studied genes changes into another as a product of random events and to describe a mechanism to give the estimated number of generations needed so one given DNA chain could become another given one.

ARTICLE HISTORY
Received: March 25, 2021
Accepted: April 08, 2021
Published: April 15, 2021

KEYWORDS
Phylogenetics; Molecular evolution; Mathematical model; Stochastic process; Markov chain

Introduction

Several attempts to study molecular evolution as a stochastic process have been made across time, each of them using different types of parameters and different hypotheses. One of the first ones was [1], which assumed a uniform probability for all the events; I consider this to be a good first hypothesis, but it is well known there exists several ways in which a mutation can be made, e.g. a transition or a trans version, each of them with different probabilities of occurrence and therefore, making this supposition needed of a change.

Some works which do consider non-uniform probability distribution for the mutation of nucleotides are [2,3], which used a molecular clock of mitochondrial DNA and comparison of a pair of nucleotide sequences. Respectively the former used a molecular clock of mitochondrial DNA to estimate the number of generations between two species, but even the authors recognized the results were not accurate with respect to the evidence found on fossils. On the other hand, the latter found a simple formula to estimate the "evolutionary distance per site" which is K=-1/2 ln (1-2 P-Q), where P and Q are respectively the fractions of nucleotides showing differences due to transitions and Trans versions. Leaving apart the fact that it is possible for 1-2 P-Q to be negative while the logarithm is only defined for positive values, this formula also has the problem that the weight it gives to the proportion depends on the proportion itself, leading to wrong estimations for the evolutionary distance. To give an example of what I mean by this, let's suppose we have one nucleotide chain α_1=(A,A,A,A,A) which becomes into α_2=(T,A,A,A,A) on the next generation and α_3=(T,T,A,A,A) after three generations; with this formula, the evolutionary distance between α_1 and α_2 would be around 0.11, but the estimation for α_1 and α_3 would be of 0.25, showing that the proportion the estimation changes is not the same given the same amount of mutations occurred, so I will propose another formula which takes this into consideration.

One last work I would like to address since it has a similar approach to what I did on my previous paper is [4], where a Markov chain with a 61 entries per side squared transition matrix was used to study codon based mutations and concluded on a maximum likelihood of α_ and β-goblin genes. The main difference I would like to remark between the paper from Goldman and Yang and what I did is that they obtained the transition matrix as a solution from a differential equation they supposed; meanwhile, I got my transition matrix as a result of data analysis, meaning I made no suppositions on the way it should behave so

Contact: Joel Valdivia Ortega ✉ joelvaltega@ciencias.unam.mx Faculty of Science from the National Autonomous University of Mexico, Mexico

my result would be a reflection of the empirical data.

Stochastic Analysis

Since the mutations do not depend on the previous ones, I can think of the process of evolution as a Markov chain by definition, so I propose to take the nucleotides on a DNA chain as a random variable which has for range the set of all possible nucleotides. Given this, I tried to calculate the transition matrix P^' which connects the wolf's genome to the genome of a boxer, a poodle, a yorkshire and a beagle [5-9]. To do this, I define

$$\$N:=\{N_1 = adenine, N_2 = cytosine, N_3 = guanine, N_4 = thymine \}$$

and D:={boxer,poodle,yorkshire,beagle} so I can build the function in figure 1.

$$f_d(i, j, n) = \begin{cases} 1 \; if \; w(n) = N_i \; and \; d(n) = N_j \\ 0 \; otherwise \end{cases}$$

Where d is inD, w(n) is the nth nucleotide of the wolf's genome and d(n) is the n-th nucleotide of the given dog's genome. Therefore I can write

$$P'_{ij} = \left(\sum_{d \; in \; D} \sum_{n=1}^{10000} f_d(i,j,n) \right) / \left(\sum_{d \; in \; D} \sum_{k=1}^{4} \sum_{n=1}^{10000} f_d(i,j,n) \right)$$

since this is the probability that a given type of nucleotide on the wolf's genome ends up being another given type of nucleotide on the genome of any of the dogs and where I took only 10000 nucleotides so I can be sure the probability distribution for a mutation to happen is the same in the whole chain.

According to the domestication of the wolf began 30000 years ago and since then, the evolution of the dog began. Supposing a new generation of dogs is born each year on average and if P is the transition matrix of the Markov chain I am working with, therefore I have P^'=P^30000, so I just have to calculate P=(P^')^(1/30000) and after taking the real part from the entries, I obtained the matrix, where P is the transition matrix which describes the probability to see a mutation on a certain nucleotide on the next generation [10].

$$P = \begin{bmatrix} 0.999 & 3.69 . 10^{-5} & 2.81 . 10^{-5} & 4.64 . 10^{-5} \\ 5.07 . 10^{-5} & 0.999 & 3.62 . 10^{-5} & 4.01 . 10^{-5} \\ 4.01 . 10^{-5} & 3.26 . 10^{-5} & 0.999 & 4.47 . 10^{-5} \\ 4.24 . 10^{-5} & 3 . 10^{-5} & 3.16 . 10^{-5} & 0.999 \end{bmatrix}$$

Figure 2: Obtained transition matrix for the Markov chain

It is worth noticing that the described calculation of P did not use any assumption of any kind and is only based on the empirical data I used, so this result should be independent of any type of scientific scheme but the mathematical one.

Now that I have the transition matrix, I should be able to begin from the wolf's genome and arrive at each dog's genome by simulating the random variables describing the nucleotides from each chain. To do so, I made an algorithm of artificial selection in which I picked a dog's genome and simulated the wolf had had 50 children from which I chose the two of them which had the closest genome to the selected dog and simulated they had a breed; after that, I used the transition matrix again to recreate 50 children from the breed and repeat the process until the desired genome was reached; this process was repeated 100 times for each dog. In order to know if this process could lead me only to the genomes from the dogs I began the work with or if I could reach any other genome, I generated a random one and applied the artificial selection algorithm choosing the random genome; same as the dogs, this process was repeated 100 times.

Given the 100 simulations for each dog and the random genome, I calculated the average amount of generations taken to reach the target, the standard deviation between the amount of generations taken to conclude the simulation and the standard deviation as a percentage of the average but I could not find a significant difference between them, meaning this algorithm would lead to any genome, not only to those the transition matrix was calculated with. Rather than a failure on the model, I see the origin of this incapability to distinguish the real genomes from the random one on both reasons: the first one is the artificial selection algorithm is very strict and unrealistic because of the very selective way the children are chosen and because it was designed only to be sure I was capable of connecting the genomes from the wolf and the dogs using only the transition matrix; the second reason is that a phenomena described by a Markov chain, such as the process of mutations itself, depends only on some given probabilities and random events and does not have some planned target nor logical decisions.

Estimating Number of Generations between Two Genomes

Let q_n be some nucleotide chain obtained from the wolf's genome after using n times the transition matrix calculated before and

$$U(n) = \sum_{m=1}^{10000} \sum_{i=1}^{4} \sum_{j=1}^{4} f_{q_n}(i,j,m)$$

which represents the number of identical nucleotides between q_nand w. In order to find a differential equation which describesU, I propose the following hypothesis:

1. The number of new differences depends on the quantity of nucleotides on the genome, and then dU/dn is proportional to|U|.
2. The probability of a mutation on a specific entry of the genome does not depend on the number of generations without change, which means dU/dn is proportional to -U [11].
3. It is possible that a nucleotide from q_n mutates in such way that it becomes the same as in w, therefore dU/dn is proportional toU(0)-U.

Therefore I propose the differential equation

dU/dn=-aU+b(U(0)-U)

Which, after taking q_0=w as an initial condition, has for solution

$$U(n) = \frac{U(0)}{a+b}(ae^{-(a+b)n} + b)$$

where a is the probability that if some nucleotide on position m on the chain q_n is the same as in w, this nucleotide changes for the next iteration so that it is no longer true for $q_{(n+1)}$, while b represents the probability that if a nucleotide on a given position m on the chain represented by q_n is not the same as in w, the corresponding nucleotides are the same for $q_{(n+1)}$.

Remembering I am understanding the molecular evolution as a Markov chain and using the values I obtained from its transition matrix, I conclude the expression I have been looking for is

$$U(n) = \frac{3U(0)}{4}(e^{-4an/3} + \frac{1}{3}),$$

Which makes clear the tendency of the nucleotides chain q_n to become independent from w at a probabilistic level, being this a successful simulation of the genetic diversity.

Now we can answer the question about the amount of necessary generations to get from one genome to another. Being U the number of differences between this two given genomes, one can solve the previous expression for n and get

$$n = \frac{1}{a+b} ln(\frac{aU(0)}{(a+b)U - bU(0)}),$$

Which for the empirical data we got here, we have that the number of generations between the wolf and the dogs is 34054±70, where the second number is the marginal error. If one remembers the transition matrix was made with the assumption that the number of generations between those species were 30000, one can say this is a very good result. It is worth mentioning that the way the transition matrix was made is not the source of the accuracy of the estimation I obtained because I proposed the differential expression independently and the only information I used from the matrix is the values of a and b.

Conclusion

Let us notice again that the described process and the presented results does not depend on any kind of hypothesis other than being a Markov chain and following the rules I proposed for the construction of the differential equation, so this can be extended not only to other sequences on the genome of these species or other species itself, but also to non-nuclear DNA or even RNA.

References

[1] Jukes TH, Cantor CR. Evolution of protein molecules. In H. N. Munro, editor, Mammalian Protein Metabolism. Academic Press. New York. 1969; 21–123.

[2] Hasegawa M, Yano T, Kishino H. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. Proceedings of the Japan Acade. 1984; 60(4):95–98.

[3] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16(2):111–120.

[4] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994; 11(5):725–736.

[5] ASM325472v1. Genome. Assembly. NCBI. Ncbi.nlm.nih.gov.2019.

[6] CanFam3.1. canFam3. Genome. Assembly. NCBI. Ncbi.nlm.nih.gov.2019.

[7] ASM18141v1. Genome. Assembly. NCBI. Ncbi.nlm.nih.gov.2019.

[8] ASM208743v1. Genome. Assembly. NCBI. Ncbi.nlm.nih.gov.2019.

[9] Beagle. Genome. Assembly. NCBI. Ncbi.nlm.nih.gov.2019.

[10] Handwerk B. How Accurate Is Alpha's Theory of Dog Domestication?. Smithsonian. 2019.

[11] Krane K. Introductory nuclear physics. Wiley New York N.Y. 1988.