Estimating Mutation Rate and Evolutionary Distance By Two Markov Chains Models

626 final project —— Xu Yuanzhuo

ABSTRACT

In this study, we developed a Markov chain-based mathematical framework to analyze DNA mutation processes. The investigation commenced with the formulation of the Jukes-Cantor models and Kimura models to characterize the evolutionary changes in nucleotide sequences. We applied these models to a specific segment of the unc-22 gene extracted from two distinct Caenorhabditis elegans specimens, providing an empirical basis for assessing the predictive accuracy of the models. Comparative simulations of both models were conducted, with a focus on their convergence towards a stationary distribution. The results demonstrate the models' effectiveness in capturing the dynamics of genetic mutation, offering insights into the underlying mechanisms of DNA evolution. These conclusions underscore the models' varying sensitivities to the underlying mutation rates and their types. Such insights are crucial for the appropriate application of these models in genetic research, particularly in the realms of phylogenetics and evolutionary biology, where accurate depiction of genetic distances is paramount. Our study advocates for a nuanced application of these models, taking into consideration the specific genetic context of the sequences under investigation to ensure the precision of evolutionary interpretations.

Keywords: Jukes-Cantor Model, Kimura Model, Evolutionary Distance

INTRODUCTION

DNA, the carrier of genetic information, consists of four nucleotides—adenine (A), guanine (G), cytosine (C), and thymine (T)—which sequence together to form genes, each coding for a specific protein. Although the replication process of DNA is highly accurate, mutations resulting from base substitutions occasionally occur, altering the genetic code. While many mutations are neutral, occurring in non-coding regions, others can be significant, causing diseases like cystic fibrosis or, rarely, beneficial adaptations that may be favored by natural selection. This paper examines the mutation rate's role in evolutionary distance and disease, providing a quantitative analysis of its implications for evolutionary biology. Continuous-time Markov chains are used to study the evolution of DNA sequences. Numerous models have been proposed for the evolutionary changes on the genome as a result of mutation.

MATHEMATICAL DERIVATION

Markov chains are used as mathematical models for DNA mutations because they capture the stochastic nature of genetic changes over time, where the probability of a future state (or mutation) depends only on the current state, not the sequence of events that preceded it.Such models are often specified in terms of transition rates between base nucleotides A, G, C, and T at a fixed chromosome location.

Model 1: Jukes-Cantor model

The Jukes-Cantor model provides a foundational framework for understanding molecular evolution. It assumes that nucleotide substitutions are equally probable and occur independently across nucleotide sites at a constant rate. Under this model, we can derive the probability of observing a particular nucleotide substitution over evolutionary time. The rate matrix Q under the Jukes-Cantor model is given by:

$$Q = \begin{pmatrix} -3r & r & r & r \\ r & -3r & r & r \\ r & r & -3r & r \\ r & r & r & -3r \end{pmatrix},$$

where r is the mutation rate per nucleotide site. The generator matrix is diagonalizable with linearly independent eigenvectors:

$$S = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}, \quad S^{-1} = \frac{1}{4}\begin{pmatrix} 1 & -1 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

corresponding to eigenvalues $-4r, -4r, -4r$, and 0. This yields the transition probability matrix:

$$P(t) = e^{tQ} = Se^{tD}S^{-1} = \frac{1}{4}\begin{pmatrix} 1+3e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} \\ 1-e^{-4rt} & 1+3e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} \\ 1-e^{-4rt} & 1-e^{-4rt} & 1+3e^{-4rt} & 1-e^{-4rt} \\ 1-e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} & 1+3e^{-4rt} \end{pmatrix},$$

where e^{tD} is the diagonal matrix of eigenvalues. We start with the equation that gives the fraction of sites that are different between two sequences:

$$f = 1 - \left(\frac{1}{4} + \frac{3}{4}(1 - \frac{4r}{3})^k\right)$$

We solve this equation for k to get:

$$k = \frac{\ln(1 - \frac{4f}{3})}{\ln(1 - \frac{4r}{3})}$$

Using the approximation given in the statement above, we can replace the denominator and solve for kr to get the Jukes-Cantor distance d:

$$d = kr \approx -\frac{3}{4}\ln(1 - \frac{4f}{3})$$

Model 2: Kimura model

Nucleotides A and G are classified as purines, while C and T are pyrimidines. Substitutions within purines or pyrimidines are termed transitions, whereas substitutions between purines and pyrimidines are known as transversions. The Kimura model, incorporating two parameters, r and s, effectively differentiates between these transitions and transversions.Moreover, because the difference in chemical structure between purine and pyrimidine is greater than that between different purines or different pyrimidines, in nature, the probability of a purine mutating into another purine or a pyrimidine mutating into another pyrimidine is significantly higher than the conversion between purines and pyrimidines. Therefore, the Kimura model is generally considered to be more realistic. article amsmath The rate matrix Q under the Kimura model is given by:

$$Q = \begin{pmatrix} a & -a+2\beta & \alpha & \beta & \\ \alpha & g & -\alpha+2\beta & \beta & \\ \beta & \beta & c & -\alpha+2\beta & \\ \beta & \beta & \alpha & t & -\alpha+2\beta \end{pmatrix},$$

and the diagonal matrix D and the matrix S are given by:

$$D = \begin{pmatrix} -2(\alpha+\beta) & 0 & 0 & 0 \\ 0 & -2(\alpha+\beta) & 0 & 0 \\ 0 & 0 & -4\beta & 0 \\ 0 & 0 & 0 & -2(\alpha+\beta) \end{pmatrix}, \quad S = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The transition function $P(t) = e^{tQ} = Se^{tD}S^{-1}$ is

$$P_{xy}(t) = \begin{cases} (1+e^{-4\beta t} - 2e^{-2(\alpha+\beta)t})/4, & \text{if } xy \in \{ag, ga, ct, tc\} \\ (1-e^{-4\beta t})/4, & \text{if } xy \in \{ac, at, gc, gt, ca, cg, ta, tg\} \\ (1+e^{-4\beta t} + 2e^{-2(\alpha+\beta)t})/4, & \text{if } xy \in \{aa, gg, cc, tt\} \end{cases}.$$

Given that p and q are the observed proportions of transitions and transversions, the expected proportions of unchanged sites (those without a transition or a transversion) and sites with transversions can be represented with the following equations:

$$P = e^{-4\alpha t/3},$$

$$Q = e^{-2(\alpha+\beta)t/3},$$

The estimated genetic distance K is then derived from the proportions of observed differences (p and q) using these equations. The formulae to estimate αt and βt from p and q are:

$$\alpha t = -\frac{3}{4}\ln\left(1-\frac{4p}{3}\right),$$

$$\beta t = -\frac{3}{4}\ln\left(1-\frac{2q}{3}-\frac{p}{3}\right).$$

Combining these, the overall genetic distance $\hat{K}$ is calculated by:

$$\hat{K} = \alpha t + \beta t$$

$$\hat{K} = -\frac{3}{4}\ln\left(1-\frac{4p}{3}\right) - \frac{3}{4}\ln\left(1-\frac{2q}{3}-\frac{p}{3}\right).$$

APPLICATION OF TWO MODELS

In this study, we selected two distinct 1080-nucleotide sequences from the unc-22 gene of Caenorhabditis elegans to explore genetic variations. The choice of C. elegans as a model organism is underpinned by its fully sequenced and well-characterized genome, short generational span, and amenability to genetic experimentation. The unc-22 gene provides a robust framework for examining the correlation between genetic mutations and their phenotypic manifestations. Considering the two chosen DNA sequences, we initiate by aligning the fragments and conducting a comparative analysis to identify the variant base pairs. Following this, we compute the ’f’ value, which serves as a quantitative measure of these differences. Subsequently, we ascertain the probabilities associated with transitions and transversions among the distinct base pairs. Concluding the process, we apply the previously derived distance formula to quantify the Evolutionary Distance, thereby elucidating the genetic divergence as depicted by the two models under consideration.

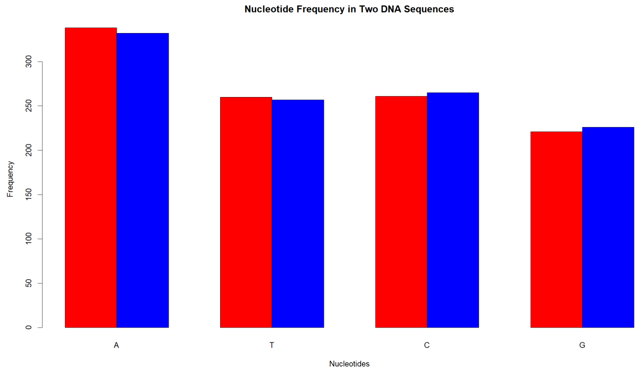

Figure 1. Blue—DNA sequence 1
Red—DNA sequence 2

Total	Difference	Transitions	Transverstions
1080	259	172	87
Probability	f	p	q
(six decimal)	0.239815	0.159259	0.080556

Table 1. Result for analysis of two DNA sequences

$$Jukes-Cantor\,Model: \hat{d}1 \approx -\frac{3}{4}\ln(1-\frac{4f}{3}) = 0.288975$$

$$Kimura\,Model : \hat{d}2 \approx -\frac{3}{4}\ln\left(1-\frac{4p}{3}\right)-\frac{3}{4}\ln\left(1-\frac{2q}{3}-\frac{p}{3}\right)=0.263722$$

Observations indicate that applying both the Jukes-Cantor and Kimura models to estimate the evolutionary distance between two distinct sequences from C. elegans yields relatively consistent results, with no marked disparity. The question then arises: under which specific conditions would the distances computed by these two models exhibit a significant divergence? Identifying and understanding these conditions is the pivotal objective of our forthcoming research efforts.

SIMULATION

Part1:analysis of difference between two models

Given the challenges in acquiring comparable DNA fragments from the same or similar species, this study employs a simulation approach to examine the conditions that lead to substantial disparities in the evolutionary distances computed by the Jukes-Cantor and Kimura models. Initially, we establish the divergence by considering the difference between the distance formulas of the two models. It is noteworthy that within the Jukes-Cantor model, the variable f corresponds to the sum of p and q in the Kimura model. In nature, transitions occur more frequently than transversions, with a typical ratio of 2:1. Based on this, our study maintains a constant p:q ratio of 2:1 to assess how variations in the f value influence the disparity between the Jukes-Cantor and Kimura models.First we calculate the difference between two models:

$$-\frac{3}{4}\ln(1-\frac{4f}{3}) = -\frac{3}{4}\ln\left(1-\frac{4p}{3}\right)-\frac{3}{4}\ln\left(1-\frac{2q}{3}-\frac{p}{3}\right)$$

Known that $f = p + q$ and set p:q=2:1:

$$1-\frac{4f}{3} < \left(1-\frac{4p}{3}\right)\left(1-\frac{2q}{3}-\frac{p}{3}\right) exist\,for any\,p : q = 2 : 1$$

We can observe that the distance calculated by the Jukes-Cantor Model are always little larger than Kimura Model. Moreover, the gap would become larger as the value of f increases.

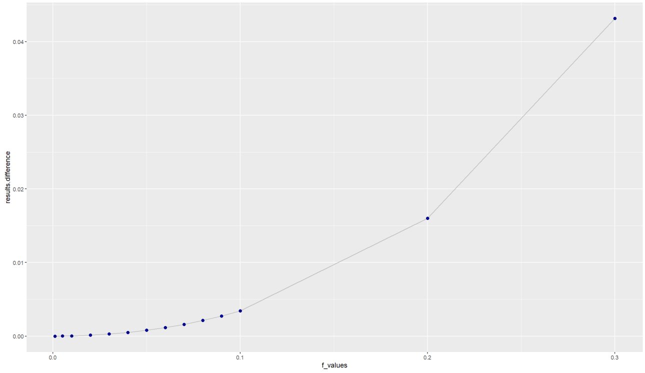

Figure 2. Influence of f Value on the Discrepancy between Evolutionary Distance Models

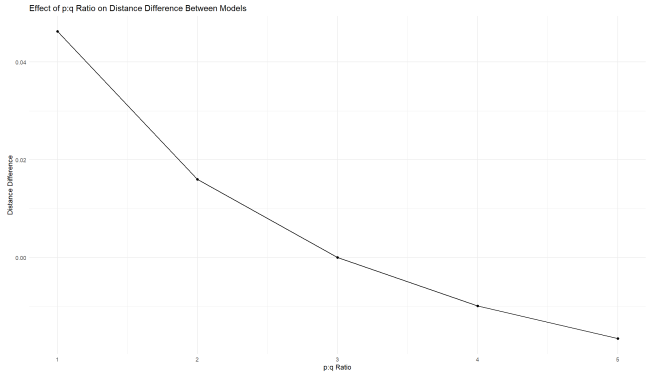

Figure 3. Effect of p:q Ratio on Distance Difference Between Models

Then,after analyzing the influence of f Value on the discrepancy between two models, we would use different p:q ratio to see the distance difference between models.Based on the results of our simulations, we observed that at a p:q ratio of 3:1—indicating a threefold higher probability of transitions compared to transversions—the distances calculated by both the Jukes-Cantor and Kimura models converge to same. Furthermore, when the p:q ratio exceeds 3:1, the distance estimated by the Kimura model is consistently greater than that estimated by the Jukes-Cantor model, with the disparity in distances increasing as the p:q ratio grows larger. Conversely, when the p:q ratio is less than 3:1, the Kimura model yields distances that are smaller than those produced by the Jukes-Cantor model, and this difference becomes more pronounced as the p:q ratio decreases. This pattern underscores the sensitivity of the Kimura model to the balance between transitions and transversions, highlighting its potential for more nuanced insights into evolutionary processes under specific genetic conditions.

Part2:analysis of stationary distribution of two models

Next, we will use base pairs as units to explore the steady-state distribution of the two models through simple matrix operations and simulations. By using matrix and formulas, we consider the continuous-time Markov chain represented by a rate matrix Q. The matrix describes the rates at which one nucleotide changes to another. Given the symmetry of the Jukes-Cantor model, the stationary distribution can be derived by examining the balance of the flow into and out of each state. The rate matrix Q for the Jukes-Cantor model is given by:

$$Q = \begin{bmatrix} -\mu & \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & -\mu & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & -\mu & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} & -\mu \end{bmatrix}$$

The stationary distribution π satisfies $\pi Q = 0$ with π being:

$$\pi = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$$

The rate matrix Q for the Kimura two-parameter model is given by:

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \beta & \alpha & \beta \\ \beta & -(\alpha+2\beta) & \beta & \alpha \\ \alpha & \beta & -(\alpha+2\beta) & \beta \\ \beta & \alpha & \beta & -(\alpha+2\beta) \end{bmatrix}$$

Assuming uniform base composition, the stationary distribution π is:

$$\pi = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$$

The final stationary distribution of both models approaches uniformity. We conducted simulations using the aforementioned C. elegans gene in R, observing that the four nucleotides converged towards a value of 0.25 at varying rates. Due to space constraints in this paper, we omit the detailed simulation process. For more information, please refer to the corresponding chart in the presentation. Notably, our simulations revealed that deviations from the 0.25 mark tend to accelerate the convergence rate.

CONCLUSION

In the comparative analysis of evolutionary distance models, our investigation has yielded several key findings. Both the Jukes-Cantor (JC) and Kimura (K) models exhibit a tendency to approach a steady-state distribution, affirming the theoretical underpinnings of these models in reflecting genetic stability over time.

Moreover, my results have shown that distances estimated by the Jukes-Cantor model are marginally larger than those by the Kimura model. This trend becomes more evident as the value of f—the fraction of differing nucleotides—rises, suggesting a propensity for the Jukes-Cantor model to amplify perceived genetic divergence at higher mutation rates.

We observed distinct behaviors in the models' performance based on the transition to transversion ratio (p:q). Specifically, when the p:q ratio surpasses 3:1, the Kimura model consistently predicts greater distances than the Jukes-Cantor model. This discrepancy intensifies as the p:q ratio increases, indicating a sensitivity of the Kimura model to higher transition probabilities. In contrast, with a p:q ratio below 3:1, the Kimura model generates smaller distances compared to the Jukes-Cantor model, and this differential is accentuated with decreasing p:q ratios.

These findings underscore the models' varying sensitivities to the underlying mutation rates and their types. Such insights are beneficial for the appropriate application of these models in genetic research, particularly in the realms of phylogenetics and evolutionary biology, where accurate depiction of genetic distances is paramount. My project advocates for a nuanced application of two models, taking into consideration the specific genetic context of the sequences under investigation to ensure the precision of evolutionary interpretations.

REFERENCE

genome.sph.umich.edu/wiki/SNP_Call_Set_Properties
Erickson, K. (2010). The Jukes-Cantor Model of Molecular Evolution. PRIMUS, 20(5), 438–445. https://doi.org/10.1080/10511970903487705
Holmquist, R. Transitions and transversions in evolutionary descent: An approach to understanding. J Mol Evol 19, 134–144 (1983). https://doi.org/10.1007/BF02300751
Nishimaki, T., Sato, K. An Extension of the Kimura Two-Parameter Model to the Natural Evolutionary Process. J Mol Evol 87, 60–67 (2019). https://doi.org/10.1007/s00239-018-9885-1
ROBERT P. DOBROW. (2016). Introduction to Stochastic Processes With R .