Model of Molecular Evolution and Its Applications in Real Gene Sequence

Victoria Nian, Eric Zhang

April 2024

1 Introduction

The study of molecular evolution models is essential for understanding the fundamental processes that drive genetic diversity and evolution. These models provide crucial insights into the patterns and rates of mutations, which have significant implications for fields ranging from evolutionary biology to medical genetics. By simulating the processes of molecular evolution, researchers can predict evolutionary outcomes, aiding in the conservation of species and informing medical research on genetic diseases.

In this project, we employed the Kimura model of molecular evolution, which distinguishes between transitions and transversions in nucleotide substitutions, offering a realistic depiction of molecular evolution. We developed a computational simulation to track the time evolution of DNA, using this model to observe random substitutions at the base or amino acid level. This simulation not only generated a detailed dataset of DNA sequences but also allowed us to rigorously test the accuracy and validity of the Kimura model by estimating its parameters and comparing them with the known true values used in our simulations.

Furthermore, to demonstrate the practical utility of our model, we applied it to analyze real-world DNA sequences from scientific literature. This step was critical in showcasing how the theoretical model can be applied to real-world biological data, providing insights into evolutionary trends and validating the model's effectiveness in practical scenarios. This project highlights the importance of integrating theoretical models with empirical data, bridging the gap between simulation and application in the study of molecular evolution.

2 Kimura Model

The Kimura model is a fundamental concept in the study of molecular evolution, providing a basis for understanding nucleotide substitutions over time. This model differentiates between two types of base pair substitutions: transitions and transversions. Transitions refer to interchanges of two-ring purines $A \leftrightarrow G$

or of one-ring pyrimidines $C \leftrightarrow T$, which are generally more common due to the chemical similarity and hence are given a different rate, often denoted by α. Transitions only involve the modifications on side groups but not the ring structure. Transversions, on the other hand, are substitutions between a purine and a pyrimidine $A \leftrightarrow T$, $A \leftrightarrow C$, $G \leftrightarrow T$, $G \leftrightarrow C$, and are less likely to occur because this mutation involves modifications in the ring structure, designated by the rate β. The Kimura model accounts for these differences by allowing separate rates for transitions and transversions, aiming to reflect the true nature of molecular changes over time.

The substitution matrix P for the Kimura model is given by:

$$P = \begin{bmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{bmatrix}$$

3 Simulation

3.1 Simulation Testing Logic

The basic logic of our simulation is based on Markov Chain. The substitution matrix has four rows, from the top to bottom, referring to A, G, C, T. The columns have the same order. Each value in the substitution matrix refers to the mutation probability from the row nucleotide to column nucleotide. In practicality, generate a random number between 0 and 1. Using the substitution matrix, determine which nucleotide does the current one changes to. Repeat this process for every nucleotide, and every generation.

Before applying our simulation on real-world DNA sequences, we tested our simulation using a simple set of parameters in order to better monitor its performance. We set α as 0.02, β as 0.01, given the fact that transition occurs more frequently than transversions. The starting sequence is $ATCG$. The number of generation is 500, and the number of simulation is 100. We assume each generation to be 20 years, and in this 20 years, each nucleotide will undergo substitution determined by the substitution matrix only once. Therefore, with 500 generations, our entire simulation time span is 10000 years. Then, using these testing parameters, we went on to prove that our simulation is both accurate and valid.

3.2 Proof of Accuracy

To prove that our simulation is accurate, we were trying to reproduce the substitution matrix solely based on the frequency of substitutions in this 500 generations and take average from the 100 simulations. The following two tables are the original substitution matrix and the reproduced substitution matrix.

	A	**G**	**C**	**T**
A	0.96	0.02	0.01	0.01
G	0.02	0.96	0.01	0.01
C	0.01	0.01	0.96	0.02
T	0.01	0.01	0.02	0.96

Table 1: The original substitution matrix set by α equals 0.02, and β equals 0.01

	A	**G**	**C**	**T**
A	0.9596	0.0209	0.0103	0.0089
G	0.0212	0.9579	0.0099	0.0108
C	0.0099	0.0107	0.9593	0.0203
T	0.0093	0.0106	0.0206	0.9599

Table 2: The reproduced substitution matrix obtain from 100 simulation

The reproduced substitution matrix is very accurate. The differences from the original one are all around 0.001. If we carry out the simulation for more times, more generations, and with longer DNA sequence, this reproduced matrix will be more accurate.

3.3 Proof of Validity

To prove that our simulation is valid, we applied an established equation called the Compact Expression for Distance, which can estimate α and β based on substitution frequencies. The equations are:

$$\alpha t = -\frac{1}{2}\, log(1-2P-Q) + \frac{1}{4}\, log(1-2Q)$$

$$\beta t = -\frac{1}{4}\, log(1-2Q)$$

In these two equations, P denotes the fraction of transitions, and Q denotes the fraction of transversions. These two values can be obtained using the substitution frequencies generated after 100 simulations. αt and βt in this equation is the same with α and β in our case because α and β are the raw rate of substitutions, αt and βt are the expected number of substitution occurred per site over a time interval t. In our case, t is a constant number, so αt and α are interchangeable in our simulation. The following two figures are the histogram of the estimated αt and βt.

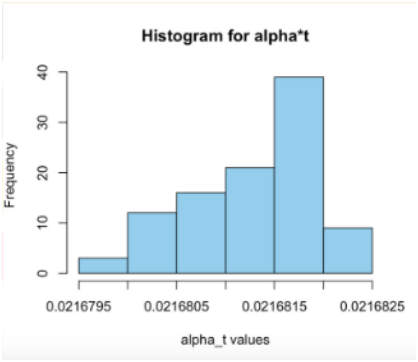

Figure 1: Histogram of estimated αt. Mean is 0.0211.

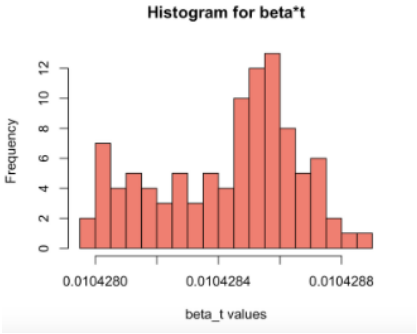

Figure 2: Histogram of estimated βt. Mean is 0.0103.

As presented in the two histograms, α has a mean of 0.0211, and β has a mean of 0.0103, which are very close to the real value. This reflects that the two equations can be applied to our simulation, and our simulation is valid. But one interesting thing to note is that both of the histograms seem to skew to the left, and the means are slightly higher than the real value. This may indicate that our simulation tend to slightly overestimate α and β, but this overestimation is within the acceptance range.

4 Application in Real Gene Sequence

Transitioning from theory to practice, we apply our rigorously tested simulation model to actual genetic sequences, specifically targeting the insulin gene. Insulin stands as a beacon in molecular evolution studies, its gene conserved across many species, yet subtly varied—offering a window into evolutionary mechanisms. By applying our model to the insulin gene, we expect to illuminate the nuances of molecular evolution, showcasing our model's potential to trace the intricate pathways of genetic change and adaptation that have shaped this critical protein throughout evolutionary history.

4.1 Choice of Parameter

4.1.1 Constant α and β

In the pursuit of discerning stochastic variations within genetic sequences, our study extends into the domain of noise detection. With fixed evolutionary rates—specifically, an alpha (α) of 0.02 and a beta (β) of 0.01—we employed the Kimura model to simulate the mutational process of the insulin gene across 100 separate simulations. The table below delineates the mutation frequencies from 100 simulations of the insulin gene using the Kimura model.

	A	**G**	**C**	**T**
A	46493	996	471	499
G	1000	49028	520	509
C	485	499	48486	996
T	509	521	981	47007

Table 3: Mutation frequencies of the insulin gene after 100 simulations with constant α and β

It is evident from the data that transitions (A $\leftrightarrow$ G and C $\leftrightarrow$ T) are markedly more frequent, with counts of 996 for A $\leftrightarrow$ G and 996 for C $\leftrightarrow$ T, as opposed to the transversions, which exhibit lower frequencies, ranging from 471 to 521 for A $\leftrightarrow$ C, A $\leftrightarrow$ T, G $\leftrightarrow$ C, and G $\leftrightarrow$ T. This empirical evidence supports the Kimura model's assertion that transitions are more probable than transversions, reflecting the chemical and structural propensities of nucleotide substitutions in DNA replication and repair mechanisms.

4.1.2 Random α and β

Having established the mutation frequencies under constant rates of evolution, our investigation progresses to assess the robustness of our model under varied evolutionary pressures. To this end, we introduce randomness into the values of α (alpha) and β (beta), simulating a spectrum of evolutionary scenarios. This approach allows us to explore the variability inherent in the evolutionary process and to gauge the sensitivity of the insulin gene to changes in substitution rates. During this stage of the study, random α and β values are generated by

$$
\begin{aligned}
z_{\text{score}} &\sim \mathcal{N}(0,1) \\
\alpha &= \sigma_\alpha \cdot z_{\text{score}} + \mu_\alpha \quad \text{rate of transition} \\
\beta &= \sigma_\beta \cdot z_{\text{score}} + \mu_\beta \quad \text{rate of transversion}
\end{aligned}
$$

first obtaining Z-scores from a standard normal distribution $N(0,1)$. These Z-scores are then scaled using the means and standard deviations calculated from the initial simulations that used constant α and β. This process ensures that each generation within the simulation is characterized by a distinct set of evolutionary parameters, introducing variability and allowing us to observe the resulting mutation patterns in the insulin gene. The table below delineates the mutation frequencies from 100 simulations of the insulin gene using the Kimura model with random α and β.

	A	G	C	T
A	48236	1070	461	478
G	1042	51288	542	507
C	514	487	45207	1011
T	417	426	983	46177

Table 4: Mutation frequencies of the insulin gene after 100 simulations with random α and β

Based on the results in Table 4, we also observe that the frequencies of transitions (A $\leftrightarrow$ G and C $\leftrightarrow$ T) are markedly more frequent.

4.1.3 Noise Detection

Transitioning into the comparative analysis, we delve into the significance of randomness and variability. It is through these lenses that we can understand the broader implications of stochastic effects on evolutionary processes, revealing how intrinsic noise contributes to genetic diversity. We plotted the distribution of constant and random α and β on two figures.

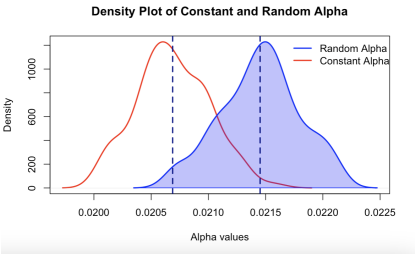

Figure 3: Comparison of constant and random alpha distributions.

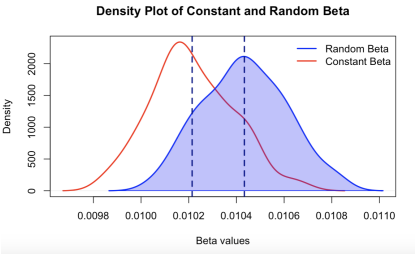

Figure 4: Comparison of constant and random beta distributions.

The density plot in Figure 3 presents a visual comparison between the distributions of constant and random alpha values. The curve for the constant alpha is depicted as a narrow peak, indicating little variation around a central value. Conversely, the random alpha distribution is broader, displaying a significant spread, which suggests a higher degree of variability.

The density plot for beta values in Figure 4, much like the one for alpha, contrasts the fixed parameter approach with stochastic variation. The constant beta curve is sharp and centered, signifying a focused range of values, while the random beta exhibits a wider curve, indicating diversity in the evolutionary process.

In summary, the comparison of constant and random distributions for alpha and beta elucidates the inherent variability in evolutionary dynamics. The

broader curves for random parameters suggest a substantial degree of stochastic influence, supporting our hypothesis that variability plays a critical role in molecular evolution.

To determine the presence and quantify the extent of this variability, we next turn to standard errors derived from our simulation data. The standard error serves as a statistical measure of precision for estimated parameters, providing insight into the reliability of observed variability within our simulations.

	Random α	**Constant α**	**Random β**	**Constant β**
Mean	0.0214517	0.0206877	0.0104334	0.0102143
SD	0.0003412	0.0003151	0.0001751	0.0001723

Table 5: Summary statistics for constant and random α and β

The standard deviations (SD) for both random α and β are larger than those for their constant counterparts in Table 5, which suggests the presence of randomness in the parameter values. This observation corroborates the hypothesis that variability is a substantial component in the simulation model, reflecting the inherent stochastic nature of molecular evolution.

4.1.4 Different Noise Level

In the previous section, we obtained the Z-score from a normal distribution with mean equals 0 and standard deviation equals 1. In real-world, this Z-score represents the level of environmental influence on the substitution probabilities. In this simulation, we set the Z-score as 50, which represents a situation in which there is higher environmental stress that increases the mutation rate. The following table is the mean and standard deviation of this 50 Z-score α and β.

	50 Z-score α	**Constant α**	**50 Z-score β**	**Constant β**
Mean	0.0221752	0.0211592	0.0107519	0.0102986
SD	0.0005054	0.0003151	0.0003066	0.0001723

Table 6: Summary statistics for constant and 50 Z-score α and β

With this higher noise level, the mean of both α and β increases, indicating a more frequent substitution. The standard deviation also increases more drastically than the previous random α and β. This drastic change from the constant α and β reflects the influence of environmental stress on molecular evolution. In real-world, such noise will become more complex, for example, only influencing the probability of transition but not transversion or vice versa, and fluctuating level of influence. Our simulation here only assumes a simple case of high positive noise level, but the effect presented is already very obvious and drastic.

5 Conclusion and Discussion

We established an R-based simulation of molecular evolution using the Kimura model. We also tested that the simulation is accurate and valid. We then applied this simulation on insulin sequence to observe how randomness and noise affect molecular evolution.

One advantage of our simulation is that we allow the customization of many settings and parameters. Users can set the starting sequence, substitution matrix, number of generation etc. More importantly, users can apply their own molecular evolution model other than the Kimura model, and even carry out a comprehensive comparison among different molecular evolution models.

Another advantage is that our simulation is proven to be accurate and valid based on the reproduction of the substitution matrix, and the estimation of α and β using an established equation.

But there are also disadvantages. Our simulation uses nested for loop to carry out the substitution process, which is not very efficient. Our code is not optimized as the running time is long. Given the fact that real-world DNA sequences can be even longer than the insulin, which is around 400 nucleotide, our code needs to be optimized to handle such amount of data.

In the future, more simulation can be done to observe the effect of high positive or negative noise level to better represent real-world situations.