Int. J. Math. And Appl., 9(1)(2021), 9–18
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

# Mathematical Model for Predicting Time of a Shared Common Ancestor

**Vedanti Bhargava**[1,*]

1 Lexington High School, Lexington, Massachusetts, United States.

**Abstract:** Accurate and efficient model development in the field of evolutionary biology is becoming increasingly important to study phylogenetic relationships. Studying these phylogenetic relationships can provide greater insight into areas such as drug development, analyzing disease transmission, and studying the host-pathogen evolutionary relationship. This study uses the idea of a transition matrix to predict how long ago two species shared a common ancestor. Using this concept and various linear algebraic ideas, a mathematical model that simulated amino acid mutations was developed, and it was implemented in Python to perform trials. The main result of the development of the model was a matrix containing the expected values for the number of years ago two amino acids were the same. Simulations of the model can be run in Python multiple times and the average of the values obtained gives an estimate of how many years ago two amino acids were the same. These represent the number of years ago the species shared a common ancestor. The result varies based on the species data analyzed. This model provides a foundation for development of more complex models that enhance phylogenetic tree development. As stated previously, this can have substantial impacts on biotechnology and biomedical sciences, which often rely on analyzing these evolutionary relationships.

**Keywords:** Transition matrix, Amino Acids, Probability, Mutations, Sequences, Simulations, Expected value.


## 1. Introduction

The development of models to further study evolutionary biology is an idea that has been gaining increasing traction in recent years. Accurate models can help to simplify understanding of various evolutionary processes, and it is becoming increasingly clear how dependent phylogenetic analysis is on strong model development [1]. In this study, we propose the development of a such a mathematical model that helps to predict the time of species divergence. The results from the study provide a better understanding of species relatedness and as a result, make the development of phylogenetic trees much easier.

Over the years there have been many approaches to attempt to predict how long ago two species shared a common ancestor. The very first idea that was used was the molecular clock, which uses the mutation rate of various molecules to when two species diverged. The molecular clock hypothesis, states that protein sequence differences accumulate at a constant rate over time [2]. This technique was soon abandoned due to it not accounting for the complexity of the evolutionary process, since evolutionary rates are most likely not constant over time, and also its time consuming nature [3]. However, the idea of using mutation rates of bio-molecules to predict how long ago two species shared a common ancestor is used in many studies, including this study.

In a later study, PAM and log odds matrices are used to predict species divergence times [4]. In a PAM matrix, an element of

* *E-mail: vedantib05@gmail.com*

the matrix $M_{ij}$ gives the probability that amino acid in column j will be replaced by the amino acid in row i, through a series of one or more point accepted mutations during a specified evolutionary interval. This type of matrix can be transformed into a log odds matrix. The log odds matrix is then used as a scoring matrix (i.e. it can distinguish between significant relationships and insignificant relationships between species) to detect distant relationships between proteins. They are both applications of substitution matrices, matrices of scores representing the mutability of amino acids, ultimately used to measure similarity between sequence alignments [5].

Although the past study uses effective ways to create a mathematical model, which not only helps to predict how long ago two species shared a common ancestor, but also takes into account the evolution process itself (they try to distinguish significant mutations rather than coincidences by using log odds matrices, and also takes into consideration the mutability of the amino acid), there is an aspect where it falls short. Due to not having a robust enough mathematical foundation, it is harder to implement it in a computational setting.

Therefore, instead of using specifically a log odds matrix, the proposed model uses the more general concept of a transition matrix (stochastic matrix), making the model easier to implement computationally. Transition matrices are used to describe the transitions of a Markov chain (a model describing various possible outcomes that can be predicted based on the current state, with the outcomes being fixed). with each of the columns as probability vectors. They will be described in more detail later on in the paper. The reason for this is that it provides a much simpler and more direct approach to predicting the time of species divergence.

The impact of this study, and additional evolutionary biology studies, in the real world is substantial. They offer an effective way to apply genetic information to biomedical sciences and biotechnology. This is because they give us information about the structure of various molecules (nucleic acids, proteins, etc). As a result, processes such as drug design, which largely involve the analysis of these molecules, can be enhanced from the results of such studies.

# 2. Methods and Results

## 2.1. Modeling amino acids and proteins

We now describe how we model the mutation of amino acids. We assume a simplified world with only three amino acids that we call $A_1$, $A_2$ and $A_3$. In this world, a protein is a finite sequence of these three amino acids. Both the order and number of the amino acids in a protein matter. For example, $S_1$ and $S_2$, where $S_1 = A_2A_3A_3A_2A_1$ and $S_2 = A_1A_3A_3A_2A_2$, are two different proteins.

The first question we seek to answer is: Given the sequence of amino acids of the protein that has a certain function in a species, what will be the sequence the protein with the same function in the descendent species several (hundreds of thousands) generations later?

For example, assume we start with the protein $S = A_2A_3A_3A_2A_1$. After a million years, the protein with the same function in a descendant species is $S = A_2A_3A_1A_2A_1$ (only the middle amino acid changed from $A_3$ to $A_1$). After an other million years the protein with the same function is $S = A_2A_3A_1A_2A_1$ (the protein did not change). After an other million years the protein with the same function is $S = A_1A_3A_1A_2A_3$ (two amino acids changed), and so on. Even though these are proteins in different organisms separated by several generations, in fact, these organisms belong to different species, we think of $S$ as being the same protein that has changed over time because it has the same function over time.

Note that the protein $S$ is a function of time, denoting time by $t$, we have $S = S(t)$. For example, taking the unit of time to be one million years, we have that, in the example of the above paragraph, $S(0) = A_2A_3A_3A_2A_1$, $S(1) = A_2A_3A_1A_2A_1$, $S(2) = A_2A_3A_1A_2A_1$ and $S(3) = A_1A_3A_1A_2A_3$. In practice, the unit of time used is called evolutionary unit, not one

million years. While we will use the evolutionary unit of time, we will not go into any detail about the meaning of this unit of time. We only have to keep in mind that one evolutionary unit of time is a very long period of time.

## 2.2. Modeling Amino Acid Mutations over time

Our model assumes that the mutations of each amino acid in the chain is independent of the other amino acids in the chain. We denote by $p_{ij}$, the probability that an amino acid $A_j$ mutated and is an amino acid $A_i$ after one unit of time. Note that $p_{jj}$ is the probability that amino acid $A_j$ is still an amino acid $A_j$ after one unit of time. In our world of three amino acids, this leads to a $3 \times 3$-matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}. \tag{1}$$

The above matrix $P$ is called the transition matrix. Note that the $\mathrm{j}^{\mathrm{th}}$ column of $P$ gives the probability of mutations of an $A_j$ amino acids in a unit of time. As an example, consider

$$P = \begin{bmatrix} 0.8 & 0.1 & 0 \\ 0.2 & 0.7 & 0.3 \\ 0 & 0.2 & 0.7 \end{bmatrix} \tag{2}$$

In this case, $p_{32} = 0.2$ and thus, the probability that an $A_2$ amino acid mutates and becomes an $A_3$ amino acid after one unit of time is 0.2.

## 2.3. Number of amino acids of different types in the model

There is a rule that states $N_E = N * P(E)$, where $N_E$ is the number of times E happens in N experiments, and P(E) is the probability of E occurring. A consequence of this rule is the following:

**Observation 2.1.** *Assume $N$ is a large integer. Given $N$ $A_j$ amino acids, after one unit of time, about $Np_{ij}$ of those amino acids are $A_i$ amino acids.*

We set $t = 0$ to be a time a very large number of units of times ago. We assume that, at times later than $t = 0$, i.e. $t \geq 0$, amino acids mutate, but they are not created nor they cease to exist in any other way. This simple means that the total number of amino acids of all the types together remains the same for all times $t >\geq 0$. We introduce the three following sequences

$$\begin{aligned} x_1^{(n)} &= \text{ number of } A_1 \text{ amino acids in our whole imaginary world at time } t = n \\ x_2^{(n)} &= \text{ number of } A_2 \text{ amino acids in our whole imaginary world at time } t = n \\ x_3^{(n)} &= \text{ number of } A_3 \text{ amino acids in our whole imaginary world at time } t = n. \end{aligned} \tag{3}$$

We are interested in times $t \gg 1$. Given the discussion of this paragraph, we have the following observation:

**Observation 2.2.** *Let $N$ be the total number of amino acids in our imaginary world at $t = 0$. Then, for all $n \geq 0$, we have*

$$x_1^{(n)} + x_2^{(n)} + x_3^{(n)} = N. \tag{4}$$

Note that the number $N$ in the above observation is very large, for example, think of $N$ as $N = 10^{20}$.

We denote by $\mathbf{x}^{(n)}$ the vector with components $x_1^{(n)}$, $x_2^{(n)}$ and $x_3^{(n)}$

$$\mathbf{x}^{(n)} = \begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \\ x_3^{(n)} \end{bmatrix}. \tag{5}$$

Observation 2.3 and the definition of the vector $\mathbf{x}^{(n)}$ (see the above equation and Equation (3)) imply the following observations:

**Observation 2.3.**

*(1). Out of the $x_1^{(n)}$ amino acids $A_1$ there are at time $t = n$, at time $t = n + 1$, $p_{11}x_1^{(n)}$ of them are still $A_1$ amino acids, $p_{21}x_1^{(n)}$ of them are $A_2$ amino acids and $p_{31}x_1^{(n)}$ of them are $A_3$ amino acids.*

*(2). Out of the $x_2^{(n)}$ amino acids $A_2$ there are at time $t = n$, at time $t = n + 1$, $p_{12}x_2^{(n)}$ of them are $A_1$ amino acids, $p_{22}x_2^{(n)}$ of them are still $A_2$ amino acids and $p_{32}x_2^{(n)}$ of them are $A_3$ amino acids.*

*(3). Out of the $x_3^{(n)}$ amino acids $A_3$ there are at time $t = n$, at time $t = n + 1$, $p_{13}x_3^{(n)}$ of them are $A_1$ amino acids, $p_{23}x_3^{(n)}$ of them are $A_2$ amino acids and $p_{33}x_3^{(n)}$ of them are still $A_3$ amino acids.*

The above observation can be summarized in the following table showing the number of amino acids of different type at times $t = n$ and $t = n + 1$:

| amino acid | number of amino acid at $t = n$ | number of amino acid at $t = n + 1$ |
|---|---|---|
| $A_1$ | $x_1^{(n)}$ | $p_{11}x_1^{(n)} + p_{12}x_2^{(n)} + p_{13}x_3^{(n)}$ |
| $A_2$ | $x_2^{(n)}$ | $p_{21}x_1^{(n)} + p_{22}x_2^{(n)} + p_{23}x_3^{(n)}$ |
| $A_3$ | $x_3^{(n)}$ | $p_{31}x_1^{(n)} + p_{32}x_2^{(n)} + p_{33}x_3^{(n)}$ |

Recalling the meaning of $\mathbf{x}^{(n+1)}$ and the definition of matrix-vector multiplication we have that the following vector equation is valid for all $n \geq 0$:

$$\mathbf{x}^{(n+1)} = P\mathbf{x}^{(n)}. \tag{6}$$

## 2.4. Numerical simulations of the total number of amino acids over time

Once the vector of initial number of amino acid, $\mathbf{x}^{(0)}$, is given, the number of amino acids at any future time $t = n$, which is the vector $\mathbf{x}^{(n)}$, can be computed for all $n$ using the vector Equation (21). In fact, since that equation is valid for all non-negative integers $n$, replacing $n$ by 0 we get that $\mathbf{x}^{(1)} = P\mathbf{x}^{(0)}$. Replacing $n$ by 1 we get that $\mathbf{x}^{(2)} = P\mathbf{x}^{(1)}$, and so on. Let $N$ be the total number of amino acids, i.e. $N = x_1^{(0)} + x_2^{(0)} + x_3^{(0)}$. In Figure 1, we plot the components of $\mathbf{x}^{(n)}/N$ as a function of $n$ in the case when the transition matrix $P$ is given by Equation (2). The dashed lines are the plot of $x_1^{(n)}$, the dotted lines the plot of $x_2^{(n)}$ and the solid lines the plot of $x_3^{(n)}$. The red lines correspond to the initial conditions $x_1^{(0)}/N = 0.1$, $x_2^{(0)}/N = 0.4$ and $x_3^{(0)}/N = 0.5$. The black lines correspond to the initial conditions $x_1^{(0)}/N = 0.8$, $x_2^{(0)}/N = 0.2$ and $x_3^{(0)}/N = 0$. Note that, as the time $t = n$ increases, the number of each type of amino acid approaches certain values. These values are the same for both initial conditions (i.e. for both values of $\mathbf{x}^{(0)}/N$). In other words, both dashed lines approach the same value, both dotted lines approach the same value, and both solid lines also approach the same value. These values are approximately 0.23, 0.46 and 0.31. In other words, after a long time, the number of $A_1$ amino acids is

$0.23N$, the number of $A_2$ amino acids is $0.46N$, and the number of $A_3$ amino acids is $0.31N$. In fact, we find that, no matter the vector of initial number of amino acids, the number of $A_1$ amino acids approaches $0.23N$, the number of $A_2$ amino acids approaches $0.46N$, and the number of $A_3$ amino acids approaches $0.31N$ as the time $t = n$ increases. In mathematical language, we write

$$\lim_{n\to\infty} \frac{\mathbf{x}^{(n)}}{N} = \mathbf{z} \text{ where } \mathbf{z} = \begin{bmatrix} 0.23 \\ 0.46 \\ 0.31 \end{bmatrix} \text{ no matter the value of } \mathbf{x}^{(0)}. \tag{7}$$

The above equation reads the limit of $\mathbf{x}^{(n)}/N$ as $n \to \infty$ is $\mathbf{z}$. In particular, if we start with the initial conditions $\mathbf{x}^{(0)}/N = \mathbf{z}$, we still have that $\mathbf{x}^{(n)}/N$ approaches $\mathbf{z}$, but $\mathbf{x}^{(0)}$ is already $\mathbf{z}$, so we expect that $\mathbf{x}^{(n)}/N = \mathbf{z}$ for all $n$. This is in fact verified in Figure 2, where we plotted the components of $\mathbf{x}^{(n)}/N$ in the case that $\mathbf{x}^{(0)}/N = \mathbf{z}$. Note that the components of $\mathbf{x}^{(n)}/N$, and thus the vector $\mathbf{x}^{(n)}/N$, remain constant and equal to $\mathbf{z}$. In particular, for each $n$ we have that $\mathbf{x}^{(n)}/N = \mathbf{z}$ and also $\mathbf{x}^{(n+1)}/N = \mathbf{z}$. This fact, in addition to Equation (21) implies

$$\mathbf{z} = P\mathbf{z}. \tag{8}$$

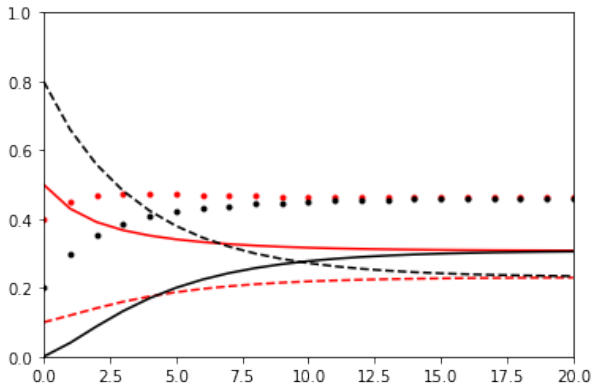


Figure 1. Evolution of the proportion of number of amino acids of different types for two different vectors of initial numbers of amino acids.

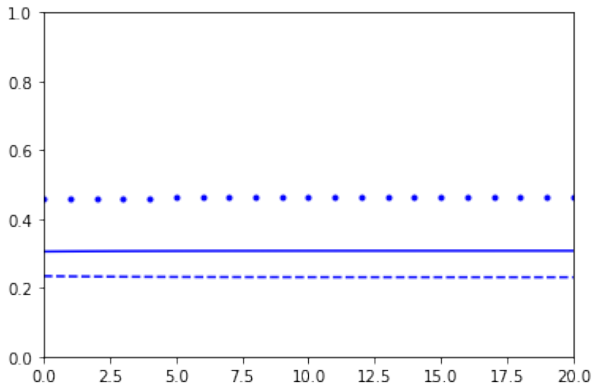


Figure 2. Evolution of proportion of number of amino acids each type when the vector of initial numbers of amino acids is $N\mathbf{z}$.

The findings of this section are a mathematical fact that we summarize here:

**Observation 2.4.** *No matter the transition matrix, there exists at least one vector* $\mathbf{z}$ *such that* $P\mathbf{z} = \mathbf{z}$ *and* $z_1 + z_2 + z_3 = 1$. *In most cases, there is exactly one such vector* $\mathbf{z}$. *We assume this is the case in the rest of this article. This vector* $\mathbf{z}$ *depends only on the matrix* $P$. *Additionally, no matter the initial condition* $\mathbf{x}^{(0)}$,

$$\lim_{n\to\infty} \frac{\mathbf{x}^{(n)}}{N} = \mathbf{z}. \tag{9}$$

Instead of computing $\mathbf{x}^{(n)}/N$ for large values of $n$ to find an approximation of $\mathbf{z}$, we find $\mathbf{z}$ by solving the system of equations

$$\begin{aligned} \mathbf{z} &= P\mathbf{z} \\ z_1 + z_2 + z_3 &= N. \end{aligned} \tag{10}$$

This is a system of four linear equations with three unknowns. Nevertheless, it always has at least one solution. In fact, in most cases, it has exactly one solution. As previously mentioned, we assume the unique solution case in this paper.
As an example, assume that the transition matrix $P$ is given by Equation (2). Then, Equations (10) become

$$\begin{aligned} 0.8z_1 + 0.1z_2 &= z_1 \\ 0.2z_1 + 0.7z_2 + 0.3z_3 &= z_2 \\ 0.2z_2 + 0.7z_3 &= z_3 \\ z_1 + z_2 + z_3 &= 1. \end{aligned} \tag{11}$$

Systems of linear equations such as System (10) can be solved using the Gaussian elimination method.

## 2.5. Probability of an amino acid of being of type $A_i$ for each $i$

We go back to finding $\mathbf{z}$, the solution of the system (10). In the particular case of the matrix $P$ given by Equation (2), we have that $\mathbf{z}$ is the solution of Equations (11). Using the algorithm we described in the previous section, we find that

$$\mathbf{z} = \begin{bmatrix} 0.230769230769 \\ 0.461538461538 \\ 0.307692307692 \end{bmatrix}, \tag{12}$$

which agrees with the value found in Equation (9) by computing $\mathbf{x}^{(n)}/N$ with large values on $n$.
Note that $z_i$ is the proportion of the amino acids that are of type $A_i$ for times $t = n$ with $n$ large. We are interested in this time regime of $t = n$ with large $n$. Let $P(A_i)$ be the probability of an amino acid being of type $A_i$. By definition, $P(A_i)$ is the proportion of the amino acids that are of type $A_i$. Thus, we have that

$$P(A_1) = z_1, \qquad P(A_2) = z_2, \qquad P(A_3) = z_3. \tag{13}$$

In vector notation

$$\mathbf{P}(\mathbf{A}) = \mathbf{z} \text{ where } \mathbf{P}(\mathbf{A}) = \begin{bmatrix} P(A_1) \\ P(A_2) \\ P(A_3) \end{bmatrix}. \tag{14}$$

## 2.6. Probability of an amino acid of type $A_i$ was of type $A_j$ a unit of time earlier

Let $1 \leq i, j \leq 3$. We define $b_{ij}$ = probability that an amino acid of type $A_i$ was of type $A_j$ a unit of time earlier. These coefficients give us the $3 \times 3$-matrix

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}. \tag{15}$$

**Observation 2.5.** *The coefficients of the matrix $B$ defined above satisfy*

$$b_{ij} = p_{ij} \frac{P(A_j)}{P(A_i)} \tag{16}$$

*Proof.* We have an amino acid. We define the events $E_1$ and $E_2$ as follows: $E_1$ = the amino acid is of type $A_i$ at time $t$. $E_2$ = the amino acid is of type $A_j$ at time $t-1$. Note that $P(E_1) = P(A_i)$, $P(E_2) = P(A_j)$, $P(E_1 \,|\, E_2) = p_{ij}$ and $b_{ij} = P(E_2 \,|\, E_1)$. Thus, applying Bayes formula, we have that

$$b_{ij} = p_{ij} \frac{P(A_j)}{P(A_i)}. \tag{17}$$

In the example when $P$ is given by Equation (2), we obtain

$$B = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.1 & 0.7 & 0.2 \\ 0 & 0.3 & 0.7 \end{bmatrix}. \tag{18}$$

□

## 2.7. Expected value of the most recent time that an amino acid of type $A_i$ and an amino acid of type $A_j$ were of the same type

Assume that we have an amino acid $\alpha$. Assume that amino acid is now of type $A_i$. The matrix computed in the previous section allow us to run computer simulations to obtain sequences $i, i_1, \ldots, i_k, \ldots$ that give us that the amino acid $\alpha$ was of type $A_{i_k}$ $k$ years ago. More precisely, we select $i_1$ randomly as follows: $i_1 = 1$ with probability $b_{i1}$, $i_1 = 2$ with probability $b_{i2}$, $i_1 = 3$ with probability $b_{i3}$. Once we have $i_1$, we select $i_2 = 1$ with probability $b_{i_1 1}$, $i_2 = 2$ with probability $b_{i_1 2}$, $i_2 = 3$ with probability $b_{i_1 3}$. We can continue computing the numbers in the sequence $i, i_1, \ldots, i_k, \ldots$ one by one.

Similarly, assume we have a second amino acid that we call $\beta$. Assume that $\beta$ is of type now of type $A_j$. We can also run numerical simulations to obtain sequences $j, j_1, \ldots, j_k, \ldots$ that give us that the amino acid $\beta$ was of type $A_{j_k}$ $k$ years ago. Assume $i \neq j$. We can now compare the sequences $i, i_1, \ldots, i_k, \ldots$ and $j, j_1, \ldots, j_k, \ldots$ to find the smallest integer $f_{ij}$ such that $i_{f_{ij}} = j_{f_{ij}}$. This gives us that the latest that the two amino acids were of the same type was $f_{ij}$ years ago. Of course, the sequences $i, i_1, \ldots, i_k, \ldots$ and $j, j_1, \ldots, j_k, \ldots$ were randomly generated. Like flipping a coin or throwing a dice, when we run random numerical simulations, we are likely to obtain different outcomes. In other words, if we run this computational experiment again, most likely we would obtain different sequences $i, i_1, \ldots, i_k, \ldots$ and $j, j_1, \ldots, j_k, \ldots$ and thus, a different value of $f_{ij}$. Regardless, the quantity of interest is $E[f_{ij}]$, the expected value of $f_{ij}$. This discussion leads to the definition of the matrix

$$\mathbf{E}(\mathbf{f}) = \begin{bmatrix} E[f_{11}] & E[f_{12}] & E[f_{13}] \\ E[f_{21}] & E[f_{22}] & E[f_{23}] \\ E[f_{31}] & E[f_{32}] & E[f_{33}] \end{bmatrix}. \tag{19}$$

Note that $f_{ii} = 0$ for all $i$ ($\alpha$ and $\beta$ were of the same type 0 years ago, which is now, because they are both of type $A_i$ now). Thus, $E[f_{ii}] = 0$ for all $i$. The other values of the matrix can be computed running many simulations as described in this section and taking averages of the values obtained.

As an example, we run the simulations in the case when the transition matrix is given by Equation (2), we obtained

$$\mathbf{E(f)} = \begin{bmatrix} 0 & 5.0591 & 6.1511 \\ 5.0591 & 0 & 3.5163 \\ 6.1511 & 3.5163 & 0 \end{bmatrix}. \tag{20}$$

The above matrix tell us that, for example, that if we have an amino acid of type $A_1$ and an other one of type $A_2$, the expected time when they were last of the same type is 5.0591 unit of time ago.

### 2.8. Expected value of the most recent time that a sequence of amino acids of type $A_i^{(1)}, A_i^{(2)}, \ldots, A_i^{(k)}$ and a sequence of amino acids of type $A_j^{(1)}, A_j^{(2)}, \ldots, A_j^{(k)}$ were of the same type

Assume that $\alpha$ and $\beta$ are now two sequences of amino acids of type $A_i^{(1)}, A_i^{(2)}, \ldots, A_i^{(k)}$ and $A_j^{(1)}, A_j^{(2)}, \ldots, A_j^{(k)}$ respectively. We define the vectors $\mathbf{i}$ and $\mathbf{j}$ such that $\mathbf{i}^T = [i^{(1)} i^{(2)} \ldots i^{(k)}]$ and $\mathbf{j}^T = [j^{(1)} j^{(2)} \ldots j^{(k)}]$.

In a similar fashion as in the last section, we can run numerical simulations that produce vectors $\mathbf{i}_s$ and $\mathbf{j}_s$ such that $\mathbf{i}_s^T = [i_s^{(1)} i_s^{(2)} \ldots i_s^{(k)}]$ and $\mathbf{j}_s^T = [j_s^{(1)} j_s^{(2)} \ldots j_s^{(k)}]$ with the meaning that the sequences $\alpha$ and $\beta$ were of type $A_i^{(1)}s, A_i^{(2)}s, \ldots, i^{(k)}s)$ and $(A_s^{(1)}, A_s^{(2)}, \ldots, j_s^{(k)})$ $s$ years ago.

Assume that $\mathbf{i} \neq \mathbf{j}$. We define $f_{\mathbf{ij}}$ to be the smallest $s$ such that $\mathbf{i}_s \neq \mathbf{j}_s$. This gives us that the latest that the two sequences of amino acids were of the same type was $f_{\mathbf{ij}}$ years ago. By running several numerical simulations with the same $\mathbf{i}$ and $\mathbf{j}$ and taking averages of the results obtained, we compute $E[f_{\mathbf{ij}}]$.

### 2.9. Computational Implementation

In order to allow easier testing of the model and an unlimited amount of simulations, the model was implemented in Python.

## 3. Discussion

There are a few outcomes of this study that seem to be the most notable. The first outcome that is notable is equation 6:

$$\mathbf{x}^{(n+1)} = P\mathbf{x}^{(n)}. \tag{21}$$

This equation highlights the importance of the transition matrix in this study, as it is how the amino acids of the next generation can be predicted. Additionally, this equation provides a foundation for the rest of the model development.

The second notable result is that no matter the values in the transition matrix the number of each amino acid, and the vector of initial amino acids approaches a certain value. In this case, the numbers were $0.23 * N$, $0.46 * N$, and $0.31 * N$, where N is the total number of amino acids at t=0. For other transition matrices, the constants will be different. This is significant because it allows the calculation of a certain type of amino acids to be easier. Additionally, if the model is extended to a more realistic world of 20 amino acids, it will be interesting to observe if this observation still holds.

The third important outcome is the development of the matrix B. The values of the matrix $b_{ij}$ represent the probability that an amino acid of type $A_i$ was of type $A_j$ a unit of time earlier. It's significance lies in the fact that it provides the base for the development of the last significant result, the most important of the whole study.

Lastly, the development of the expected value matrix is the most important outcome of this model. Each element provides the expected value of how many years ago amino acid $A_i$ and $A_j$ were the same. It is done through repeated running of the model and averaging the values that are obtained. This outcome is vital to the study, as it answers the original question that was to be answered: how long ago did two species share a common ancestor? Although it does not necessarily confirm that two species shared a common ancestor, simply because it predicts how long ago two amino acids were the same, it does provide a foundation for extension of the model. Once the model is extended, and made more robust, this question can be answered much more easily.

This study and phylogenetic studies in general have a significant impact in other scientific fields. Perhaps the largest impact extends to biotechnology and biomedical sciences. The information gained from predicting the time of species divergence can provide insight into the structure of various biological molecules, such as proteins, nucleic acids, etc. As an example, the analysis of various pathogenic species allows the tracing of infectious disease transmission, and this is enhanced by better phylogenetic tree development. Furthermore, the analysis of host and pathogen relationships can be enhanced after gaining more insight into their genetic relationships by observing these phylogenetic trees. This can allow for better drug development, since the analysis of structural and functional relationships of the two is enhanced.

# 4. Conclusion

There are two takeaways from this study that seem to be the most important. The first is the observation that at any time, there are $0.23 * NA_1$ amino acids, $0.46 * NA_2$ amino acids, and $0.31 * NA_3$ amino acids, where N is the total number of amino acids in the imaginary world at t=0. These numbers are specific to the transition matrix that we came up with in equation 2. However, through this observation we can conclude that there must be additional such coefficients for different transition matrices, and $\mathbf{x}^{(n)}/N$ always approaches a certain value. The second major takeaway is how to obtain the expected value of $f_{ij}$, the greatest number of years that amino acid $A_i$ and $A_j$ were of the same type. This can be obtained by calculating the values of the matrix

$$\mathbf{E}(\mathbf{f}) = \begin{bmatrix} E[f_{11}] & E[f_{12}] & E[f_{13}] \\ E[f_{21}] & E[f_{22}] & E[f_{23}] \\ E[f_{31}] & E[f_{32}] & E[f_{33}] \end{bmatrix}. \tag{22}$$

through running many simulations and taking averages of the resulting values. Not only is this result important because it answers the question of the study, but it also highlights how it is important for the model to be easily implementable into code (to allow for easier testing and simulation).

In the introduction we motivated the problem of developing a mathematical model that uses the idea of a transition matrix to predict how long ago two species shared a common ancestor. We decided to this through the analysis of amino acid mutations in protein sequences of various species. The main goal of using a transition matrix was to allow easier implementation in a programming language, namely Python. That goal was met through this study, since a robust mathematical foundation of the model allowed for easy translation to Python. As a result, testing the model was made much more efficient.

This study focused solely on proteins containing 3 amino acids for simplification of the model. The next logical step for extension of this study would be to create a model that accommodates all 20 amino acids, since a 3-amino-acid world is not representative of the real world. Once the model is extended to more amino acids, one can choose and analyze a specific protein that has been conserved across various species through this model. Another direction that this study could be extended is to account for insertion and deletion mutations. This model considers only substitutions. However, in reality,

insertions and deletions are common mutations and need to be considered to develop a fully accurate model. There are many different future directions that this study can act as a foundation for, and the results of these future studies can be extended to various scientific phenomena.

## References

[1] N. Goldman and Z. Yang, *A codon-based model of nucleotide substitution for protein-coding DNA sequences*, Molecular Biology and Evolution, 11(5)(1994), 725-736.

[2] E. Zuckerkandl and L. Pauling, *Molecular disease, evolution, and genetic heterogeneity*, Academic Press, New York, (1962).

[3] S. Kumar and S. B. Hedges, *Advances in time estimation methods for molecular data*, Molecular Biology and Evolution, 33(4)(2016), 863-869.

[4] M. Dayhoff, R. Schwartz and B. Orcutt, *22 a model of evolutionary change in proteins*, Atlas of Protein Sequence and Structure, 5(1978), 345-352.

[5] S. F. Altschul, *Substitution Matrices*, In: Encyclopedia of Life Sciences (ELS), John Wiley & Sons, Ltd., Chichester, (2008).