Int. J. Math. And Appl., 9(2)(2021), 209–218
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

Modeling Common Ancestors Through Mutations in Protein Sequences

Anish Gangavaram[1,*]

1 William Mason High School, 6100 Mason Montgomery Rd, Mason, OH, 45040, USA.

Abstract: We consider a mathematical model of different amino acid types. This model uses probability in order to find the predicted length of time that each amino acid type and strand have differed from each other. We analyze the model by setting the proportion for an amino acid to mutate to another. This constant is then used inside of the model and outputs an average that we can use to predict the length of time since two organisms had a common ancestor.

Keywords: Common Ancestors, Protein Sequences, Probability of Mutations, Expected Time of Common Ancestor.

1. Introduction

Deoxyribonucleic acid, or DNA for short, is made up of nucleotides. In each nucleotide there is a phosphate group, a deoxyribose sugar, and one of the following four nitrogenous bases: Adenine(A), Thymine(T), Cytosine(C), and Guanine(G). A sequence of nucleotides makes up a gene [4]. In the process of transcription and translation, genes code for proteins that are produced by the cell in order to achieve a specific function. Different genes lead to different proteins, and different proteins accomplish different functions and with enough differences in the protein, new species are created. For example, a dog has different genes than a human because humans lack some of the proteins a dog has, while dogs lack some of the proteins humans possess [3].

A sequence of nucleotides makes up genes, and a sequence of three nucleotides produces a specific amino acid in the cell. There are a total of 20 different amino acids, yet millions of combinations of proteins can be formed due to the differences in the sequence and length leading to many different proteins from a small number of amino acids. The 3-nucleotide sequences present in genes code for a specific amino acid which is a part of the protein formed. Consequently, differences in the nucleotide sequence can lead to the production of different proteins, resulting in different species of organisms.

Mutations are changes in the DNA of an organism and they can lead to changes in proteins and species. These mutations are passed down and as time progresses, a single species evolves leading to the birth of another species. By looking at the DNA sequence, we can back track and find the length of time for which two different species shared the same ancestor [1].

In this article, we consider two stands of DNA and construct a mathematical model in order to estimate the time since these organisms shared a common ancestor. These strands of DNA are different and we assume that natural selection had no impact, due to the effect of natural selection on organisms' DNA. The model is relative in time.

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

This article is constructed as listed. In Section 2, the basic mathematics necessary for the model are explained. In Sections 3-10, the model is constructed. In Section 11, the two protein sequences are compared. In Section 12, the conclusions and real world applicability of this method are explored.

2. Preliminary Mathematics

In order to model the length of time since two organisms shared a common ancestor, we need to understand components of linear algebra [2] and probability. To start we will look at vectors and matrices.

In the following vector, $\mathbf{x}$, we see a column of numbers starting with x_1 and ending with x_n. We say that vector $\mathbf{x}$ has n components and that $\mathbf{x}$ is an n-vector. We can further classify components in the vector as x_i to indicate the i$^{\text{th}}$ component of the vector for example, x_{17} is the 17$^{\text{th}}$ component of vector $\mathbf{x}$.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \tag{1}$$

In the following matrix, F, we notice that it has d rows and c columns. Thus, matrix F is $d \times c$ matrix. Components of matrix F can be identified as f_{dc}. For example, f_{35} indicates the component in the 3$^{\text{rd}}$ row and the 5$^{\text{th}}$ column of matrix F.

$$F = \begin{bmatrix} f_{11} & \dots & f_{1c} \\ & \vdots & \\ f_{d1} & \dots & f_{dc} \end{bmatrix}, \tag{2}$$

Now we start to perform basic function with vectors and matrices. For addition, we can start with two vectors, $\mathbf{z}$ and $\mathbf{y}$.

$$\mathbf{z} = \begin{bmatrix} 4 \\ 7 \\ 10 \end{bmatrix} \tag{3}$$

$$\mathbf{y} = \begin{bmatrix} 3 \\ 2 \\ 6 \end{bmatrix} \tag{4}$$

In order to add them, we will add each corresponding component so we will add z_1 and y_1, z_2 and y_2, and z_3 and y_3. The sum of z_1 and y_1 gives us t_1 provided that $\mathbf{z} + \mathbf{y} = \textit{vector}\ \mathbf{t}$. Thus, vector $\mathbf{t}$ is equal to

$$\mathbf{t} = \begin{bmatrix} 7 \\ 9 \\ 16 \end{bmatrix} \tag{5}$$

This same process holds true for subtraction, except instead of adding the components, subtract the corresponding components from each other. For example, subtracting z_1 from y_1 gives us g_1 provided that $\mathbf{y} - \mathbf{z} = \textit{vector}\ \mathbf{g}$. Thus, vector $\mathbf{g}$

equals to

$$\mathbf{g} = \begin{bmatrix} -1 \\ -5 \\ -4 \end{bmatrix} \tag{6}$$

Now, we look at how to multiply two matrices.

$$C = \begin{bmatrix} 4 & 3 & 7 \\ 1 & 7 & 2 \\ 9 & 4 & 8 \end{bmatrix} \tag{7}$$

$$D = \begin{bmatrix} 5 & 6 & 7 \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \tag{8}$$

Provided that the product of matrix C and D is equal to matrix B, we are going to find the components of matrix B. In order to do this, we must first make sure we can multiply these two matrices. Matrix C and matrix D have 3 rows and 3 columns. We can multiply these matrices because the number of columns in matrix C, 3, is equal to the number of rows in matrix D. Matrix B will have dimensions 3×3, because 3 columns from matrix C and 3 rows from matrix D. For matrix B, b_{11} is equal to $c_{11} * d_{11} + c_{12} * d_{21} + c_{13} * d_{31}$; b_{12} is equal to $c_{11} * d_{12} + c_{12} * d_{22} + c_{13} * d_{32}$; b_{13} is equal to $c_{11} * d_{13} + c_{12} * d_{23} + c_{13} * d_{33}$. This pattern continues on where each matching row from matrix C is multiplied by each matching component from the columns of matrix D. Thus, matrix B:

$$B = \begin{bmatrix} 51 & 65 & 79 \\ 20 & 30 & 40 \\ 81 & 102 & 123 \end{bmatrix} \tag{9}$$

Next, we must understand concepts related to probability [5]. If a selected event occurs 10 times in every 20 times, we say that the probability of the event happening, $P(E)$, is $10/20 = 1/2$.

From there, we will calculate the probability of two events happening and either event happening. For example, we have a dice and we want to know the probability that the number it lands on is both $A = \{\text{odd}\}$ and $B = \{\text{greater than } 4\}$. In order to do this, we look at the intersection, $\cap$. However, if we want to look at the chance of $A = \{\text{landing even}\}$ or $B = \{\text{a number greater than } 4\}$, we look at the union, $\cup$. Therefore, $A \cup B = \{2, 4, 5, 6\}$ and $A \cap B = \{5\}$.

Observation 2.1. *$P(B)+P(A) = P(A\cap B)+P(A\cup B)$. Additionally, if A and B are disjointed events, then $P(A\cap B) = 0$. Thus, we can say that $P(A \cap B) = 0$ and $P(A \cup B) = P(A) + P(B)$.*

Provided that $P(A \cap B) = P(A)P(B)$, then we can say that these events A and B are independent events. For example, the results from flipping a coin and rolling a die are independent events because the results never coincide.

Definition 2.2. *Given that C and D are two events, we can carry out an experiment L number of times. L_C represents the number of times event C occurs in the L number of times the experiment was run. $L_{D\cap C}$ represents the number of times event $D \cap C$ occurs in all of the L number of times the experiment was run. From there, we can say that the conditional probability of D provided that C occurs is*

$$P(D|C) = \frac{L_{D\cap C}}{L_C}. \tag{10}$$

Observation 2.3. *Given that F and G are two events we can say that*

$$P(F|G) = \frac{P(F \cap G)}{P(G)}. \tag{11}$$

Proof. The probability of F occurring, given that G occurred is the number of times F and G occurred divided by the number of times G occurred. Additionally, L_F is the number of times F happened during the whole L number of times the experiment was run. Thus, we can say

$$P(F|G) = \frac{L_{F \cap G}}{L_G} = \frac{P(F \cap G)}{P(G)}. \tag{12}$$

□

Observation 2.4. *Given that F and G are two independent events, don't have an intersection, we can say that*

$$P(F|G) = P(F). \tag{13}$$

Proof. From Observation 2, we see that $P(F|G) = P(F \cap G)/P(G)$. Since F and G are independent events, we say that $P(F \cap G) = P(G)P(F)$. By replacing $P(F \cap G)$ with $P(G)P(F)$ in Observation 2, we can prove this observation. □

Observation 2.5. *Given that F and G are two events, according to Bayes formula we can say that*

$$P(F|G) = P(G|F)\frac{P(F)}{P(G)}. \tag{14}$$

Proof. From Observation 2.3, $P(F \cap G) = P(F|G)P(G)$. Then, we can reverse the roles of F and G to get that $P(G \cap F) = P(G|F) * P(F)$. Thus, $P(F|G)P(G) = P(G|F) * P(F)$. From there we divide by $P(G)$, in order to prove this observation. □

3. Modeling Amino Acid Mutations in Proteins

In order to model proteins, we must look at amino acids because amino acids are the primary structure of a protein. In our model, we assume there are only 3 amino acids that we call A_1, A_2, A_3. We start with Protein $Q = A_3A_2A_1A_2A_2A_1A_3A_3A_1A_3$ which is from Species Q and we compare Protein Q with Protein $Z = A_2A_2A_3A_2A_3A_1A_1A_3A_1A_2$ which is from Species Z. Protein Q and Z both perform the same function and that is why we are able to compare these 2 proteins. We notice that there are differences in the 1^{st}, 3^{rd}, 5^{th}, 7^{th} and 10^{th} amino acid. These changes take millions of years to occur and the time for each amino acid to mutate will be represented as one evolutionary unit. Additionally, we can say each amino acid is independent of any other amino acids in the sequence. Lastly, in this comparison of amino acids we presume that the Sequence Q is the original sequence and Sequence Z is the evolved sequence.

Since each amino acid is independent in Sequences Q and Z, we need a matrix to represent the probability of a mutation resulting in A_1 from A_1, A_2, A_3. These same probabilities are necessary for A_2 and A_3 and thus we use a transition matrix that gives the probability of mutation from one amino acid type to another.

For example, R is a transition matrix:

$$R = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.7 & 0.8 \\ 0.2 & 0 & 0.1 \end{bmatrix} \tag{15}$$

In transition matrix R, r_{dc} represents the probability that amino acid A_c is now amino acid A_d. For example, after one evolutionary unit of time, r_{12} indicates the probability that amino acid A_2 mutated to amino acid A_1 after one unit of time. From here, we use an initial number of amino acids in order to determine a new frequency of amino acids, given the original frequency of amino acids and t units of time.

For example, if $\mathbf{r}_0$ is the frequency of amino acids at time, $t = 0$, then we can calculate the frequency of amino acids at time, $t = n$ by multiplying $\mathbf{r}_0$ by the transition matrix to the n$^{\text{th}}$ power, i.e. $\mathbf{r}_n = \mathbf{r}_0 R^n$.

4. Total Number of Different Types of Amino Acids in Model

With the assumption that the frequency of an event, multiplied by the total number of events, N, gives the number of times an event occurred. This assumption can be used to make the following observation:

Observation 4.1. *Starting with Nr_d amino acids, after one unit of time, about Nr_{cd} of those amino acids are A_c amino acids with the assumption that N is a large integer.*

With $t = 0$, we start with an initial number of amino acids and as $t \geq 0$, amino acids aren't created or destroyed but rather mutated. Therefore, we can say that the number of amino acids in the sequence remains the same over time.

$$\begin{aligned} x_1^{(n)} &= \text{ number of } A_1 \text{ amino acids in our whole imaginary world at time } t = n \\ x_2^{(n)} &= \text{ number of } A_2 \text{ amino acids in our whole imaginary world at time } t = n \\ x_3^{(n)} &= \text{ number of } A_3 \text{ amino acids in our whole imaginary world at time } t = n. \end{aligned} \tag{16}$$

Given this information, we can conclude the following:

Observation 4.2. *Given that N is the total number of amino acids in our model at $t = 0$. Then, for all $n \geq 0$, we have*

$$x_1^{(n)} + x_2^{(n)} + x_3^{(n)} = N. \tag{17}$$

In the above observation N is very large. In order to quantify this, we can say think of N as $N = 10^{100}$. The vector $\mathbf{x}^{(n)}$ has the following components $x_1^{(n)}$, $x_2^{(n)}$ and $x_3^{(n)}$.

$$\mathbf{x}^{(n)} = \begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \\ x_3^{(n)} \end{bmatrix} \tag{18}$$

Given Observation 4.1 and vector definition of $\mathbf{x}^{(n)}$, we can create the following observation:

Observation 4.3.

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

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

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

The following observation is represented in the Table 1 for simplicity where number of different types of amino acids at times $t = n$ and $t = n + 1$ are shown:

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

Table 1. **Evolution of Number of Amino Acids**

Using matrix-vector multiplication and meaning of $\mathbf{x}^{(n+1)}$ the following vector equation can be created which is valid when $n \geq 0$:

$$\mathbf{x}^{(n+1)} = R\mathbf{x}^{(n)} \tag{19}$$

This vector equation can be used to calculate the number of A_1, A_2, and A_3 given the vector $\mathbf{x}^{(n)}$ and the transition matrix R.

5. Convergence of Amino Acid Frequency

Let N be the total number of amino acids and let the vector $\mathbf{r}^{(n)}$ represent the frequency of each of 3 types of amino acids. The components of the vector $\mathbf{r}^{(n)}$ are $r_1^{(n)}$, $r_2^{(n)}$ and $r_3^{(n)}$. This means that $r_1^{(n)}$, $r_2^{(n)}$ and $r_3^{(n)}$ are the frequencies of A_1, A_2 and A_3 amino acids at time $t = n$, respectively, and thus, $r_1^{(n)}/N$, $r_2^{(n)}/N$ and $r_3^{(n)}/N$ are the proportions of A_1, A_2 and A_3 amino acids at time $t = n$. Note that $r_1^{(n)}/N + r_2^{(n)}/N + r_3^{(n)}/N = 1$ for all n.

In Figure 1 we plot $r_1^{(n)}/N$, $r_2^{(n)}/N$ and $r_3^{(n)}/N$ for two different initial values of the proportions. More precisely, we plot $r_1^{(n)}/N$, $r_2^{(n)}/N$ and $r_3^{(n)}/N$ with the red, blue and green circles, respectively, when $r_1^{(0)}/N = 0$, $r_2^{(0)}/N = 0$ and $r_3^{(0)}/N = 1$, and we plot $r_1^{(n)}/N$, $r_2^{(n)}/N$ and $r_3^{(n)}/N$ with plus the red, blue and green plus signs, respectively, when $r_1^{(0)}/N = 5/9$, $r_2^{(0)}/N = 2.5/9$ and $r_3^{(0)}/N = 1.5/9$.

As the time $t = n$ increases, the proportions of each of amino acids A_1, A_2 and A_3 approaches $z_1 = 0.403$, $z_2 = 0.507$ and $z_3 = 0.09$, respectively, regardless of the initial values of the proportions. It is impossible to distinguish the plus sings from the circles in the graph. This is a mathematical fact, i.e.

$$\lim_{n\to\infty} \frac{r_1^{(n)}}{N} = z_1 = 0.403, \quad \lim_{n\to\infty} \frac{r_2^{(n)}}{N} = z_2 = 0.507, \quad \lim_{n\to\infty} \frac{r_3^{(n)}}{N} = z_3 = 0.09,$$

regardless of the values of $r_1^{(0)}/N$, $r_2^{(0)}/N$ and $r_3^{(0)}/N$. In vector notation

$$\lim_{n\to\infty} \frac{\mathbf{r}^{(n)}}{N} = \mathbf{z} \text{ where } \mathbf{z} = \begin{bmatrix} 0.403 \\ 0.507 \\ 0.09 \end{bmatrix} \text{ no matter the value of } \mathbf{r}^{(0)}. \tag{20}$$

The above equation reads the limit of $\mathbf{r}^{(n)}/N$ as $n \to \infty$ is $\mathbf{z}$. This means that, after a long time, the number of A_1 amino acids is $0.403N = z_1N$, the number of A_2 amino acids is $0.507N = z_2N$, and the number of A_3 amino acids is $0.09N = z_3N$. 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.

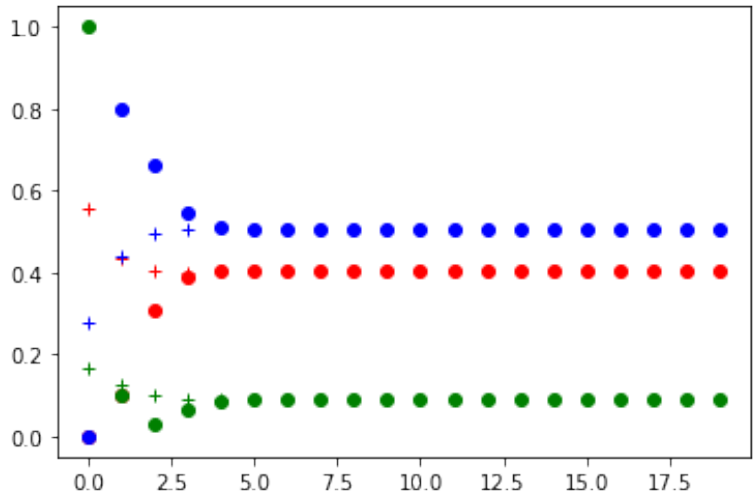

Figure 1. **Plot of $r_1^{(n)}/N$, $r_2^{(n)}/N$ and $r_3^{(n)}/N$ with two different starting values**

6. Total Number of Different Types of Amino Acids Over Time

Given the initial number of amino acids, the components of $\mathbf{x}^{(0)}$, is given, the number of amino acids at any future time $t = n$ can be computed for all values of n using vector Equation (19). This equation is valid for where n is a non-negative integer so plugging in n as 0 we get that $\mathbf{x}^{(1)} = R\mathbf{x}^{(0)}$ and from there we can use that value to plug in n as 1 to get $\mathbf{x}^{(2)} = R\mathbf{x}^{(1)}$, and so on. Therefore, this process can then be used to continually calculate the number of amino acids during each successive unit of time.

To summarize this section we make the following observation:

Observation 6.1. *Regardless of the transition matrix, there is at least one vector* $\mathbf{z}$ *such that* $R\mathbf{z} = \mathbf{z}$ *and* $z_1 + z_2 + z_3 = 1$. *In most cases, there is exactly one such vector* $\mathbf{z}$ *and this assumption will be the case for the remainder of the article. This vector* $\mathbf{z}$ *depends only on the matrix* R. *Furthermore, no matter the initial condition of* $\mathbf{x}^{(0)}$,

$$\lim_{n\to\infty} \frac{\mathbf{x}^{(n)}}{N} = \mathbf{z}. \tag{21}$$

7. Probability of Each Amino Acid

Here the limit of Equation (21) with the numerical values that we are using is

$$\mathbf{z} = \begin{bmatrix} 0.40298507462 \\ 0.50746268656 \\ 0.0895522388 \end{bmatrix} \tag{22}$$

This simulation looks at the evolutionary perspective and n has to be a very large number to represent the large evolutionary units of time that have passed causing mutations in amino acids. For example z_1 is the proportion of A_1 amino acids, z_2 is the proportion of A_2 amino acids , and z_3 is the proportion of A_3 amino acids for large values of n. $P(A_d)$ is the probability of an amino acid being of type A_d and thus we can say that

$$P(A_1) = z_1, \qquad P(A_2) = z_2, \qquad P(A_3) = z_3. \tag{23}$$

This can also be expressed in vector notation:

$$\mathbf{P}(\mathbf{A}) = \mathbf{z} \text{ where } \mathbf{P}(\mathbf{A}) = \begin{bmatrix} P(A_1) \\ P(A_2) \\ P(A_3) \end{bmatrix}. \tag{24}$$

8. Probability Amino Acid of Type A_d was A_c

In our limited world, we presumed that there are only 3 amino acids and thus d and c can only be $1, 2$, or 3. b_{dc} can be defined as the probability that an amino acid of type A_d was of type A_c a unit of time earlier. With this definition, we can create the following 3×3-matrix to represent b

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}. \tag{25}$$

Observation 8.1. *The coefficients of the matrix B can be solved with the following equation*

$$b_{dc} = r_{dc} \frac{P(A_c)}{P(A_d)} \tag{26}$$

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_d at time t.

E_2 = the amino acid is of type A_c at time $t-1$.

Note that $P(E_1) = P(A_d)$, $P(E_2) = P(A_c)$, $P(E_1 \,|\, E_2) = r_{dc}$ and $b_{fg} = P(E_2 \,|\, E_1)$. Thus, applying Bayes formula, we have that

$$b_{dc} = r_{dc} \frac{P(A_c)}{P(A_d)}. \tag{27}$$

In the example when R is given by Equation (15), we obtain

$$B = \begin{bmatrix} 0.600 & 0.282 & 0.203 \\ 0.213 & 0.700 & 0.173 \\ 0.985 & 0 & 0.100 \end{bmatrix} \tag{28}$$

□

9. Expected Value of Latest Time that A_d and A_c were Alike

Assume we have an amino acid γ and that this amino acid is of type A_d. Matrix B located in the previous sections simulates the passing of evolutionary units that result in sequences $d, d_1, \ldots, d_y, \ldots$. These sequences predict how long ago that amino acid γ was of type A_{d_y}, y evolutionary units of time ago.

To do this we select d_1 randomly as follows: $d_1 = 1$ with probability b_{d1}, $d_1 = 2$ with probability b_{d2}, $i_1 = 3$ with probability b_{d3}. After obtaining d_1, we select $d_2 = 1$ with probability $b_{d_1 1}$, $d_2 = 2$ with probability $b_{d_1 2}$, $d_2 = 3$ with probability $b_{d_1 3}$. From there, we continue to find the numbers in the sequence $d, d_1, \ldots, d_y, \ldots$ one by one.

Assume we have a second amino acid that we call δ and that δ is of type A_c. Similarly to the previous scenario, we run numerical simulations with the corresponding matrix values that result in sequences $c, c_1, \ldots, c_y, \ldots$ that give us that the amino acid δ was of type A_{c_y} y evolutionary units of time ago.

Using both sequences and with the assumption that $d \neq c$ we can compare the sequences $d, d_1, \ldots, d_y, \ldots$ and $c, c_1, \ldots, c_y, \ldots$ to find the smallest integer, f_{dc}, so that $d_{f_{dc}} = c_{f_{dc}}$. This gives us the maximum amount of time that the two amino acids were of same type f_{dc} evolutionary units of time ago.

Note that the sequences $d, d_1, \ldots, d_y, \ldots$ and $c, c_1, \ldots, c_y, \ldots$ were created using statistical probabilities so these sequences were random simulations. Thus, if we perform the simulation again, it is highly likely that the sequences will be different resulting in a different value of f_{dc}, which alters the maximum amount of time that the two amino acids were of the same type.

Nevertheless, we are looking for $E[f_{dc}]$, the expected value of f_{dc}. The matrix $E[f_{dc}]$ can be approximated by running the simulation many times and averaging the results in order to best approximate the values of $E[f_{dc}]$. In this case, we ran the simulation 1,000,000 times and averaged the results.

The process described above leads to the definition of the matrix

$$\mathbf{E}(\mathbf{f}) = \begin{bmatrix} E[f_{11}] & E[f_{12}] & E[f_{13}] \\ E[f_{21}] & E[f_{22}] & E[f_{23}] \\ E[f_{31}] & E[f_{32}] & E[f_{33}] \end{bmatrix} \tag{29}$$

We recognize that $f_{dd} = 0$ for all d (γ and δ were of the same type 0 evolutionary units of time ago. This means that presently they are both of the same type A_d, so $E[f_{dd}] = 0$ for all d. Since this is an expected value, we can run the simulation many number of times and averaging the results to best approximate the expected values as described above. Running the simulations as described above with the transition matrix given by Equation (15), we obtained

$$\mathbf{E}(\mathbf{f}) = \begin{bmatrix} 0 & 2.897 & 2.080 \\ 2.897 & 0 & 3.216 \\ 2.080 & 3.216 & 0 \end{bmatrix} \tag{30}$$

The above matrix tells us that if we have an amino acid of type A_d and another amino acid of type A_c, the expected time that they were of the same type is given by $E[f_{dc}]$. For example, if we have amino acid of type A_3 and another amino acid of type A_2, the expected time where these amino acids were of the same type is given by $E[f_{32}] = 3.216$ evolutionary units of time ago.

10. Expected Value of Latest Time that Sequences were Alike

Assume that ζ and η are sequences of amino acids of type $A_d^{(1)}, A_d^{(2)}, \ldots, A_d^{(y)}$ and $A_c^{(1)}, A_c^{(2)}, \ldots, A_j^{(y)}$ respectively. From this, we define the vectors $\mathbf{d}$ and $\mathbf{c}$ such that $\mathbf{d}^T = [d^{(1)} d^{(2)} \ldots d^{(y)}]$ and $\mathbf{c}^T = [c^{(1)} c^{(2)} \ldots c^{(y)}]$.

Similar to the last section, we use numerical simulations to produce vectors $\mathbf{d}_q$ and $\mathbf{c}_q$ such that $\mathbf{d}_q^T = [d_q^{(1)} d_q^{(2)} \ldots d_q^{(y)}]$ and $\mathbf{c}_q^T = [c_q^{(1)} c_q^{(2)} \ldots c_q^{(y)}]$. This means that the sequences ζ and η were of type $(d_q^{(1)}, d_q^{(2)}, \ldots, d_q^{(y)})$ and $(c_q^{(1)}, c_q^{(2)}, \ldots, c_q^{(y)})$ q evolutionary units of time ago.

With the assumption that $\mathbf{d} \neq \mathbf{c}$, we define $f_{\mathbf{dc}}$ to be the smallest q such that $\mathbf{d}_q \neq \mathbf{c}_q$. Therefore, the latest that the sequences of amino acids were the same is $f_{\mathbf{dc}}$ evolutionary units of time ago. Similarly, we run numerical simulations with the same $\mathbf{d}$ and $\mathbf{c}$ and average the results to compute $E[f_{\mathbf{dc}}]$, the expected value of latest possible time that the sequences were the same.

11. Comparison of Protein Q and Protein Z

As described earlier in the article, Protein Q is a sequence of amino acids from Species Q and Protein Z is a sequence of amino acids from Species Z. Using the simulation described above we calculate $E[f_{\mathbf{ih}}] = 5298.34$. This means that 5298.34 evolutionary units of time ago, these two species had the same protein sequence meaning that at this time there was a common ancestor. This method can be applied to other sequences and the less time the process yields, means that the two ancestors had a common ancestor very long ago as 5298.34 evolutionary units of time is very long. On the contrary, if the method yields a small value, then this means that the two ancestors had a common ancestor closer to the present.

12. Conclusions

In the paper we described a mathematical model to estimate the time when two species had their most recent common ancestor by comparison of protein sequences. In our simulation, we only took into count three amino acids, but this simulation could easily be expanded to 20 amino acids with the expansion of the transition matrix. We used basic programming in Python for the numerical simulations in addition to linear algebra and probability. Using these skills, we were able to compare two sequences and this method of comparison could be used to compare protein sequences from real proteins made by animals.

References

[1] Carl Ivar Branden and John Tooze, *Introduction to protein structure*, Garland Science, (2012).

[2] Werner H Greub, *Linear algebra*, Volume 23, Springer Science & Business Media, (2012).

[3] Jonathan Pevsner, *Bioinformatics and functional genomics*, John Wiley & Sons, (2015).

[4] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, J. B. Reece and N. A. Campbell, *Campbell biology*, Eleventh edition, Pearson Education Inc., (2017).

[5] Sheldon M Ross, *Introduction to probability models*, Academic press, (2014).