Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 7, No. 1, January 2025, pp: 13-26; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received August 3, 2024; revised September 2, 2024; accepted September 18, 2024; date of publication November 11, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i4.457

How to cite: Nasrin Irshad Hussain, Kuntala Boruah and Adil Akhtar, "Predicting Evolutionary importance of Amino Acids through Mutation of Codons Using K-means Clustering", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 1, pp. 13-26, January 2025.

Predicting Evolutionary Importance of Amino Acids through Mutation of Codons Using K-means Clustering

Nasrin Irshad Hussain[1], Kuntala Boruah[1], and Adil Akhtar[2]

[1] Department of Computer Application, Sibsagar University, Sivasagar-785665, India

[2] Department of Mathematics, Golaghat Engineering College, Assam-785621, India

Corresponding author: Kuntala Boruah (e-mail: kuntala17@gmail.com)

ABSTRACT Mutation is a random biological event that may cause permanent (long term) change in living organism induced by several structural or composition alteration in the proteins. During mutation, genetic materials such as nucleotide bases in the codons are changed which potentially contributed to the alteration in the codons and consequently the amino acid that new codon encodes. In this study mutation at different nucleotide base positions within the codons is analyzed to understand the evolutionary importance of amino acids. By creating hypothetical mutations at the first, second and third positions of all 61 codons (excluding stop codons) and using K-means clustering, we categorized the resulting amino acids. Our analysis reveals that mutations at the second base position generate the highest number of distinct amino acids, indicating greater evolutionary significance compared to first and third position mutations. We applied the proposed framework on COVID-2 SARS-CoV-2(Homo sapiens) amino acid sequence and are able to deduce several significant findings about the mutation patterns. The clustering analysis revealed that amino acids such as Glycine (G), Alanine (A), Proline (P), Valine (V) and one polar amino acid are recurrent in the combined centroids of the clusters. These amino acids, predominantly hydrophobic, play a crucial role in stabilizing protein structures. This framework not only gives the insight understanding of mutation patterns and their biological significance but also underscores the importance of specific amino acids in the evolutionary process.

INDEX TERMS Amino acid, Codons, Machine learning, K-means clustering, Mutation, COVID-2 SARS CoV-2.

I. INTRODUCTION

Proteins are pivotal macromolecules in living organisms, virtually accountable for all biological activities. The genetic code of an organism, stored in DNA, is transcribed into messenger RNA (mRNA) during protein synthesis. mRNA codons guide the assembly of amino acids into polypeptides. Due to the immense biological significance of protein, a substantial amount of research is dedicated to understanding its behavior. In parallel with biological studies, artificial intelligence based algorithms such as deep learning (DL) and machine learning (ML) are widely employed in various aspects of protein research, such as Protein Structure Prediction, Protein-Protein Interaction (PPI) Prediction, Protein Function Annotation, Protein Folding, Proteomics, Protein design, Neurodegenerative Diseases study etc.,[1],[2],[3],[4],[5],[6],[7],[8].

State-of-the-art technologies such as machine learning (ML) and deep learning (DL) offer predictive and decision-making capabilities across various fields. Machine learning focuses on the development of algorithms and statistical models that can learn from the input data provided to the model and make predictions on unseen data. The performance of these models depends on the quality and quantity of the data they are trained on; the more diverse the input data, the more generalized the model will be. Machine learning models can be supervised or unsupervised based on the learning strategy. In supervised learning, both the data and the associated labels are provided during training, whereas unsupervised learning is trained on unlabeled datasets. As we are interested in exploratory analysis, we prefer an unsupervised learning approach for which we employed the K-Means Clustering

algorithm because of the evidence of its efficiency in several biological network-related applications [9],[10],[11],[12].

In recent years there have been increasing interests in the use of machine learning to investigate complex biological systems. Mutation of Amino acids in the spike protein plays a vital role in SARS-CoV-2 evolution process. Mutations in Receptor Binding Domain, the spike protein have led to a change in spike-ACE2 recognition, have resulted in viral immune evasion and the inability of neutralizing antibodies [13],[14]. A new method, D614G mutation developed in which the replacement of amino acid G (Gly) for amino acid D (Asp), is common in the spike protein during the early stages of a pandemic and conservative across all major forms [15], [16]. In point base mutations in the spike protein, corona viruses frequently undergo viral genomic recombination, particularly in the late pandemic phase when several genotypes co-circulate[17]. The evolution of SARS-CoV-2 follows the mutation-selection-adaptation theory of Darwinian within the population itself. The evolutionary trend leads to improve the transmissibility of variations while decreasing pathogenicity, which keeps the virus in human hosts for an extended period of time[18].The impact of mutation of amino acids may lead to changes in the folding and stability of the protein[19], [20]. Protein sequence mutations can change the native structure supplied by the sequence of wild-type[21], [22].There have been many research going on protein structure and its stability, proteins are unexpectedly robust to site mutations bearing significant numbers of substitutions with few alterations in structure, stability, or function[21].Mutation impact on protein stability which may cause different disorder or disease. For example, Sickle-Cell Anemia (SCA), is a group of genetically passed down blood disorders. When a single amino acid substitution occurs the glutamate (hydrophobic) presents in SCA is replaced by valine(hydrophilic). This phenomenon leads to sickle like shape which cause sickle cell anemia. Graph mining techniques are also used to analyse biological networks. In amino acid network based on property similarity analysed through centrality measures to find out the amino acids which play an evolutionary important role [42].

In the following sections, we have discussed some theoretic concepts which are used in our research work.

A. AMINO ACID AND CODONS

Amino acids are the building blocks of proteins, consisting of a central carbon atom, a hydrogen atom, an amino group, a carboxyl group and a variable R group (side chain). A protein is formed when a chain of such amino acid folds together to give a specific three-dimensional structure. There are twenty amino acids in nature which are responsible to create any protein in living organisms. Each amino acid is encoded by 3 base long sequences called codon. These bases can be one of four types: Adenine (A), Cytosine (C), Guanine (G), and Uracil (U). In each position of codon i.e., at 1st base position,2nd base position and 3rd base position there are four possibilities of bases (A or U or G or C) which results in 64 codons ($4^3 = 64$).The three triplets UAA, UAG and UGA are known as stop codons or nonsense codons and their role is to stop the biosynthesis. The mapping between 61 codon and 20 amino acids is many to one i.e., one amino acid can be encoded by multiple codons.

B. MUTATION

The process of mutation is one that causes a permanent change in DNA or RNA sequence. Mutations may happen for various reasons as the process is random. The changing from one amino acid to any other amino acids may cause alteration in corresponding gene sequence. In mutation the genetic code may be inherited and transferred to the next generation. Certain mutations have no impact on evolution since they cannot be transferred to descendants. The special mutations that matter to large-scale evolution are those which can be passed on to the offspring or descendants. There may be neutral mutation also where after change in the nucleotide in a codon, the new nucleotide also encode the original amino acid resulting in no change of amino acid sequence. Variations of mutations can arise in sequences of DNA or RNA. Following are the various kinds of mutations:

C. SUBSTITUTION MUTATION

A mutation that switches one base to another one is called substitution mutation. Additionally; there are two kinds of substitution mutations: transversion and transition. Mutations classified as transitions happen when pyrimidine bases (C ↔ T) are switched for purine bases (A ↔ G) or pyrimidine bases (C↔T) are switched for purine bases. The transversions mutation happen when pyrimidines or purines are switched around. Silent mutation is the substitution of a codon with one that encodes the same amino acid without changing the protein that is produced. An example of this mutation is given below:

CTG GA G
CTG GG G

D. INSERTION MUTATION

Insertion mutation occurs when one or more additional bases are inserted into the DNA sequence. In this structure, the synthesized protein could not perform as desired. Genetic disorders can occur based on the part of the gene in which the insertion takes place. The following sequence shows the insertion phenomena.

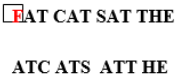
CTGGAG
CTGGUGG AG

E. DELETIONS MUTATION

Deletion mutation may happen when a part of the sequence is lost, or deleted. Then the genetic material is lost through this removal. The deleted nucleotide may alter the purpose of the resulting protein or proteins. This leads to a various genetic disease. The following example shows the deletion phenomena.

F. FRAMESHIFT MUTATION

In this type of mutation, the entire codon may be changed due to insertion or deletion. This resulting a new sequence of codons, which may change the translated polypeptide chain. For example, consider the sequence, "FAT CAT SAT THE". Here, each word represents a particular codon. If we eliminate the first letter and rewrite the sequence then it does not carry any sense or meaning. The following example shows the phenomena clearly.



G. AMINO ACID NETWORK (AAN)

Protein is one of the most important components of a biological being as it is directly or indirectly responsible for all cellular activity. Therefore, several researchers devised various stat of the art technique to study protein and its behaviors. An amino acid is the building block of protein i.e., a protein is a chain of amino acid. The interaction (biochemical or electrostatic) between the amino acids of a protein or between different proteins is called as amino acid network (AAN) also known as protein-protein interaction (PPI) network. These networks are essential for understanding the complex biological processes and functions of proteins.

H. K-MEAN CLUSTERING

K-means clustering is an effective unsupervised machine learning technique. This process allows training the model by using unlabeled, unclassified data and enables the algorithm to operate on that data without supervision. This technique is for partitioning the data into a certain number of clusters such that grouping is done based on underlying patterns or structures in the data. The K-means algorithm initiates with randomly chosen centroids which serve as the starting points for each cluster, to process the data set. It then carries out iterative (repetitive) calculations to optimize the centroids' positions. The K centers change their locations until no more changes are done or in other words centers do not move any more. At last the algorithm minimize the objective function which is known as square error function given Eq. (1) [23].

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2 \quad (1)$$

where, $||x_i - v_j||$ is the Euclidean distance between x_i and v_j and C is the number of cluster centers where C_i is the number of data points in i^{th} cluster in the equation (1)[23].

Despite the popularity of *K*-means clustering, it is difficult to choose number of clusters (or *K*) before the algorithm has been implemented. To address this issue, two quantitative methods are used- elbow plot and silhouette score. When implementing an elbow plot, look for the section of the line that looks similar to an elbow. The elbow is the point where the decrease begins to plateau. Here we will use elbow plot method to find out the optimal *K* value.

This method uses the concept of *wcss* means within cluster sum of squares that defines the total variations within a cluster. The Eq. (2) to calculate the value of *wcss* for 3 clusters[24].

$$wcss = \sum p_{i\ in\ cluster1}\ Distance(\mathrm{P_i}\ C_1)^2 + \sum p_{i\ in\ cluster2}\ Distance(\mathrm{P_i}\ C_2)^2 + \sum p_{i\ in\ cluster3}\ Distance(\mathrm{P_i}\ C_3)^2 \quad (2)$$

Here, $\sum p_{i\ in\ cluster1}\ Distance(\mathrm{P_i}\ C_1)^2$ is the sum of square of the distances between each data point and its centroid within a cluster 1 and same for cluster2 and cluster3 in the equation (2) [24].

The uniqueness of the proposed framework lies in the application of K-means clustering to the mutated results of base positions within codons of amino acids. Traditionally, machine learning methods are often integrated with biological techniques to achieve meaningful insights, as seen in the work of Ali et al., who explored amino acid networks based on mutations from a graph-theoretic perspective. They developed a Distance Matrix for amino acid networks by analyzing transition and transversion mutations of codons[25]. Similarly, Lee et al. utilized point mutations as a data augmentation technique to enhance the performance of Deep Neural Networks (DNNs) in genomic data analysis[26].

The remainder of this paper is structured as follows: Section II briefly discusses related works published in recent years. Section III covers the materials and methods used in this study. Section IV presents the results and findings, highlighting data generation through mutation and processing using the K-Means Clustering approach, along with the corresponding biological inferences. Section V is dedicated to the discussion, supported by bar graphs and heat maps. Finally, Section VI concludes the paper, summarizing the key findings and suggesting directions for future research.

II. RELATED WORK

The related works reviewed in this section attempt to highlight significant insights into various aspects of computational biology, particularly in the encoding of protein sequences, prediction models and the use of graph-based techniques. Most machine learning techniques in

computational biology involve converting the symbolic data of protein sequences into numeric vector representations.

Graphs are frequently used in biology to represent chemical compounds and protein sequences [27]. Numerous research teams have used methods from computational geometry and computer vision to tackle the challenge of identifying spatial themes. The difficulty of finding spatial motifs for pairs of molecules can be treated as the Largest Common Point set (LCP) problem when a protein is represented as a set of points in R3. This problem involves determining the greatest common subset between two sets of points.

Zamani and Kremer[28] investigate the efficiency of various encoding techniques by using substitution scoring matrices and artificial neural networks. To evaluate the effectiveness of an amino acid encoding scheme by comparing it to the actual biological roles played by the amino acids. Their proposed encoding scheme was based on the genetic codon, reflecting the coding process during protein synthesis.

Azadani et al. [29] proposed an innovative graph-based summarization approach that makes use of domain-specific knowledge and an efficient data mining method named frequent item-set mining. They construct a concept-based model of the source and mapping documents that discovers correlations to find a similarity function to represent the graph. The summarizer then uses a clustering approach based on minimum-spanning trees to identify the document's numerous sub-themes. The results of the experiment show how a summarization system performs on various baselines and benchmark approaches. The evaluation of the results shows that the given approach can significantly improve the performance in the biomedical domain of the summarization systems.

In the year 2022, Hou et al.[30] proposed a model based on machine learning, Fourier-transform infrared spectroscopy (FTIR) raw spectra and first derivative data to predict the amino acids content. Techniques such as Partial Least square regression, decision trees, and radial basis artificial neural networks were used in the prediction. Compared to using raw spectra, model performances were enhanced for a few amino acids when utilizing the first derivative.

Rafieezade and Fazeli[31] estimated the acid dissociation constant (pKa) of the amino group associated with 52 amino acids using the quantitative structure-property relationship (QSPR) method. Four distinct regression models are used in this study: Decision Tree (DT), PSO-SVM (Particle Swarm Optimization), FFNN (Feedforward Neural Network) and Genetic Algorithm-Multiple Linear Regression (GA-MLR). Recently, Yuan et al.[32] proposed a deep graph-based network for protein-protein interacting site prediction by converting the prediction problem into a graph classification task solved using deep learning techniques such as initial residual and identity mapping techniques, which demonstrated performance enhancement compared to structure-based methods.

Thangavel et al. [33] explore the importance of Network Analysis and Graph Theory, looking at their historical evolution, key ideas, and applications in a range of fields. They examine the applications of the mathematical framework to real-world issues, ranging from computer networks to social networks and beyond. The crucial role that Network Analysis and Graph Theory play in the current era of computer science and offer insights into its possible [34]formulate a novel model for the Corona virus (COVID-19), which may classify the different Corona virus types and identify SARS-CoV-2 from other Corona viruses, reducing the number of features to enhance the performance of the model. For evaluating the model, they used machine learning techniques for checking the accuracy, precision, sensitivity and specificity.

III. MATERIALS AND METHODS

The proposed framework aims to predict the evolutionary importance of amino acids. In Phase I, mutations of 61 different codons are generated by systematically altering the base positions within each codon. This work focuses on mutations at the 1st, 2nd, and 3rd positions of the codon because these positions play distinct roles in determining the impact of mutations. The decision to focus on these specific positions is grounded in their biological relevance and the differential impact they have on protein structure and function. Several researchers have studied the positional importance of bases in codons, noting that the frequency of errors in codons decreases from the third base, followed by the first base, with the second base exhibiting the least error frequency[35]. Furthermore, the polarity property of amino acids reveals that the second base position of a codon is associated with the hydrophobicity of the resulting amino acids. Amino acids with a U at the second position of their corresponding codons are hydrophobic and have low polarities according to the Grantham polarity scale. Those with an A at the second position are hydrophilic (polar amino acids). Amino acids with a C at the second position of their codons have intermediate polarities, while those with a G in the second position do not follow any regular pattern in their polarities[36].

In phase II K-means clustering technique is applied to group the derived amino acids based on mutation data. The elbow method was used to determine the optimal number of clusters. This method involves plotting the Sum of Squared Errors (SSE) against the number of clusters and selecting the point where the reduction in SSE starts to plateau (the elbow point). This approach ensures that the number of clusters selected provides a balance between capturing the data's variance and avoiding overfitting. For this analysis, the optimal number of clusters was found to be three (K=3) for

each mutation case (1st, 2nd and 3rd base mutations). The consistency in the number of clusters across different mutation positions adds robustness to the results. In phase III, the proposed framework is validated on some well known sequence and the biological significance is analyzed. FIGURE 1 illustrates the methodology adopted in the study.

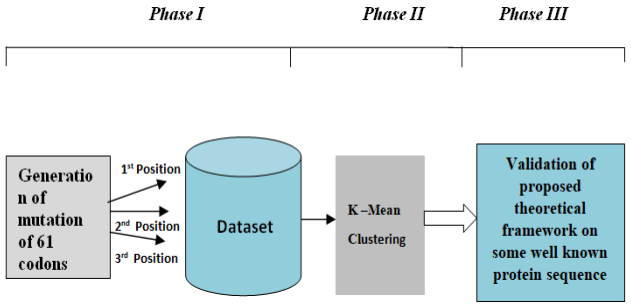


FIGURE 1. Methodology of the study

The study utilized Python programming in a Google Colab environment to implement the K-means clustering algorithm. Libraries such as scikit-learn were used for clustering, and matplotlib was employed for visualizing the clustering results. These tools were chosen based on their widespread use and reliability in bioinformatics research. The framework was validated by applying it to the SARS-CoV-2 spike glycoprotein sequence. The results from this real-world dataset were consistent with theoretical expectations, particularly highlighting the significant impact of second base mutations. There are two quantitative methods commonly used to determine the optimal number of clusters: the elbow plot and the silhouette score. Many researchers recommend the elbow method due to its compatibility with a wide range of situations and large-scale data. In our framework, the elbow plot method was used to determine the optimal number of clusters, or the "k" value. The elbow method involves calculating the Sum of Squared Errors (SSE) and inertia to visualize the data in a plot. The number of clusters is displayed on the x-axis, with SSE on the y-axis. SSE refers to the tendency, or inertia, of data points to cluster around their nearest cluster center (i.e., the centroid). As the k value increases, inertia gradually decreases. When interpreting the elbow plot, the "elbow" point is identified as the point where the line begins to flatten, indicating the optimal number of clusters.

In recent years, Ali et al. [25] developed a distance matrix of amino acids based on mutation and base position, using a graph-theoretic approach with various centrality measures to illustrate the flow of evolutionary messages in amino acid networks. They used mutations as a relationship between different amino acids. Akhtar & T. Ali [37]constructed hydrophobic and hydrophilic networks based on the mutation of codons within amino acids. They discussed the degree of distribution and skewness within the network to investigate the importance of amino acids. The findings of this research works are expected to contribute in the following ways:

a. Helps to understand the effect of mutations at different base positions within a codon.
b. The application of K-means clustering on mutation data is expected to unveil underlying biochemical and evolutionary principles, such as hydrophobic vs. hydrophilic characteristics and structural constraints on protein folding.
c. The application of the framework to the SARS-CoV-2 spike glycoprotein sequence to check whether the proposed framework align with known biological patterns, particularly highlighting the critical role of second base mutations in influencing protein function.

IV. RESULT

The genetic code is a series of codons that specify which amino acids are required to make up specific proteins. The sequence of amino acid is very specific and crucial for the synthesis of a particular type of protein, so much that even a single change in a codon may result in a completely different protein. The proposed framework progresses in three phases:

PHASE I: GENERATION OF MUTATION FOR ALL OF THE 61 CODONS

Tough in reality the mutation is a random process however, for this work we are considering a hypothetical situation of controlled mutation where we are focusing on position-wise single substitution mutation in a codon. A single nucleotide change in any codon in the amino acid chain can result in change of the entire property of the protein. As a codon is three nucleotides long (triplet), considering the possibility of mutation at the 1st, 2nd and 3rd positions, a maximum of 9 different codons may be generated. However, mutations do not always cause a change in the original amino acid. Sometimes the mutated codon may encode the same amino acid as before (neutral mutation) so the number of newly derived amino acid after mutation may be between zero to nine.

Initially, we generate all possible changes in codons due to mutations at the first position or the leftmost nucleotide of the codon (Case I). In the second case, we induced mutations in the second position of the codon (Case II). Finally, we consider the changes due to mutations in the rightmost base i.e., the third position of the codon (Case III). By examining these position-specific mutations, we aim to understand their impact on the properties and behavior of the amino acid and consequently, the protein. The pseudocode representation of the generation of mutation for 1^{st},2^{nd} and 3^{rd} base of codon is shown below:

Mutating Codons to Identify Distinct Amino Acids

```
Initialize:
    GeneticCodeTable = {codon:
    amino_acid}
    Nucleotides = ['A', 'U', 'C', 'G']
    CodonList = generate_all_codons()
Function mutate_codon(codon, pos):
    Mutated = []
      For n in Nucleotides:
      If n ≠ codon[pos]:
      Mutated.append(codon[:pos] + n +
      codon[pos+1:])
      Return Mutated
Function translate (codon):
      Return GeneticCodeTable[codon]
Analyze Mutations:
Results = []
For codon in CodonList:
      Mutations = []
      For pos in [0, 1, 2]:
      For mutant in mutate_codon(codon, pos):
      Mutations.append((mutant,
      translate(mutant)))
            Results.append((codon,
            Mutations))
Compile Results:
    For (codon, Mutations) in Results:
      orig_amino = translate(codon)
        Neutral, NonNeutral, Unique = 0, 0,
        set()
          For (mutant, amino) in Mutations:
          If amino == orig_amino: Neutral += 1
          Else: NonNeutral += 1;
          Unique.add(amino)
          Results.append((codon, Neutral,
          NonNeutral, list(Unique)))
Output Results:
      Print Results
```

GeneticCodeTable: This dictionary maps each codon (a sequence of three nucleotides) to its corresponding amino acid. It is initialized with predefined codon-to-amino-acid mappings.

Nucleotides: A list of the four possible nucleotides ('A', 'U', 'C', 'G') that can make up a codon.

CodonList: A list that contains all possible combinations of three nucleotides. It is generated using the generate_all_codons() function.

Function mutate_codon(codon, pos): This function generates all possible single-point mutations of a given codon at a specified position.

Function translate(codon): This function translates a codon into its corresponding amino acid using the GeneticCodeTable.

For better understanding each of the case of mutations are explained in details:

CASE-I: MUTATION AT THE 1ST POSITION

In this case it is assumed that the 1st position of the codon is changed to three other nucleotides due to mutation. We checked the 1st position mutation of all 61codons. Sometime newly generated codons are again mapped to same amino acids. For example: the amino acid 'M' represent by the codon AUG, base 'A'is changed to U, G, C which represent the codons UUG, GUG, CUG and their corresponding amino acids are L, V, L respectively i.e., we can deduce that 1st position mutation of AUG results in two distinct amino acids. In the following figure (FIGURE 2), we have shown a tree diagram for the 1st base mutation of amino acid M.

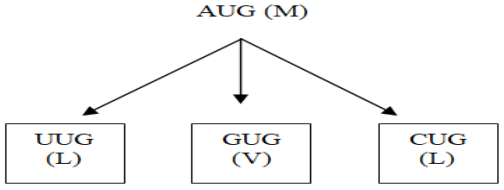


FIGURE 2.First position Mutation of AUG

TABLE I records the details of 1st base mutation for 61 codons with special emphasison the amino acid derived during the mutation.

CASE-II: MUTATION AT THE 2ND POSITION

Same process is repeated while generating mutation at 2nd position of codons. For example: AUG, in the 2nd base mutation U can be changed to A, G or C which represent the codons AAG, ACG, AGG and their corresponding amino acids are K, T, R respectively. In the following figure (FIGURE 3), we have shown a tree diagram for the mutation of amino acid M based on 2nd base mutation.

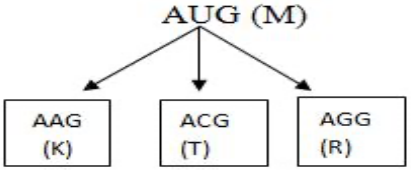


FIGURE 3. Second position Mutation of AUG

TABLE1 records the details of 2nd base mutation for 61 codons with special emphasis on the number of new amino acid derived during the mutation.

CASE-III: MUTATION AT THE 3RD POSITION

The 3rd base mutations of all cordons are generated following the same strategy as earlier. For example: AUG, in the 3rd base mutation G can be changed to A, U or C which represent the codons AUA, AUC, AUU and their corresponding amino

acid is I. In the following figure (FIGURE 4), we have shown a tree diagram for the 3rd base mutation of amino acid M.

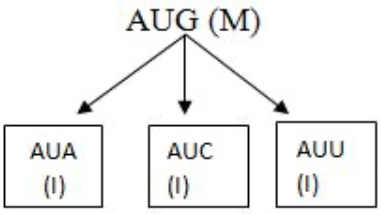


FIGURE4.Third position Mutation of AUG

Similarly, we have calculated 3rd base mutation for other 61 codons of different amino acids. In TABLE 1the number of amino acids derived due to mutated codons at 3rd base position is illustrated.

TABLE 1
Mutation record of 1st , 2nd and 3rd base

Amino Acids	Original Codons	Derived Distinct Amino acid after mutation		
		1st base	2nd base	3rdbase
Glycine(G)	GGU, GGC, GGA, GGG	S,W, R,C	D,A, V,E	0
Alanine(A)	GCU,GCC, GCA, GCG	T,P,S	D,G, V,E	0
Valine(V)	GUU,GUC, GUA, GUG	I,L, M,F	D,A, G,E	0
Leucine(L)	UUA,UUG, CUU, CUC, CUA, CUG	I,L,F, V,M	S,W, H,P, R,Q	F
Isoleucine(I)	AUU, AUC, AUA	L,V,F	N,T,S,K, R	M
Methionine(M)	AUG	L,V	K,T,R	I
Phenylalanine(F)	UUU, UUC	I,L,V	Y,S,C	L
Tryptophan(W)	UGG	R,G	S,L	C
Proline(P)	CCU, CCC, CCA, CCG	T,A,S	H,R,L,Q	0
Tyrosine(Y)	UAU, UAC	N,H,T	S,C,F	0
Serine(S)	AGU, AGC, UCU, UCC, UCA, UCG	R,G,C, T,P,A	N,T,I,Y, C,F,L,W	R
Threonine(T)	ACU, ACC, ACA, ACG	P,A,S	N,S,I,K, R,M	0
Glutamicacid (E)	GAG, GAA	K,Q	A,G,V	D
Cysteine(C)	UGU, UGC	S,R,G	Y,S,F	W
Asparagine(N)	AAU, AAC	H,D,Y	T,S,I	K
Glutamine(Q)	CAA, CAG	K,E	P,R,L	H
Aspartic acid(D)	GAU, GAC	N,H,Y	A,G,N,V	E
Lysine(K)	AAA, AAG	Q,E	T,R,I,M	N
Arginine(R)	AGG, AGA, CGU, CGC, CGA, CGG	R,G,W ,S,C	K,T,M,I, H,P,L,Q	S
Histidine(H)	CAU, CAC	N,D,Y	P,R,L	Q

PHASE II: K MEAN CLUSTERING

We employed the elbow method to find the optimal value of '***K***' for K-means clustering based on the provided dataset. In our problem, we input the 'Number of codons' and 'Number of derived amino acids' from TABLE 1 (1st position, 2nd position and 3rd position mutation). The method involves plotting the explained variance as a function of the number of clusters and selecting the elbow point of the curve as the optimal number of clusters.

The explained variance is measured using the Sum of Square Error (SSE). FIGURE 5(a) demonstrates the graph of SSE against the number of clusters. The point where the graph starts to decrease more slowly is considered the elbow point. For the 1st, 2ndand 3rd position mutations, the elbow point is found to be at K=3.

For the 1st, 2ndand 3rd position mutations, the elbow point is found to be at K=3. Therefore, the optimal '***K***' value for K-means clustering is set to 3. Next K-mean clustering is applied on the dataset derived from TABLE 1. For 1st position mutation, the derived unique amino acids are grouped into 3 clusters namely C0, C1and C2 as shown in FIGURE 5(b).

Each of the three group contains exclusive sets of amino acids : C, D, Q, E, F, K, N, M, H, Y and W lies into cluster C0, whereas amino acids S, R, L and T, A, P, V, G, I lies in cluster C1 and cluster C2 respectively. Also, the centroid of C0, C1 and C2 are [1.8, 2.5], [6.0, 4.6] and [3.8, 3.3] respectively. Further, we have calculated the combine centroid of the clusters C0, C1 and C2 which is [3.8, 3.4].An observation is made regarding the combined centroid is that it is equivalent to centroid of the cluster C2 indicating its dominance over other clusters.

Following the similar strategy, K-mean clustering is applied on 2nd base mutation derived amino acids. FIGURE 5(c) represents the clusters of derived amino acids.

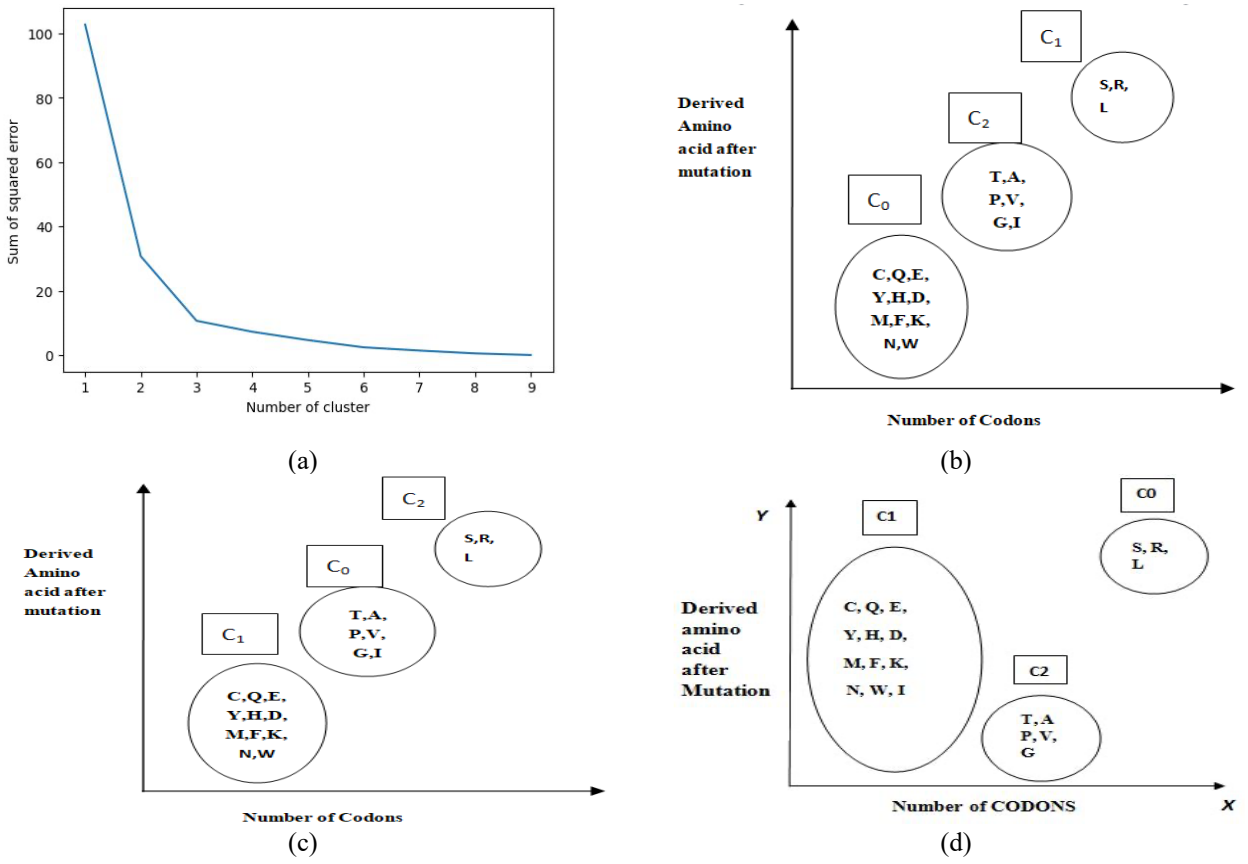


FIGURE 5. Cluster of amino acid (a) K-means cluster technique (elbow plot method), (b) Cluster of 1st position mutation, (c) Cluster of 2nd position mutation (d)Cluster of 3rd position mutation.

The amino acids T, A, P, V, G, I lies into the cluster C0, whereas the amino acids C, D, Q, E, F, K, N, M, H, Y,W and S, R, L lies into cluster C1 and cluster C2 with centroids [3.8, 4.5], [1.8, 3.0] and [6.0, 7.3] respectively. Next, we have calculated the combine centroid of the clusters C0, C1 and C2 which is [3.8, 4.9]. From here we observed that the combine centroid of the clusters C0, C1 and C2 is equivalent to centroid of the cluster C0indicating its dominance.

FIGURE 5(d) demonstrated the clusters obtained by applying K-mean clustering on 3rd base mutation data (obtained from TABLE 1.).The amino acids S, R, L lies into the cluster C0, where as the amino acids C, D, Q, E, F, K, N, M, H, Y, I ,W and T, V, G, A, P lies into cluster C1 and cluster C2 respectively with centroids: [6.0, 1.0], [1.9, 9.1] and [4.0, -1.1].Next, we have calculated the combine centroid of the clusters C0, C1 and C2 which is found to be [3.9, 3]. From here we observed that the combine centroid of the clusters C0, C1 and C2 is equivalent to centroid of the cluster C2.

We observed that mutations in the 2nd base position of 61 codons result in the maximum number of distinct amino acids compared to the 1st and 3rd base positions. Therefore, mutations in the 2nd base position are more significant than those in the 1st and 3rd positions, as they potentially may change both the genotype and phenotype of protein structures. On applying K-means clustering to the cases, we found that the combined centroids of clusters in case I is equivalent to C2, in case II is equivalent to C0 and caseIIIis equivalent to C2. Surprisingly, amino acids in these centroids are the same: Alanine (A), Proline (P), Valine (V), Glycine (G), Isoleucine (I), and Threonine (T). All these amino acids are hydrophobic in nature except Threonine (T). Hydrophobic amino acids play a crucial role in protein stability[38]. Therefore, we may conclude that the amino acids in the centroids of the clusters are important for stabilizing protein structures, with the hydrophilic amino acid being polar.

From the above study, we may also conclude that these amino acids (A, P, V, G, I, T) are more important in the evolutionary process of amino acids [39].The study also indicates that the amino acids A, P, G, and V have four codons each, whereas Threonine (T) has three codons. This suggests that amino acids with three or four codons may play an important role in the mutation process, impacting the evolution of amino acids.

PHASE III: APPLICATION OF THE PROPOSED FRAMEWORK ON COVID- SARS-COV-2 SPIKE GLYCOPROTEIN SEQUENCE

The proposed framework is evaluated using the amino acid or nucleotide sequence in FASTA format available in

National Centre for Biotechnology Information (NCBI) repository. The amino acids or nucleotides in sequences are represented by codons, which are sets of three nucleotides. The FASTA file provides the protein sequences along with their accession numbers and virus types. There are numerous coronavirus sample datasets available such as Alpha coronaviruses, Bat coronaviruses, MERS-CoV, SARS-CoV and SARS-CoV-2 etc. For this study, we randomly selected the viral protein COVID-19 SARS-CoV-2 (Homo sapiens) Dissociated S1 domain of SARS-CoV-2 Spike bound to ACE2 (Non-Uniform Refinement). This dataset was published in NCBI on 1 December 2020. The Covid-19 dataset have protein sequence where minimum and maximum lengths of protein sequence are 21 and 7097 amino acids respectively.

TABLE 2
Number of derived amino acid after mutation

Amino acid	Frequency Of amino acid in the sequence	Original codon	1st Base Mutation	2nd Base Mutation	3rd Base Mutation
G	49	196	196	196	0
A	37	148	111	148	0
V	59	236	236	236	0
L	62	372	310	372	372
I	33	99	99	165	99
M	6	6	12	18	6
F	49	98	147	147	98
W	7	7	14	14	7
P	40	160	120	160	0
Y	35	70	105	105	0
S	56	336	336	448	336
T	58	232	174	348	0
E	24	48	48	72	48
C	21	42	63	63	42
N	54	108	162	162	108
Q	27	54	54	81	54
D	31	62	93	93	62
K	30	60	60	120	60
R	29	174	145	232	174
H	9	18	27	27	18
Total	**716**	**2526**	**2512**	**3207**	**1484**

For further analysis, we developed a table based on the provided sequence using our proposed framework and the data from TABLE 1. To determine the "Original Codons" for TABLE 2, we multiplied the number of codons for each amino acid (from the amino acid codon chart) by its frequency of appearance in the sequence. For example, Glycine (G) has 4 codons and appears 49 times in the sequence, resulting in 196 original codons (4 × 49 = 196).Similarly, to calculate the number of derived amino acids in the sequence for 1st, 2nd and 3rd base mutations, we multiplied the frequency of each amino acid by the number of distinct amino acids resulting from mutations at that base position (as indicated in 1st,2nd and 3rdbase mutation TABLE1). For instance, Glycine (G) has a frequency of 49, and the number of derived amino acids for 1st, 2nd and 3rd base mutations is 4, 4, and 0, respectively. Therefore, the number of derived amino acids for Glycine is 196 (4 × 49 = 196) for both 1st and 2nd base mutations, and 0 (0 × 49 = 0) for the 3rd base mutation. We calculated the number of derived amino acids after mutations for all amino acids in the COVID-19 sequence, as summarized in TABLE 2.

From the above table (TABLE 2) we have observed that the total derived amino acids is 2512(1st base mutation), 3207(2nd base mutation) and 1484(3rd base mutation) respectively. The biologically most significant base i.e., the second base mutation induces the highest number of derived amino acid whereas the least biological significant base i.e., the third base induces least number of derived amino acid after mutation. As mention above the second base mutation has the highest changes or probability of affecting the protein and thus the phenotype. Therefore, we may successfully validate that the second base position is biologically most significant using this COVID-19 sequence. In the next step we have used machine learning technique: K-means clustering to cluster the amino acids of the sequence. The cloud environment of Google Colab is used to execute the python programme for k-mean clustering. To cluster the dataset, the number of original codons considers in X-axis and number of drive amino acids consider in Y-axis. This clustering was designed to elucidate patterns of mutational stability and evolutionary significance among the amino acids (shown in FIGURE 6-8.). As visible from the FIGURE 6, three clusters are formed to group the derived amino acids with centroids [46.6,57.5],[169, 161.37] and[354,323]respectively. Cluster 0 contains: M,W,H,C,E,Q,K,D,Y,I; Cluster 1 contains: F,N,A,P, R,G,J,V and Cluster 2 contains S,L. The average or combined coordinate of the centroids is [189.86, 180.62] which falls in cluster 1 indicating importance of amino acids in that cluster.

Similarly centroids obtained on applying K-means clustering on 2nd base mutation are [40.77, 65.88], [313.33, 389.33] and [152.37, 180.75] (shown in Fig 10). Cluster 0 contains amino acids: W,H,M,E,C,Q,D,Y,K; Cluster 1 contains: I,F,N,A,P,R,G,V and Cluster 2 contains: J,S,L. The combine centroid of the clusters is [168.82, 211.98] which falls in cluster 2. The 3rd base mutation, centroids are [56, 50.16], [354, 354] and [191,29] respectively. The combined (Average) centroid is[200,144] which fall in the cluster 2.In case of 1stbase mutation cluster 1 is the combined cluster where as in 2nd base mutation and 3rd base mutation combined centroid falls in cluster 2. The cluster 1 in 1st base mutation contains the amino acids F, N, G, V, T, R, A, P. In 2nd and 3rd base mutation C2 cluster contains amino acids I, N, F, R, G, A, P, V and A, P, G, V, R, T respectively. We observed that the amino acids Glycine (G), Alanine (A), Proline (P), Valine (V) and Arginine (R) are common in all groups. All these amino acids are hydrophobic in nature except Arginine (R). So it may conclude that the amino acids in the centroids of the combined clusters are important for stabilizing protein structures as these are hydrophobic in nature and with the hydrophilic amino acid which is polar. The amino acids A, P, G, and V have four codons each and they are non-polar in physico-chemical nature, whereas

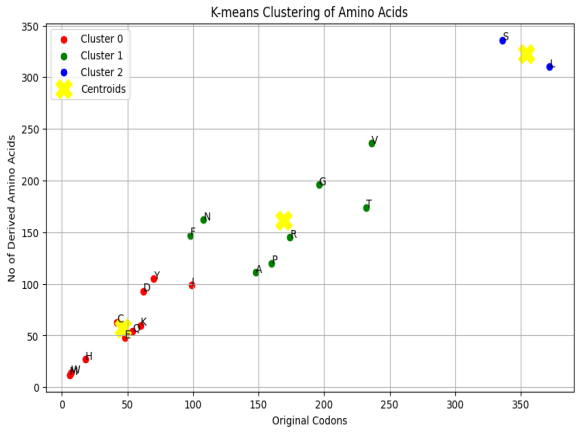


FIGURE 6.*K-Means* clustering of 1st base mutation

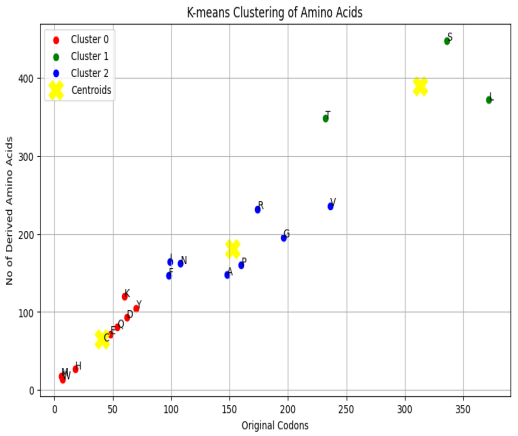


FIGURE 7. K Mean clustering of 2nd base mutation

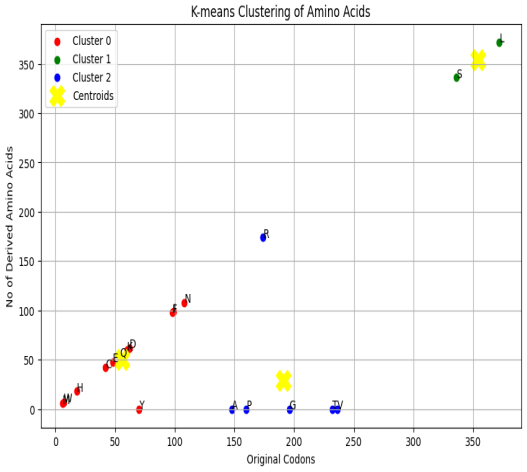


FIGURE 8. K Mean clustering of 3rd base mutation

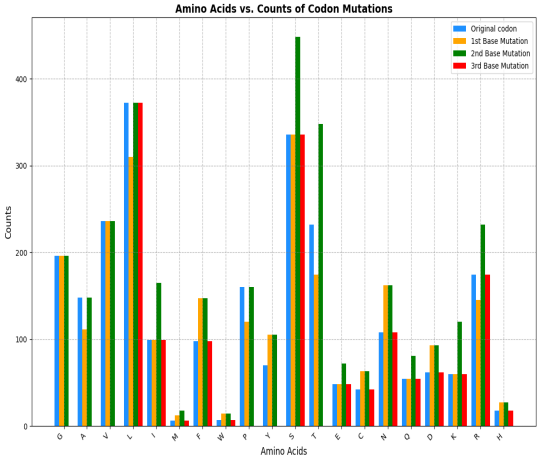


FIGURE 9. Newly derived unique amino acids vs. original codon count

Arginine (R) has six codons. These amino acids with four or six codons may play an important role in the mutation process in this sequence[39].Thus, the results of the COVID-19 sequence validate our proposed assumption.

V. DISCUSSION

On applying the proposed framework on COVID19 spike protein sequence we recorded the results in TABLE 2. To understand the trend of distinct amino acids derived after mutation in all the possible cases i.e., 1st base, 2nd base and 3rd base is represented in the form of bar graph for better visualization and understanding (FIGURE 9). It can be observed Glycine (G) and Valine (V) maintain their codon count after mutations at the 1st and 2nd base positions. However, mutations at the 3rd base position result in no valid codon leading to a count of 0. In total six (6) amino acids namely G,A,V,P,Y and T are found to be resilient to 3rd base mutation. This stability can be attributed to the redundancy of the genetic code, where different codons often encode the same amino acid due to wobble pairing. This redundancy ensures that proteins remain functionally stable despite genetic variations at the third base position. However in case of 1st and 2nd base mutation no such resilience or neutral mutation is observed. Majority of amino acids demonstrated a substantial increase in count of distinct codons after 2nd base mutataion compared to 1st or 3rd base mutation indicating that changes here are more likely to alter the protein's structure and function. This highlights the evolutionary importance of the second base in codon sequences. The first base mutations show an intermediate effect between the second and third base mutations. While they lead to significant variability in derived amino acids, the impact is not as pronounced as with second base mutations. This suggests that first base mutations can cause changes in protein structure, but their effects are less extensive compared to second base mutations. Methionine (M) shows an increase in derived amino acids when mutations occur at the first and second base positions, but it remains stable with third base mutations. This indicates a moderate sensitivity to mutations, which could lead to functional changes in proteins but within a constrained range. Phenylalanine (F) exhibits significant variability, particularly with first and second base mutations, suggesting a higher sensitivity to mutations that could impact protein function more dramatically. The graph also shows that hydrophobic amino acids such Glycine (G), Alanine (A), Proline (P), and Valine (V) are highly stable in response to third base mutations, with no derived amino acids resulting from such mutations.

FIGURE 10. Heat map amino acid counts for different Amino acids

This stability likely plays a key role in maintaining protein structure, as these amino acids are critical for forming the hydrophobic cores that stabilize protein folding. For better visualization we are reprinting the heatmap of amino acids counts for different type of mutations (in FIGURE 10). For most amino acids, the counts remain relatively stable across the Original Codon, 1st Base Mutation and 2nd Base Mutation. This stability is evident in amino acids such as Glycine (G), Valine (V), Serine (S), where the color intensity remains similar across these categories. This suggests that mutations at the first and second bases often preserve the amino acid or have a limited impact on its overall occurrence. The 3rd Base Mutation column shows significantly lower counts (often zero) for many amino acids. This is particularly noticeable in Glycine (G), Alanine (A), Valine (V), Proline (P), Tyrosine (Y), Threonine (T), and Cysteine (C), where the count drops to zero.This drop indicates that third-base mutations are often synonymous, meaning they do not change the amino acid. The genetic code's redundancy allows for variations at the third base without altering the encoded amino acid, which explains the low or zero counts. Amino acids like Leucine (L) and Serine (S) exhibit significant changes in counts when mutations occur at the first and second bases. The color intensity is much higher in these categories compared to others, indicating that these amino acids are more susceptible to changes in these positions. Methionine (M) and Tryptophan (W) show unique patterns where their counts increase in the 1st and 2nd base mutations compared to the original codon. This might indicate that mutations in these positions lead to increased occurrences of these amino acids, suggesting a higher tolerance or adaptability to such mutations.

Our study aimed to predict the biological significance of amino acids in protein structures through mutation analysis, focusing on how mutations at different base positions impact protein structure and function. The results reveal that mutations at the 2nd base position are the most biologically significant, as they produce the highest number of distinct amino acids, which may affect protein stability and function. This aligns with the evolutionary importance of the 2nd base, where nucleotide substitutions are more likely to lead to structural changes in proteins. In contrast, mutations at the 3rd base position showed minimal impact, often resulting in synonymous mutations due to the redundancy of the genetic code. This provides stability in protein structures, as the amino acid sequence remains unchanged despite these mutations. The clustering analysis of the mutated amino acids shows that hydrophobic amino acids, such as Glycine (G), Alanine (A), Proline (P), and Valine (V), consistently cluster together. This suggests that hydrophobic amino acids play a critical role in maintaining protein structure, as they form the hydrophobic core essential for protein stability. The analysis further demonstrates that these hydrophobic amino acids are resilient to mutations at the 3rd base, reinforcing their stabilizing role in protein folding.

Our findings regarding the impact of 2nd base mutations are consistent with studies like Ali et al. [39] who analyzed amino acid networks based on mutations and found that 2nd base mutations have significant implications on protein behavior, particularly in hydrophobic interactions. Additionally, research by Zamani and Kremer [28] using artificial neural networks and substitution scoring matrices showed that mutations at the 2nd base have a more pronounced effect on protein function than 1st and 3rd base mutations. However, in contrast to our findings, Yewdell [17] emphasized that both 1st and 2nd base mutations in viral proteins can result in significant structural alterations, especially in immune-evading mutations. This contrast suggests that the significance of base position mutations may vary depending on the protein and biological system under study. Another related study by Nagar et al. [40]developed a model to predict site-specific amino acid substitutions and identified that 1st base mutations, while less frequent, can still impact protein function in certain contexts, particularly in pathogenic settings. This aligns with our observation that 1st base mutations cause intermediate variability in derived amino acids but less than 2nd base mutations.

While our framework provides valuable insights into the impact of base-specific mutations on amino acids, there are several limitations. First, the study is based on a controlled, hypothetical mutation model, which may not fully capture the complexity and randomness of natural mutations. Real-world mutations often involve interactions with other biological factors, such as environmental influences or the presence of epistatic interactions, which are not considered in our model.

Second, the study focuses solely on single-point mutations within codons and does not account for other types of mutations, such as insertions, deletions, or frame-shift mutations, which could have significant biological impacts. Moreover, the study's reliance on K-means clustering, while effective, may not capture the full range of relationships

TABLE 3
COMPARISON WITH OTHER WORK

Author	Mutation type	Used Machine learning	Discussion/Result
J. W. Yewdell [17]	Point base mutation	Yes	Antigenic drift understanding Covid 19 evolutionary accumulation of amino acid mutation in viral proteins.
Nagar et al. [40]	Substitution mutation	Yes	EvoRator2 model design to predict per-site sets of tolerated amino acids and diverse applications in biomedicine such as identification of pathogenic missense mutations, Drug design etc.
Zamani et al.[28]	Substitution mutation	Yes	Investigate the efficiency of number of common amino acid used in encoding by using artificial neural networks and substitution scoring matrices.
Ali et al. [39]	Transition and transversion mutations	No	Construct amino acid networks based on mutations from a graph-theoretic perspective. They developed a Distance Matrix for amino acid networks by analyzing transition and transversion mutations of codons.
Chen et al. [41]	Deletion Mutation	No	Role of KLF6 in prostate cancers, particularly who have high grade, they examined KLF6 for deletion, mutation, and loss of expression in 96 prostate cancer samples including 21 xenografts/cell lines.
Proposed Framework	Base position mutation	Yes	Mutation of codons of amino acid to extract new patterns of amino acids in clusters using k-means clustering technique such as elbow plot. Conclude with the biological significance using this pattern prediction.

between mutated amino acids, as it simplifies the underlying biological complexity into distinct clusters. Despite these limitations, the findings have important implications for evolutionary biology, protein structure research, and practical applications in fields like drug discovery and disease prediction. The identification of 2nd base mutations as biologically significant suggests potential target sites for drug design, especially in cases where structural changes in proteins are critical for disease progression. Understanding the role of hydrophobic amino acids in maintaining protein stability can also aid in the design of more resilient protein-based therapeutics.

Additionally, the study demonstrates the potential of machine learning techniques, such as K-means clustering, to reveal underlying patterns in mutational data. Future research could expand this approach to incorporate more advanced machine learning models and larger, more diverse datasets to further validate the framework's findings. From the literature review, there are some similar works mentioned in Table 3 to compare with our work. There are different machine learning techniques used from various perspective to study the amino acids mutation.

VI. CONCLUSION

Our research aims to predict the biological significance of amino acids in protein structures through mutation analysis and machine learning. In this interdisciplinary field, machine learning offers new opportunities to uncover insights from complex biological networks. Specifically, we applied K-means clustering to amino acids to derive informative patterns from position-based mutation clusters. In this study, we developed a novel framework to predict the evolutionary importance of amino acids through controlled mutation analysis and machine learning. By generating mutations for all 61 codons of essential amino acids, we used K-means clustering to group the resulting amino acids into three clusters. Our results show that mutations at the 2nd base position have the greatest biological significance, as indicated by the higher number of derived amino acids after mutation. This suggests that when a sequence experiences frequent 2nd base mutations, there is a high probability of significant impact on protein structure and function. Clustering analysis revealed that amino acids such as Glycine (G), Alanine (A), Proline (P), Valine (V), and one polar amino acid frequently appear in the centroids of the clusters. These amino acids, mainly hydrophobic, play a crucial role in stabilizing protein structures. Our framework was validated on the COVID-19 SARS-CoV-2 sequence, further supporting our findings and demonstrating the method's potential in understanding protein behavior and evolutionary dynamics. While the framework successfully identifies mutational impact patterns, it is essential to acknowledge its limitations. The study uses a controlled, hypothetical mutation model, which may not fully capture the complexity of natural mutations, limiting the ability to reveal more intricate patterns. Future research could build on this work by exploring additional machine learning models and incorporating larger, more diverse datasets to further validate and refine the framework. Practical applications, such as drug discovery and disease prediction, could benefit from this approach. Furthermore, the use of alternative machine learning techniques to analyze gene sequences in gene banks could significantly enhance the framework's broader utility.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

RESEARCH FUNDING

This research received no external funding.

REFERENCES

[1] Y. Liu, Y. Liu, and Z. Li, "Protein–Protein Interaction Prediction via Structure-Based Deep Learning," *Proteins*, p. prot.26721, Jun. 2024, doi: 10.1002/prot.26721.

[2] S. Ohno, N. Manabe, and Y. Yamaguchi, "Prediction of protein structure and AI," *J Hum Genet*, Jan. 2024, doi: 10.1038/s10038-023-01215-4.

[3] D. Listov, C. A. Goverde, B. E. Correia, and S. J. Fleishman, "Opportunities and challenges in design and optimization of protein function," *Nat Rev Mol Cell Biol*, Apr. 2024, doi: 10.1038/s41580-024-00718-y.

[4] P. Notin, N. Rollins, Y. Gal, C. Sander, and D. Marks, "Machine learning for functional protein design," *Nat Biotechnol*, vol. 42, no. 2, pp. 216–228, Feb. 2024, doi: 10.1038/s41587-024-02127-0.

[5] S. Patil, J. Seth, and A. Ojha, "Investigating the Role of HPC in AI-based Protein-Protein Interaction Analysis," in *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, Jabalpur, India: IEEE, Apr. 2024, pp. 1003–1009. doi: 10.1109/CSNT60213.2024.10545781.

[6] K. H. Sumida *et al.*, "Improving Protein Expression, Stability, and Function with ProteinMPNN," *J. Am. Chem. Soc.*, vol. 146, no. 3, pp. 2054–2061, Jan. 2024, doi: 10.1021/jacs.3c10941.

[7] Q. Zhang, B. Liu, G. Cai, J. Qian, and Z. Jin, "Application of the AlphaFold2 Protein Prediction Algorithm Based on Artificial Intelligence," *JTPES*, vol. 4, no. 02, pp. 58–65, Feb. 2024, doi: 10.53469/jtpes.2024.04(02).09.

[8] Y. Zhou, K. Tan, X. Shen, Z. He, and H. Zheng, "A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration," in *2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, Shanghai, China: IEEE, Mar. 2024, pp. 749–753. doi: 10.1109/ICAACE61206.2024.10548253.

[9] Z. He, X. Shen, Y. Zhou, and Y. Wang, "Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering," in *Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing*, Beijing China: ACM, Jan. 2024, pp. 468–473. doi: 10.1145/3665689.3665767.

[10] X. Liu, J. Xing, H. Fu, X. Shao, and W. Cai, "Analyzing Molecular Dynamics Trajectories Thermodynamically through Artificial Intelligence," *J. Chem. Theory Comput.*, vol. 20, no. 2, pp. 665–676, Jan. 2024, doi: 10.1021/acs.jctc.3c00975.

[11] Y. Jiang, Y. Dang, Q. Wu, B. Yuan, L. Gao, and C. You, "Using a k-means clustering to identify novel phenotypes of acute ischemic stroke and development of its Clinlabomics models," *Front. Neurol.*, vol. 15, p. 1366307, Mar. 2024, doi: 10.3389/fneur.2024.1366307.

[12] L. Chen, D. R. Roe, M. Kochert, C. Simmerling, and R. A. Miranda-Quintana, "k-Means NANI: An Improved Clustering Algorithm for Molecular Dynamics Simulations," *J. Chem. Theory Comput.*, vol. 20, no. 13, pp. 5583–5597, Jul. 2024, doi: 10.1021/acs.jctc.4c00308.

[13] N. Magazine, T. Zhang, Y. Wu, M. C. McGee, G. Veggiani, and W. Huang, "Mutations and evolution of the SARS-CoV-2 spike protein," *Viruses*, vol. 14, no. 3, p. 640, 2022.

[14] Z. Chen *et al.*, "Emerging Omicron subvariants evade neutralizing immunity elicited by vaccine or BA.1/BA.2 infection," *Journal of Medical Virology*, vol. 95, no. 2, p. e28539, Feb. 2023, doi: 10.1002/jmv.28539.

[15] T. M. Wassenaar, V. Wanchai, G. Buzard, and D. W. Ussery, "The first three waves of the Covid-19 pandemic hint at a limited genetic repertoire for SARS-CoV-2," *FEMS Microbiology Reviews*, vol. 46, no. 3, p. fuac003, 2022.

[16] T.-J. Chang *et al.*, "Genomic analysis and comparative multiple sequences of SARS-CoV2," *Journal of the Chinese Medical Association*, vol. 83, no. 6, pp. 537–543, 2020.

[17] J. W. Yewdell, "Antigenic drift: understanding COVID-19," *Immunity*, vol. 54, no. 12, pp. 2681–2687, 2021.

[18] E. Goldman, "How the unvaccinated threaten the vaccinated for COVID-19: A Darwinian perspective," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, no. 39, p. e2114279118, Sep. 2021, doi: 10.1073/pnas.2114279118.

[19] M. Lorch, J. M. Mason, A. R. Clarke, and M. J. Parker, "Effects of Core Mutations on the Folding of a β-Sheet Protein: Implications for Backbone Organization in the I-State," *Biochemistry*, vol. 38, no. 4, pp. 1377–1385, Jan. 1999, doi: 10.1021/bi9817820.

[20] M. Lorch, J. M. Mason, R. B. Sessions, and A. R. Clarke, "Effects of Mutations on the Thermodynamics of a Protein Folding Reaction: Implications for the Mechanism of Formation of the Intermediate and Transition States," *Biochemistry*, vol. 39, no. 12, pp. 3480–3485, Mar. 2000, doi: 10.1021/bi9923510.

[21] D. M. Taverna and R. A. Goldstein, "Why are proteins so robust to site mutations?," *Journal of molecular biology*, vol. 315, no. 3, pp. 479–484, 2002.

[22] N. Tokuriki and D. S. Tawfik, "Stability effects of mutations and protein evolvability," *Current opinion in structural biology*, vol. 19, no. 5, pp. 596–604, 2009.

[23] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, Dec. 2012, doi: 10.1016/j.phpro.2012.03.206.

[24] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," *Accounting, Auditing and Finance*, vol. 1, no. 1, pp. 5–8, Oct. 2020, doi: 10.23977/accaf.2020.010102.

[25] T. Ali and C. Borah, "Analysis of amino acids network based on mutation and base positions," *Gene Reports*, vol. 24, p. 101291, 2021.

[26] H. Lee, U. Ozbulak, H. Park, S. Depuydt, W. De Neve, and J. Vankerschaver, "Assessing the reliability of point mutation as data augmentation for deep learning with genomic data," *BMC Bioinformatics*, vol. 25, no. 1, p. 170, Apr. 2024, doi: 10.1186/s12859-024-05787-6.

[27] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha, "Comparing Graph Representations of Protein Structure for Mining Family-Specific Residue-Based Packing Motifs," *Journal of Computational Biology*, vol. 12, no. 6, pp. 657–671, Jul. 2005, doi: 10.1089/cmb.2005.12.657.

[28] M. Zamani and S. C. Kremer, "Amino acid encoding schemes for machine learning methods," in *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Atlanta, GA: IEEE, Nov. 2011, pp. 327–333. doi: 10.1109/BIBMW.2011.6112394.

[29] M. Nasr Azadani, N. Ghadiri, and E. Davoodijam, "Graph-based biomedical text summarization: An itemset mining and sentence clustering approach," *Journal of Biomedical Informatics*, vol. 84, pp. 42–58, Aug. 2018, doi: 10.1016/j.jbi.2018.06.005.

[30] Y. Hou *et al.*, "Fourier-transform infrared spectroscopy and machine learning to predict amino acid content of nine commercial insects," *Food Sci. Technol*, vol. 42, p. e100821, 2022, doi: 10.1590/fst.100821.

[31] M.-R. Rafieezade and A. Fazeli, "Predicting the amino group pKa of amino acids using machine learning-QSPR methods," *Journal of Molecular Liquids*, vol. 408, p. 125355, Aug. 2024, doi: 10.1016/j.molliq.2024.125355.

[32] Q. Yuan, J. Chen, H. Zhao, Y. Zhou, and Y. Yang, "Structure-aware protein–protein interaction site prediction using deep graph convolutional network," *Bioinformatics*, vol. 38, no. 1, pp. 125–132, Dec. 2021, doi: 10.1093/bioinformatics/btab643.

[33] P. Thangavel, P. Shyamala Anto Mary, G. Kavitha, and K. Deiwakumari, "Graph Theory And Network Analysis: Exploring Connectivity In Computer Science," vol. 35, pp. 4884–4899, 2023.

[34] W. Alkady, K. ElBahnasy, V. Leiva, and W. Gad, "Classifying COVID-19 based on amino acids encoding with machine learning algorithms," *Chemometrics and Intelligent Laboratory Systems*, vol. 224, p. 104535, May 2022, doi: 10.1016/j.chemolab.2022.104535.

[35] R. Sanchez, E. Morgado, and R. Grau, "Gene algebra from a genetic code algebraic structure," *J. Math. Biol.*, vol. 51, no. 4, pp. 431–457, Oct. 2005, doi: 10.1007/s00285-005-0332-8.

[36] R. Grantham, "Amino Acid Difference Formula to Help Explain Protein Evolution," *Science*, vol. 185, no. 4154, pp. 862–864, Sep. 1974, doi: 10.1126/science.185.4154.862.

[37] A. Akhtar and T. Ali, "Networks in Amino Acids Based on Mutation," *Studies in Microeconomics*, vol. 3, no. 2, pp. 89–100, Dec. 2015, doi: 10.1177/2321022215588863.

[38] K. M. Biswas, D. R. DeVido, and J. G. Dorsey, "Evaluation of methods for measuring amino acid hydrophobicities and interactions," *Journal of Chromatography A*, vol. 1000, no. 1–2, pp. 637–655, Jun. 2003, doi: 10.1016/S0021-9673(03)00182-1.

[39] T. Ali, A. Akhtar, and N. Gohain, "Analysis of amino acids network based on distance matrix," *Physica A: Statistical Mechanics and its Applications*, vol. 452, pp. 69–78, Jun. 2016, doi: 10.1016/j.physa.2016.01.074.
[40] N. Nagar, J. Tubiana, G. Loewenthal, H. J. Wolfson, N. Ben Tal, and T. Pupko, "*EvoRator2*: Predicting Site-specific Amino Acid Substitutions Based on Protein Structural Information Using Deep Learning," *Journal of Molecular Biology*, vol. 435, no. 14, p. 168155, Jul. 2023, doi: 10.1016/j.jmb.2023.168155.
[41] C. Chen *et al.*, "Deletion, Mutation, and Loss of Expression of *KLF6* in Human Prostate Cancer," *The American Journal of Pathology*, vol. 162, no. 4, pp. 1349–1354, Apr. 2003, doi: 10.1016/S0002-9440(10)63930-2.
[42] Hussain NI, Boruah K. "Analysis of amino acids network based on graph mining," *Network Biology*. 2024 Sep 1;14(3):242.

AUTHOR BIOGRAPHY

Nasrin Irshad Hussain is a research scholar in the Department of Computer Applications at Sibsagar University, Sibsagar, Assam, India. She earned her M.Sc. in Information Technology from Assam Kaziranga University in 2015. Her research interests encompass Computational Biology, Network Biology, and Graph Theory. Nasrin's work focuses on exploring the intersection of computational techniques and biological systems, particularly in understanding complex biological networks and structures. She is currently engaged in research that integrates machine learning and graph-based methods to analyze biological data, aiming to contribute to advancements in bioinformatics and network analysis.

Dr. Kuntala Boruah is an Assistant Professor in the Department of Computer Applications at Sibsagar University, previously serving as Assistant Professor at Assam Rajiv Gandhi University of Cooperative Management (ARGUCOM), Assam. She completed her B.Sc. in Physics from Gauhati University in 2008, followed by an MCA from Dibrugarh University in 2011. In 2019, she earned her Ph.D. from Tezpur University. Her research focuses on DNA Computing, Deep Learning, Bioinformatics, and IoT.

Dr. Adil Akhtar is an Assistant Professor in the Department of Mathematics at Golaghat Engineering College, Golaghat, Assam, India. He completed his M.Sc. in Mathematics from Tezpur University in 2011 and earned his M.Phil. and Ph.D. from Dibrugarh University in 2012 and 2016, respectively. His research focuses on Graph Theory, Mathematical Modeling, and Network Biology. Dr. Akhtar is dedicated to exploring complex networks and developing mathematical models that contribute to understanding biological systems and their underlying structures. His work aims to bridge theoretical mathematics with real-world applications in biology and related fields.