Applied Unsupervised machine Learning in Bioinformatics Sequences

Esraa Abdul Hussein Alwan [1], Hassan Nima Habib [2*], Salma A Mahmood [3]
[1] University of Basrah, College of Sciences, Iraq
[2] University of Basrah, College of Agriculture, Iraq
[3] University of Basrah, College of Information Technology and Computer Science, Iraq

* Corresponding Author: **Hassan Nima Habib**

Article Info

ISSN (Online): 3107-6580
Volume: 01
Issue: 05
September - October 2025
Received: 15-08-2025
Accepted: 17-09-2025
Published: 16-10-2025
Page No: 20-25

Abstract
In recent years, bioinformatics has begun to develop in finding the type of disease, finding vaccines and treatment, forensic medicine, etc. Due to the abundance of data and obtaining accurate results as quickly as possible, and based heavily on machine Learning algorithms which executed on these huge data to do different tasks, such Predictions, Classification, Outler Detection, Model Discovery and Description, and many other tasks.
In this study a type of unsupervised machine learning algorithms, clustering was used to classify the DNA sequences. The Clustering Methods is very useful in Biomedical data, al different levels, DNA, RNA, and proteins, it used to predicate and identify unknown sequences depending on known ones, classify different sequences in groups, and build a hieratical structure that represents the genealogical tree, which is very useful in knowing genealogy and detecting crimes.
In this study, we used two types of clustering algorithms, K-mean and Hierarchical clustering, use elbow algorithm to find optimal value of K and different similarity measurements. found to give similar results. The used Dataset consist of 160 amino acids sequences that was collected from gene bank and the agriculture collage of Basrah university. It is stored in different extension such as (fasta, txt, docx).

DOI: https://doi.org/10.54660/.IJECA.2025.1.5.20-25

Keywords: Bioinformatics, DNA Computation, K-Mean Clustering, Hierarchical Clustering.

1. Introduction

Bioinformatics It is the management, analysis, classification and storage of Biological information such as DND, RNA and proteins, by using computers. In recent years, reliance has been on DNA greatly for all living organisms. These data have important features that make them very difficult to processing, such as, huge, unstructured and heterogeneous data. Therefore, the greatest reliance was on computer algorithms, as machine learning, data mining, and artificial intelligence algorithms [1].
As the use of machine learning in bioinformatics data has the ability to infer from data, simulate different data, formulation of drugs and vaccines, genetic prediction, mutations and disease prevention [2] recognizing new patterns of input data, and developing algorithmic systems for computers that improve with experience [3].
Many studies were done on such an unsupervised learning, especially during the spread of the Corona epidemic 2019-2021, which benefited greatly from these algorithms to classify different versions of Covid-19 disease and trying to find new ways to limit its spread. The following has review of some of them.
Onno Eberhard (2022) [4] He presented a study to find out the relationship between two organisms in terms of comparing their genomes and calculating the degree of kinship.
Juhyeon (2022) [5] SARS-COV-2 sequences data were compared their data with the original data, then find the similarity between the target and the original sequence using clustering

Mantu Bera (2021) [6]. This study includes the identification of large conserved masses within a large number of sequences and their study and their immunological capabilities to determine the masses that can serve as the vaccine.

E. Banjarnahor *et al*. (2021) [7] This research takes samples of cov-2 DNA sequences from 20 infected countries. The research uses Euclidean distance to determine the distance matrix. They used the Needleman-Wunsch algorithm to limit the spread of this virus, it is necessary to identify the kinship of this virus, and the most used way to know the kinship of this virus is to build a tree or a cluster, so the researcher was interested in applying the hierarchical assembly method in analyzing the genetic relationship on the SARS-COV-2 DNA sequence.

Yawei Li *et al*. (2021) [8] used clustering methods in phylogenetic analysis to group a total of 16,873 publicly available SARS-CoV-2 strains, to determined value of k use elbow method. To improve the accuracy, we use a state-of-the-art deep learning clustering algorithm

E. Banjarnahor *et al*. (2021) [9] The DNA sequences of SARS-CoV-2 were collected from many affected countries, this research uses the K-Mean clustering method in the assembly and uses the multiple coding vectors in the sequence analysis, use elbow method to determined value of k. The results were that the DNA sequence of SARS-COVID-2 consisted of two groups, and the second group had the largest number of members.

Heba Saadeh *et al*. (2020) [10] In her study, a k-mean gene expression microarray data algorithm that measures the expression levels of human genes in four erythroid stages was applied. After cleaning filtering and normalizing the data, where the Elbow algorithm is used in order to determine the number of aggregates in a K-mean.

In our research, we present two types of unsupervised machine learning method k-means clustering, Hierarchical clustering with different similarity measures. The main objective is to develop an application for classification, unknown identification, and built a hieratical dendram tree.

This study is divided as follow; the next section describes the algorithms used, section 3 discuss the used algorithms and their results, finally, conclusions was presented.

2. Method: Clustering and Bioinformatics

The following diagram, figure 2, describe the general steps of this work. The general algorithm can be described in the following steps:

1. Determine the required processes, classification or hieratical tree building.
2. Input DNA sequences that needs to be processed.
3. Execute elbow algorithm to find the best number of K-cluster to classify the input data.
4. Find similarity between data in order to classify them, using different similarity measurements, to find best measure to be used.
5. Execute repetitive k-mean clustering algorithm that result unsupervised classification of input data.

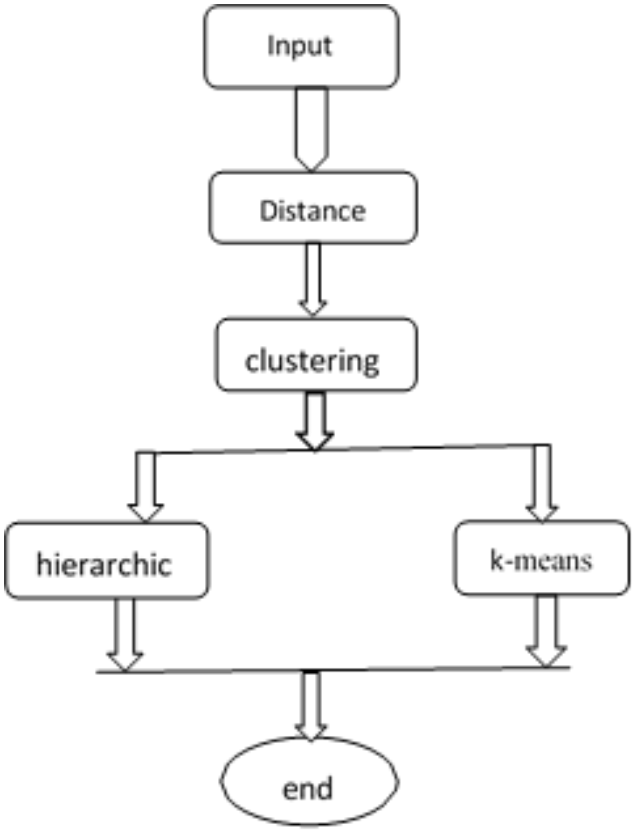


Fig 1:

3. Result:

3.1. Data set

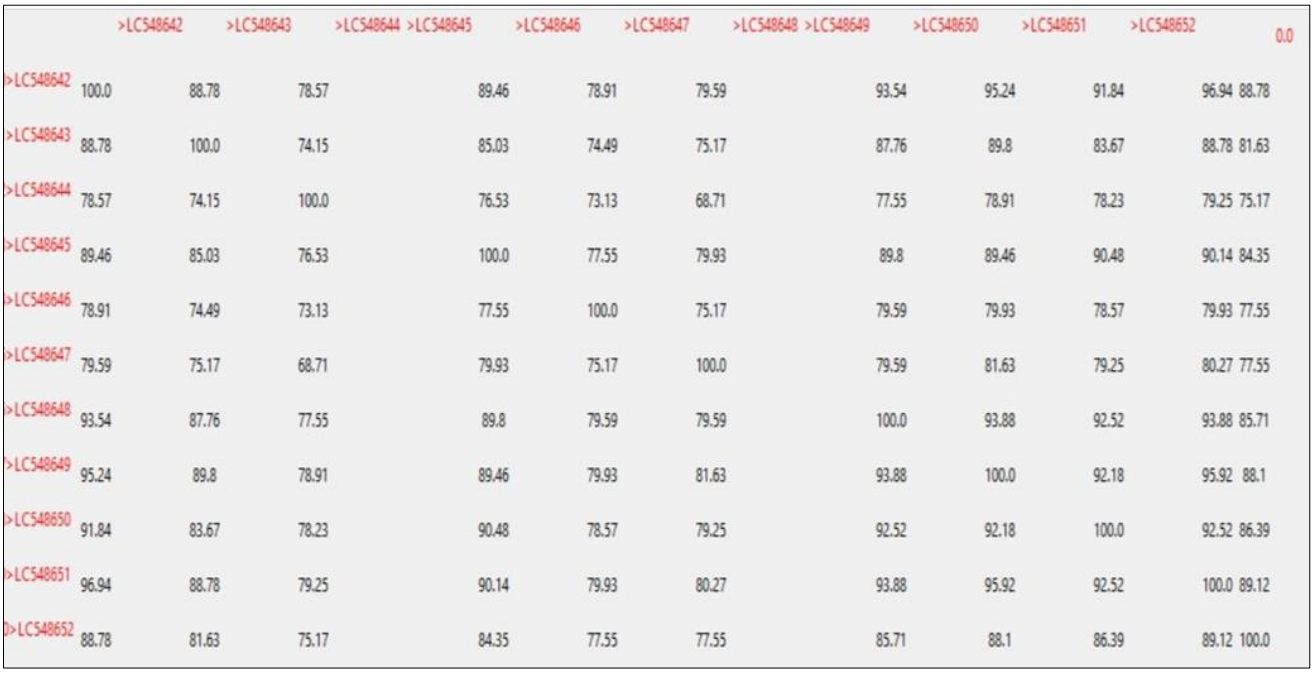
Data are collected from the gene bank as well as samples from the College of Science [11] and Agriculture, the University of Basra in Iraq. It was stored a database that includes 200 sequences. These sequences were about diseases, (HAMP, HBB, Thalassemia, Type1(AL022723), Type2(BC018404)). In this study, DNA sequences of any type were entered (Fasta, Text, word), not just FASTA.

Similarity Matrix

The similarity matrix calculates the similarity ratio between each DNA sequences with the unknown of the DNA sequences, which are of different lengths.

In this research, a method was proposed to find the similarity matrix and it was applied to a number of similarity measures, such as, (Hamming, Levenshtein Distance and Spearman's Rank Correlation)

The time was calculated to find the optimal one and our proposed was approved that it is fastest.



	>LC548642	>LC548643	>LC548644	>LC548645	>LC548646	>LC548647	>LC548648	>LC548649	>LC548650	>LC548651	>LC548652	0.0
>LC548642	100.0	88.78	78.57	89.46	78.91	79.59	93.54	95.24	91.84	96.94	88.78	
>LC548643	88.78	100.0	74.15	85.03	74.49	75.17	87.76	89.8	83.67	88.78	81.63	
>LC548644	78.57	74.15	100.0	76.53	73.13	68.71	77.55	78.91	78.23	79.25	75.17	
>LC548645	89.46	85.03	76.53	100.0	77.55	79.93	89.8	89.46	90.48	90.14	84.35	
>LC548646	78.91	74.49	73.13	77.55	100.0	75.17	79.59	79.93	78.57	79.93	77.55	
>LC548647	79.59	75.17	68.71	79.93	75.17	100.0	79.59	81.63	79.25	80.27	77.55	
>LC548648	93.54	87.76	77.55	89.8	79.59	79.59	100.0	93.88	92.52	93.88	85.71	
>LC548649	95.24	89.8	78.91	89.46	79.93	81.63	93.88	100.0	92.18	95.92	88.1	
>LC548650	91.84	83.67	78.23	90.48	78.57	79.25	92.52	92.18	100.0	92.52	86.39	
>LC548651	96.94	88.78	79.25	90.14	79.93	80.27	93.88	95.92	92.52	100.0	89.12	
>LC548652	88.78	81.63	75.17	84.35	77.55	77.55	85.71	88.1	86.39	89.12	100.0	

Fig 2: Similarity Matrix

3.2. K-means algorithm

Iterative, numerical, unsupervised, and non-deterministic describe the K-means clustering method. For combining enormous dataset collections, it is often used in data mining. It is a clustering method that partitions the given datasets into k unique clusters using an iterative process that achieves a local minimum.

Initial cluster centres for the K-means method are picked at random from the dataset using the parameter k, which is a user-specified value. Each point in a given dataset is allocated to the nearest cluster center throughout each iteration. The new centroid of each cluster is recalculated as the mean of all cluster points once all data points have been grouped into clusters. The process is repeated until either the maximum number of iterations is achieved or the centroids of freshly generated clusters stay constant [12– 13]

This algorithm was implemented on a certain number of DNA sequences, as well as on the database that was created, which contains (160) sequences.

Algorithm k-mean

Step1: entered raw data.

Step2: Call function input data.

Step3: Call function the similarity matrix uses the proposed algorithm.

Step4: from the similarity matrix the highest value is taken from each row, which indicates the most similarity between two series, and the value of the row, column and value similarity is stored in new matrix, like the following example [1].

Step5: calculate the value in k according the step-in algorithm 3.8.1.1

Step6: run k-mean algorithm, new matrix is input the algorithm

Step7: print the output

To execute K-means algorithm, we must determine the best K value by using elbow algorithm each time, before clustering. This step is very important to find best k value depending on data distribution. We used different data for different types of diseases, and by using the similarity matrix to find DNA sequences that are more similar.

3.2.1. K value calculation

In most studies that use the algorithm (k-mean) the value of k is determined at the beginning into two groups, three or maybe four, for example.

When giving a value of k = 2, the data is divided into two groups, and this does not give accurate results (meaning that the entered data are two types of disease), It would be good if the research was on one type of disease, but calculating the k value based on the similarity matrix values shows the correct results.

In some studies, it relied on the elbow algorithm, and this algorithm has been applied and gives value an approach to the method used to calculate the k value based on the values. In the elbow method, the value of k is found from graph and then the value is entered into an algorithm k-mean

In this study, we use the DNA sequence, and the difference in the sequence of a nitrogenous base leads to a difference in the similarity percentage, as the similarity values in 90% >= differ from 89%, and so on for lower values.

For this, the value of k is calculated based on the similarity matrix and finding the most similarity ratio in the row as we mentioned earlier, then the number of values within the ranges is calculated.

Algorithm

Step1: Listmax is matrix result from similarity matrix,in similarity matrix selecte max value in each row and save in listmax

Step2: Repeat until end listmax

Step3: If the value in determined range then k=k+1

Step4: If the value in same range then k=k

Step5: End

When applying the elbow algorithm and the proposed method for calculating the k value, we notice that the results are convergent, for the following input data.

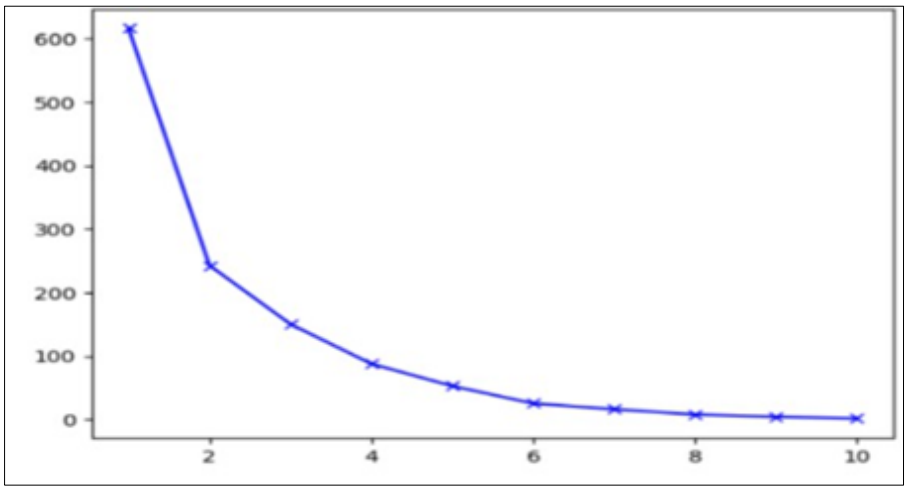


Fig 3: Elbow method with best value of k=3 and for use our method the value of k =3

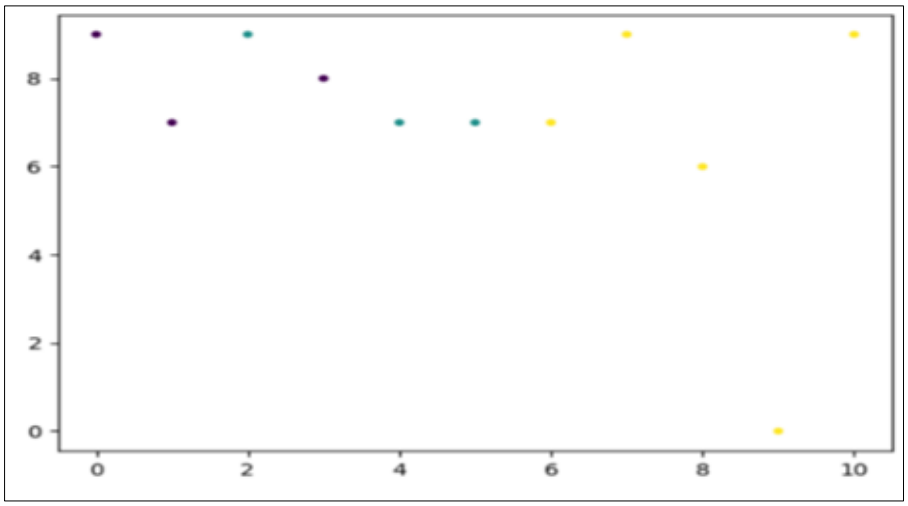


Fig 4: K-Means for input of 11 DNA sequences

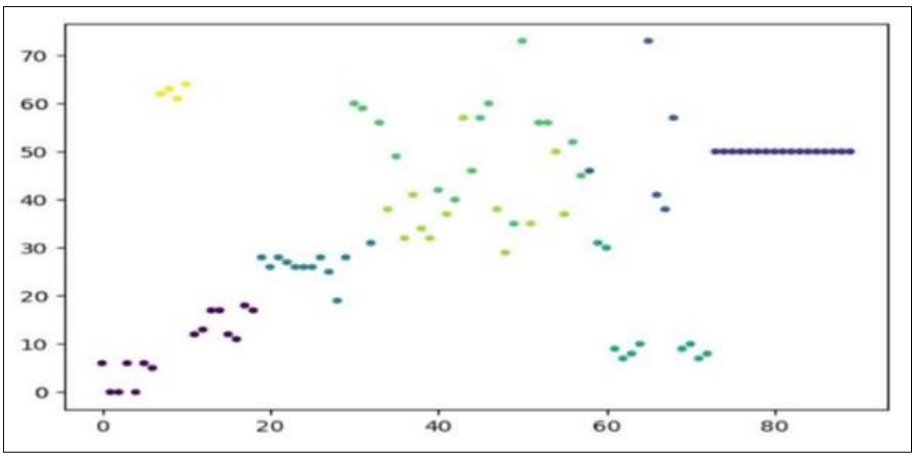


Fig 5: K-Means for database contented use 90 DNA sequences

Agglomerative Hierarchical Clustering

This algorithm was applied on sequences in collected database. The similarity matrix was used to apply the algorithm, through this algorithm, the relationship between the DNA sequences can be observed as shown in the following figure.

Algorithm

Step1: Input raw data
Step2: Call function input
Step3: Call similarity matrix
Step4: Run hierarchy.dendrogram('singl')
Step5: plt.show()

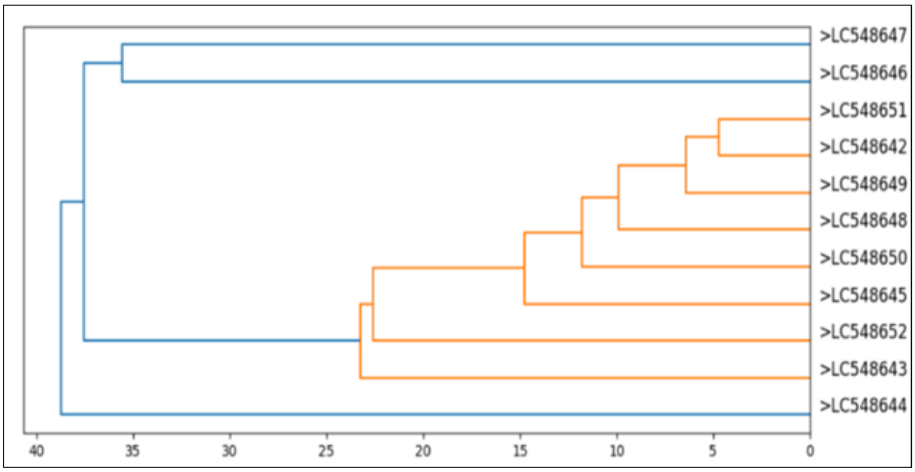


Fig 6: A Hierarchical Clustering with 11 sequences use Single linkage

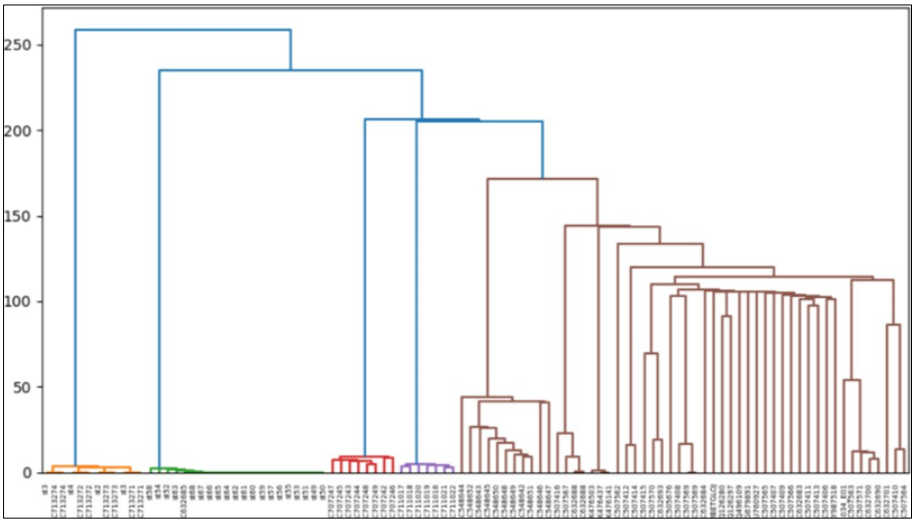


Fig 7: A Hierarchical Clustering For 100 sequences use Single linkage

4. Conclusions

The clustering methods are very important in classifying and analyzing the entered data, the entered data is only a DNA sequence. This research examines the sequences of all organisms, and many diseases like (HAMP, HBB, Thalassemia, Type1(AL022723), Type2(BC018404))

In this research, we used a method to calculate the best value of k and then used it in K-means clustering algorithm based on the results of the similarity matrix. Also, the similarity matrix was used to implement the algorithm Hierarchical Clustering, to find similar series for the purpose of classification.

5. References

1. Malik M, Undavia JN. Trials, skills, and future standpoints of AI based research in bioinformatics. Int J Recent Technol Eng. 2020;9(1):968-972. doi:10.35940/ijrte.A1920.059120
2. Hanif W, *et al.* Artificial intelligence in bioinformatics. 2019.
3. V S. An empirical science research on bioinformatics in machine learning. J Mech Contin Math Sci. 2020;spl7(1). doi:10.26782/jmcms.spl.7/2020.02.00006
4. Eberhard O. Growing phylogenetic trees using hierarchical clustering. 2022. doi:10.1101/2022.02.08.479565
5. Kim J, Cheon S, Ahn I. NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. BMC Bioinformatics. 2022;23(1). doi:10.1186/s12859-022-04718-7
6. Bera M. Artificial intelligence in bioinformatics. 2021. Available from: www.ijisrt.com
7. Banjarnahor E, Bustamam A, Mangunwardoyo W, Sarwinda D. Implementation of hierarchical clustering method in analyzing genetic relationship on DNA SARS-CoV-2 sequences. J Phys Conf Ser. 2021;1811(1):012074. doi:10.1088/1742-6596/1811/1/012074
8. Li Y, Liu Q, Zeng Z, Luo Y. Unsupervised clustering analysis of SARS-CoV-2 population structure reveals six major subtypes at early stage across the world. 2020. doi:10.1101/2020.09.04.283358
9. Saadeh H, al Fayez RQ, Elshqeirat B. Application of K-means clustering to identify similar gene expression patterns during erythroid development. Int J Mach Learn Comput. 2020;10(3):452-457. doi:10.18178/ijmlc.2020.10.3.956
10. Al-Atbee BMAK, Al-Hmudi HAM, Al-Salait SKA. Molecular and serological detection of Parvovirus B19 infection in patients with sickle cell and thalassemia disorders in Basrah province/Iraq. 2005.
11. Sarkies G, Ohannesian O. Epileptic detection from EEG recordings based on machine learning techniques.
12. Raheem SF, Alabbas MAS. Dynamic artificial bee

colony algorithm. 2021.
13. Pérez-Ortega J, Almanza-Ortega NN, Vega-Villalobos A, Pazos-Rangel R, Zavala-Díaz C, Martínez-Rebollar A. The K-means algorithm evolution. Available from: www.intechopen.com

How to Cite This Article
Alwan EA, Habib HN, Mahmood SA. Applied Unsupervised Machine Learning in Bioinformatics Sequences. Int J Eng Comput Appl. 2025;1(5):20–25. doi:10.54660/IJECA.2025.1.5.20-25

Creative Commons (CC) License
This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.