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