An Applied Program for DNA Sequences Alignment

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

* Corresponding Author: **Esraa Abdul Hussein Alwan**

Article Info

P-ISSN: 3051-3383
E-ISSN: 3051-3391
Volume: 06
Issue: 02
July - December 2025
Received: 14-10-2025
Accepted: 17-11-2025
Published: 11-12-2025
Page No: 164-169

Abstract
DNA is the identity of a living organism through which we know the characteristics of an organism. DNA is a big data that need to be processed, so as to obtains accurate results of a prediction and classification, a new trends of Machine learning processing technique. This study is part of applied application of bioinformatics contained many applications, sequences alignments, clustering and classification. This application is very important to bioinformatics scientists for disease detection and tracking, mutations Detection, Forensic medicine, agriculture, and proof of lineage.
This study is devoted to process Sequences Alignment and find the similarity matrix. An algorithm was proposed and compared with three algorithms in terms of time (hamming, Levenshtein, and rank). The proposed algorithm was the fastest in terms of time.

DOI: https://doi.org/10.54660/IJAIET.2025.6.2.164-169

Keywords: Bioinformatics, DNA Processing, Machine Learning

1. Introduction

Bioinformatics is a science to manage, dealing with and exploitation of vital information using computer methods. It is a merger between life sciences and computer sciences. It is related to molecular sciences, and thus it can be called computational life science.

It also includes operations, such as Bio- simulation and Bio-imaging that lead to the management and analysis of vital information. Bioinformatics comes from the analysis of giant biomolecules such as DNA and RNA and what results from the activities taking place in them, which is the production of proteins [1, 2]. On the other hand, the computer arranges and classifies this information for the purpose of finding patterns for analysis and exploiting the information that in it, what is known as DNA Data Mining. It helps reveal new biological Bioinformatics has great potential for analysis in different areas such as genomics, proteomics, drugs, discovery and development, protein structure, cell biology, molecular modeling, and gene expression, processes that help unify the life sciences [3].

The applications of bioinformatics began to expand day by day. It is used in the alignment and analysis of amino acids and proteins for the purpose of classifying them and predicting their functions. It was used in the manufacture of medicines and in the fields of forensic medicine, as well as in the field of manufacturing, biotechnology and agriculture. As well as in the field of health care, by analyzing the patient's genome sequences and using bioinformatics to identify harmful mutations, which helps in early treatment [1].

Recent developments in this field that use the machine learning have led to ease in bioinformatics, where large data is prospected, processed and stored. DNA Data mining is defined as searching for hidden data within large groups of DNA data and extracting patterns from it.

In recent years, due to the abundance of DNA data, it has been widely used in the field of medical sciences in the study of DNA and genetics to study and understand DNA sequences to susceptibility to disease [4].

With the emergence of the Covid-19 virus, the greatest reliance has been on bioinformatics in studying the disease and finding treatment and prevention from it, Entedar A. J. Alsaadi et. al. (2019) [5] presented study to identify its strongholds for the purpose of find a cure or limitation of its spread, Onno Eberhard (2022) [6] presented study to find the relationships between two organisms in terms of comparing their genomes and calculating the degree of kinship. Mantu Bera(2021) [6] devoted on 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] takes samples of cov-2 DNA sequences from 20 infected countries. They used Euclidean distance to determine the distance matrix and Needleman-Wunsch algorithm in try to limit the spread of this virus, it is necessary to identify the kinship of this virus, and the most used method to know the kinship of this virus is into build a tree or a cluster, so the researcher was interested in applying the hierarchical assembly method by 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 improve the accuracy, we use a state-of-the-art deep learning clustering algorithm.

M. Saqib Nawaz et al.(2021) [9] where an algorithm is designed to find locations in the genome sequence where it changes Nucleotide bases and mutation rate calculation results obtained indicate that SPM and mutagenesis analysis techniques can reveal interesting information and patterns in the COVID-19 genome sequencing to examine the evolution and differences in COVID-19 strains respectively. Marwa A. Abd Elwahaab et al. (2019) [10] a representative of both of the three groups from protein sequences is presented. The similarity/difference vector is evaluated rather than the normal similarity/difference matrix based in the representative of the group.

In this paper, we will discuss the algorithms that were used in calculating the similarity matrix, and then the unsupervised learning algorithms.

2. Sequence Alignment

Analyzing differences and similarities in biological sequences is one of the most basic and most important processes. By analyzing and comparing differences and similarities in biological sequences, biological sequences show structure and function information. The structure is determined by the sequence and the function is determined by the structure. One of the goals is to find structures and functions similar through similarities and differences

Sequence Alignment is an effective method for analyzing the position and types of mutations hidden in biological sequences and allowing precise comparisons to be made. In computational, and bioinformatics, studying the similarity of DNA sequences is essential. In almost all studies that explore evolutionary relationships, analysis of gene function, and prediction of protein structure and sequencing, it is necessary to calculate the similarity

Four algorithms were used, which are (Hamming,Levenshtein, Spearman rank correlation distance) and the fourth one was proposed called standerd algorithm) and the time was compared. It was noted that the proposed algorithm is faster, and after finding the alignment, the similarity matrix is created

The algorithms used are:

it is an algorithm for calculating the similarity ratio between two strings of equal length,

The hamming distance can be mathematically defined as: [11]

$$A_h(x,y)= \sum_{n}^{i=1} com(xi, yi)$$

$$x = y \Longrightarrow \text{com} = 0$$

$$x \neq y \Longrightarrow \text{com} = 1$$

where A_{ham} is the Hamming distance between the two sequences (x, y) and i is the index of the compared variables in the total number of variables n, Its function is similar to a gateway XOR.

To find the similarity ratio:

$$sim_{ratio} = A_{ham}(x,y)/n$$

Spearman's Rank Correlation

Linear correlation is the coefficient that expresses the strength and direction of the relationship between two phenomena only, and the relationship is either negative or positive on the one hand, and weak or strong on the other hand

Otherwise That Spearman is used for the correlation from ranks if we assume that the variable A has the rank (RA) and that the variable B has the rank (RB), and assuming that (d) represents the difference through the two ranks, meaning (d = RA − RB). The Spearman coefficient from ranks correlation is given via the following formula:

$$P=1-\frac{6 \sum di2}{n(n2-1)}$$

where n is the number from ordered pairs. [12]

p = Spearman's rank correlation coefficient

di = Difference through the two ranks of all observation

n = Number from observations

The Spearman Rank Correlation can pick a value from +1 to -1 where,

- A value from +1 means a perfect association from rank
- A value from 0 means that there is no association through ranks
- A value from -1 means a perfect negative association from rank

Levenshtein Distance (LD)

The Levenshtein distance between two strings a, b of length $|a|$ and $|b|$ respectively, is given by

$lev(a, b)$ where

$leva,b\ (i,j) = \{ \max(i,j)$

$if\ \min(i,j) = 0,\ min\ \{ leva,b\ (i - 1,j) + 1$

$leva,b\ (i,j - 1) + 1$

$leva,b\ (i - 1,j - 1) + 1(ai \neq bj)\ otherwise$ and i is the terminal character position of string a and j is the terminal character position of string b. The weights assigned to all error types is 1 [11].

Stander Algorithm

A, B two sequences and length A=n,length B= m ,to finde the similarity between to sequence ,see the stepe:

L= result test length A, B

$$St(A,B)= \begin{cases} A(i) = B(j) & , sum = +1 \\ Else & , sum = sum \end{cases}$$

Sim_{AB}= (St/L) *100

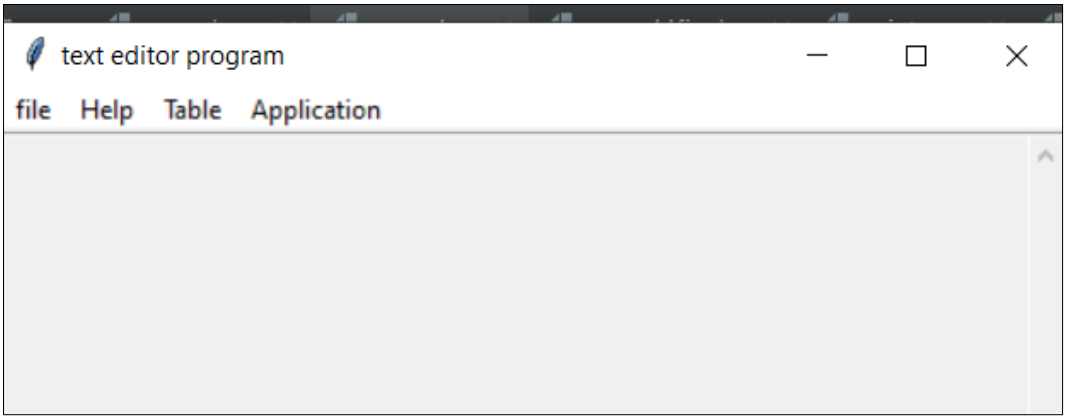
Converting nitrogenous bases into amino acids
An algorithm was created to convert the nitrogenous bases, DNA and RNA, into amino acids according to the table below (1), where every three nitrogenous bases were taken and converted into an amino acid and displayed on the screen with its location shown.

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	Phe (F)	TCT	Ser (S)	TAT	Tyr (Y)	TGT	Cys (C)	T
	TTC		TCC		TAC		TGC		C
	TTA		TCA		TAA	STOP	TGA	STOP	A
	TTG		TCG		TAG		TGG	Trp (W)	G
C	CTT	Leu (L)	CCT	Pro (P)	CAT	His (H)	CGT	Arg (R)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln (Q)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile (I)	ACT	Thr (T)	AAT	Asn (N)	AGT	Ser (S)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys (K)	AGA	Arg (R)	A
	ATG	Met (M)	ACG		AAG		AGG		G
G	GTT	Val (V)	GCT	Ala (A)	GAT	Asp (D)	GGT	Gly (G)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu (E)	GGA		A
	GTG		GCG		GAG		GGG		G

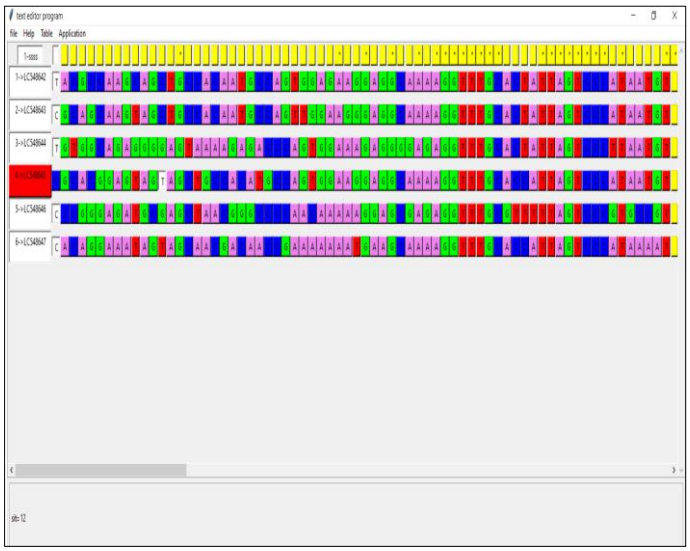
(a) Group

For RNA, the same table (1) is used, but the presence of the nitrogen base (U) uracil instead of the nitrogen base (T) thymine.

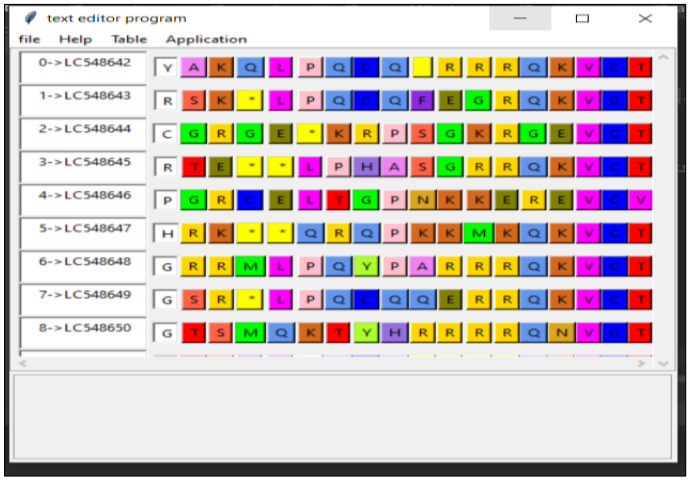
Results
Data set
Data are collected from the genebank as well as samples from the department of Bioinformatics, College of Agriculture, the University of Basra in Iraq [13]
It was stored on the computer, as well as creating a database that includes (200) sequences
This sequence is about diseases (HAMP, HBB, Thalassemia, Type1(AL022723), Type2(BC018404))
Whereas, in this study, DNA sequences of any type are entered (Fasta, Text, word), not just FASTA
Down display the program:



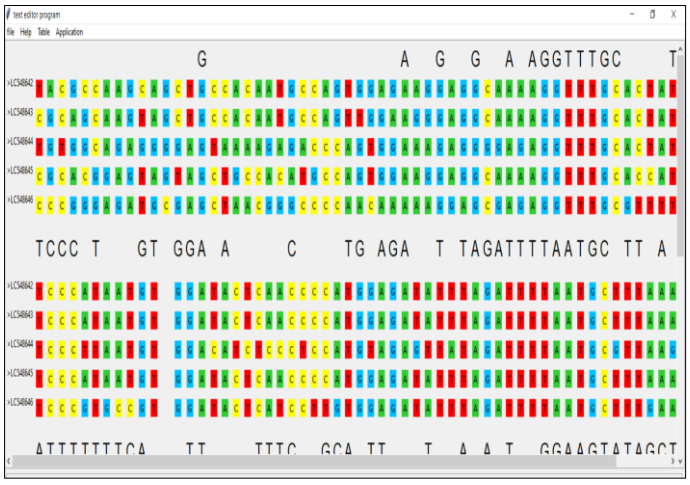
alignment from numbers of DNA sequences
an algorithm was created to find out the alignment between the DNA sequences, and it can display the location of the different DNA, which can indicate the occurrence of a mutation, as well as the length of the DNA sequence, which can be known, by alignment, mutations can be found



Analyses to Amino Acid



Display



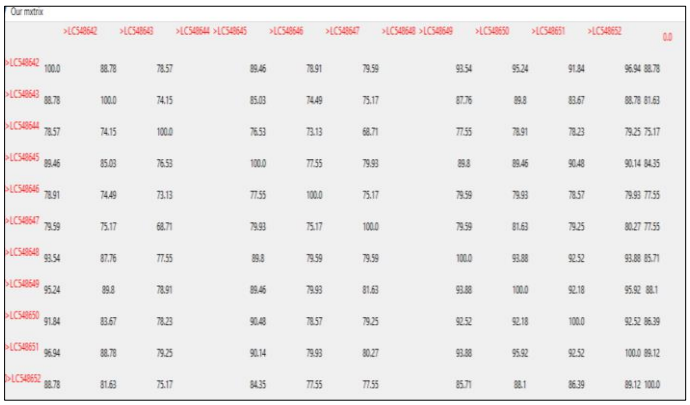
Similarity matrix

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

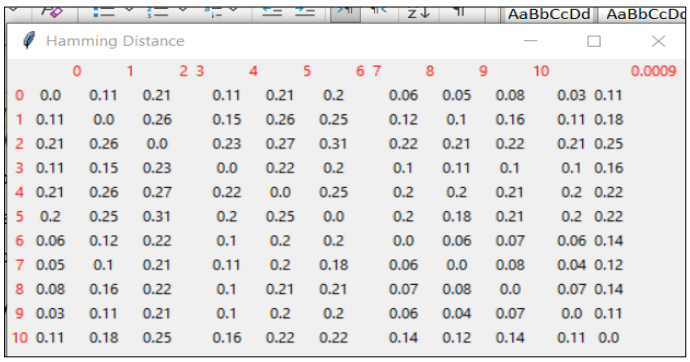
In this research, a method was proposed to find the similarity matrix and it was applied to a number of data**,** the time was calculated and then it was compared with (hamming, Levenshtein, and rank) algorithms and the time calculation for each algorithm was found that the proposed algorithm is the fastest in time

Three algorithms (hamming, Levenshtein, and rank) were compared in addition to the proposed algorithm and were applied to a different number of data and the execution time was calculated. It was noted that the proposed algorithm is

the fastest in execution. it runs on computer11th Gen Intel(R) 2.80GHz, windows 10 Core (TM) i7-1165G7 @

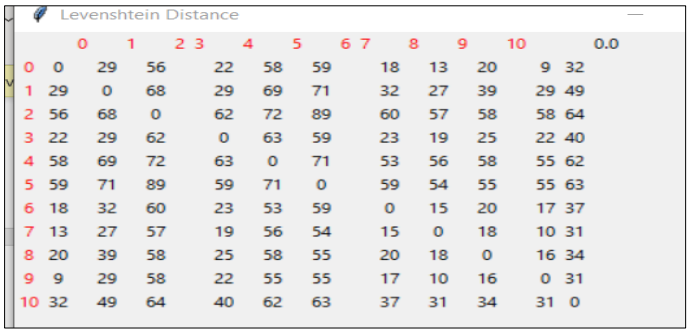


Using Algorithm proposed (called Standard Algorithm)



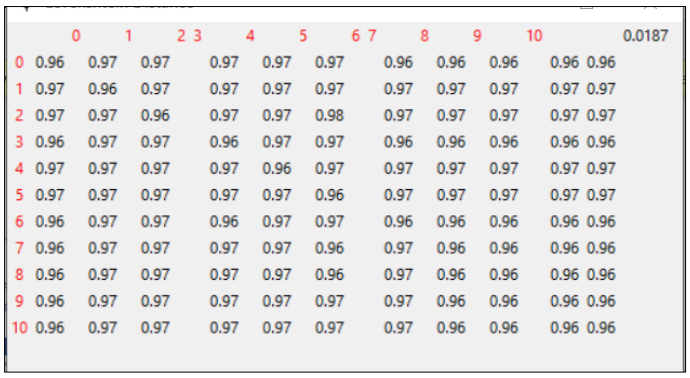
	0	1	2	3	4	5	6	7	8	9	10	0.0009
0	0.0	0.11	0.21	0.11	0.21	0.2	0.06	0.05	0.08	0.03	0.11	
1	0.11	0.0	0.26	0.15	0.26	0.25	0.12	0.1	0.16	0.11	0.18	
2	0.21	0.26	0.0	0.23	0.27	0.31	0.22	0.21	0.22	0.21	0.25	
3	0.11	0.15	0.23	0.0	0.22	0.2	0.1	0.11	0.1	0.1	0.16	
4	0.21	0.26	0.27	0.22	0.0	0.25	0.2	0.2	0.21	0.2	0.22	
5	0.2	0.25	0.31	0.2	0.25	0.0	0.2	0.18	0.21	0.2	0.22	
6	0.06	0.12	0.22	0.1	0.2	0.2	0.0	0.06	0.07	0.06	0.14	
7	0.05	0.1	0.21	0.11	0.2	0.18	0.06	0.0	0.08	0.04	0.12	
8	0.08	0.16	0.22	0.1	0.21	0.21	0.07	0.08	0.0	0.07	0.14	
9	0.03	0.11	0.21	0.1	0.2	0.2	0.06	0.04	0.07	0.0	0.11	
10	0.11	0.18	0.25	0.16	0.22	0.22	0.14	0.12	0.14	0.11	0.0	

Using Hamming Algorithm

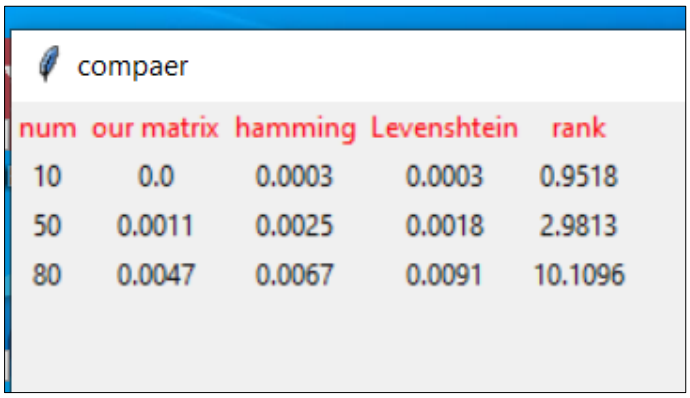


	0	1	2	3	4	5	6	7	8	9	10	0.0
0	0	29	56	22	58	59	18	13	20	9	32	
1	29	0	68	29	69	71	32	27	39	29	49	
2	56	68	0	62	72	89	60	57	58	58	64	
3	22	29	62	0	63	59	23	19	25	22	40	
4	58	69	72	63	0	71	53	56	58	55	62	
5	59	71	89	59	71	0	59	54	55	55	63	
6	18	32	60	23	53	59	0	15	20	17	37	
7	13	27	57	19	56	54	15	0	18	10	31	
8	20	39	58	25	58	55	20	18	0	16	34	
9	9	29	58	22	55	55	17	10	16	0	31	
10	32	49	64	40	62	63	37	31	34	31	0	

Using Levenshtein Algorithm



	0	1	2	3	4	5	6	7	8	9	10	0.0187
0	0.96	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.96	
1	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	
2	0.97	0.97	0.96	0.97	0.97	0.98	0.97	0.97	0.97	0.97	0.97	
3	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.96	0.96	0.96	0.96	
4	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97	
5	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	
6	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.96	0.96	0.96	0.96	
7	0.96	0.97	0.97	0.97	0.97	0.96	0.97	0.96	0.96	0.96	0.96	
8	0.96	0.97	0.97	0.97	0.97	0.96	0.97	0.96	0.96	0.96	0.96	
9	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	
10	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	



Using Spearman rank correlation distance

Conclusions

We benefit from this research in various fields of life, and that the alignment method is necessary to know the characteristics of DNA, for example in viruses, to know the areas of strength and weakness, and that the similarity matrix resulting from the proposed algorithm is necessary as it was used in clustering algorithms, and the results proved that the proposed methods are the best in terms of The accuracy of the results

This application is for bioinformatics and is easy to use by anyone other than (the system designer), as the help part has an explanation in English and Arabic

References

1. Al-Khafaji ZM, Ibrahim AA. Bioinformatics. Baghdad: University of Baghdad; 2012.
2. Ohied BM, Al-Badran AI. Mitochondrial DNA (hypervariable region I) diversity in Basrah population – Iraq. Genomics. 2020;112(5):3560-4. doi: 10.1016/j.ygeno.2020.04.004.
3. Singh P, Singh N. Role of data mining techniques in bioinformatics. Int J Appl Res Bioinformatics. 2021;11(1):51-60. doi: 10.4018/ijarb.2021010106.
4. Bagga S, Singh GN. Applications of data mining.
5. Alsaadi EAJ, Neuman BW, Jones IM. A fusion peptide in the spike protein of MERS coronavirus. Viruses. 2019;11(9):825. doi: 10.3390/v11090825.
6. Bera M. Artificial intelligence in bioinformatics [Internet]. 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. Preprint. 2020. doi: 10.1101/2020.09.04.283358.
9. Nawaz MS, Fournier-Viger P, Shojaee A, Fujita H. Using artificial intelligence techniques for COVID-19 genome analysis. Appl Intell. 2021;51(5):3086-103. doi: 10.1007/s10489-021-02193-w.
10. Abd Elwahaab MA, Abo-Elkhier MM, Abo El Maaty MI. A statistical similarity/dissimilarity analysis of protein sequences based on a novel group representative vector. Biomed Res Int. 2019;2019:8702968. doi: 10.1155/2019/8702968.
11. Logan RL IV. Optimized Levenshtein distance for clustering third-generation sequencing data.
12. Ali K, Al-Hameed A. Spearman's correlation coefficient in statistical analysis. Int J Nonlinear Anal Appl. 2022;13:2008-22. doi: 10.22075/ijnaa.2022.6079.
13. 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.

How to Cite This Article

Alwan EA, Mahmood SA, Habib HN. An Applied Program for DNA sequences alignment. Int J Artif Intell Eng Transform. 2025;6(2):164–169. doi:10.54660/IJAIET.2025.6.2.164-169.

Creative Commons (CC) License