**Int. J. Nonlinear Anal. Appl. 12 (2021) No. 2, 1531-1539**
**ISSN: 2008-6822 (electronic)**
http://dx.doi.org/10.22075/ijnaa.2021.5281

# Using alignment-free methods as preprocessing stage to classification whole genomes

Najah Abed Alhadi Shanan[a,*], Hussein Attya Lafta[a], Sura Z. Al_Rashid[b]

[a] *Computer Department, Science College for Women, University of Babylon, Babylon, Iraq*
[b] *College of Information Technology, University of Babylon, Babylon, Iraq*

*(Communicated by Madjid Eshaghi Gordji)*

## Abstract

In bioinformatics systems, the study of genetics is a popular research discipline. These systems depend on the amount of similarity between the biological data. These data are based on DNA sequences or raw sequencing reads. In the preprocessing stage, there are several methods for measuring similarity between sequences. The most popular of these methods is the alignment method and alignment-free method, which are applied to determine the amount of functional matching between sequences of nucleotides DNA, ribosome RNA, or proteins. Alignment-based methods pose a great challenge in terms of computational complexity, In addition to delaying the time to search for a match, especially if the data is heterogeneous and its size is huge, and thus the classification accuracy decreases in the post-processing stage. Alignment-free methods have overcome the challenges of alignment-based methods for measuring the distance between sequences, The size of the data used is 1000 genomes uploaded from National Center for Biotechnology Information (NCBI), after eliminating the missing and irrelevant values, it becomes 860 genomes, ready to be segmented into words by the k-mer analysis, after which the frequency of each word is counted for each query. The size of a word depends on a value of k. In this paper we used a value of k =3 . . . .8, for each iteration will count times of frequencies words.

*Keywords:* 16S RNA, DNA, k-mers

[*] Corresponding author

*Email addresses:* najahhadi78@gmail.com (Najah Abed Alhadi Shanan ), hzazmk@yahoo.com (Hussein Attya Lafta), sura_os@itnet.uobabylon.edu.iq (Sura Z. Al_Rashid)

*Received:* March 2021 *Accepted:* May 2021

## 1. Introduction

The huge volume of biological data, which began to increase in the last twenty years, necessitated the development of bioinformatics systems that analyze this type of data such as DNA, RNA or proteins. The main objective of data analysis is to find the amount of similarity between the sequences. The similarity between the sequences is expressed by measuring the distance between two sequences or a group of sequences of the same type [3]. The amount of the distance between the sequences is one of the main measures used to distinguish between types through their characteristics[9]. This stage is considered as a pre-processing for solving many research problems as handling missing value and relegation irrelevant and noise gene [15] diagnosing several diseases such as cancer[12], detecting viruses[25], classification taxonomy bacteria[1], detecting sites that contain functions in protein regions or prediction It is carried out through known sites[13], regression diseases through mutations at genes of mouse [4]and other medical and biological research. The methods for measuring the similarity of sequences is based on alignment and alignment-free methods The best method is an alignment-free method, k-mer analysis is a common method of these methods. The results of k-mer analysis can be adapted as inputs to several algorithms, whether they are clustered such as Latent Dirichlet Allocation method [14], Convolutional Neural Network [8], Naïve Bayes classifier[6], support vector machine[21].

The paper is arranged in the following sections: The second section presents basic concepts of biological data and clarification of alignment and alignment-free methods. The third section describes the research methodology and data set used in the research. The fourth section presents the results of the experiments. Finally, the fifth section concludes the conclusion and future works. This work represents the preprocessing of genome sequencing, which is the first section of a second paper, in which preprocessing and post-processing is applied, where bacteria are classified based on their strains and families by adapting the k-mer method to the Latent Dirichlet Allocation method, which is one of the topic modelling methods that deals with text mining, The top ten words with the highest frequency are selected to be ready for the post-processing stage. on the other hand, the Naïve Bayes algorithm is applied to classify each class of bacteria (class, order, family, genus) and then the model's performance is evaluated by calculating the degree of accuracy for each class[1].

## 2. Methods

### *2.1. Background of dataset*

The gene is the primary component of physical and functional heredity. Genes are the building blocks of DNA. The DNA data were saved as a code consisting of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Base pairs are formed when two DNA bases, A with T and C with G. and vice versa are paired together. Each base has sugar and phosphate molecules attached to it. The term ”nucleotides” refers to the basic building blocks of DNA. Figure 1 show that

The shape of the double helix is similar to a ladder, as it consists of a pair of nucleotides as in Figure 1 above, and this connection is represented as the rungs of the ladder. The sugar and phosphate molecules surround the nucleotide parts from the outside. The double helix sequence of the base can be replication through each strand of nucleic acids in the double helix. The protein production process takes place through the single strand of RNA, which is part of the double helix of DNA. The tRNA strand is responsible for the transcription of the nucleic acids to the ribosome, after which it binds the ribosome RNA (rRNA) to make the protein [16]. Figure 2 shows the process of replication, transcription and protein production.

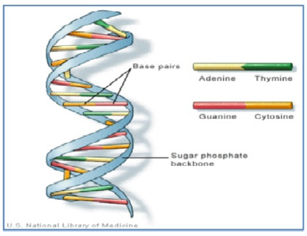


Figure 1: structure of DNA

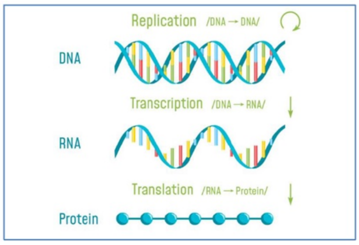


Figure 2: the mechanism of DNA replication

There are several differences between DNA and RNA in terms of form, functions and components. As we mentioned above, the RNA strand is single and is part of the double strand of DNA.DNA provides the code for the cell 's activities, while RNA converts that code into proteins to carry out cellular functions.RNA is the nucleic acid that makes proteins from the code provided by DNA through the processes of transcription and translation. It contains oxygen, unlike DNA, which is hypoxic, where one of its components is Uracil (U) and not thymine(T) as in DNA, and the remaining components A, G, C are the same in both. The function of RNA is regulated to protein production in the cell and the transport of hereditary material [2].

### *2.2. Sequence alignment methods*

Alignment methods are the most common algorithms in the field of bioinformatics systems, their functions are determined by measuring the amount of functional similarity between the sequences, whether the sequences are DNA, RNA or protein sites, this similarity includes function, structure or evolutionary relations between the sequences [19]. Alignment methods are divided into two classes: global alignment and local alignment. Global alignment is the most improved and developed method that works along the full length of all sequences. In the case of local alignment, it is preferred over global alignment through its function in identifying areas of similarity within long sequences that are usually different in a wide range. The limitation in local alignment is the computational complexity in the stages of determining regions of similarity between sequences.[18]. In general, sequence alignment methods are slow in computations, but they were developed through the introduction of dynamic programming or probabilistic methods that search a large database. However, these updates to alignment methods do not give very similar results.

### *2.3. Alignment-free sequence*

Because of the large volume of biological data, which increases annually, in addition to their heterogeneity, alignment methods are unable to analyze this type of data. Therefore, it is necessary to develop new analysis systems such as alignment-free methods to reduce the high dimensionality and the short implementation time[23]. The K-mer analysis is one of the most famous alignment-free methods that deal with sub-sequences or words extracted from sequences[20].

### *2.3.1. k-mer /word method*

k-mers are enriched with pieces of sequences (called words) that may be DNA and RNA sequences. The K-mer functions as a source for DNA sequence assembly [22] improving heterogeneous gene expression [24] [11] diagnosis of diseases and viruses and production of metagenomics vaccines [17, 7]. In general, the K-mer has been applied in genome computation and sequence analysis where nucleotides (eg, A, T, G, C) are composed of k-mer. Here, the ordering methodology is used to convert the sequences into words and the latter is stored in a bag of words Within each sample. In Figure 6, for example, k = 8, all overlapping k-mers, represented by words, can be extracted from the gene sequence, by sliding fixed windows.

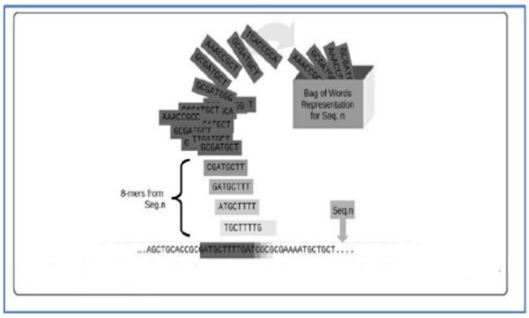
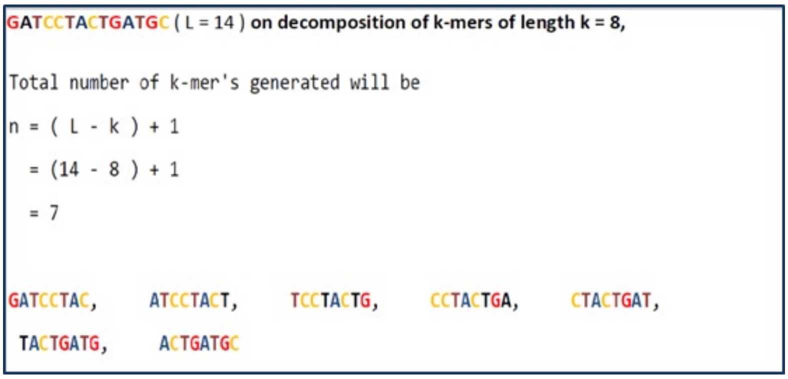


Figure 3: k-mer method

k-mer counted by the equation

$$n = (l - k) + 1 \tag{2.1}$$

where n is times of k-mer, L is the length of the sequence, k is the width of the word.
for example to show processing of sequences by k-mer analysis.



**GATCCTACTGATGC ( L = 14 ) on decomposition of k-mers of length k = 8,**

```
Total number of k-mer's generated will be
n = ( L - k ) + 1
  = (14 - 8 ) + 1
  = 7

GATCCTAC,     ATCCTACT,     TCCTACTG,     CCTACTGA,     CTACTGAT,
 TACTGATG,     ACTGATGC
```

## 3. Methodology

This methodology includes the pre-processing stage and it has three phases: the first phase is to upload the raw data from the NCBI database by the amount of 1000 genome, the second phase deals with missing values and gets rid of irrelevant values and noise data to getting 860 genome having only nucleotides (A, G, C, T) , and the third phase is to fragment the sequences into sub-sequences called (words)by alignment-free methods with k-mer/wo

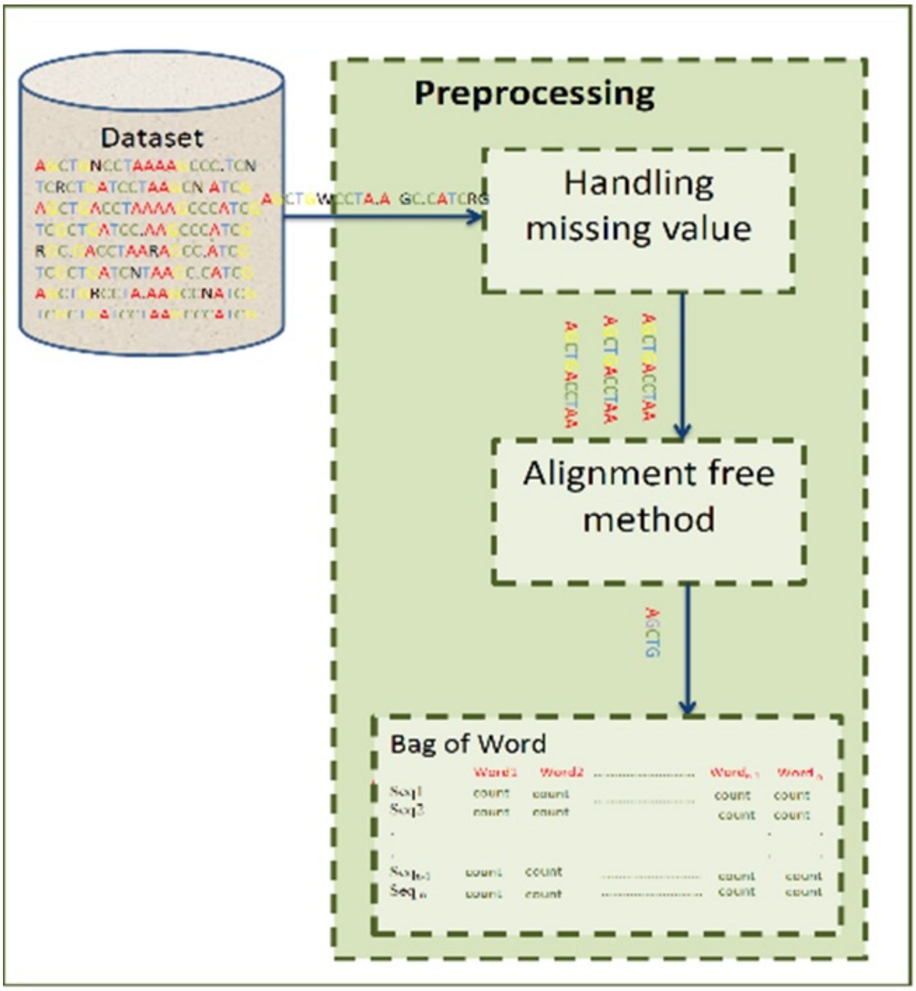


Figure 4: The preprocessing methodology

### 3.1. *Preprocessing stage*

#### 3.1.1. *Dataset*

The dataset used in the methodology is the 16 ribosome RNA with the size of 1000 sequences, which was uploaded from the NCBI Bio Encyclopedia [10, 5].

#### 3.1.2. *Handling missing value*

The missing values and noise data for 1000 sequences are dealt with by an algorithm that is designed to search in each sequence, the search condition is the presence of only the components of DNA (A, G, C, T). The sequencing was rejected. Thus, only 860 sequences containing only the four nucleic acid components were obtained. Algorithm (1) below shows the work steps.

**Algorithm (1) The steps of Handling Missing Values**

**Input** : two dimensional array of nucleotides (AGCT), $a_{(ij)}$, where i is number of sequences and j is number of nucleotides (AGCT) $DNA_i$ is array of DataSet (AGCT) for one dimensional.
**Output** : one dimensional $B_i$ new dataset after handling missing value.

```
Begin
    For $i$ in $N$, where $N$ is no. of sequences
        Flag = True
        For $j$ in $seq_i$ $_{fromDNAi}$
            If $a_{ij}$ NOT in [A, G, C, T]
                Flag = FALSE
        End for
        If flag = TRUE
            B. append = $seq_i$
    End for
End
```

### *3.1.3. k-mer analysis*

The function of k-mer analysis is the fragmentation of the sequences which have 860 sequences after handling missing value, into words by the style of overlap (the sliding window) and the shifting by one in each iteration. The word width based on the value of k.

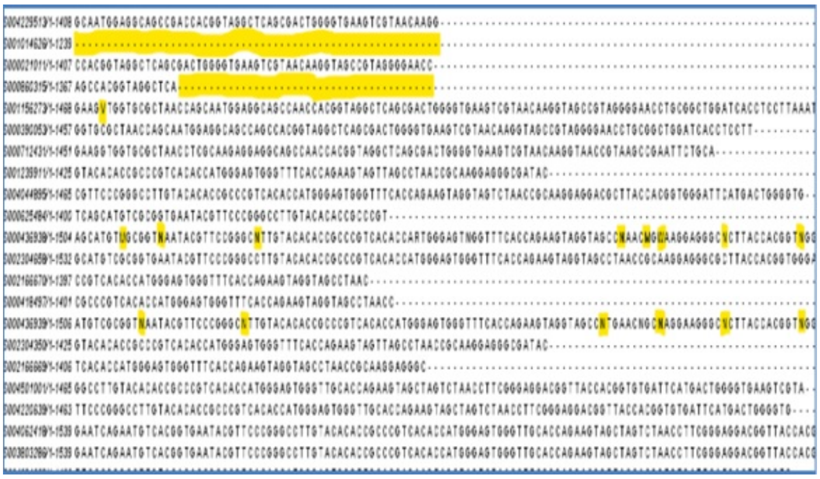
**Algorithm (2) k-mer Analysis**

```
Input: L,K,n: where L is length of sequence and k is size of word and n is number of sequences. $DS(B_1)$ out of algorithm (1).
Output: array1 [m,n] where m is index of no. of size of dictionary and n is index of no. of sequences.
k_mers (seq.: string, k: integer)
L= length (seq.)
array1 = new array of $L - k + 1$ empty strings.
// Repeat for the number of k_mers in the sequence.
// Save $n^{th}$ k_mer in the array output.
for n=0 to $L - k + 1$ do.
    array1 [n] = words of seq. from letter n to letter $n + k - 1$.
return array1.
```

After segmentation sequences to words, the second stage counts each word with query (sequence) for each cycle of k-mer and then store them on the Bag Of Word.

## 4. Results

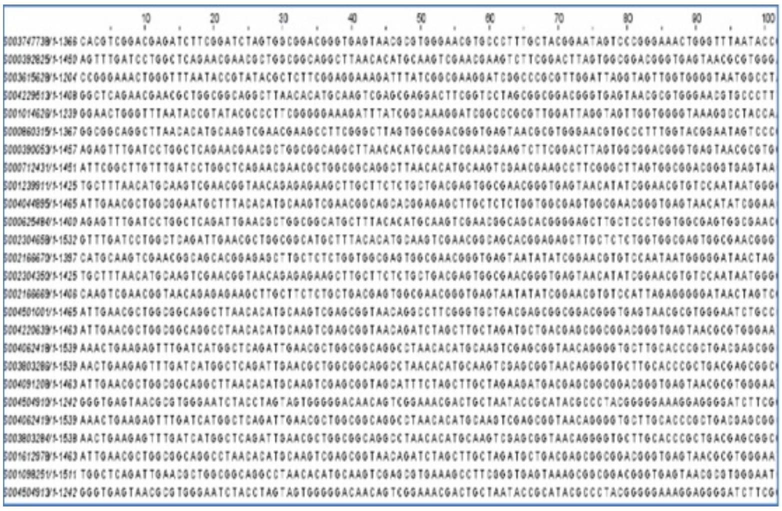
### *4.1. Row Data Set Before Handling Missing Value*



```
S004229513/1-1408 GCAATGGAGGCAGCCGACCACGGTAGGCTCAGCGACTGGGGTGAAGTCGTAACAAGG............................................
S001014629/1-1239 .......................................................................................................
S000021011/1-1407 CCACGGTAGGCTCAGCGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACC...............................................
S000860315/1-1367 AGCCACGGTAGGCTCA.......................................................................................
S001156273/1-1468 GAAGVTGGTGCGCTAACCAGCAATGGAGGCAGCCAACCACGGTAGGCTCAGCGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTGCGGCTGGATCACCTCCTTAAAT
S000390053/1-1457 GGTGCGCTAACCAGCAATGGAGGCAGCCAGCCACGGTAGGCTCAGCGACTGGGGTGAAGTCGTAACAAGGTAGCCGTAGGGGAACCTGCGGCTGGATCACCTCCTT.........
S000712431/1-1451 GAAGGTGGTGCGCTAACCTCGCAAGAGGAGGCAGCCAACCACGGTAGGCTCAGCGACTGGGGTGAAGTCGTAACAAGGTAACCGTAAGCCGAATTCTGCA.............
S001239911/1-1425 GTACACACCGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGTTAGCCTAACCGCAAGGAGGGCGATAC......................................
S004044895/1-1465 CGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGGTAGTCTAACCGCAAGGAGGACGCTTACCACGGTGGGATTCATGACTGGGGTG..
S000625484/1-1400 TCAGCATGTCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGT.................................................................
S004369389/1-1504 AGCATGTUGCGGTNAATACGTTCCGGGCNTTGTACACACCGCCCGTCACACCARTGGGAGTNGGTTTCACCAGAAGTAGGTAGCCNAACMGDAAGGAGGGCYCTTACCACGGTNGG
S002304659/1-1532 GCATGTCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGGTAGCCTAACCGCAAGGAGGGCGCTTACCACGGTGGGA
S002166670/1-1397 CCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGGTAGCCTAAC....................................................................
S004184937/1-1401 CGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGGTAGCCTAACC................................................................
S004369393/1-1506 ATGTCGCGGTNAATACGTTCCCGGGCNTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGGTAGCCNTGAACNGCNAGGAAGGGCYCTTACCACGGTNGG
S002304350/1-1425 GTACACACCGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTAGTTAGCCTAACCGCAAGGAGGGCGATAC......................................
S002166669/1-1406 TCACACCATGGGAGTGGGTTTCACCAGAAGTAGGTAGCCTAACCGCAAGGAGGGC..........................................................
S004501001/1-1465 GGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTGCACCAGAAGTAGCTAGTCTAACCTTCGGGAGGACGGTTACCACGGTGTGATTCATGACTGGGGTGAAGTCGTA--
S004220639/1-1463 TTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTGCACCAGAAGTAGCTAGTCTAACCTTCGGGAGGACGGTTACCACGGTGTGATTCATGACTGGGGTG----
S004062418/1-1539 GAATCAGAATGTCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTGCACCAGAAGTAGCTAGTCTAACCTTCGGGAGGACGGTTACCACG
S003803286/1-1539 GAATCAGAATGTCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTGCACCAGAAGTAGCTAGTCTAACCTTCGGGAGGACGGTTACCACG
```

Figure 5: screen shoot of 1000 sequence row dataset

### *4.2. Data Set After Handling Missing Value*

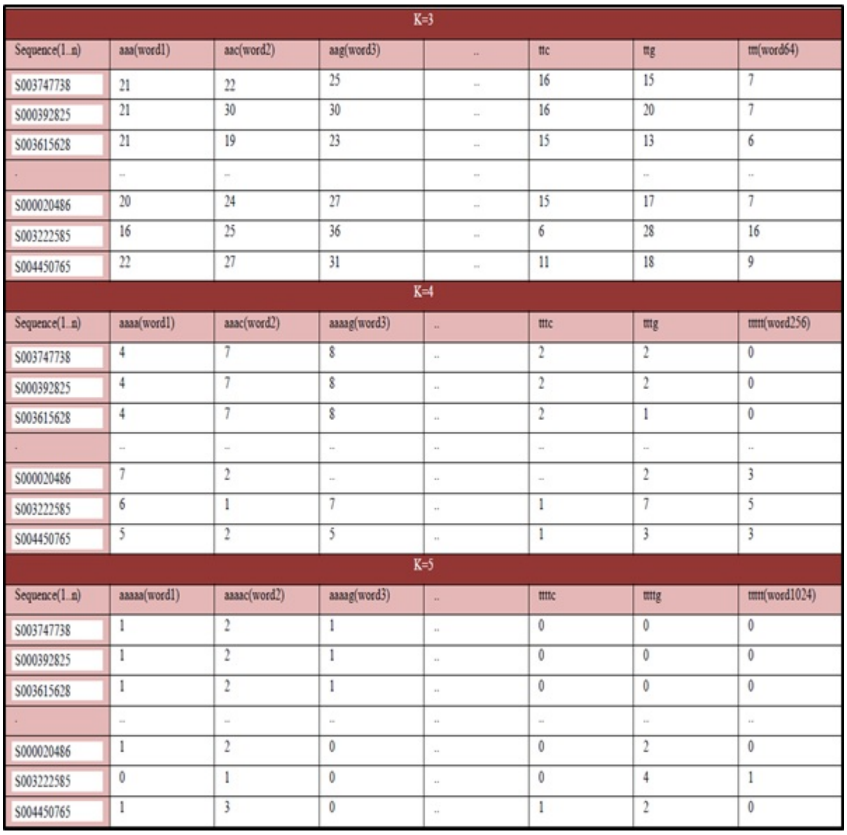


```
                           10        20        30        40        50        60        70        80        90       100
S003747739/1-1366 CACGTCGGACGAGATCTTCGGATCTAGTGGCGGACGGGTGAGTAACGCGTGGGAACGTGCCCTTTGCTACGGAATAGTCCCGGGAAACTGGGTTTAATACC
S000392825/1-1450 AGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAACGAAGTCTTCGGACTTAGTGGCGGACGGGTGAGTAACGCGTGGG
S003615629/1-1204 CCGGGAAACTGGGTTTAATACCGTATACGCTCTTCGGAGGAAAGATTTATCGGCGAAGGATCGGCCCGCGTTGGATTAGGTAGTTGGTGGGGTAATGGCCT
S004229513/1-1408 GGCTCAGAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGAGGACTTCGGTCCTAGCGGCGGACGGGTGAGTAACGCGTGGGAACGTGCCCTT
S001014629/1-1239 GGAACTGGGTTTAATACCGTATACGCCCTTCGGGGGAAAGATTTATCGGCAAAGGATCGGCCCGCGTTGGATTAGGTAGTTGGTGGGGTAAAGGCCTACCA
S000860315/1-1367 GGCGGCAGGCTTAACACATGCAAGTCGAACGAAGCCTTCGGGCTTAGTGGCGGACGGGTGAGTAACGCGTGGGAACGTGCCCTTTGGTACGGAATAGTCCC
S000390053/1-1457 AGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAACGAAGTCTTCGGACTTAGTGGCGGACGGGTGAGTAACGCGTG
S000712431/1-1451 ATTCGGCTTGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAACGAAGCCTTCGGGCTTAGTGGCGGACGGGTGAGTAA
S001239911/1-1425 TGCTTTAACATGCAAGTCGAACGGTAACAGAGAGAAGCTTGCTTCTCTGCTGACGAGTGGCGAACGGGTGAGTAACATATCGGAACGTGTCCAATAATGGG
S004044895/1-1465 ATTGAACGCTGGCGGAATGCTTTACACATGCAAGTCGAACGGCAGCACGGAGAGCTTGCTCTCTGGTGGCGAGTGGCGAACGGGTGAGTAACATATCGGAA
S000625484/1-1400 AGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCTTTACACATGCAAGTCGAACGGCAGCACGGGGAGCTTGCTCCCTGGTGGCGAGTGGCGAAC
S002304659/1-1532 GTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCTTTACACATGCAAGTCGAACGGCAGCACGGAGAGCTTGCTCTCTGGTGGCGAGTGGCGAACGGG
S002166670/1-1397 CATGCAAGTCGAACGGCAGCACGGAGAGCTTGCTCTCTGGTGGCGAGTGGCGAACGGGTGAGTAATATATCGGAACGTGTCCAATAATGGGGATAACTAG
S002304350/1-1425 TGCTTTAACATGCAAGTCGAACGGTAACAGAGAGAAGCTTGCTTCTCTGCTGACGAGTGGCGAACGGGTGAGTAACATATCGGAACGTGTCCAATAATGGG
S002166669/1-1406 CAAGTCGAACGGTAACAGAGAGAAGCTTGCTTCTCTGCTGACGAGTGGCGAACGGGTGAGTAATATATCGGAACGTGTCCATTAGAGGGGGATAACTAGTC
S004501001/1-1465 ATTGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGGTAACAGGCCTTCGGGTGCTGACGAGCGGCGGACGGGTGAGTAACGCGTGGGAATCTGCC
S004220639/1-1463 ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGATCTAGCTTGCTAGATGCTGACGAGCGGCGGACGGGTGAGTAACGCGTGGGAA
S004062418/1-1539 AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGGGGTGCTTGCACCCGCTGACGAGCGG
S003803286/1-1539 AACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGGGGTGCTTGCACCCGCTGACGAGCGGC
S004091208/1-1463 ATTGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGGTAGCATTTCTAGCTTGCTAGAAGATGACGAGCGGCGGACGGGTGAGTAACGCGTGGGAA
S004504910/1-1242 GGGTGAGTAACGCGTGGGAATCTACCTAGTAGTGGGGGACAACAGTCGGAAACGACTGCTAATACCGCATACGCCCTACGGGGGAAAGGAGGGATCTTCG
S004062419/1-1539 AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGGGGTGCTTGCACCCGCTGACGAGCGG
S003803284/1-1538 AACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGGGGTGCTTGCACCCGCTGACGAGCGGC
S001612978/1-1463 ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGATCTAGCTTGCTAGATGCTGACGAGCGGCGGACGGGTGAGTAACGCGTGGGAA
S001098251/1-1511 TGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGTGAAAGCCTTCGGGTGAGTAAAGCGGCGGACGGGTGAGTAACGCGTGGGAAT
S004504913/1-1242 GGGTGAGTAACGCGTGGGAATCTACCTAGTAGTGGGGGACAACAGTCGGAAACGACTGCTAATACCGCATACGCCCTACGGGGGAAAGGAGGGATCTTCG
```

Figure 6: screen shoot of 860 sequence dataset after filtering having only (A,G,C,T)

### *4.3. k–mer*

Table 1: Word counting in each query is calculated based on the width of k in each cycle



| K=3 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Sequence(1..n) | aaa(word1) | aac(word2) | aag(word3) | .. | ttc | ttg | ttt(word64) |
| S003747738 | 21 | 22 | 25 | .. | 16 | 15 | 7 |
| S000392825 | 21 | 30 | 30 | .. | 16 | 20 | 7 |
| S003615628 | 21 | 19 | 23 | .. | 15 | 13 | 6 |
| . | .. | .. | | .. | | .. | .. |
| S000020486 | 20 | 24 | 27 | .. | 15 | 17 | 7 |
| S003222585 | 16 | 25 | 36 | .. | 6 | 28 | 16 |
| S004450765 | 22 | 27 | 31 | .. | 11 | 18 | 9 |
| K=4 | | | | | | | |
| Sequence(1..n) | aaaa(word1) | aaac(word2) | aaaag(word3) | .. | tttc | tttg | tttt(word256) |
| S003747738 | 4 | 7 | 8 | .. | 2 | 2 | 0 |
| S000392825 | 4 | 7 | 8 | .. | 2 | 2 | 0 |
| S003615628 | 4 | 7 | 8 | .. | 2 | 1 | 0 |
| . | .. | .. | .. | .. | .. | .. | .. |
| S000020486 | 7 | 2 | .. | .. | .. | 2 | 3 |
| S003222585 | 6 | 1 | 7 | .. | 1 | 7 | 5 |
| S004450765 | 5 | 2 | 5 | .. | 1 | 3 | 3 |
| K=5 | | | | | | | |
| Sequence(1..n) | aaaaa(word1) | aaaac(word2) | aaaag(word3) | .. | ttttc | ttttg | ttttt(word1024) |
| S003747738 | 1 | 2 | 1 | .. | 0 | 0 | 0 |
| S000392825 | 1 | 2 | 1 | .. | 0 | 0 | 0 |
| S003615628 | 1 | 2 | 1 | .. | 0 | 0 | 0 |
| . | .. | .. | .. | .. | .. | .. | .. |
| S000020486 | 1 | 2 | 0 | .. | 0 | 2 | 0 |
| S003222585 | 0 | 1 | 0 | .. | 0 | 4 | 1 |
| S004450765 | 1 | 3 | 0 | .. | 1 | 2 | 0 |

## 5. Conclusion and future works

Our methodology is to find similarity in genome sequences, where it is possible to compare sequences of different lengths to find the amount of matching distance between them or the absence of a match. k-mer gives approximate results for finding distances based on frequency profile or characters pattern. k-mer used counting for the number of times the word is present in the queries. In the results Table 1, the higher the value of k, the lower the similarity between the sequences, in addition to the complexity of the comparison process and the delay in execution time. In future work, these results can be considered as inputs to the clustering algorithms. Or classification, such as a topic modelling or Naïve Bayes classifier . Where words are considered to attribute and counting them are considered values.

## References

[1] N. Abed, A. Shanan, H. A. Lafta and S. Z. Al Rashid, *Bacteria taxonomic classification using machine-learning models*, Solid State Tech. 64 (2021) 1091–1112.

[2] S. Aggarwal, *Using Mutual Information for extracting Biclusters from Gene Expression Data*, New Delhi, 2013.

[3] A.K. Al-Mashanji and S.Z. Al-Rashi, *Computational Methods for Preprocessing and Classifying Gene Expression Data- Survey*, 4th Sci. Int. Conf. Najaf, SICN 2019, March (2020) 121–126.

[4] S.Z. Al-Rashid and N.H. Al-Aaraji, *Bayesian Models with Coregionalization to Model Gene Expression Time Series for Mouse Model for Speed Progression of ALS Disease*, Eur. J. Sci. Res. 1 (2015) 1–20.

[5] J. R. Cole et al., *The Ribosomal database project: Improved alignments and new tools for rRNA analysis*, Nucleic Acids Res. 1 (2009) 141-–145.

[6] M. El Kourdi, A. Bensaid and T. Rachidi, *Automatic Arabic document categorization based on the Naïve Bayes algorithm*, Proc. Workshop on Comput. Approaches to Arabic Script-based Languages, (2004) 51–58.

[7] K. Eschke, J. Trimpert, N. Osterrieder and D. Kunec, *Attenuation of a very virulent Marek's disease herpesvirus (MDV) by codon pair bias deoptimization*, PLoS Pathog. 14 (2018) 1-–24.

[8] A. Fiannaca et al., *Deep learning models for bacteria taxonomic classification of metagenomic data*, BMC Bioinf. 19 (2018) 61–76.

[9] G. Gamage, N. Gimhana, A. Wickramarachchi, V. Mallawaarachchi and I. Perera, *Alignment-free whole genome comparison using k-mer forests*, 19th Int. Conf. Adv. ICT Emerg. Reg. ICTer 2019 - Proc. 2019.

[10] L. Y. Geer, N. Gimhana, A. Wickramarachchi, V. Mallawaarachchi, and I. Perera, *The NCBI BioSystems database*, Nucleic Acids Res. 38 (2009) 492-–496.

[11] C. Gustafsson, S. Govindarajan, J. Minshull, and M. Park, *Codon bias and heterologous protein expression. [Trends Biotechnol. 2004]- PubMed result*, Trends Biotechnol., 2004.

[12] S. J. Kho, M. L. Raymer, H. B. Yalamanchili, and A. P. Sheth, *A novel approach for classifying gene expression data using topic modeling*, ACM-BCB 2017 - Proc. 8th ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Inf. (2017) 388—393.

[13] J. M. Kirk et al., *Functional classification of long non-coding RNAs by k-mer content*, Nat. Genet. 10 (2018) 1474—1482.

[14] M. La Rosa, A. Fiannaca, R. Rizzo and A. Urso, *Probabilistic topic modeling for the analysis and classification of genomic sequences*, BMC Bioinformatics, 6 (2015) 1-–9.

[15] P.A. Mundra and J.C. Rajapakse, *Gene and sample selection using T-score with sample selection*, J. Biomed. Inf. 59 (2016) 31—41.

[16] A. Nair, *Computational biology & bioinformatics: a gentle overview*, Commun. Comput. Soc. India 5 (2007) 1—13.

[17] S.C. Perry and R.G. Beiko, *Distinguishing microbial genome fragments based on their composition: Evolutionary and comparative genomic perspectives*, Genome Biol. Evol. 2 (2010) 117—131.

[18] V.O. Polyanovsky, M.A. Roytberg and V.G. Tumanyan, *Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences*, Algorithms Mol. Biol. 6 (2011) 1—12.

[19] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak and F. Herrera, *A survey on data preprocessing for data stream mining: Current status and future directions*, Neurocom. 239 (2017) 39-–57.

[20] A. Sievers, F. Wenz, M. Hausmann and G. Hildenbrand, *Conservation of k-mer composition and correlation contribution between introns and intergenic regions of animalia genomes*, Genes. (Basel) 9 (2018) 1—19.

[21] K. Simek et al., *Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data*, Eng. Appl. Artif. Intell., 4 (2004) 417—427.

[22] G.Z. Valenci, M. Rubinstein, R. Afriat, Z.D. Shira Rosencwaig, E. Rorman and I. Nissan, *Draft Genome Sequences of Cronobacter muytjensii Cr150 , Cronobacter turicensis Cr170, and Cronobacter sakazakii Cr611 Gal*, Microbiology Resource Announ. 9(44) (2020) 9—11.

[23] S. Vinga and J. Almeida, *Alignment-free sequence comparison - A review*, Bioinf. 4 (2003) 513-–523.

[24] M. Welch et al., *Design parameters to control synthetic gene expression in Eschorichia coli*, PLoS One, 9 (2009).

[25] R. Yin, Z. Luo and C. K. Kwoh, *Alignment-free machine learning approaches for the lethality prediction of potential novel human-adapted coronavirus using genomic nucleotide*, bioRxiv, (2020) 1–18.