# Bacteria taxonomic classification using Machine learning models

**Article** *in* Solid State Technology · January 2021

**2 authors:**

Sura Alrashid
Babylon University/ information technology college
**23** PUBLICATIONS **23** CITATIONS
SEE PROFILE

Hussein Lafta
University of Babylon\College of Science for Women
**39** PUBLICATIONS **57** CITATIONS
SEE PROFILE

**Solid State Technology**
**Volume: 64 Issue: 1**
**Publication Year: 2021**

# Bacteria taxonomic classification using Machine-learning models

Najah Abed Alhadi Shanan[1], Hussein Attya Lafta[2], Sura Z. Al_Rashid[3]

[1]College for Women, University of Babylon, Babylon, Iraq. E-mail: najahhadi78@gmail.com
[2]Hussein Attya Lafta works at Computer Department, Science College for Women, University of Babylon, Babylon, Iraq. Email: hzazmk@yahoo.com
[3]Sura Z. Al_Rashid works at College of Information Technology, University of Babylon, Babylon, Iraq.
*Email:sura_os@itnet.uobabylon.edu.iq*

***Abstract*—**Classification of taxonomic for genomic sequences is commonly depended on evolutionary distance acquired by alignment methods. alignment-free method introduced based on probabilistic topic modeling out of clustering or Naïve Bayes algorithm through classification. Using a k-mer (fractions of length k) fragmentations of DNA sequences and the Latent Dirichlet Allocation algorithm(LDA), a clusters are built for 16S RNA bacterial sequences for different number of topics, or adopt classifiers through Naïve Bayes based on k-mer fragments .The classification model is evaluated during the cross-validation procedure, taking into account the bacterial data set of 1000 sequences belonging to the majority numeric bacteria phyla : class, order, family and genus. To test the efficiency of the proposed model. The results, in terms of accuracy scores and for four categories, range from "100%", at the "class level", to "98%" at the "genus level", taking into account k-mers of length 8. The robustness of the proposed model indicates these results.

***Keywords***- *16S RNA . DNA* k-*mers . LDA , Naïve Bayes*

## I. INTRODUCTION

The taxonomic classification of bacteria has become of major importance in bioinformatics systems. Not all bacteria are known, only approximately 27% of bacterial populations have species that can be grown in biological laboratories[1]. taxonomy is" naming, defining and classifying groups of biological organisms on the basis of shared characteristics". Organisms are grouped together into taxonomies and these groups are given taxonomic rank; Donated rank groups can be grouped to form a higher rank higher group, thus creating a taxonomic hierarchy. Recently, the major ranks are used: Class, Order, Family, Genus, and Species. Due to the large amount of biological data generated by highly efficient sequencing techniques, the essential need to design and implement bioinformatics systems for the purpose of analyzing and processing this type of biological data has arisen. Taxonomic studies of bacterial species are based on analysis of 16S RNA genome sequences [2,3], which can be shown as a species barcode. The aim of the analysis for 16S sequences is to find similarities between sequences, in terms of evolutionary distance, using alignment algorithms for example such as BLAST [4]. By measuring the similarity between the sequences, it is possible to classify different species of bacteria belonging to similar strains or with known species belong to same strain . Most of the methods used to classify bacteria whose main goal is to find a match between groups and taxonomic classes (taxa), by using tools to measure the distance of similarity between sequences. The clustering method was performed taking into account both evolutionary distances [5,6] and compression-based distances [7,8] depended on measures of global similarity [9]. And also apply compression-based methods to study barcode sequences [10,11]. The classification algorithms for the 16 RNA sequences were comprehensively compared

[3]. The authors obtained the conclusion that the Free-alignment tool relied on the proposed naive Bayesian workbook [12]. In addition, the Simrank search algorithm provides molecular ecologists with a high-throughput, open source choice for comparing large sequence sets to find similarity[13]. Best classification results of 16S RNA gene sequences were produced. The classification approaches described above can be adapted to the implementation of k-mer sequencing of DNA, which uses small fragments of DNA sequences of fixed length k. In addition, the algorithm is implemented in [12]. Adaptation of k-mer to DNA sequences to train a Bayesian classifier. Then, the resulting form is used to allocate a taxonomic label to the query sequence. An alignment-free approach is provided to handle taxonomic classification of 16S gene sequences. Probability topic models support for 16S sequence classifier mode. Probabilistic topic Models "is a statistical algorithm that allows word distribution in documents, and is able to abstract a group of persistence meaningful radical, called topics , that may be used to labeling documents with a semantic aspect." The main concept is to strip the topics of data for their DNA sequences and then characterize those sequences that share the same topic and belong to the taxonomic group.

## II. METHODS

- **Biological concept:**

A gene is the fundamental unit of inheritance physical and functional. DNA consists of genes. The DNA data was stored in the form of a code consisting of four chemical bases: adenine (A), guanine (G), cytosine (C) and thymine (T). DNA bases paired, A with T and C with G, into so-called base pairs. Both the molecules of sugar and phosphate are linked to each base. The nucleotide is commonly referred to as a base, sugar and phosphate. Two long strands, which shape the spiral called the double helix are structured into nucleotides as Figure (1).

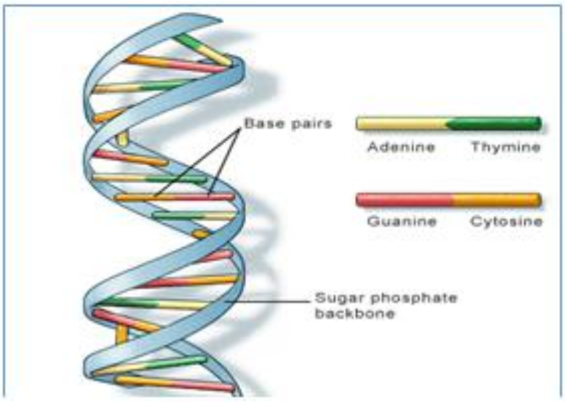


Figure 1: structure of DNA

The double helix structure is rather like the ladder, which consists of the basic pairs that form the ladder's rungs, and Sugar and phosphate molecules that form ladder vertical lateral components. The duplicate sequence of the basis can be repeated by each strand of DNA in the double helix . A universal mechanism in which RNA molecules guide the synthesis of proteins on Ribosomes is one of these active processes. The RNA (tRNA) molecules used for transmitting amino acids to the ribosome are then bound by ribosomal RNA (rRNA) and are then coded proteins in Figure(2).

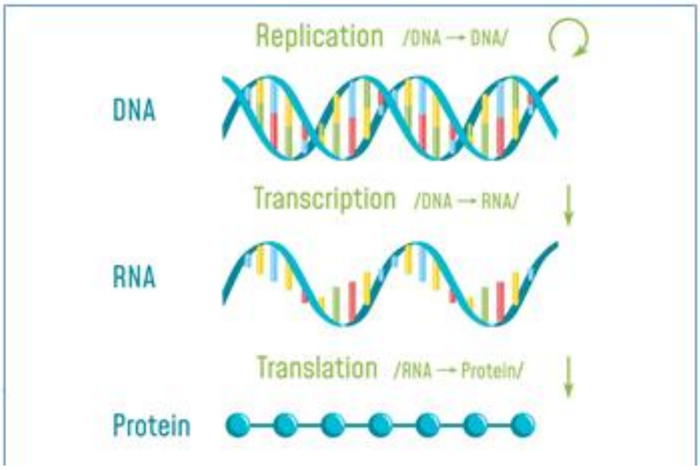


Figure 2: The mechanism of DNA replication

This is important when cells are divided since the exact copies of the DNA present in the old cell are needed for each new cell. Protein-coding genes are just about 1% of DNA; the other 99% are non-coding. A polymer module essential for the coding, decoding, regulation and gene expression of various biochemical functions, ribonucleic acid (RNA). Nucleic acids are RNA and DNA. Nucleic acids are one of four key macromolecules necessary for all known types of life, along with lipids, protein and carbohydrates. Like DNA, RNA is assembled as a nucleotide chain but, compared to DNA, RNA is found in nature, rather than a double strand. to relay genetic information (using guanine, uracil, adenine and cytosine nitrogen bases denoted by letters G, U, A, and C), cellular organisms use messenger RNA (mRNA) to regulate synthesis for particular protein , Figure 3.

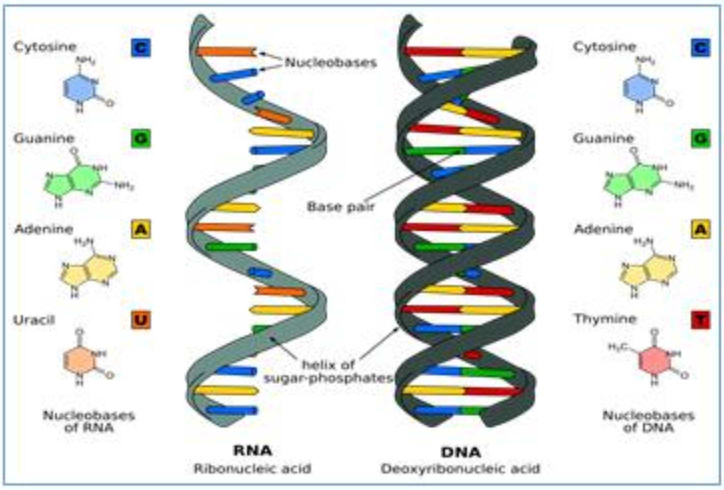


Figure 3: Comparing Strucure DNA and RNA

Bacteria are a type of biocell. The bacteria have various types, from rods to spirals. They have several different shapes. Life forms land bacteria, earth, water, acidic heat, radioactive waste as well as the Earth's crust's deep biosphere.[14].Bacteria also exist in symbiotic and parasite connection to plants and animals.In order to classify bacteria, each species must be allocated to the genus (binary nomenclature) which is in turn the lower level of the hierarchy of a specific species.(phylum, class, order, family).[15][16].as Figure: (4).

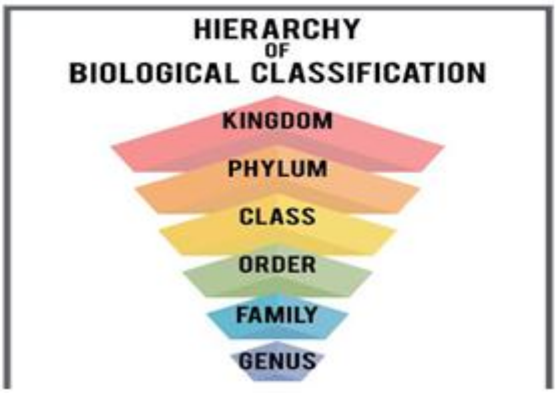
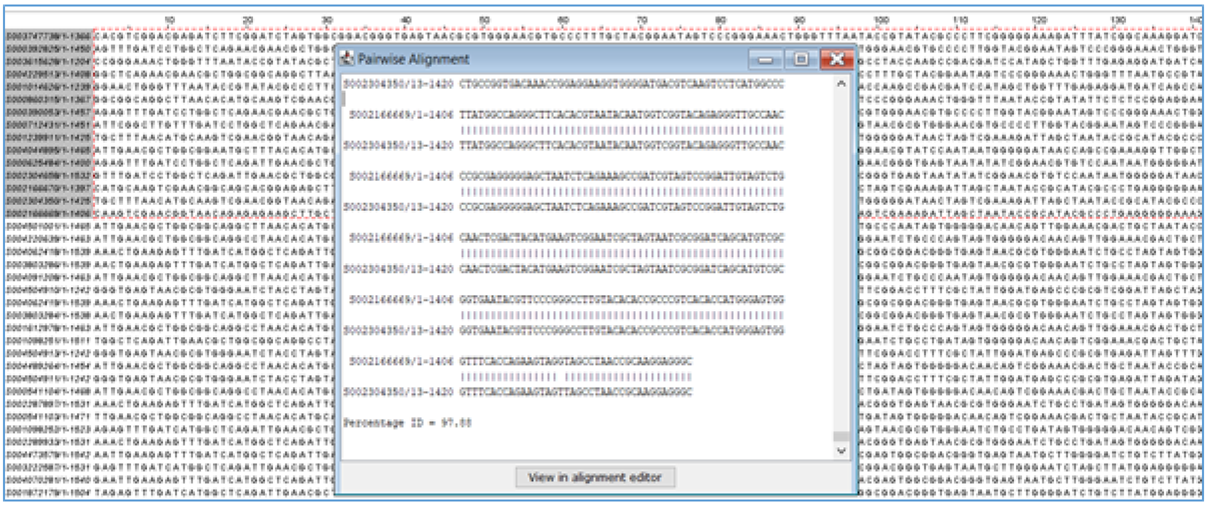


Figure 4. HIERACHY of BIOLOGICAL CLASSIFICATION

- **Sequence alignment**

In the field of bioinformatics, sequence alignment is a way of organizing DNA sequences, RNA or proteins for identification of similarity regions that can be the product of the function, structure or evolutionary relations between the sequences [17]. Gaps between the residues are placed in aligning identical or related characters into successive columns, such as Figure (5).



```
S002304350/13-1420 CTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCCTCATGGCCC

 S002166669/1-1406 TTATGGCCAGGGCTTCACACGTAATACAATGGTCGGTACAGAGGGTTGCCAAC
                   |||||||||||||||||||||||||||||||||||||||||||||||||||||
S002304350/13-1420 TTATGGCCAGGGCTTCACACGTAATACAATGGTCGGTACAGAGGGTTGCCAAC

 S002166669/1-1406 CCGCGAGGGGGAGCTAATCTCAGAAAGCCGATCGTAGTCCGGATTGTAGTCTG
                   |||||||||||||||||||||||||||||||||||||||||||||||||||||
S002304350/13-1420 CCGCGAGGGGGAGCTAATCTCAGAAAGCCGATCGTAGTCCGGATTGTAGTCTG

 S002166669/1-1406 CAACTCGACTACATGAAGTCGGAATCGCTAGTAATCGCGGATCAGCATGTCGC
                   |||||||||||||||||||||||||||||||||||||||||||||||||||||
S002304350/13-1420 CAACTCGACTACATGAAGTCGGAATCGCTAGTAATCGCGGATCAGCATGTCGC

 S002166669/1-1406 GGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGG
                   |||||||||||||||||||||||||||||||||||||||||||||||||||||
S002304350/13-1420 GGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGG

 S002166669/1-1406 GTTTCACCAGAAGTAGGTAGCCTAACCGCAAGGAGGGC
                   |||||||||||||||| |||||||||||||||||||||
S002304350/13-1420 GTTTCACCAGAAGTAGTTAGCCTAACCGCAAGGAGGGC

Percentage ID = 97.88
```

Figure 5: showing example of sequence  pairwise alignment

BLAST (Basic Local Alignment Search Tool) is a technology used in sequence alignment. This provides excellent results if the sequences are the same and are aligned precisely, but if the sequences are different, it is not possible to achieve a consistent alignment, which means that there is no implementation of the sequence alignment. Alignment-based methods limit their computational complexity and time-consuming, so when dealing with large-scale sequencing data is restricted [18].

- **Alignment free sequence**

The growth and demand for analysis of various types of data generated by biological research has provided a boost in the field of bioinformatics [19]. The genomic sequence and structure data on DNA, RNA and proteins, gene expression profiles or microarray data, metabolites, are data based on genomic sequences. Sequencing data grows in the next generation sequencing methods at an exponential rate. Sequence analysis, with a broad range of applications for database research, genomic comment, comparative genomics, molecular evolution and genetics prediction, has remained a major research area since the development and growth of bioinformatics. The k-mer analysis, which is most widely used in large and heterogeneous sequences, occurs in many free alignment sequence techniques [20].

- **k-mer analysis**

k-mers are fragments of sequences(called words) of length included  a DNA,RNA sequence. K-mer is used as a source for assembling DNA sequences,[21] improves the heterologous gene expression[22][23] identifying and generating attenuated vaccines in metagenomics vaccines[24][25]. K-mer was originally employed in the computational genomics and sequence analysis whereby nucleotides (ie., A, T,G, and C) are composed of k-mer. Here, we used a ranking approach to transform subsequences values into a bag-of-words. Within each sample, subsequences were sorted in ascending order based on their frequency values. In Figure 6, as example k=8, all overlapping k-mers, represented by words, can be extracted from the gene sequence, by sliding windows fixed.

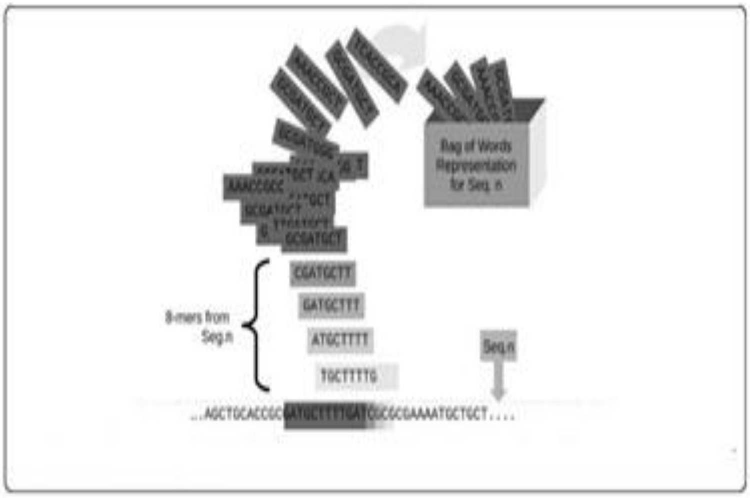

Bag of Words Representation for Seq. n
AAACCGCC
GCGATGCT
CGATGCTT
GATGCTTT
ATGCTTTT
TGCTTTTG
8-mers from Seq.n
Seq.n
...AGCTGCACCGCGATGCTTTTGATCGCGCGAAAATGCTGCT....

Figure 6 :k-mer analysis

- **Topic Modeling**

Topic models are based on the idea that documents are a confusion of topics, while topics are a confusion of words. These models provide us with a simple probabilistic technique for generating documents. Topical models are generative models for text combinations. To create a document based on a topic model, first choose a topics distribution. Second, a random topics is chosen according to the distribution and a word is generated according to the distribution of the words. This process is replicated so that each word in the document is created and a full document is developed. By reversing this process, it is possible to infer the collection of topics that led to the generation of the document corpus [26].

- **Latent Dirichlet Allocation(LDA)**

The Latent Dirichlet Allocation (LDA) is a probabilistic obstetrical model used to model any combination of data that has been segregated [27]. A corpus of text documents may be viewed as a combination of separate data, so LDA can be used for modeling a corpus of documents. Implemented LDA for the first time. They described LDA as a three-stage hierarchical Bayesian model in which each element of a combination is modeled on a finite mix of a sub-state set of topics. In turn, each topic is modeled on a sub-state set of topic probabilities as an infinite mix." A document may be viewed as a combination of topics and each topic shall be viewed as a combination of words to match a document corpus into this concept.

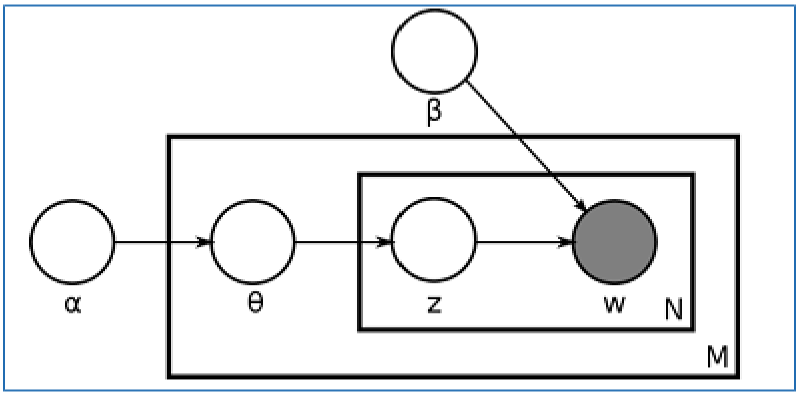


Figure 7: LDA representation graphical model.

Figure (7) display Replicates are shown by the 'plate' boxes. The outer plate is used to documents, while the inside plate explains the iterative choice of topics and words [28]. The LDA steps are shown in algorithms (1). LDA is used to group the data un supervision as a clustering algorithm (1) [27]. Specifies that for each Document w in Corpus D, Algorithm (1) of LDA shall take the following generative procedure.

1. Choose N ~ Poisson (£).

2. Choose θ ~ Dirichlet (α).

3. For each of the N words wn in document w:

(a) Choose a topic zn ~ Multinomial (θ).

(b) Choose a word wn from p(wn | zn,β), a multinomial probability conditioned on the topic zn.

The following are some of the assumptions make about achieving a generative probabilistic corpus LDA model. Firstly, we assume that the dimensionality k of the distribution of Dirichlet is known and therefore defined by the dimension of the topic variable z. In the second case, we assume the k*v matrix β in which β ij = p(w j =1|zi = 1) is parameterized in the word probabilities, which is regarded as a fixed amount to be calculated. The following probability is the k-dimensional random variable Dirichlet θ the same as the (k-1) simple:

$$p(\theta\,|\,\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}, \qquad ..(1)$$

where α is a k-vector with components α i > 0, and the Gamma function is Γ. corpus level parameters are α and β, which are sampled only once during the process of corpus generation. θd is a vector at the document level and is sampled once by document (θd is the topic of distribution for document d). N is separate from all other variables (α and z) producing results. It is also an auxiliary variable and its randomness is usually overlooked. A joint distribution of the mixture of topics β as well as a set of N topics z and N words w as can be calculated for given values of α and β as:

$$p(\theta,\mathbf{z},\mathbf{w}\,|\,\alpha,\beta) = p(\theta\,|\,\alpha)\prod_{n=1}^{N}p(z_n\,|\,\theta)p(w_n\,|\,z_n,\beta), \qquad ..(2)$$

In the above equation p(zn |θ) is θi for a single i such that $z_n^i$=1. If we integrate over θ and then sum over z , then we get:

$$p(\mathbf{w}\,|\,\alpha,\beta) = \int p(\theta\,|\,\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n\,|\,\theta)p(w_n\,|\,z_n,\beta)\right)d\theta. \qquad ..(3)$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(\mathcal{D}\,|\,\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d\,|\,\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}\,|\,\theta_d)p(w_{dn}\,|\,z_{dn},\beta)\right)d\theta_d. \qquad ..(4)$$

Figure 5 offers a probabilistic graphical model of LDA. The figure shows that there are three stages of LDA. Finally, zdn and wdn variables are word-level variables sampled once in each document for each word. Thus documents may be connected to multiple topics under LDA, and multiple words may be combined with each topic.

- **Classification**

A classification function in machine learning is characterized as the method by which a label (or category) from several categories is assigned to one instance. supervised learning is a machine learning class that is very common with classification tasks.

- NAÏVE BAYES

Theorem states that the probability of an event can be expressed as a function of related events [29].Naive Bayes classifies can be easily built with no complicated iterative parameters .This makes it suitable for bioinformatics system, Infer interaction between genes and proteins of TFs via gene expression datasets and TFs where the structure of the gene-regulatory network is understood [30][31]. The study is continuing.Even though is easy to implement , it produces very good results in prediction and estimation systems, especially in Biological data , Example prediction for taxa of unknown bacteria based on 16 ribosomal RNA data. This theorem is expressed in equation(5):

$$P(y/x) = \frac{p(y) * p(\frac{x}{y})}{p(x)} \qquad ..(5)$$

Where p(y/x) is the probability of y given x . p(y) the probability of y.p(x) is the probability of x. p(x/y) is the probability of x given y .The final probability is a result of chaining input feature of a class. The equation's parameters show how to calculate a posterior probability p(x/y) this theorem assumes that the value of predictor (p(x)) is independent from the value of other predictors, which called the class conditional independence [32].this method uses the probability theory to classify input data that don’t have class labels . Its superiority show with categorical data, but it is limited with numerical data.

## III. METHODOLOGY

This work contains three stages: preprocessing (missing value, free alignment), clustering and classification, as shown in (Figure 8):

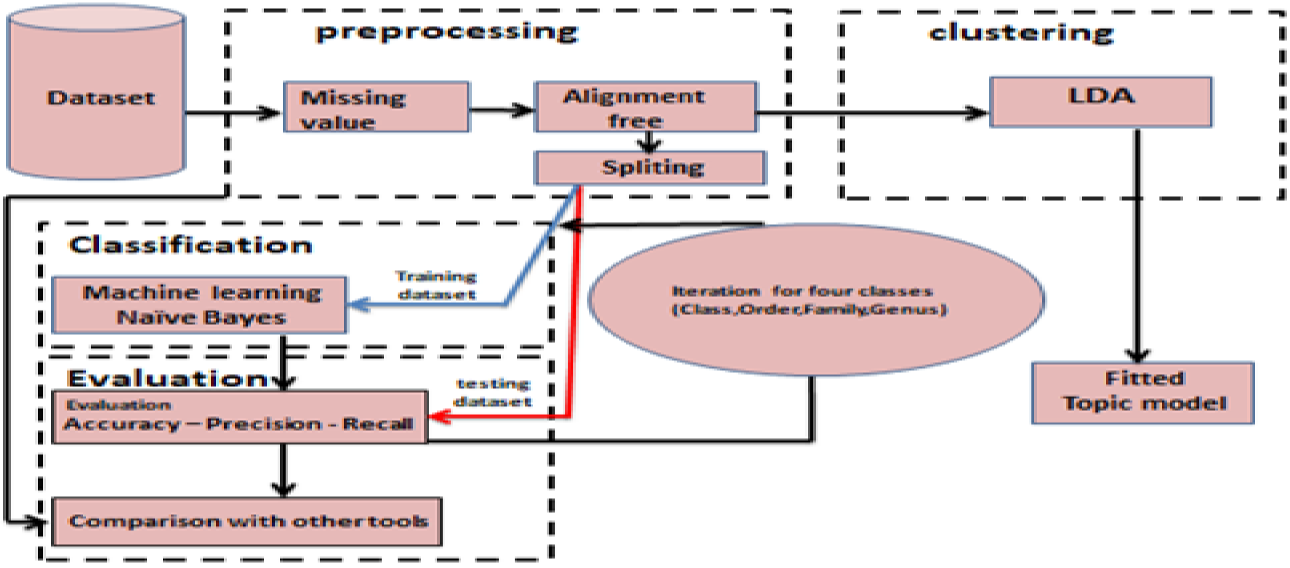


Figure 8: Proposed methodolgy

- **Bacteria Dataset**

RDP Ribosomal Database II (RDP-II) version 10.27 has downloaded the 16S RNA sequences [6]. The strains are the very best sample species, with an average length of 16S of approximately 1200 to 1400 bp; good quality [meaning a quality check by the RDP system was performed in the selected sequences]. NCBI taxonomy reflects their taxonomic class, from class to genus, distinguished NCBI Biosystems’ taxonomic nomenclature [33] .The downloaded sequences with regard to only those classes containing at least ten elements, in order to obtain a well-balanced training set .The (Figure 9) shows taxonomy each class to sub class. In each classification cycle many subclasses are obtained from the four hierarchical classes (class, order, family, genus), taxa (class) have 3 sub classes (Alpha, Beta,Gamma).Taxa (order) have 20 subclass (aeromonadales, alteromomnadal, ... xanthomomnadal).Taxa(family) class have 38 subclass and the genus class have hundreds of subclass .

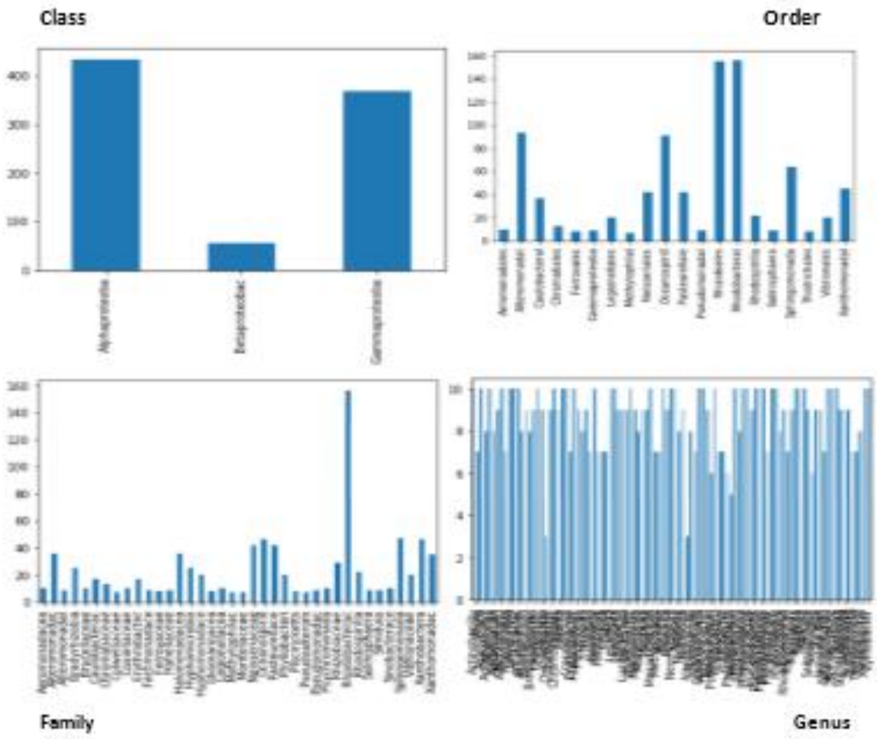


Figure 9:types for each categoriy of bacteria strains

- **Pre-processing stage**

Data preprocessing methods are important to extract useful information from original datasets. These techniques include data cleaning tasks as handling missing value, duplicate data, and noise removal and avoid high dimensionally. Data preprocessing techniques are applied to convert the row data into an understandable format, and adapted it to fit the analysis methods as pipeline methodology.

1. **Missing value handling**

A missing value in DNA sequence is an empty letter in sequence (AGCT) or called "gap" between two letters. On other hand, Irrelevant values as (N M W R), it considered are missing values, as figure (8). Missing values for a dataset have an impact on the accuracy of results in the clustering and classification models.

2. **k-mer analysis**

The L-length sequence will contain L-K + 1 k-mers and the total possible nk-mers, where n is the number of potential monomers (say four in the case of DNA). Suppose p is the k-mer and s is a string. The frequency of p in s can be defined as the number of iterations p shows as a substring of

$$\text{freq}\ (p,s)= \frac{count\,(p,s)}{|s|-k+1} \qquad ..(6)$$

Let's get K ⊆ Z +. The K-mer combination for the s sequence defined as the frequencies of all possible K-mers for k ∈ K. for the four-letter alphabet and K = {k | 1≤k ≤n}, the origin of the K-mer formation is given by 4k $=\frac{4}{3}$

$$\sum_{k=1}^{n} 4^{k} = \frac{4}{3}(4^{n}-1) \qquad ..(7)$$

s divided by the number of sequences with length k in s [11].

3.3 Clustering and classification stages

In this time , The LDA was used as clustering method is explained in section (2), then we used the Naïve Bayes as classification stage , it is explained in section(2) .

- **Evaluation:-**

To obtain a good model fit, the dataset is typically split into two disjoint subsets For the purpose of evaluating the model's performance .These subsets are referred to as the training data and the test data. Percentage of splitting the data set to "70%" of training data and "30%" of test data .Two datasets should be independent from each other and represent the same seen by the model. The test data set should therefore only be utilized once by the algorithm. The training data are used to fit the model, and contain the samples from which the algorithm learns on its own. The model knows which label each sample within the training set has, and iteratively updates its parameters accordingly in order for its predicted outcome to come closer to the expected one. The algorithm which using for validation is k-fold when k is number of iterations validation . A performance metric Mi is given to and retained for each iteration i, and the model discarded so a new model can be trained on the next division of training and test sets. After k iterations have passed, the average over all the performance scores is determined to yield an overall performance score [34].

$$M = \frac{1}{K} \sum_{i=1}^{k} M \qquad ..(8)$$

## IV. Results

### 4.1 Data before handling missing value

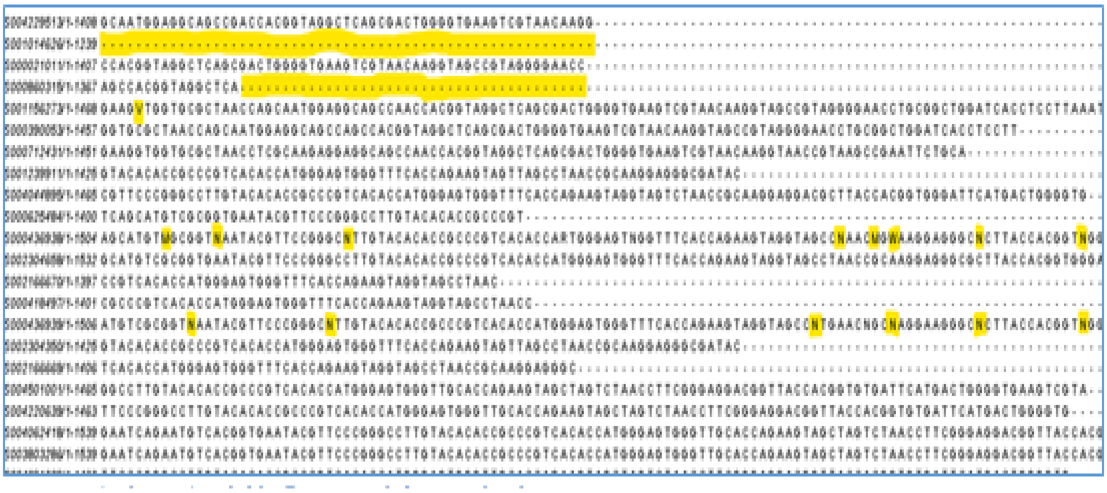


Figure 10: The presence of Irrelevant values as (N M W R..), as well as there are gaps between the letters.the number of sequence are 1000.

## 4.2 Data after handling missing value

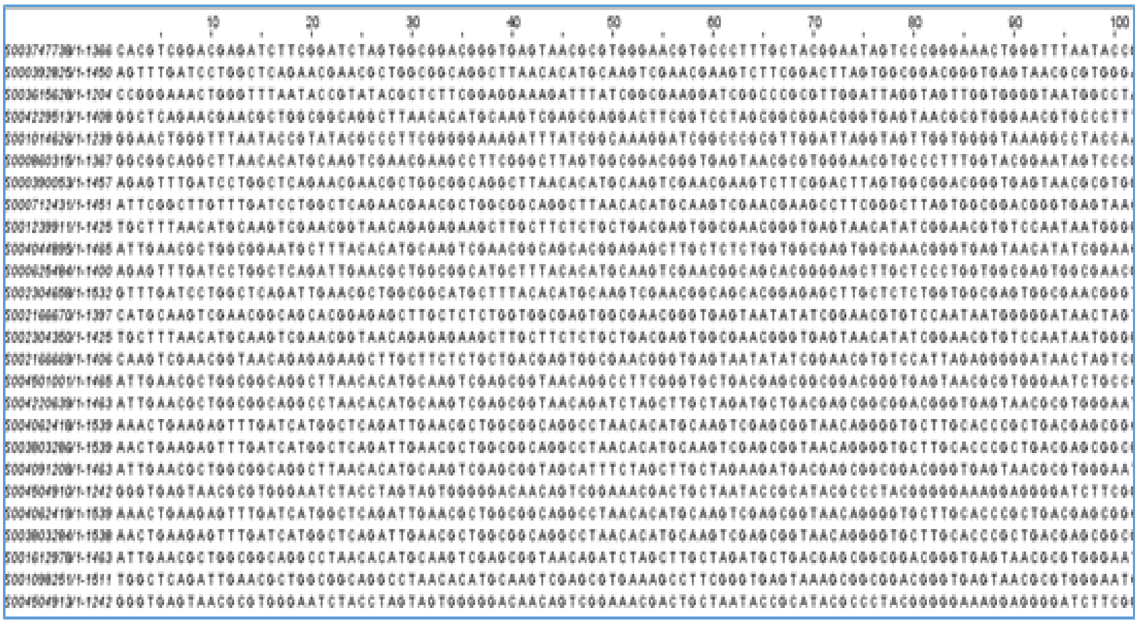


Figure 11: Missing and irrelevant values are handled , thus the data size reduced from 1000 to 860 sequence.each sequence have length (1..n) ,n =number of letters(AGCT).

- **K-mer:**

k-mer refers to all of a sequence's subsequences of length Ķ, such that the sequence AGAT would have four monomers (A, G, A, and T), three 2-mers (AG, GA, AT), two 3-mers (AGA and GAT) and one 4- mer (AGAT), After that, all the repeated words in each document are counted and placed in a bag of words, the most frequent words have the highest rank in the bag using the rank method .

Table 1:Word frequency in each sequence is calculated based on the size of k in each iteration.

| 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 | ttttt(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 |
| K=6 | | | | | | | |
| Sequence(1..n) | aaaaaa(word1) | aaaaac(word2) | aaaaag(word3) | .. | tttttc | tttttg | tttttt(word4096) |
| S003747738 | 0 | 0 | 1 | .. | 0 | 0 | 0 |
| S000392825 | 0 | 0 | 1 | .. | 0 | 0 | 0 |
| S003615628 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| . | .. | .. | .. | .. | .. | .. | .. |
| S000020486 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S003222585 | 0 | 0 | 0 | .. | 0 | 1 | 0 |
| S004450765 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| K=7 | | | | | | | |
| Sequence(1..n) | aaaaaaa(word1) | aaaaaac(word2) | aaaaaag(w3) | .. | ttttttc | ttttttg | ttttttt(word6348) |
| S003747738 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S000392825 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S003615628 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| . | .. | .. | .. | .. | .. | .. | .. |
| S000020486 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S003222585 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S004450765 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| K=8 | | | | | | | |
| Sequence(1..n) | aaaaaaaa(word1) | aaaaaaac(word2) | aaaaaaag | .. | tttttttc | tttttttg | tttttttt(word65536) |
| S003287966 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S003808864 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S003808859 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| .. | .. | .. | .. | .. | .. | .. | .. |
| S000020486 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S003222585 | 0 | 0 | 0 | .. | 0 | 0 | 0 |
| S004450765 | 0 | 0 | 0 | .. | 0 | 0 | 0 |

- **Clustering**

All the k-mers obtained by modeling the text (with high probability) in LDA As combination of features. (Table 2) indicates the mechanism by which a test or a LDA functionality training document. The number is count to represent a document Number of occurrences from each LDA topic for each term (k-mer). For an LDA model, then there are over 50 features represented by 5 topics with 10 words (k-mer).

Table 2: Results of clustering stage, Subsequences (words) which have high probability with 3 topics when k=(3..8).

| K=3 | Top ten words probability | K=4 | Top ten words probability |
|---|---|---|---|
| Topic 1 | ('0.028*"ggg" + 0.026*"gaa" + 0.025*"gga" + 0.025*"aag" + 0.024*"aga" + 0.024*"tgg" + 0.024*"gag" + 0.023*"agc" + 0.022*"cgg" + 0.022*"gtg"') | Topic 1 | '0.011*"gtgg" + 0.010*"ggtg" + 0.009*"tggg" + 0.009*"cagc" + 0.009*"ggaa" + 0.009*"ggga" + 0.009*"gcaa" + 0.009*"cggg" + 0.009*"gggg" + 0.008*"ggag"' |
| Topic2 | 0.037*"ggg" + 0.032*"gga" + 0.030*"tgg" + 0.028*"gag" + 0.023*"agc" + 0.023*"gtg" + 0.023*"gcc" + 0.022*"cag" + 0.021*"cgg" + 0.021*"gca" | Topic2 | '0.012*"tggg" + 0.010*"gggg" + 0.010*"ggag" + 0.009*"cgca" + 0.009*"ggtg" + 0.009*"gtga" + 0.008*"gtgg" + 0.008*"aagc" + 0.008*"ggaa" + 0.008*"tgag"' |
| Topic3 | '0.032*"tgg" + 0.028*"ggg" + 0.027*"ggt" + 0.026*"acg" + 0.024*"gcg" + 0.023*"gtg" + 0.023*"aag" + 0.023*"cgg" + 0.023*"ggc" + 0.022*"gcc"' | Topic3 | '0.012*"gggg" + 0.012*"ggaa" + 0.012*"tggg" + 0.011*"gtga" + 0.010*"ggga" + 0.010*"ggtg" + 0.010*"ctgg" + 0.009*"cgga" + 0.009*"acgg" + 0.008*"gtgg"' |
| Topic4 | '0.031*"ggg" + 0.030*"gga" + 0.028*"gaa" + 0.026*"ggc" + 0.024*"ggt" + 0.022*"gca" + 0.022*"tgg" + 0.021*"aat" + 0.021*"gtg" + 0.020*"agg"' | Topic4 | '0.012*"ggga" + 0.012*"tggg" + 0.011*"gggg" + 0.011*"ggaa" + 0.009*"gaat" + 0.008*"gaag" + 0.008*"ggag" + 0.008*"ggtg" + 0.008*"agaa" + 0.007*"gtag"' |
| Topic5 | '0.038*"ggg" + 0.028*"cgg" + 0.028*"gtg" + 0.024*"aag" + 0.024*"gga" + 0.024*"gaa" + 0.024*"tgg" + 0.022*"agc" + 0.022*"gcg" + 0.021*"gag"' | Topic5 | '0.011*"ggtg" + 0.011*"ggga" + 0.011*"gggg" + 0.009*"gaag" + 0.009*"tggg" + 0.009*"ctgg" + 0.008*"gtgg" + 0.008*"cgga" + 0.008*"cctt" + 0.007*"gcta"' |
| K=5 | Top ten words probability | K=6 | Top ten words probability |
| Topic 1 | 0.006*"ggaat" + 0.006*"ttcgg" + 0.005*"ggtga" + 0.004*"gggga" + 0.004*"gaagg" + 0.004*"tgggg" + 0.004*"tcgga" + 0.004*"cacac" + 0.004*"gttgg" + 0.003*"gcaac"' | Topic 1 | '0.003*"tgggga" + 0.002*"gtgggg" + 0.002*"ggtgaa" + 0.002*"ttcgga" + 0.002*"actggg" + 0.002*"gtgcta" + 0.002*"taatac" + 0.002*"gccgcg" + 0.002*"gctaac" + 0.002*"ggggaa"' |
| Topic2 | '0.006*"ttcgg" + 0.005*"tgggg" + 0.004*"ggtga" + 0.004*"ggaat" + 0.004*"tcgga" + 0.004*"cagca" + 0.004*"aggtg" + 0.004*"gcaac" + 0.004*"gatga" + 0.004*"gttcg"' | Topic2 | '0.003*"tgggga" + 0.003*"gtgggg" + 0.002*"ggggat" + 0.002*"gaagaa" + 0.002*"cggtga" + 0.002*"gaggaa" + 0.002*"gtgaaa" + 0.002*"gcggtg" + 0.002*"gaaggc" + 0.002*"gtaaag"' |
| Topic3 | '0.006*"gggga" + 0.005*"tgggg" + 0.004*"aaagc" + 0.004*"ggaat" + 0.004*"gtgaa" + 0.004*"gggag" + 0.004*"cctgg" + 0.004*"aagcg" + 0.004*"ctggg" + 0.004*"gtggg"' | Topic3 | '0.003*"gttcgg" + 0.003*"cacacc" + 0.002*"caaccc" + 0.002*"ggtgaa" + 0.002*"ttcgga" + 0.002*"gaggtg" + 0.002*"ggaatt" + 0.002*"caatgg" + 0.002*"gtaaag" + 0.002*"tgccag"' |
| Topic4 | '0.006*"ggtga" + 0.005*"gggga" + 0.005*"cgcaa" + 0.004*"gtggg" + 0.004*"gttgg" + 0.004*"tgggg" + 0.004*"aagcc" + 0.004*"gggag" + 0.004*"agttg" + 0.004*"ggaat"' | Topic4 | '0.003*"tgggga" + 0.003*"gtgggg" + 0.003*"ggtgaa" + 0.003*"agttgg" + 0.002*"cgcaag" + 0.002*"cagtgg" + 0.002*"cggtga" + 0.002*"cgaagg" + 0.002*"gccctt" + 0.002*"gctaac"' |
| Topic5 | '0.005*"tgggg" + 0.005*"gggga" + 0.004*"gtggg" + 0.004*"gggaa" + 0.004*"gtgaa" + 0.004*"gggag" + 0.004*"ggtga" + 0.004*"aagtc" + 0.004*"ggaat" + 0.004*"ggagg"' | Topic5 | '0.003*"tgggga" + 0.003*"gtgggg" + 0.002*"cttcgg" + 0.002*"gaagaa" + 0.002*"ttcggg" + 0.002*"cggtga" + 0.002*"taatac" + 0.002*"cgagcg" + 0.002*"gtgaaa" + 0.002*"tgtgaa"' |
| K=7 | Top ten words probability | K=8 | Top ten words probability |
| Topic 1 | '0.002*"gtgggga" + 0.002*"cttcggg" + 0.002*"ccttcgg" + 0.001*"gaagaag" + 0.001*"agcggtg" + 0.001*"cggtgaa" + 0.001*"tacacac" + 0.001*"tcggaat" + 0.001*"tggggag" + 0.001*"acaatgg"' | Topic 1 | '0.001*"agaggtga" + 0.001*"ggtgaaat" + 0.001*"tgccagca" + 0.001*"acaatggg" + 0.001*"cagttcgg" + 0.001*"cgtaaagc" + 0.001*"gccagcag" + 0.001*"gacaatgg" + 0.001*"ccttcggg" + 0.001*"attagata"' |
| Topic2 | '0.002*"acaatgg" + 0.001*"gaggtga" + 0.001*"ggaggaa" + 0.001*"caaccct" + 0.001*"tgttcgg" + 0.001*"aactgcc" + 0.001*"agaggtg" + 0.001*"cgggagg" + 0.001*"ggtgaaa" + 0.001*"tcggaat"' | Topic2 | '0.002*"ccttcggg" + 0.001*"gccttcgg" + 0.001*"gcggtgaa" + 0.001*"gtgaaatg" + 0.001*"gaggaagg" + 0.001*"cttcggga" + 0.001*"tgccagca" + 0.001*"tgaagtcg" + 0.001*"cctacggg" + 0.001*"ggtggagc"' |
| Topic3 | '0.002*"gtgggga" + 0.001*"tggggag" + 0.001*"tacacac" + 0.001*"ggtgaaa" + 0.001*"acaatgg" + 0.001*"ggattag" + 0.001*"gttcgga" + 0.001*"gaaggcg" + 0.001*"gatcagc" + 0.001*"gaggcag"' | Topic3 | '0.001*"aaagcgtg" + 0.001*"gcggtgaa" + 0.001*"gcccttat" + 0.001*"cccttatg" + 0.001*"tgaagtcg" + 0.001*"tgttgggt" + 0.001*"cggaggaa" + 0.001*"gaatacgt" + 0.001*"ccgcaagg" + 0.001*"agatgttg"' |
| Topic4 | '0.002*"gtgggga" + 0.001*"acaatgg" + 0.001*"tcggaat" + 0.001*"ggaggaa" + 0.001*"tgttggg" + 0.001*"agtgggg" + 0.001*"tggggag" + 0.001*"gggatga" + 0.001*"agcagtg" + 0.001*"ggtgggg"' | Topic4 | '0.001*"agaggtga" + 0.001*"ggtgaaat" + 0.001*"ggaggcag" + 0.001*"cgaaggcg" + 0.001*"gccgcggt" + 0.001*"tgccagca" + 0.001*"aggtgaaa" + 0.001*"gaggtgaa" + 0.001*"atgccgcg" + 0.001*"gaggcagc"' |
| Topic5 | '0.002*"gtgggga" + 0.002*"acaatgg" + 0.002*"gaggtga" + 0.002*"agaggtg" + 0.001*"gttcgga" + 0.001*"ggtgaaa" + 0.001*"tggggag" + 0.001*"tacacac" + 0.001*"gtgaaat" + 0.001*"gcgtggg"' | Topic5 | '0.002*"ccttcggg" + 0.001*"gtgaaatg" + 0.001*"cctacggg" + 0.001*"tgccagca" + 0.001*"gccttcgg" + 0.001*"gaggaagg" + 0.001*"aaagcgtg" + 0.001*"gttaagtc" + 0.001*"ctgagaca" + 0.001*"cttcggga"' |

The topic distribution is used as features for each topic in LDA. Table 3: shows the LDA features can be shown as probability of query sequences for five topics with difference k values . The topic distribution of each topic from the LDA model represents a document. The number of sequence(features) is equal to number of topics ,not equal to number of words.

Table 3: Results of clustering stage, three query sequences (documents) with five topics when k= (3-8)

| **K=3** | Squence_1 | Squence_2 | Squence_3 | **K=4** | Squence_1 | Squence_2 | Squence_3 |
|---|---|---|---|---|---|---|---|
| **Topic 1** | 0.1991 | 0.1993 | 0.1993 | **Topic 1** | 0.1068 | 0.1005 | 0.308 |
| **Topic2** | 0.1991 | 0.1975 | 0.1975 | **Topic2** | 0.3219 | 0.2769 | 0.1067 |
| **Topic3** | 0.206 | 0.2071 | 0.2071 | **Topic3** | 0.3131 | 0.1696 | 0.1084 |
| **Topic4** | 0.1998 | 0.1998 | 0.1998 | **Topic4** | 0.1544 | 0.1149 | 0.2499 |
| **Topic5** | 0.1960 | 0.1960 | 0.1960 | **Topic5** | 0.1035 | 0.3378 | 0.2259 |
| **K=5** | Squence_1 | Squence_2 | Squence_3 | **K=6** | Squence_1 | Squence_2 | Squence_3 |
| **Topic 1** | 0.3591 | 0.5353 | 0.5629 | **Topic 1** | 0.0000 | 0.0000 | 0.5507 |
| **Topic2** | 0.0177 | 0.4266 | 0.0200 | **Topic2** | 0.4913 | 0.9997 | 0.1122 |
| **Topic3** | 0.5379 | 0. 2352 | 0.0864 | **Topic3** | 0.0000 | 0.0000 | 0.0000 |
| **Topic4** | 0.03674 | 0.0000 | 0.3266. | **Topic4** | 0.0000 | 0.0000 | 0.2288 |
| **Topic5** | 0.4831 | 0.0000 | 0.5106 | **Topic5** | 0.5085 | 0.0000 | 0.1081 |
| **K=7** | Squence_1 | Squence_2 | Squence_3 | **K=8** | Squence_1 | Squence_2 | Squence_3 |
| **Topic 1** | 0.0000 | 0.0000 | 0.6429 | **Topic 1** | 0.9853 | 0.0000 | 0. 4899 |
| **Topic2** | 0.0000 | 0.0000 | 0.1564 | **Topic2** | 0.0000 | 0.9987 | 0.0678 |
| **Topic3** | 0.0000 | 0.9998 | 0.0328 | **Topic3** | 0.0145 | 0.0000 | 0.3654 |
| **Topic4** | 0.9998 | 0.0000 | 0.1465 | **Topic4** | 0.0000 | 0.0000 | 0.5177 |
| **Topic5** | 0.0000 | 0.0000 | 0.0211 | **Topic5** | 0.0000 | 0.0000 | 0.0000 |

- **Classification**

The classification results was obtained using the Naïve Bayes algorithm with four metrics(Accuracy _Precission_Recall_F1) when k=[1..8] , the Accuracy increases in the class column when k = (5,6,7,8), as The best accuracy was obtained in the from  class 100% to an accuracy of 98% in the genus column when k = 7,8, as Figure 12 .

Table 4: Results of classification the Naïve Bayes algorithm

| Naïve Bayes | Class | | | | Order | | | | Family | | | | Genus | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | precision | recall | F1 | Accuracy | precision | recall | F1 | Accuracy | precision | recall | F1 | Accuracy | precision | recall | F1 |
| K=3 | 0.969 | 0.972 | 0.969 | 0.970 | 0.829 | 0.855 | 0.829 | 0.822 | 0.86 | 0.889 | 0.86 | 0.857 | 0.93 | 0.947 | 0.93 | 0.929 |
| K=4 | 0.93 | 0.947 | 0.93 | 0.929 | 0.926 | 0.941 | 0.926 | 0.928 | 0.93 | 0.939 | 0.93 | 0.932 | 0.965 | 0.975 | 0.965 | 0.965 |
| K=5 | 1 | 1 | 1 | 1 | 0.961 | 0.966 | 0.961 | 0.961 | 0.961 | 0.973 | 0.961 | 0.963 | 0.981 | 0.987 | 0.981 | 0.98 |
| K=6 | 1 | 1 | 1 | 1 | 0.965 | 0.97 | 0.965 | 0.965 | 0.996 | 0.996 | 0.996 | 0.996 | 0.984 | 0.989 | 0.984 | 0.984 |
| K=7 | 1 | 1 | 1 | 1 | 0.981 | 0.982 | 0.981 | 0.981 | 0.996 | 0.996 | 0.996 | 0.996 | 0.981 | 0.986 | 0.981 | 0.98 |
| K=8 | 1 | 1 | 1 | 1 | 0.996 | 0.996 | 0.996 | 0.996 | 0.996 | 0.996 | 0.996 | 0.996 | 0.984 | 0.989 | 0.984 | 0.984 |

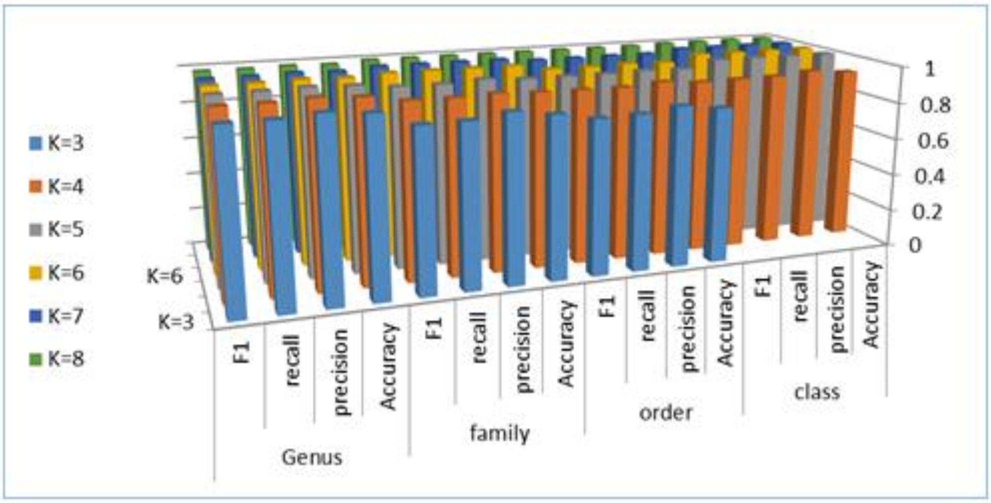


Figure12: A graphic chart showing the classification metrics for the four categories(,Order,Family,Genus) according to k value.

## V. Conclusion

Through experiments in the pre-processing stage, we use full genome 16sRNA ,the size of the data (number of sequences) was reduced from 1000 to 860 sequences, where the irrelevant values were eliminated and the missing value handling, Then the sequences break into subsequences (words) using the free alignment method. The words most frequent are depending on the ranking method. These most frequent words are inputs to the clustering stage by the LDA algorithm to extract a fitted topic model. On the other hand, it is possible to consider the output of preprocessing as input to the classification phase by the Naïve Bayes algorithm, and lastly, the evaluation stage is verified. The correctness of the classification results using the cross validation method. We notice that the score of accuracy increases in eac h of the four levels (class, order, family, genus) with increasing the value of k until accuracy reaches at ("100 %", at "class level", to "98%" at "genus level" when the value of k = 8).

## Reference

[1] N. K. Dudek *et al.*, “Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome,” *Curr. Biol.*, vol. 27, no. 24, pp. 3752-3762.e6, 2017.

[2] M. Drancourt, P. Berger, and D. Raoult, “Systematic 16S rRNA Gene Sequencing of Atypical Clinical Isolates Identified 27 New Bacterial Species Associated with Humans,” *J. Clin. Microbiol.*, vol. 42, no. 5, pp. 2197–2202, 2004.

[3] Z. Liu, T. Z. Desantis, G. L. Andersen, and R. Knight, “Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers,” *Nucleic Acids Res.*, vol. 36, no. 18, pp. 1–11, 2008.

[4] B. B. Salgaonkar, M. Kabilan, and J. M. Braganca, “Basic local alignment search tool.,” *World journal of microbiology & biotechnology*, vol. 27, no. 4. pp. 403–410, 2011.

[5] M. La Rosa, A. Fiannaca, R. Rizzo, and A. Urso, “Genomic sequence classification using probabilistic topic modeling,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8452 LNBI, pp. 49–61, 2014.

[6] M. La Rosa, R. Rizzo, and A. Urso, “Soft topographic maps for clustering and classifying bacteria using housekeeping genes,” *Adv. Artif. Neural Syst.*, vol. 2011, 2011.

[7] M. La Rosa, S. Gaglio, R. Rizzo, and A. Urso, “Normalised compression distance and evolutionary distance of genomic sequences: comparison of clustering results,” *Int. J. Knowl. Eng. Soft Data Paradig.*, vol. 1, no. 4, p. 345, 2009.

[8] M. La Rosa, R. Rizzo, A. Urso, and S. Gaglio, “Comparison of genomic sequences clustering using normalized compression distance and evolutionary distance,” in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 2008, pp. 740–746.

[9] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi, “The similarity metric,” *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[10] L. E. Peterson, F. Masulli, and G. Russo, “A Study of Compression–Based Methods for the Analysis of Barcode Sequences,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7845 LNBI, no. January, 2013.

[11] M. La Rosa, A. Fiannaca, R. Rizzo, and A. Urso, “Alignment-free analysis of barcode sequences by means of compression-based methods,” *BMC Bioinformatics*, vol. 14, no. SUPPL7, 2013.

[12] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, “Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy,” *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, 2007.

[13] T. Z. DeSantis *et al.*, “Simrank: Rapid and sensitive general-purpose k-mer search tool,” *BMC Ecol.*, vol. 11, 2011.

[14] J. K. Fredrickson *et al.*, “Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the Hanford Site, Washington State,” *Appl. Environ. Microbiol.*, vol. 70, no. 7, pp. 4230–4241, 2004.

[15] A. K. Verma and S. Prakash, “Status of Animal Phyla in Different Kingdom Systems of Biological Classification,” *Int. J. Biol. Innov.*, vol. 02, no. 02, pp. 149–154, 2020.

[16] C. R. Woese, O. Kandler, and M. L. Wheelis, “Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 12, pp. 4576–4579, 1990.

[17] D. W. Mount, “Sequence and genome analysis,” *Bioinforma. Cold Spring Harb. Lab. Press Cold Spring Harb.*, vol. 2, 2001.

[18] C. Kemena and C. Notredame, “Upcoming challenges for multiple sequence alignment methods in the high-throughput era,” *Bioinformatics*, vol. 25, no. 19, pp. 2455–2465, 2009.

[19] J. Rothberg, B. Merriman, and G. Higgs, “FOCUS: EDUCATING YOURSElF IN BIOINFORMATICS,” *Yale J. Biol. Med.*, vol. 85, pp. 305–308, 2012.

[20] S. Vinga and J. Almeida, “Alignment-free sequence comparison - A review,” *Bioinformatics*, vol. 19, no. 4, pp. 513–523, 2003.

[21] G. Z. Valenci, M. Rubinstein, R. Afriat, a Z. D. Shira Rosencwaig, E. Rorman, and I. Nissan, “Draft Genome Sequences of Cronobacter muytjensii Cr150 , Cronobacter turicensis Cr170, and Cronobacter sakazakii Cr611 Gal,” no. June, pp. 9–11, 2020.

[22] M. Welch *et al.*, “Design parameters to control synthetic gene expression in Eschorichia coli,” *PLoS One*, vol. 4, no. 9, 2009.

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

[24] S. C. Perry and R. G. Beiko, “Distinguishing microbial genome fragments based on their composition: Evolutionary and comparative genomic perspectives,” *Genome Biol. Evol.*, vol. 2, no. 1, pp. 117–131, 2010.

[25] 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.*, vol. 14, no. 1, pp. 1–24, 2018.

[26] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, “Topics in semantic representation,” *Psychol. Rev.*, vol. 114, no. 2, pp. 211–244, 2007.

[27] F. P. Johnson and D. M. Robinson, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res. 3,* 2003.

[28] D. M. Blei, K. Franks, M. I. Jordan, and I. S. Mian, “Statistical modeling of biomedical corpora: Mining the Caenorhabditis Genetic Center Bibliography for genes related to life span,” *BMC Bioinformatics*, vol. 7, pp. 1–19, 2006.

[29] Y. Chien, “Pattern classification and scene analysis,” *IEEE Trans. Automat. Contr.*, vol. 19, no. 4, pp. 462–463, 1974.

[30] M. K. Titsias, A. Honkela, N. D. Lawrence, and M. Rattray, “Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison,” *BMC Syst. Biol.*, vol. 6, 2012.

[31] A. K. AL-Mashanji and S. Z. AL-Rashi, “Computational Methods for Preprocessing and Classifying Gene Expression Data-Survey,” in *2019 4th Scientific International Conference Najaf (SICN)*, 2019, pp. 121–126.

[32] K. Vembandasamy, R. Sasipriya, and E. Deepa, “Heart Diseases Detection Using Naive Bayes Algorithm,” *Int. J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, pp. 441–444, 2015.

[33] L. Y. Geer *et al.*, “The NCBI BioSystems database,” *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp. 492–496, 2009.

[34] G. De Clercq, “Deep learning for classification of DNA,” 2019.