Classifying Various Types of Influenza Viral Strains

Davisdon Fleurantin
Computer Science
Georgia State University
Atlanta, Georgia
dfleurantin1@student.gsu.edu

Tien Tran
Computer Science
Georgia State University
Atlanta, Georgia
tientran119@student.gsu.edu

***Abstract*—. The influenza virus is one of the most studied viral infections. This virus has a high rate of mutations, evolving into different strains. There are human and non-human strains, and the human strains are further divided. The different strains have been classified using HMM models. In the project, different models were trained and tested using a text processing system to classify the different strains or serotypes of the virus. The system selects the most frequent words in the protein sequences of each serotype to classify them. The random forest classify predicts the classification at 97% accuracy. The Naïve Bayes model classify the HA protein of the virus against the NA protein at 100% accuracy.**

I. Introduction

Bioinformatics research has been growing exponentially and researchers in computer science, biology and others, have gained insightful information using its tool. This age of information has produced tons of data, especially in genomics. Using bioinformatics tools, these biological data can be stored and help develop new methods and software tools for elucidating what is contained in these data [1]. Biological data, in the form of nucleic acids, pose many questions and carry enormous challenges in the scientific community. These nucleic acids are deoxyribonucleic acids (DNA) and ribonucleic acids (RNA). According to the Central dogma of molecular biology, the flow of genetic information starts at DNA to RNA, then proteins [2]. The building blocks of life is protein, and its basic structure is a chain of amino acids. There exist 20 amino acids. The amino acids perform various functions such as metabolism regulation, muscle building and others [3]. Genes possess the information to make proteins. Viruses have nucleic acids which are vital for their survival.

In the case of the influenza, the virus recognizes the gene it wants to fuse with. After fusing with the gene, the viral nucleic acid becomes parts of the host cell [4]. The influenza virus A cause contagious respiratory illnesses originating in the throat and other organs. This virus evolves rapidly because of antigenic drift and antigenic shift. Antigenic drift occurs when small changes in the nucleic structure of the virus cause surface proteins on the virus to also change. The HA and NA surface proteins are antigens or biomarkers that are recognized by the immune system. The changes do not alter the conformation of these surface proteins. Antigenic shift happens when major changes in the nucleic structure affect the conformation of the surface proteins of the virus. The surface proteins are drastically changed. When those changes occur, the influenza virus from an animal can affect humans [5] . The HA surface protein bind the viral coat to the host cell receptors and facilitate its entry. In contrast, the NA protein frees the virus from the host cell receptors [6] . There are many subtypes of the influenza A virus such as H1, H2, H3 and more; these subtypes create various strains of the virus. There are H1N1, H1N2 and more.

A. Significance

Due to the high number of strains of the influenza virus, methods are needed to classify each strain, and categorize each strain. Classical assays methods have been used in the past. Reverse transcription PCR is one of many methods used but it can become burdensome 3 . Different reagents and primers must be purchased to perform a PCR. The cost, in addition to the time it will take, would add unneeded pressure to laboratories. The BLAST search method can be used to categorize the different strains; however, the method will not divulge any new information such as whether mutation occurred [7].

B. Previous Research

To classify the different strains of the virus, profile hidden Markov models were used. First, the various protein sequences were aligned using the Clustal X program for multiple sequence alignment. In multiple sequence alignment (MSA), three or more protein sequences are aligned simultaneously based on the assumption that they share a common ancestor. Subsequent proteins are then added to grow the alignment. The program will eventually reveal the conserved region shared by all the protein sequences. Following the Clustal X, the protein sequences were analyzed using two models, a hidden Markov model and database search. The hidden Markov model uses a statistical method to determine the frequency of amino acid residues at a known position. This model reveals that at some specific position, an amino acid in the multiple alignment is more prone to insertion or deletion [8] . An HMM program is used to build profiles for each of the subtype of the virus using the pre-aligned protein sequences and generate their statistical score. An HMM database is created using known profiles of the influenza virus subtypes. To search the database for the most likely sequence alignment, a dynamic programming algorithm is implemented where an input sequence is put against an HMM profile. A score is generated that corresponds to how similar are the conserved regions of both the input sequence and the HMM profiles in the database [6].

During the pre-processing phase of large data, especially genetic data, novel techniques are implemented such as feature selection. Feature selection is the process of selective features that would have a positive influence on the predictability of the chosen model [9]. It enhances the classification accuracy by ridding the data of unwanted and unnecessary features [10].

Feature selection algorithms have been used before. A 2-gram or 2-kmer extracted protein feature was used training and test data for the classification of protein sequences into corresponding superfamilies of globin, insulin, kinase, ras, and trypsin. A SGERD-based model (steady state genetic algorithm for extracting fuzzy rules from data) was used to predict the classification. The classification accuracy was 96.45% [11].

The goal of this paper is to test if the Naïve Bayes classifier can predict the serotypes of the influenza A virus by feeding it subsets of selected protein sequences. Datasets of H1N1, H2N2, H3N2, and H5N1 of the HA proteins from the virus were fused into one larger dataset. These serotypes of the virus were from human, avian and swine. It is a multiclassification system with 11 different serotypes. Additionally, the Naïve Bayes classifier was used in a binary classification. A dataset of the human's influenza A virus HA proteins of the H1N1 versus the swine's H1N1 serotype was created.

The scientific literature has various insights on novel classification software that have performed well on text classification. The feature selection concept has performed well. Since sequences of nucleic acid or amino acids can be considered as long strings and are therefore text, the algorithm can be used in classification of the influenza virus serotypes.

This paper is organized as follows. In the introduction, background information is given about bioinformatics and a little brief anecdote of the transfer of virus to host gene. The influenza A virus is introduced. In Section 2, the materials and methods are laid out. The proposed algorithms of feature selection and classification are carefully explained. In Section 3, we describe the experimental results using the influenza A datasets. Section 4 gives a thorough discussion of the results obtained from the previous section. Section 5 concludes the paper and future work is presented.

II. Materials and Methods

The methodology process involves individual task of equal importance. It starts with the pre-processing of the data to select the most qualified parts. This selection is used in the training process. It follows sequence encoding in which a sequence is assigned a unique number. From the sequence encoding, the most prevalent features are chosen to be trained using the Naïve Bayes classifier. Fig I illustrates the process below.

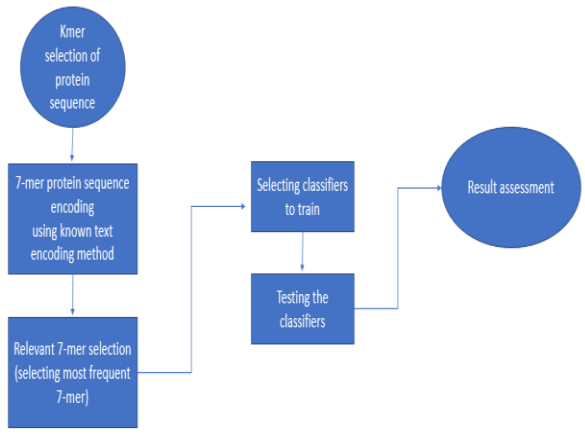


Fig. 1. The proposed methods of classifying the protein sequences

A. Selection of Data from the NCBI Database

The National Center for Biotechnology Information (NCBI) has an extensive database of the different strains of the influenza virus. The protein sequences are all arrange in a fasta format. For this research, the subtypes H1N1, H2N2, H3N2, H5N1 of the human, avian, and swine of the HA proteins of influenza A virus were chosen. Several of each subtype's peptide sequences were placed into one dataset. Using the three hosts, there are a total of 11 different subtypes of the virus Fig. II. There are 13,366 sequences divided into 11 serotypes.

	Serotype	Sequences
0	H1N1	DTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDNHNGKLCLLKG...
1	H1N1	DTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDNHNGKLCLLKG...
2	H1N1	DTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCLLKG...
3	H1N1	DTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLENSHNGKLCLLKG...
4	H1N1	DTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCLLKG...
...	...	...
13362	Av-H5N1	MERIVIALAIISIVKGDQICIGYHANNSTEQVDTIMEKNVTVTHAQ...
13363	Av-H5N1	MERIVIALAIISIVKGXXDQICIGYHANNSTEQVDTIMEKNVTVTH...
13364	Av-H5N1	MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQ...
13365	Av-H5N1	MEKIVLLLAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQ...
13366	Av-H5N1	MEKIVLLFAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQ...

Fig II. Dataset consisting of the subtype (serotype) and the different sequences.

B. Trainingand Testing the System

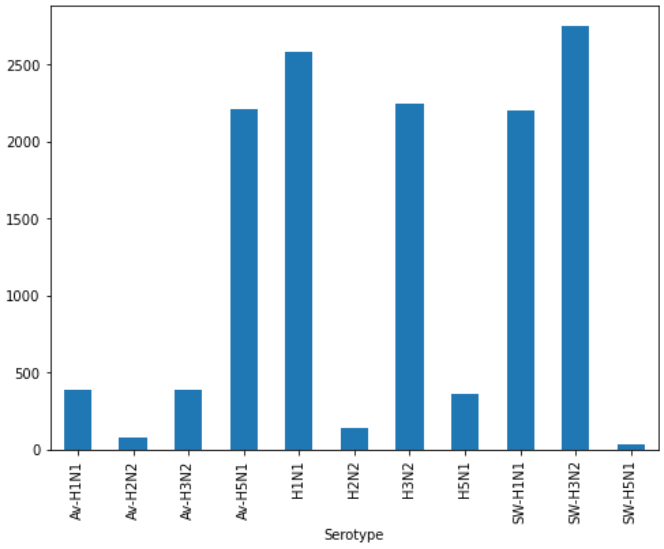
The data in Table I are trained to recognize patterns and to ensure there is a high accuracy and efficiency of algorithms which are utilized to train the data. This system uses a low percentage of datasets in the training phase. The testing phase uses the remaining datasets to gather predictions from the trained model. 80% of the datasets is used on the testing phase. These two phases utilize the N-gram or K-mer counting algorithm on the protein sequences.

C. K-mer counting of protein subsequences

As previous stated, the genetic information starts with DNA, then through the process of transcription, become RNA. The nucleic acid RNA is then translated into proteins. Proteins are composed of 20 amino acids. Consider a protein sequence B of

length Z, '$X_1X_2X_3…X_Z$', where X_z ε {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}; z

Table 1. Distribution of the influenza A virus serotypes.



= 1,2,…Z. A K-mer is a string of consecutive k amino acids in the protein sequence B [12]. The goal of K-mer or N-gram counting is to count the frequency of subsequences (k-mers) from protein sequences. The high frequency of k-mers is used to build a set of related families of peptide features. A 7-kmer is used in this paper.

There are 11 serotypes used in this paper. What makes these serotypes different than one another? The difference must be related to the frequency of 1-kmer, or 2-kmer, et., therefore from the k-mers, we would be able to classify a novel influenza A virus and place it into one of the serotypes. The 7-mers were determined using a sliding window of length k: one amino acid is shifted at each iteration from position 1 to Z-k+1, until the entire sequence terminates (Fig. III.).

For a protein sequence:

```
DTICIGYHANNSTDT
DTICIGY
 TICIGYH
  ICIGYHA
   CIGYHAN
    IGYHANN
     GYHANNS
      YHANNST
       HANNSTD
        ANNSTDT
```

Fig. III. Sliding Window generating 7-mers subsequences

. Fig. IV shows the resulting dataset when the data in Fig. II undergo 'K-merization'. Each 7-mer or feature is counted to determine its frequency. The data must be converted to numeric values to in order to be trained. To accomplish this task, the well-known 'Bag of Words' algorithm is used.

The 'Bag of Words' algorithm is used for feature extraction. It describes the frequency of words in a dataset. It involves tokenizing each word or token in a protein sequence, build vocabulary from these tokens and generate vectors. The theory behind this model is data are similar if they have similar content. We would be able to discern important insights into the meaning of the dataset.

	Serotype	words
0	H1N1	[DTICIGY, TICIGYH, ICIGYHA, CIGYHAN, IGYHANN, ...
1	H1N1	[DTICIGY, TICIGYH, ICIGYHA, CIGYHAN, IGYHANN, ...
2	H1N1	[DTICIGY, TICIGYH, ICIGYHA, CIGYHAN, IGYHANN, ...
3	H1N1	[DTICIGY, TICIGYH, ICIGYHA, CIGYHAN, IGYHANN, ...
4	H1N1	[DTICIGY, TICIGYH, ICIGYHA, CIGYHAN, IGYHANN, ...
...	...	...
13362	Av-H5N1	[MERIVIA, ERIVIAL, RIVIALA, IVIALAI, VIALAII, ...
13363	Av-H5N1	[MERIVIA, ERIVIAL, RIVIALA, IVIALAI, VIALAII, ...
13364	Av-H5N1	[MEKIVLL, EKIVLLL, KIVLLLA, IVLLLAI, VLLLAIV, ...
13365	Av-H5N1	[MEKIVLL, EKIVLLL, KIVLLLA, IVLLLAI, VLLLAIV, ...
13366	Av-H5N1	[MEKIVLL, EKIVLLF, KIVLLFA, IVLLFAI, VLLFAIV, ...

Fig. IV. 7-mer distribution of each serotype

The first step of 'Bag of Words' is to create a token of each sequence. Table II uses the last of protein sequences of serotype 10 from Fig. IV. To generate a vector, we need to determine the frequency or count of each word in all the protein sequences (Table II). Following counting the occurrence of all the words, the count needs to be sorted in decreasing order as displayed in Table III.

Table II. Bag of Words: Tokenizing the subsequences

Sequence no. 13362	Sequence no. 13363	Sequence no. 13364	Sequence no. 13364	Sequence no 13366
MERIVIA	MERIVIA	MEKIVLL	MEKIVLL	MEKIVLL
ERIVIAL	ERIVIAL	EKIVLLL	EKIVLLL	EKIVLLF
RIVIALA	RIVIALA	KIVLLLA	KIVLLLA	KIVLLFA
IVIALAI	IVIALAI	IVIALAI	IVIALAI	IVLLFAI
VIALAII	VIALAII	VLLLAIV	VLLLAIV	VLLFAIV

Using the data from Table IV, a matrix or vector is created. In our example, the words in the table are the features that will dictate the classification. In Table II, there are 5 proteins sequences which are split into 5 words (from sequence 13362 to 13366. The vector is a binary vector consisting of 0 and 1; 1 means a word exists in the dataset and 0 means the word does not exist (Table V). The algorithm selects the most frequent words in the dataset and get rid of the least common ones. The matrix model is then fed to several machine learning algorithms for the classification of each serotype.

Table III. Bag of Words. Histogram

Word	Count	Word	Count
MERIVIA	2	MEKIVLL	3
ERIVIAL	2	EKIVLLL	2
RIVIALA	2	KIVLLLA	2
IVIALAI	4	VLLLAIV	2
VIALAII	2	EKIVLLF	1
KIVLLFA	1	IVLLFAI	1
VLLFAIV	1		

Table IV. Bag of Words. Sorting in descending order

Word	Count	Word	Count
IVIALAI	4	KIVLLLA	2
MEKIVLL	3	VLLLAIV	2
MERIVIA	2	KIVLLFA	1
ERIVIAL	2	VLLFAIV	1
RIVIALA	2	EKIVLLF	1
VIALAII	2	IVLLFAI	1
EKIVLLL	2		

III. Result

This paper investigated a dataset containing fasta-formatted protein sequences of different serotypes of the influenza A virus. The serotypes were from human, avian and swine species. There were H1N1, H2N2, H3N2, H5N1 from each species. Absent from the NCBI database were the H2N2 serotypes of swine species. Therefore, there were 11 different serotypes. They were chosen because they are the most common strains of the influenza A virus during annual influenza season [6].

Table V. Vector Space

Words / Sequence no.	IVIALAI	MEKIVLL	MERIVIA	ERIVIAL	RIVIALA	VIALAII	EKIVLLL	KIVLLLA	VLLLAIV
13362	1	0	1	1	1	1	0	0	0
13363	1	0	1	1	1	1	0	0	0
13364	1	1	0	0	0	0	1	1	1
13365	1	1	0	0	0	0	1	1	1
13366	0	1	0	0	0	0	0	0	0

The 11 strains of the virus were assembled into one dataset. The sequences were used to generate a series of 7-mers using a sliding window of length 7. The 7-mer subsequences are used in the 'Bag of Words' algorithm to select the most common features or subsequences in the dataset for each serotype and they were placed into vector. The system is divided into the training and testing phases. 20% of the data were used in the training phase, and 80% were used in the testing phase. The resulting data were then fed into several classification models. The models used were Naïve Bayes, decision tree , nearest neighbor and random forest. The best classifier for this experiment is illustrated in Fig. IV.

A. Model Comparison

Fig. IV displays the cross validation mean of the models chosen to determine the classification of the serotypes. The random forest classifier is the best model for this experiment. Its cross validation mean exceeds the others as shown in Table VI. This classifier is chosen for further analysis. Additionally, 20% of the dataset were used in the training phase and the remaining datasets were used in the testing phase. The training and testing accuracy of the random forest classifier (RFC) is 99% and 97%, respectively. These percentages validate the efficiency of the random forest model. The training and testing accuracy scores of the K-Nearest neighbor model were omitted because of memory allocation issues.

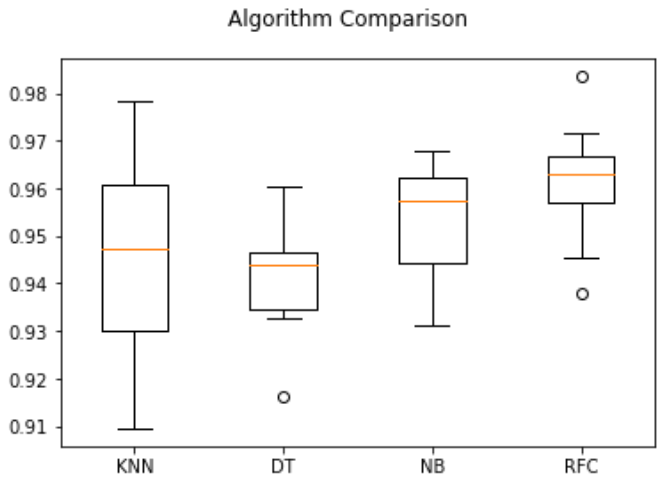


Fig. IV. Comparison of each classifier's cross validation score

B. Classification report and Model evaluation.

The goal of the classification report is to measure the efficiency of the classification model. It consists of the precision, recall, f1-score and support scores. The precision score indicates the proportion of true positives from cases that are predicted as positives. The chosen model is the random forest. In Table VII, the serotypes H5N1 (human) and Sw-H5N1 (swine) have low scores while the remaining serotypes have high scores. The recall scores display the proportion of true positives from cases that are actual positives. Again, the human and swine H5N1 serotypes have low scores. The overall accuracy of the model indicates the proportion of correct classifications from all given cases [13]. ROC is a chart that displays the number of true positives cases versus the number of false positives. The chart displays all 11 serotypes of the influenza A virus in this experiment using the random forest model (Fig V).

Table VI. Classification accuracy from generated matrix of each model

Model	Cross Validation Accuracy mean	Training Accuracy	Testing Accuracy
DT	94%	99%	96%
NB	95%	95%	95%
RFC	96%	99%	97%

Table VII. Classification report of the Random Forest model

Serotype	Precision	Recall	F1	support
H1N1	1.00	1.00	1.00	536
H2N2	1.00	1.00	1.00	27
H3N2	1.00	1.00	1.00	461
H5N1	0.72	0.53	0.61	81
Sw-H1N1	0.99	1.00	0.99	421
Sw-H3N2	0.99	0.99	0.99	547
Sw-H5N1:	0.67	0.22	0.33	9
Av-H1N1	1.00	0.98	0.99	59
Av-H2N2	1.00	1.00	1.00	17.00
Av-H3N2	0.96	0.95	0.95	74
Av-H5N1	0.90	0.96	0.93	442
Accuracy			0.97	2674
macro avg	0.93	0.88	0.89	2674
weight avg	0.97	0.97	0.97	2674

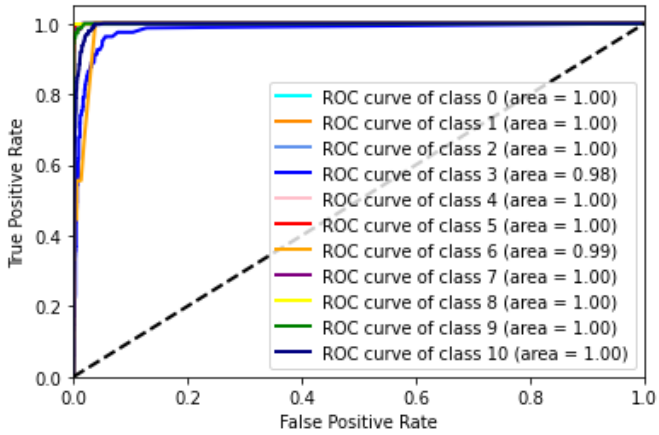


Fig V. Random Forest Model: ROC curves

IV. Discussion

The results confirm that the different models are successful in predicting the classification of serotypes of the influenza A virus. They show that the random forest model is the best model between the decision tree, naïve Bayes and K-Neighbor classifier. The random forest model classifies the data at 97% accuracy. For the 20% training phase, the accuracy was at 99%. Sherif's influenza's Profile HMMs, previously explained above, is able to classify the serotypes at an 100% accuracy, and the classification of serotypes of human versus non-human is captured at accuracies ranging from 55.5% to 97%. Our result for the classification of human's HA protein of H1N1serotype versus human's NA protein of H1N1 serotype stands at 100% accuracy. The true label 0 represents the human's HA protein of the virus and 1 represents the NA protein of the human influenza virus (Fig VI).

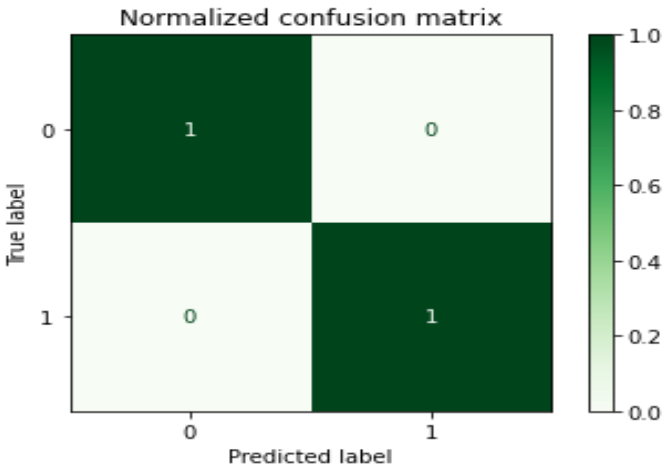


Fig. VI. Normalized Confusion matrix: Human HA H1N1 (0) Vs Human NA H1N1 (1)

References

[1] N. Luscombe, D. Greenbaum and M. Gerstein, "What is Bioinformatics? A Proposed Definition and Overview of the Field", *Methods of Information in Medicine*, vol. 40, no. 04, pp. 346-358, 2001. Available: 10.1055/s-0038-1634431.

[2]F. CRICK, "Central Dogma of Molecular Biology", *Nature*, vol. 227, no. 5258, pp. 561-563, 1970. Available: 10.1038/227561a0.

[3] J. Claverie and C. Notredame, *Bioinformatics for Dummie*, 2nd ed. 2007.

[4] B. Alberts, A. J. Lewis, et al, A. Johnson, J. Lewis and e. al, *Molecular Biology of the Cell*, 4th ed. New York: Garland Science, 2002.

[5] F. CARRAT and A. FLAHAULT, "Influenza vaccine: The challenge of antigenic drift", *Vaccine*, vol. 25, no. 39-40, pp. 6852-6862, 2007. Available: https://doi.org/10.1016/j.vaccine.2007.07.027.

[6] F. Sherif, Y. Kadah and M. El-Hefnawi, "Classification of human vs. non-human, and subtyping of human influenza viral strains using Profile Hidden Markov Models", *2011 1st Middle East Conference on Biomedical Engineering*, pp. 221-224, 2011. Available: 10.1109/mecbme.2011.5752105 [Accessed 4 April 2020].

[7] G. Lu et al., "GenomeBlast: a web tool for small genome comparison", *BMC Bioinformatics*, vol. 7, no. 4, p. S18, 2006. Available: 10.1186/1471-2105-7-s4-s18.

[8] S. Eddy, "Profile hidden Markov models", *Bioinformatics*, vol. 14, no.9, pp. 755-763, 1998. Available: 10.1093/bioinformatics/14.9.755.

[9] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013, p. 204.

[10] M. Iqbal, I. Faye, B. Samir and A. Md Said, "Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics", *The Scientific World Journal*, vol. 2014, pp. 1-12, 2014. Available: 10.1155/2014/173869.

[11] E. Mansoori, M. Zolghadri and S. Katebi, "Protein Superfamily Classification Using Fuzzy Rule-Based Classifier", *IEEE Transactions on NanoBioscience*, vol. 8, no. 1, pp. 92-99, 2009. Available: 10.1109/tnb.2009.2016484.

[12] J. Wen, R. Chan, S. Yau, R. He and S. Yau, "K-mer natural vector and its application to the phylogenetic analysis of genetic sequences", *Gene*, vol. 546, no. 1, pp. 25-34, 2014. Available: 10.1016/j.gene.2014.05.043 [Accessed 4 April 2020].

[13] D. Bužić and J. Dobša, "Lyrics classification using Naive Bayes", *Electronics and Microelectronics (MIPRO*, vol. 201841, pp. 1011-1015., 2018. [Accessed 4 April 2020].