Review on Application of Machine Learning Algorithm in DNA Sequence Classification

[1]S. Bavankumar*, [2]Dr. V. Rathikarani
[1]Research Scholar, [2]Assistant Professor
[1,2]Department of CSE,
Annamalai University, Chidambaram, India
*Corresponding Author: sbavankumarcse@smec.ac.in

Abstract

DNA (Deoxyribonucleic acid) is a large molecule found in living things. Its main purpose is to store information. As sequencing technology has improved, DNA sequence data has grown at a very fast rate. This has put the study of DNA sequences in the big data wave. Also, machine learning is a powerful method for analyzing big amounts of data and learning on its own. It has been used a lot to analyze DNA sequence data and has led to a lot of study successes. First, the review explains how sequencing technology has changed over time. It also goes into detail about DNA sequence data structure and sequence resemblance. Then look at the basic steps of data mining, summarize a few of the most important machine learning algorithms, and talk about the problems that machine learning algorithms face when mining biological sequence data, as well as some possible future answers. Then, we look at four common ways machine learning is used with DNA sequence data: DNA sequence alignment, DNA sequence classification, DNA sequence grouping, DNA pattern mining, with help of DNA in Forensics. looked at the history and importance of their biological applications, and put together a list of how the field of DNA sequence data mining has changed and what problems might come up in the future.

DNA, found in most of our cells, is unique to each person and leaves a trail everywhere we go. This aids forensic investigators who use DNA to identify crime scene victims and suspects. This review discussed genetic markers in forensics and their limitations.

Keywords: DNA sequence, machine learning, data mining, DNA sequence alignment, DNA sequence classification, DNA sequence clustering, DNA pattern mining, DNA forensics.

1. INTRODUCTION

We are living in the era of the genome, in which scientific advancements have made it possible for humans to investigate the unsolved mysteries of life. In recent decades, one of the most notable aspects of the evolution of molecular biology has been the rapid proliferation of biological data, and a huge biological information database has swiftly evolved as a result of this expansion. It became clear that we needed to get actionable insights from the vast amounts of data at our disposal, and at the same time, the field of bioinformatics emerged. The field of bioinformatics draws from many other fields of study. It makes extensive use of mathematics, the life sciences, and computer science in order to extract biological information from biological data, and it also directs the relevant research efforts of biological researchers.

The first step is to analyze the DNA sequence of the genome to learn more about the protein-coding region of the genome. The next stage is to run a simulation and create a best guess about the protein's three-dimensional structure. Using the protein's function as a starting point, the researchers develop the final drug design.

According to available data, the amount of biological information doubles approximately every 18 months. The initial nucleic acid sequence database developed by GenBank in 1982 contained 606 sequences and 680,000 nucleotide bases. As of February 2013, its database contains 162 million biological sequence data, totaling 150 billion nucleotide bases. How to extract knowledge from vast amounts of data in order to guide biological research is one of the most essential topics of study in bioinformatics.

It is important to do two things when dealing with complex biological data: (a) ensure that the data accurately reflect the true meaning of biology by finding a solution to the challenge of storing and managing enormous amounts of data; and (b) extract valuable information from the data. Application of machine learning is crucial to the development of artificial intelligence. It is utilized extensively in bioinformatics and can govern machine learning without explicit programming.

In living organisms, DNA can be considered a biomacromolecule. It directs the growth of biological development and the operation of life activities because it carries and provides the genetic information necessary for life.

Increasing data processing capacities and generating pertinent biological information are just two of the many potential applications of machine learning that are currently being investigated in the study of sequence data. This review focuses on data mining and machine learning techniques that can be applied to DNA sequences. This section provides a brief history of the evolution of sequencing technology, describes the DNA sequence data structure, and introduces several sequence encoding methods used in machine learning. In addition, we emphasize that sequence similarity is the foundation of data mining on DNA sequences. We have summarized the most prevalent machine learning algorithms and reviewed the data mining process in depth.

In DNA sequence data, common applications of machine learning include sequence alignment, sequence classification, sequence clustering, and sequence grouping, as well as pattern mining. The following generalizations are a result of our research. Parallel computation and distributed sequence alignment may soon dominate DNA sequence alignment in the laboratory. Finding the optimal method to represent sequence features and evaluate DNA sequence categorization is a significant challenge in the scientific community. This is a difficult circumstance. Effective clustering requires an understanding of how to extract distinct subsequences from a DNA sequence.

Mining DNA sequence patterns will result in a significant increase in the number of potential sequence patterns, which will consume a considerable amount of time and space. The development of an appropriate search method and the elimination of superfluous repetitions in sequence patterns will be a major research focus in the coming years.

2. BASIC KNOWLEDGE OF DNA

DNA, or deoxyribonucleic acid, is a molecule that contains the genetic information required for an organism to develop, survive, and reproduce. These instructions are encoded in the DNA of all living things and are transmitted down through the generations.

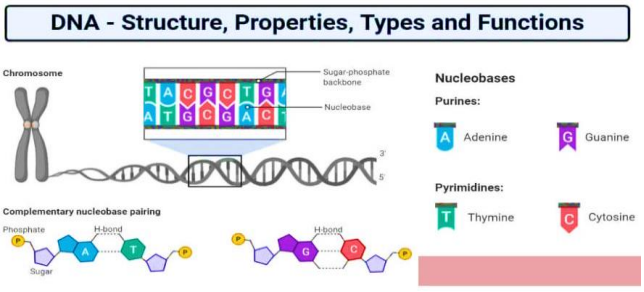


Fig 1. DNA Structure

2.1 The structure of DNA

Nucleotides are the building components utilized to create DNA. Every nucleotide's structure includes a phosphate group, a sugar group, and a nitrogen base. The four types of nitrogen bases are adenine (abbreviated as A), thymine (abbreviated as T), guanine (abbreviated as G), and cytosine (abbreviated as C).

The sequence of these bases determines the instructions for DNA, also known as the genetic code. The order of nitrogen bases in a DNA sequence can generate genes in the same manner as the order of alphabetic letters. Genes are the "words" that instruct cells on how to construct proteins; they are produced by reading the DNA sequence from top to bottom. Ribonucleic acid (RNA) is a variety of nucleic acid responsible for translating DNA's genetic information into proteins. Approximately 20,000 genes and 3 billion nucleotides constitute the human genome.

The double helix structure consists of two intertwined and interconnected elongated strands of nucleotides. If one were to imagine the structure of a double helix as a ladder, the phosphate and sugar molecules would represent the vertical sides of the ladder, while the bases would represent the horizontal rungs. The nucleotide adenine is observed to make a complementary base pair with thymine in the context of DNA, whereas guanine forms a complementary base pair with cytosine. When these nucleotides are located on opposite strands of the DNA molecule, this pairing occurs.

DNA molecules are so long that they cannot fit inside cells without undergoing a specialized packaging process. It is necessary for DNA to be organized into compact structures known as chromosomes in order for it to fit within cells. Every chromosome contains a single DNA molecule. Each human cell comprises a nucleus containing 23 chromosomal pairs, also known as chromosomal pairs.

2.2 The discovery of DNA

In 1869, a German scientist named Frederich Miescher made the groundbreaking discovery that DNA exists. However, for a long time, scientists didn't realize the molecule was so important. In 1953, James Watson, Francis Crick, Maurice Wilkins, and Rosalind Franklin deduced that DNA is shaped like a double helix. This realization led them to the conclusion that DNA may be used to relay data about biological mechanisms. The discovery of the molecular structure of nucleic acids and its

implications for the transmission of information in living beings earned Watson, Crick, and Wilkins the 1962 Nobel Prize in Medicine.

2.3 DNA sequencing

Researchers are able to determine the order in which bases appear in a DNA sequence thanks to a technological advancement known as DNA sequencing. It is possible to use this method to determine the order of bases in chromosomes, genes, or the entire genome. In the year 2000, scientists successfully completed the sequencing of the human genome for the very first time in its entirety.

2.4 DNA testing

Your DNA can provide information about the medical history of your family as well as your own potential vulnerability to specific conditions. DNA tests, which are often commonly referred to as genetic tests, are carried out for a variety of purposes, some of which include the diagnosis of genetic illnesses, the identification of carriers of genetic mutations who may pass those abnormalities on to their offspring, and the evaluation of a person's susceptibility to hereditary diseases. For instance, a genetic test can identify if an individual has a mutation in the BRCA1 or BRCA2 gene, which is associated to an increased risk of breast and ovarian cancer. Another example is that the test can determine if an individual has the BRCA1 or BRCA2 gene. Because the results of a genetic test could have an impact on a person's overall health, it is standard practice to provide genetic counseling in conjunction with these types of examinations.

2.5 DNA Sequence Similarity

Data mining for DNA sequences is routinely carried out from these vantage points, and the comparison of sequence similarities is an essential part of research carried out in these fields. There is little question that sequence similarity plays a significant role in the process of data mining on DNA sequences.

When referring to different sequences, what is indicated by the word "sequence similarity" is the occurrence of points within those different sequences that are equal to or identical to one another. Evaluating the degree of resemblance between two sequences could take the form of a numerical value or an illustrative description. Both of these approaches have their advantages and disadvantages. If the degree of similarity between two sequences is greater than thirty percent, then it is

determined that the sequences have a homologous relationship. This is done by comparing the sequences and looking at the degree of similarity between them.

Therefore, if there is a considerable degree of similarity between the two sequences, it is likely that they share an evolutionary ancestor because of the similarities that they share. At the same time, additional predictions can be made regarding the function of the unidentified sequence if a similar sequence can be found among sequences that have already been established to have a particular usage.

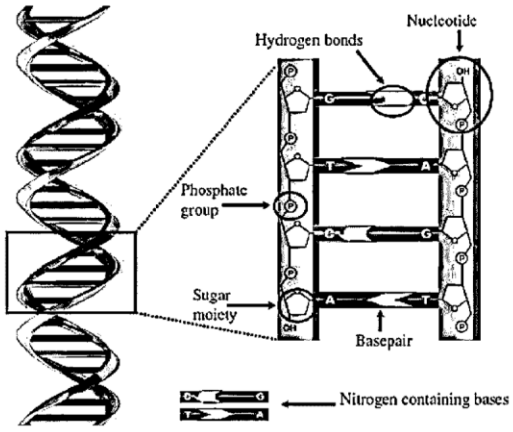


Fig 2. Double Helix of DNA

In bioinformatics and computational biology, the comparison of DNA sequences for similarities is the cornerstone of sequence analysis. Analyzing gene function, predicting protein structure, and retrieving sequences are a few examples of disciplines that require similarity computations. When selecting a method for analyzing sequence similarity, it is essential to consider both the application's specific requirements and the context of the biological data. In addition, we intend to enhance the chosen strategy so that it meets these requirements. This concept serves as the foundation and impetus for DNA sequence data extraction.

3. APPLICATION OF MACHINE LEARNING IN DNA SEQUENCE

Within the larger field of computer science, machine learning is an essential subfield that must be studied. On the one hand, machine learning paves the way for the extraction of The comparison of two or more sequences requires a fundamental process known as sequence alignment. During this process, the arrangement of the bases in each sequence is analyzed in a particular order. Changing sequences that have clearly stated goals into sequences that have unclear or hazy goals. The findings of the alignment also show how similar biological characteristics and sequences are to one another to a certain degree. Within the field of bioinformatics, the study of sequence alignment is an important and necessary subject to investigate. The application of sequence

alignment analysis can result in the acquisition of supplementary information concerning the structural and functional features of biological sequences. The presence of gene recombination and mutation in the biological world has a negative impact on the ability of DNA to repair and duplicate itself, which in turn has an effect on the dynamics of evolutionary change. The utilization of sequence alignment analysis enables the evaluation of the degree of similarity that is present among DNA sequences, which in turn enables the investigation of evolutionary links.

Other sequence alignment types include double sequence alignment and multi-sequence alignment. The simultaneous alignment of multiple sequences is known as multi-sequence alignment. Over time, aligning a growing number of sequences becomes increasingly difficult. As the study of biological sequence alignment has advanced, many other sequence alignment tools, including CLUSTAL, TCOFFEE, and MUSCLE, have been developed. We compared three comparable DNA sequences using the CLUSTAL program. There are 25 pairs of bases that are precise matches on the graph. This 46-base segment is a portion of a significantly lengthier sequence. Over 54.35 percent sequence similarity indicates that local similarities are probable between the three sequences. This is the most fundamental conceivable comparison. Comparing sequences in the real world is significantly more difficult.

Early biological sequence alignment research used dual sequence alignment. The Needleman-Wunsch algorithm uses dynamic programming to align comparable sequences. Known as the global sequence comparison method and optimization matching algorithm. Dynamic programming was developed by Smith and Waterman (1981) to locate sequence segments with high local similarity. Sadly, the Smith-Waterman algorithm compares slowly. Find the most identical nucleotides by finding the longest common substring between two DNA sequences. Dynamic programming must calculate the double sequence alignment score matrix before extracting aligned text. We compared two DNA strands of different lengths using CLUSTAL. Figure 5 shows local comparisons. Blanks must be added to maximize common bases because the DNA sequences have different base counts. Ideal matches include 25 bases and substantial local commonalities.

5. DNA SEQUENCE ALIGNMENT

Comparing two or more sequences with regard to the organization of their bases is what is involved in the process of sequence alignment. The primary objective of sequence alignment is to find a match between sequences the functions of which are unknown and sequences the functions of which are known.

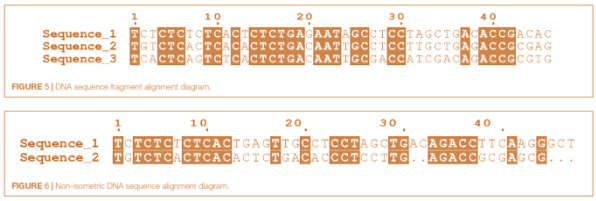
The alignment results further emphasize the biological commonalities between the sequences. When it comes to bioinformatics, sequence alignment is at the very top of the list of fundamental and important issues to be studied. Sequence alignment analysis considerably improves one's capacity to predict the structure and function of biological sequences. Genetic recombination and mutation have been shown to be the result of DNA evolution by biological investigations, however this process has not recovered or duplicated. However, in evolutionary studies, sequence alignment analysis can be used to look for genetic commonalities.

Double and multiple sequence alignment are under the umbrella term of sequence alignment. Aligning two sequences results in an alignment of several sequences. The more sequences there are to align, the more challenging the process becomes. Methods for aligning biological sequences such as CLUSTAL, TCOFFEE, and MUSCLE have emerged as a result of years of research. We compared three similar DNA sequences with CLUSTAL. Figure 5 displays the outcomes of these regional comparisons. In this diagram, the red region indicates the three sequences that are an exact match. There are 25 identical bases displayed in the diagram. The 46-base-type sequencing fragment. Local commonalities were suggested by the 54.35% similarity between the three sequences.

Dual sequence alignment was the starting point for research on biological sequence alignment. This occurred very early on. Needleman-Wunsch is a typical algorithm for sequence alignment. It is also known as the global sequence comparison method and the optimization matching algorithm. Needleman and Wunsch aligned dual sequences using techniques based on sequence similarity and dynamic programming. Smith and Waterman (1981) redesigned dynamic programming as a local optimal algorithm. This allows the algorithm to locate sequence fragments with a high degree of local similarity. A disadvantage of the Smith-Waterman algorithm is its slow pace of comparison. Finding the longest common substring is required when looking

for the greatest matching base number between two DNA sequences.

Identifying the longest common substring between two sequences will accomplish the task. To obtain the matched text, it is necessary to first ascertain the score matrix of the double sequence alignment and then employ the dynamic programming technique. We used the CLUSTAL instrument to compare two DNA sequences of vastly different lengths. Figure 6 depicts the regional distribution of the comparison's outcomes. Since the two DNA sequences are of differing lengths, gap filling is required to determine the number of shared bases. This will be accomplished by locating the maximum number of matched bases. All 25 bases must be identical, and there must be a striking degree of cellular similarity as well.



6. DNA SEQUENCE CLASSIFICATION

Classification is an essential mining operation in machine learning.

The objective is to predict the types of unobserved samples by learning a classification model from the training sample set. Data mining continues to struggle with the identification of biological sequences as a distinct data type. Due to the non-numerical characteristics of the biological sequence elements, the sequence interaction between sequence components, and the varying sequence durations of distinct events, this is a difficult problem to solve. The objective of sequence classification is to aid in the identification of genes within DNA molecules by predicting the type of DNA sequence based on structural or functional similarity, the function of the sequence, and the relationship of the sequence to other sequences.

The classification of DNA sequences using DAWGs was first proposed by Levy and Stormo (1997). Muller and Koonin (2003) proposed that DNA sequences could be classified using vector space. Ranawana and Palade (2005) proposed a multi-classifier strategy for the identification of E. coli promoter sequences in genomic DNA. He First, the sequence is encoded using one of four possible encoding methods, and then it is used to train one of four possible neural networks. Using an aggregation algorithm, the categorization results from the four individual neural networks were then combined.

For this purpose, a modified form of the logarithmic opinion pool procedure was utilized. Experiments have demonstrated that the same data provided to a neural network using various encoding methods can yield somewhat distinct results. Due to these findings, insights into the data are feasible. In contrast, by combining the outputs of multiple classifiers trained on the same input data into a multi-classifier, we can obtain better results than the performance of a single neural network. Obtaining the correct network parameters can be a significant challenge, which is a significant obstacle when constructing neural networks. You will need to deploy the neural network and fine-tune your encoding strategy to achieve your goal.

Recent years have seen the emergence of convolutional neural networks as one of the most popular deep learning models. Convolutional neural networks excel at the extraction of abstract features from data. Nguyen et al. (2016) proposed a new method for DNA sequence classification using convolutional neural networks. As a result of using DNA sequences as text data in their investigation, they were able to devise their technique.

This method employs a vector with a singular endpoint to represent the sequence used as the model's input. The information associated with each nucleotide in the fundamental position sequence is therefore preserved. Twelve distinct DNA sequence data sets were utilized in the evaluation of the model. According to the findings, the predictive potential of the model across all of these data sets has increased significantly.

7. DNA SEQUENCE CLUSTERING

Cluster analysis, one of the numerous machine learning techniques, is widely employed. This differs from the classification in that the specific categories are unknown beforehand. Cluster analysis is a technique for identifying relationships between unsupervised data. The DNA clustering method is founded on an analysis of sequence similarity. Cluster analysis is a technique for investigating the biological functions of DNA sequences by clustering sequences with similar characteristics. The central problem in DNA sequence classification is determining whether or not two sequences are comparable. The results of categorizing DNA sequences are influenced by a variety of parameters, and much of the current research on clustering DNA sequences relies on the local properties of DNA. The development of a clustering algorithm that takes into account

the global characteristics of DNA sequences would significantly improve the accuracy of clustering and the subsequent study of DNA sequence clusters.

Krause et al. (2000) and Enright et al. (2002), pioneering international scholars, developed the SYSTERS method and the GENERAGE algorithm, respectively. The two algorithms share the fundamental concept of completing sequence clustering using a hierarchical clustering method based on the calculated similarity between sequences. Based on graph theory, Gerhardt et al. (2006) proposed an approach for clustering DNA sequences. This technique employs a triplet network to discover the route structure of the genome.

This network model is comprised of DNA triplets as its vertices. Physical proximity between two vertices on the genome indicates a relationship between those vertices. Then, we examine the cluster's topology to gain a greater understanding of the network's structure. In the end, two major distinctions pique his interest: the GC content and the periodicity of the DNA sequence's base pairs. In order to accomplish this, he has created synthetic test data consisting of DNA sequences and investigated a clustering technique that employs a synthetic random network. The result demonstrates that the clustering coefficient is scientifically beneficial. Wei D. proposed a novel method for grouping non-aligned DNA sequences, which he named mBKM. This algorithm's foundation was the innovative distance measure DMk. Using this technique, the DNA sequence is converted into a feature vector. This technology enables the transformation of DNA sequences into feature vectors. The DNA sequence is described by these feature vectors in terms of the frequency, position, and ordering of k-tuples. Wei et al. (2012) discovered that the mBKM technique is effective at identifying DNA sequences with similar biological properties and locating connections between them. However, edge length was not taken into account, and the method has not been modified to address issues resulting from excessively long insertions or repeated sequences.

Extraction of characteristic subsequences from the DNA sequence and development of an efficient similarity measure based on the biological significance of the sequence are the two most important aspects of DNA sequence clustering at present. Incorporating the two previously described fundamental factors into the algorithm design for clustering DNA sequences will increase the applicability of clustering results.

8. GENOME AND FORENSIC GENETICS

Any DNA locus intended for forensic genetics should have the following properties:

- A high level of polymorphism is desirable.
- Should be easy and cheap to characterize.
- It ought to be simple to grasp and evaluate in relation to other labs.
- should have a low mutation rate.

The 3.2 billion base pairs (BPs) that make up the genome can be studied in any detail because to recent advances in molecular biology technologies. Originating in a Biological Process The three most important parts of this process are collection, characterization, and storage.

Multiple Biological Data Sets With the exception of RBCs, the vast majority of the trillions of cells that make up the human body have nuclei. Each nucleated cell has two identical copies of an individual's genome and can be used to generate a DNA fingerprint. It is expected that samples may degrade to some degree; in extreme circumstances, more cellular material may be needed to create a DNA profile.

Nucleated cell biological samples, such as: •

- Liquid blood or dry deposits, are required for forensic genetic testing.
- Saliva, sperm, or solid deposits.
- Bone and teeth are examples of hard tissues.
- Follicular hair.

8.1 Collection and Handling of Material at the Crime Scenes

It is widely believed that whole blood is one of the most common sources of DNA. For the first 5-7 days, it is stored at 4°C in ethylenediamine tetraacetic acid, which prevents blood coagulation. After processing, DNA samples are stored at either -20°C for a few weeks or -80°C for longer periods. At crime scenes, epithelial cells are collected using a sterile bud or brush. After harvesting, they are stored in a paper or plastic bag at ambient temperature and humidity. The investigation of a crime scene requires caution, including the preservation of evidence and the use of protective gear such face masks and body suits. Serious issues can arise from careless handling of evidence. Cross-contamination is a major source of evidence degradation,

which can throw off conclusions or alter findings.

9. OPEN ISSUES

The examination of DNA sequences allows researchers to learn more about the genetic diversity of living things. DNA sequence analysis has been in high demand because of the exponential development of DNA sequence data. The following issues persist at this time in DNA sequence data mining:

1. Large-scale DNA sequence data processing still faces efficiency hurdles;
2. Algorithms for mining DNA sequence data should be tailored to specific biological tasks by taking into account relevant prior information and sequence properties.
3. The ability to efficiently analyze sequence similarity requires knowledge of how to extract features from DNA sequences and develop appropriate similarity measures.
4. The "black box" aspect of machine learning makes it hard to provide a credible explanation for machine learning's output from a biological standpoint, hence restricting the model's applicability.

CONCLUSION

Hardware has come a long way in the last few decades, giving researchers in fields like omics, biological imaging, medical imaging, etc. new tools with which to collect data. Nonetheless, advances in the life sciences have resulted in a serious problem. Data mining techniques are now being investigated for their potential utility in bioinformatics analysis. Data mining architecture, machine learning techniques, and cutting-edge analytic capabilities for processing biological data are all part of this field. The growth of machine learning has also been aided by the merging of formerly separate disciplines. Artificial neural networks, deep learning, and reinforcement learning are to thank for the progress made by machines in the field of AI. As computational power, data storage speed, and costs have improved, researchers in a wide variety of sectors have been able to evaluate previously inaccessible biological data. Using machine learning and genomics in mining will produce more applicable findings, which will contribute to societal development.

We believe machine learning research in DNA sequence analysis should focus on two things:

It discusses the biological significance of DNA sequences. There are numerous algorithms that can analyze DNA patterns, but their mining outcomes are sensitive and specific, so they vary significantly. Thus,

everyone must investigate how to incorporate the biological significance of DNA patterns into data mining.

As data increases, however, conventional analysis tools become slower to calculate. Important is the study of efficient computation methods. Spread computing and parallel computing increase the efficiency of mining.

Choosing the appropriate DNA sequence coding method for a given task is essential. This can enhance the algorithm and reduce instructional time.

This review concludes with a discussion of sequencing technique, DNA sequence data structure, sequence similarity, source, and characteristics. We also discuss machine learning algorithms briefly and suggest biological sequence data.

The difficulties encountered by machine learning techniques in the mining industry, as well as potential solutions. The discussion then turned to four prevalent applications of machine learning to DNA sequence data: DNA sequence alignment, classification, clustering, and pattern mining. We reviewed the history of these techniques, discussed their significance in the biological community, and summarized the most recent research findings from around the globe. We discussed a number of pressing issues, as well as possible new developments and trends in the field of DNA sequence data extraction. Future connections between the biological field and machine learning, I believe, will enhance mining outcomes.

References:

[1] Bilofsky, H. S., Burks, C., Fickett, J. W., Goad, W. B., Lewitter, F. I., Rindone, W. P., et al. (1986). The GenBank genetic sequence databank. Nucleic Acids Res. 14, 1–4. doi: 10.1093/nar/14.1.1

[2] Bosco, G. L., and Di Gangi, M. A. (2016). “Deep learning architectures for DNA sequence classification,” in Proceedings of the International Workshop on Fuzzy Logic and Applications (Cham: Springer), 162–171. doi: 10.1007/978-3-319-52962-2_14

[3] Chen, L., and Liu, W. (2011). “An algorithm for mining frequent patterns in biological sequence,” in Proceedings of the 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) (Piscataway, NJ: IEEE), 63–68. doi: 10.1109/ICCABS.2011.5729943

[4] Choong, A. C. H., and Lee, N. K. (2017). “Evaluation of convolutionary

neural networks modeling of DNA sequences using ordinal versus one-hot encoding method," in Proceedings of the 2017 International Conference on Computer and Drone Applications (IConDA) (Piscataway, NJ: IEEE), 60–65. doi: 10.1109/ICONDA.2017.8270400

[5] Chowdhury, B., and Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 109, 419–431. doi: 10.1016/ j.ygeno.2017.06.007

[6] Chu, W. W. (2014). Data mining and knowledge discovery for Big Data. Stud. Big Data 1, 305–308. doi: 10.1007/978-3-642-40837-3 Delibas, E., and Arslan, A. (2020). DNA sequence similarity analysis using image texture analysis based on first-order statistics. J. Mol. Graph. Model. 99:107603. doi: 10.1016/j.jmgm.2020.107603

[7] Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584. doi: 10.1093/nar/30.7.1575

[8] Gerhardt, G. J. L., Lemke, N., and Corso, G. (2006). Network clustering coefficient approach to DNA sequence analysis. Chaos Solitons Fractals 28, 1037–1045.doi: 10.1016/j.chaos.2005.08.138

[9] Henikoff, S., and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919. doi: 10.1073/pnas. 89.22.10915

[10] Huo, H. W., and Xiao, Z. W. (2007). A multiple alignment approach for DNA sequences based on the maximum weighted path algorithms. Ruan Jian Xue Bao(Journal of Software) 18, 185–195. doi: 10.1360/jos180185

[11] Jangam, S. R., and Chakraborti, N. (2007). A novel method for alignment of two nucleic acid sequences using ant colony optimization and genetic algorithms. Appl. Soft Comput. 7, 1121–1130. doi: 10.1016/j.asoc.2006.11.004

[12] Junyan, Z., and Chenhui, Y. (2015). "Sequence pattern mining based on markov chain," in Proceedings of the 2015 7th International Conference on Information Technology in Medicine and Education (ITME) (Piscataway, NJ: IEEE), 234–238. doi: 10.1109/ITME.2015.49

[13] Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al.

(2006). Machine learning in bioinformatics. Brief. Bioinform. 7, 86–112. doi: 10.1093/bib/bbk007

[14] Lee, Z. J., Su, S. F., and Chuang, C. C. (2008). Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. Appl. Soft Comput. 8, 55–78. doi:10.1016/j.asoc.2006.10.012

[15] Levy, S., and Stormo, G. D. (1997). "DNA sequence classification using DAWGs," in Structures in Logic and Computer Science, eds

[16] J. Mycielski, G. Rozenberg, and A. Salomaa (Berlin: Springer), 339–352. doi: 10.1007/3540-63246-8_21

[17] Li, J., Wong, L., and Yang, Q. (2005). Guest editors' introduction: data mining in bioinformatics. IEEE Intell. Syst. 20, 16–18. doi: 10.1109/MIS.2005.108

[18] Ma, Q., Wang, J. T. L., Shasha, D., and Wu, C. H. (2001). DNA sequence classification via an expectation maximization algorithm and neural networks: a case study. IEEE Trans. Syst. 31, 468–475. doi: 10.1109/5326.983930

[19] Mao, G. (2019). "Association matrix method and its applications in mining DNA sequences," in Proceedings of the International Conference on Applied Human Factors and Ergonomics (Piscataway, NJ: IEEE), 154–159. doi: 10.1007/978-3- 030-20454-9_15

[20] Mendizabal-Ruiz, G., Román-Godínez, I., and Torres-Ramos, S. (2018). Genomic signal processing for DNA sequence clustering. PeerJ 6:4264. doi: 10.7717/peerj.4264

[21] Mondal, S., and Khatua, S. (2019). "Accelerating pairwise sequence alignment algorithm by mapreduce technique for next-generation sequencing (ngs) data analysis," in Emerging Technologies in Data Mining and Information Security,eds

[22] A. Abraham, P. Dutta, J. Mandal, A. Bhattacharya, and S. Dutta (Cham: Springer), 213–220. doi: 10.1007/978-981-13-1498-8_19 Müller, H. M., and Koonin, S. E. (2003). Vector space classification of DNA sequences.

[23] J. Theor. Biol. 223, 161–169. doi: 10.1016/S0022-5193(03)00082-1
Naznin, F., Sarker, R., and Essam, D. (2011). Vertical decomposition with genetic algorithm for multiple sequence alignment. BMC Bioinformatics 12:353. doi:10.1186/1471-2105-12-353

[24] Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R.,et al. (2016). DNA sequence

classification by convolutional neural network.

[25] J. Biomed. Sci. Eng. 9:280. doi: 10.4236/jbise.2016.95021 Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. Curr. Protoc. Bioinform. 42, 18. doi: 10.1002/0471250953.bi0301s42

[26] Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85, 2444–2448. doi: 10.1073/pnas.85.8.2444

[27] Ranawana, R., and Palade, V. (2005). A neural network based multi-classifier system for gene identification in DNA sequences. Neural Comput. Appl. 14, 122–131. doi: 10.1007/s00521-004-0447-7

[28] Rogozin, I. B., Milanesi, L., and Kolchanov, N. A. (1996). Gene structure prediction using information on homologous protein sequence. Comput. Appl. Biosci. 12, 161–170. doi: 10.1093/bioinformatics/12.3.161