Implementation of Machine Learning Technique for Fast and Reliable DNA Sequence Classification

Kshatrapal Singh[1*][0000-0002-5965-0783], Raja Sarath Kumar Boddu[2]

Primary affiliation
[1, 2,]School of Computer Engineering, Lincoln University College, Petaling Jaya, 47810, Malaysia
pdf.kshatrapal@lincoln.edu.my
Secondary affiliation
[1]Department of CSE, KCC Institute of Technology and Management, Greater Noida, 201308, India,
mekpsingh1@gmail.com
[2]Department of AI&ML, Raghu Engineering College, Vishakhapatnam, 531162, India
iamrajaboddu@gmail.com

Abstract

DNA sequence data is currently growing at an exponential rate due to advancements in sequencing methods, which has also thrust DNA sequence research into the big data revolution. A variety of strong computer algorithms known as machine learning (ML) may create predictive models by intelligently and autonomously analyzing enormous amounts of frequently unstructured data. It has achieved many experimental successes and is commonly utilized in the analysis of DNA sequence data. Since DNA contains most of an organism's genetic information, it can be used to classify DNA sequences and identify diseases at an earlier stage. This explains why biological computation places a high value on the grouping of DNA sequences. This research has proposed a method for classifying DNA sequences using data obtained from the NCBI. This work proposes a new method for feature extraction from DNA sequence that employs hot vector matrix, as well as a machine learning-based classifier. Each word pair in the hot vector that represents the DNA sequence is denoted by a binary matrix that shows where each nucleotide is located in the sequence. After that, the final matrix is fed into a conventional CNN in order to extract features.

Keywords

DNA sequence, Machine learning, Biological Computation, Decision trees, Support vector machine (SVM).

1. Introduction

In the era of the genome, scientific advancements have made it possible for people to gather information on the secrets of living things. One important aspect of the evolution of molecular biology in the past few decades has been the rapid proliferation of biological data, leading to the establishment of a vast biological information collection very quickly. These massive amounts of data must be used to extract meaningful expertise, and bioinformatics emerged at the same time. The field of bioinformatics is multidisciplinary [1]. It fully mines biological details from biological data using arithmetic, computational science, and life sciences, and it additionally directs the pertinent biological investigation of researchers. In particular, the genomic DNA sequence analysis is the first step in obtaining knowledge about the protein coding region. Then modelling and forecasting the protein's spatial structure. Lastly, the scientists create the required medication design based on the protein's functions [2].

Research on biology has demonstrated that gene sequences are not unordered, randomized strings. They are made up of smaller components arranged in a straight line. There are four different types of deoxyribonucleotides (bases) that link the DNA sequence. A factor in DNA molecule variation is base order [3]. Data pertaining to DNA sequences varies from other types of data in several ways, mostly because of:

- The letters A, T, C, and G are non-numeric in the sequence of DNA data;
- Data from DNA sequences has unique biological significance;
- Several sequences varied significantly in length. whereas certain sequences consist of just handful characters, another can reach hundreds of megabytes in length;
- Prior to data analysis, suitable data preprocessing must be carried out due to specific sequencing method errors and noises in the sequence dataset.

In order to create a matrix for inputs model training, the sequence of genes must be processed by first converting the string sequence into a numerical value. Sequential encoding, k-mer encoding, and one hot encoding are the three broad techniques for sequence encoding. Sequential encoding performs similarly to one-hot encoding in terms of efficiency, but it requires a lot less training period. Deep learning techniques frequently employ one-hot encoding, that pairs well with CNN (convolutional neural networks) techniques [4]. Furthermore, one-hot encoding performs quite consistently across various data sets; however, excellent efficiency necessitates the use of an appropriate CNN. In certain assessment data sets, ordinal codes expressed as matrices exhibit the greatest performance. The appropriate architecture of sequence encoding and representations determines how well CNN finds DNA patterns [5]. The single-point coding approach can yet be improved, as seen by the ordinal coding approach's strong performance.

2. Literature Review

The following section contains an overview of relevant work on the subject of DNA sequence categorization. Table 1 lists the research projects completed by various researchers and provides explanation to frequently asked questions about the utility of DNA sequence grouping and the various techniques that can be applied to it.

Table 1: Overview of previous work done.

Sr. No.	Approaches	Details
1.	DNA sequence analysis with an expectation–maximization algorithm and neural networks. [7, 9]	In the publication, a novel technique for identifying E. Coli supporters in DNA and determining in case a given DNA sequence contains E. Coli supporters is presented. Since it lowers the probability distribution of lengths, EM algorithm—which is superior to earlier algorithms—is used in this work to locate binding sites in the E. Coli promoter sequence. Following the identification of binding locations, each sequence's characteristics will be chosen based on the information richness at hand, and the orthogonal encoding methodology will be used to describe the results. Lastly, characteristics are added to a neural network for promoter prediction.
2.	Complementary analysis approaches for protein sequences. [8, 12]	This research examines five different protein classification techniques and applies them to associated proteins included in the PROSITE catalogue. Four of them (the block-based technique) rely on database searches, while the

		other rely on recurring patterns found within a protein group. They have come to the conclusion that the block-based technique is thought to be the most appropriate method when discussing amino acids occurring in blocks, and that utilizing all five of these strategies can produce a decent categorization outcome.
3.	Analysis of gene expression data using fuzzy logic. [10]	Microarray methods have made it possible to estimate multiple aspects of level of gene expression simultaneously. These levels that are produced can also be used to categorize tissues into prognostic or diagnostic groups. Developing a straightforward, technology-neutral classification approach can be highly helpful, since estimates from different microarray inventions are generated on different scales. This study uses two examples to demonstrate how fuzzy logic might be helpful in capturing problems related to the grouping of gene expression data. Fuzzy inference performs identically as other classifiers in terms of categorization, though it is more straightforward and understandable.
4.	Vector space classification of DNA sequences. [11]	The problems with intron and exon identifying are highlighted in this work. PCA is used to classify DNA sequences. Sequences are transformed into textual vectors in order to depict word content. Lastly, the grouping of sequences is based on these word contents. When it comes to the categorization of introns and exons, this method has been evaluated on numerous DNA data sets and has the greatest degree of precision in relation to other methods.
5.	A neural network-based multiple-classifier system for gene identification in DNA sequences [12].	With the aim of identifying the promoter sequence (E. coli), this paper suggests a neural network-based multiple-classifier scheme. This is due to the fact that each gene has a promoter sequence, hence finding the E. coli promoter is crucial for the precise determination of DNA sequence genes. This multi-classifier system outperforms previous systems that have been established thus far in terms of prediction, according to testing conducted on a variety of promoter and non-promoter genes.

3. Suggested Methodology

As far as we are aware, FASTA-based DNA sequences are text-based; yet, they are composed entirely of consecutive letters without any spaces, indicating that no words are included in them. To have it function similarly as CNN on text, we must translate the sequences into words. Every nucleotide maintains its place in the DNA sequence as a result of the conversion. The illustration below, Figure 1, helps clarify this idea. In this case, we are utilizing a window with a constant size of 4 and fixed strides of 1. Each time a window scans an ongoing DNA sequence, it is interpreted as a word.

Ultimately, word sequences have been deduced from DNA data. With this data, the textual-based CNN approach is able to be used.

As far as we are aware, there are just 4 letters in the FASTA DNA file form: A, C, G, and T. There are 256 unique terms in a dictionary if the window size of 4 is selected. This implies that a 256-size hot vector can be used to express each word. Ultimately, a 2 D matrix that preserves each nucleotide's position was created from the produced words of the DNA sequence.

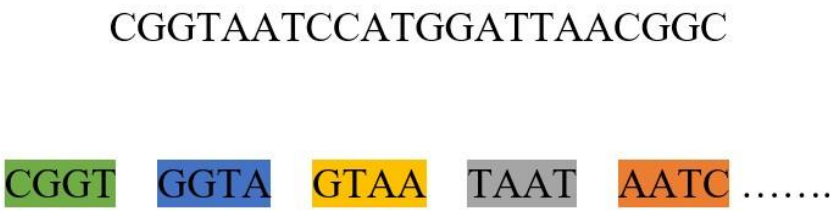

Fig. 1: Word sequence derived from DNA data.

Each time four successive words, like as CGGT, GGTA, GTAA, TAAT, and AATC, are selected with the slide of stride 1 and placed to the target sequence, as in the illustration above. With the 256 words (44) in our dictionary, we can now create a 2 D matrix of numerical values by representing each word with its appropriate vector, as seen in Figure 2.

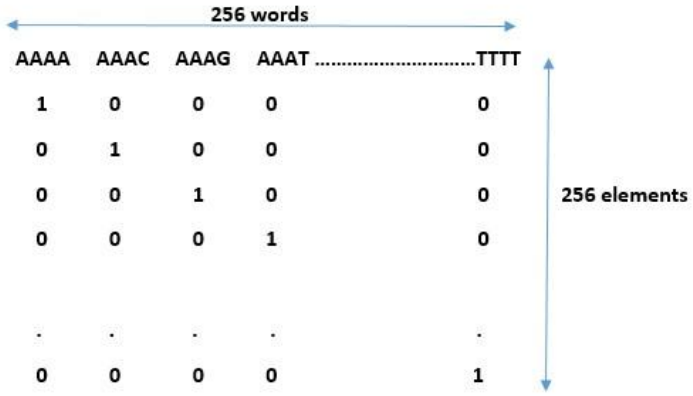

Fig. 2: Dictionary of words consists of DNA bases.

Right now, a vector is constructed for the specified DNA sequences employing this dictionary. Figure 3 shows a suitable example of vector representation for some words. Early detection of fatal viruses such as Corona can aid in the earlier development of medications and assist stop the pandemics that we are now experiencing. This explains why one of the main functions of genome sequence categorization in the field of computational biology. The present research proposes a framework wherein, upon a patient's visit to a physician, DNA samples, whether from a virus or another disease, are gathered and compared to a database to determine the illness's existence. Yet, it gets challenging to detect them as there are not enough patterns accessible for different disorders.

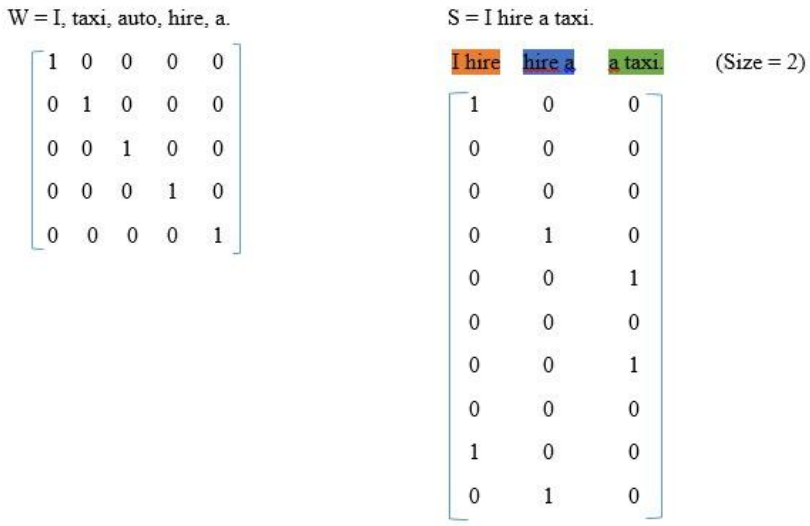

Fig. 3: Illustration of Vector representation.

Since FASTA-based genetic sequences are known to be excessively large and complicated, feature extraction cannot be performed directly on this data. Within this research, a new vector-based numerical structure is presented, where the location of nucleotide is retained by utilizing binary 0 or 1. This was necessary because the data required to be transformed into a similar numerical form. After that, the vector matrix is fed into a conventional CNN in order to extract features. Next, test data is used to train and assess the models.

4. Result with Discussion

For assessment reasons, we have employed a python driven tool that can conduct feature extraction as well as classification of DNA samples. In the present research, we analyzed five prominent classifiers based on a number of factors: decision trees, recurrent neural networks (RNN), K-Nearest Neighbor (KNN) method, support vector machines (SVM), and convolution neural networks (CNN). Eighty features were selected through the Chi-square selection approach, the Zscore method for feature normalization, and the K-mer feature descriptor along a K-mer size of three. The K-means technique was implemented to cluster the data, and the count of clusters is three. Four distinct parameters have been calculated and compared: Sensitivity (Sn), Specificity (Sp), Precision (P) and Accuracy (Acc). They are defined as follows:

$$Sn = 1 - \frac{FN}{TP} \quad (1)$$

$$Sp = 1 - \frac{FP}{TN} \quad (2)$$

$$Acc = 1 - \frac{FN + FP}{TP + TN} \quad (3)$$

$$P = 1 - \frac{FP}{TP + FP} \quad (4)$$

Where FN indicated false negatives, FP refers false positives, TP is true positives and TN gives true negatives. The following definitions apply to terminology like true positive, true negative, false positive, and false negative. "True + ve" refers to the positive (+ ve) result predictions that the model correctly made. "True -ve" refers to the accurate negative (-ve) result predictions made by the model. The model's inaccurate optimistic (+ ve) result predictions are referred to as "false + ve." The model's inaccurate negative (-ve) result predictions are referred to as "false -ve."

Accuracy in multiclass categorization is described as:

$$\text{Acc} = 1 - \frac{FN\ (i) + FP\ (i)}{TP\ (i) + TN(i)} \qquad (5)$$

The instances included in this class are represented by false negative (i), false positive (i), true positive (i), and true negative (i). One well-liked multiclass approach is the Naive Bayes method. On the basis of these factors, we have contrasted our approach with alternative test data categorization methods. For evaluation and analysis of the suggested approach with different machine learning techniques k- fold cross-validation is employed, at which k = 4.

Table 2: Parameters result with CNN.

Parameters	F0	F1	F2	F3	F4	Average
Sn	70	65	70	60	60	65
Sp	90	95	95	100	90	94
Acc	78	78.5	68	69	76	73.9
P	93.1	94.2	88.6	89.7	91.3	91.38

Table 3: Parameters result with SVM.

Parameters	F0	F1	F2	F3	F4	Average
Sn	60	60	65	60	65	62
Sp	100	95	95	100	90	96
Acc	58	52	61	64	59	58.8
P	89.2	91.2	84	86.6	87.8	87.8

Table 4: Parameters result with KNN.

Parameters	F0	F1	F2	F3	F4	Average
Sn	85	75	70	85	85	80
Sp	95	80	90	75	70	82
Acc	76.5	88	78.5	80	78	80.2
P	81.2	76.2	93.6	77.9	67.5	79.3

Table 5: Parameters result with Decision Tree.

Parameters	F0	F1	F2	F3	F4	Average
Sn	60	70	65	65	70	66
Sp	60	65	65	70	75	67
Acc	58.5	67.5	64	63.5	64.5	63.6
P	56.7	67.2	65.2	64.3	69	64.5

Table 6: Parameters result with RNN.

Parameters	F0	F1	F2	F3	F4	Average
Sn	65	55	50	48	45	52.6
Sp	100	95	85	95	90	93
Acc	74	76.5	71	66	63.5	70.2
P	79	91	92	99	98	91.8

Table 7: Parameters result with proposed method.

Parameters	F0	F1	F2	F3	F4	Average
Sn	65	60	55	50	60	58
Sp	95	100	95	90	95	95
Acc	96	96.5	95	97	97.5	**96.4**
P	94.6	93.8	98.6	97.5	96.2	**96.1**

According to the calculation above, CNN provides an accuracy with 73.9, Decision Tree provides an average accuracy with 63.6, KNN provides an accuracy of 80.2, RNN provides an accuracy of 70.2, SVM provides an accuracy of 58.8, and the suggested approach provides the highest accuracy 96.4 when compared to the other techniques for classification as listed in Tables 2, 3, 4, 5, 6 and 7.

5. Conclusion

This research has proposed a technique for the classification of DNA sequences using data obtained from the NCBI. When DNA sequences are categorized, different diseases' patterns are revealed. Such patterns are then evaluated against trials from currently infected individuals and can aid in the pre-diagnosis of disease. Such DNA sequences are too large and complicated to be provided directly for feature extraction; instead, they must be translated into a comparable numerical format. This work presents a novel vector-based numerical model in which each nucleotide's place is reserved through binary numbers 0 or 1. After that, the vector matrix is fed into a conventional CNN in order to extract features. Regarding four different parameters Specificity, Accuracy, Precision, and Sensitivity a comparison of roughly six classifiers including Convolution neural network (CNN), Support Vector Machines (SVM), K-Nearest Neighbor (KNN) approach, Decision Trees, Recurrent Neural Networks (RNN), and the suggested technique was conducted.

References

1. Ibrahim OAS, Hamed BA, El-Hafeez TAbd. A new fast technique for pattern matching in biological sequences. 2023; 79(1):367–88.
2. Roy A, Chakraborty SJRE, Safety S. Support vector machine in structural reliability analysis: A review 2023: p. 109126.
3. Jäger J, Krems RVJNC. Universal expressiveness of variational quantum classifiers and quantum kernels for support vector machines. 2023. 14(1): p. 576.
4. Dragomir MP, Calina TG, Perez E, Schallenberg S, Chen M, Albrecht T, Koch I, Wolkenstein P, Goeppert B, Roessler SJE. DNA methylation-based classifier differentiates intrahepatic pancreato-biliary tumours 2023. 93.
5. Andrade-Girón D, Carreño-Cisneros E, Mejía-Dominguez C, Velásquez-Gamarra J, Marín-Rodriguez W, Villarreal-Torres H. R.J.E.E.T.o.P.H. Meleán-Romero, and Technology, support vector machine with optimized parameters for the classification of patients with COVID-19. 2023. 9: p. e8–e8.
6. Shorabeh SN, Samany NN, Minaei F, Firozjaei HK, Homaee M, Boloorani ADJRE. Decis model based Decis tree Part swarm Optim algorithms identify optimal locations solar power plants Constr Iran. 2022; 187:56–67.

7. Manoharan A, Begam K, Aparow VR, J.J.o.E D. Artificial neural networks, gradient boosting and support Vector Machines for electric vehicle battery state estimation: a review. 2022. 55: p. 105384.
8. Zhang H, Zou Q, Ju Y, Song C, Chen DJCB. Distance-based support vector machine to predict DNA N6-methyladenine modification. 2022. 17(5): p. 473–82.
9. Costa VG, C.E.J.A.I R, Pedreira. Recent advances in decision trees: An updated survey 2023. 56(5): p. 4765–4800.
10. Lee CS, Cheang PYS, J.A.i.D M. Predictive analytics in business analytics: decision tree. 2022; 26(1):1–29.
11. Kshatrapal Singh, Ashish Kumar, Manoj Kumar Gupta, Modified k-string in composition vector method for DNA sequence comparison based on maximum entropy principle, Journal of Interdisciplinary mathematics, ISSN 0972-0502, vol. 23 (1), pp. 31-41, 2020.
12. Sarkar S, Mridha K, Ghosh A, Shaw RN. Machine Learning in Bioinformatics: New Technique for DNA Sequencing Classification, in Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022. 2022, Springer. p. 335–355.