https://doi.org/10.48047/AFJBS.6.11.2024.1285-1297

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper

Open Access

IMPROVING DNA GENETIC CODE SIMILARITY EVALUATION THAT USES MACHINE LEARNING AND THE ENHANCED LONGEST COMMON SUBSEQUENCE (ELCS) ALGORITHM

Dr. T. Sarathamani[1], Bablu Pramanik[2], Mamani Bandyopadhyay[3],

[1,2,3]Assistant Professor, Department of Computer Science and Engineering, Brainware University, Kolkata, India

Article Info

Volume 6, Issue 11, July 2024

Received: 23 May 2024

Accepted: 20 June 2024

Published: 09 July 2024

Doi:10.48047/AFJBS.6.11.2024.1285-1297

ABSTRACT:

Applications of DNA sequence analysis in biology and computer science, including gene discovery, evolution research, and genetic illness diagnosis. Finding commonalities across codes in various domains is essential. When comparing DNA sequences, conventional techniques like the subsequence (LCS) algorithm are frequently employed. But in order to clarify the difficulties in assessing DNA similarity sequence, this research study presents a method that makes use of machine learning techniques and the Enhanced Longest Common Subsequence Algorithm (ELCS). The suggested ELCS algorithm combines the strength of data-driven models with the efficiency of sequence alignment. It predicts alignment scores using a trained machine learning model, which lessens workload while keeping accuracy high. The Enhanced LCS Algorithm (ELCS) was implemented and deployed with Support Vector Machines using the NCBI GenBank nucleotide sequence dataset.

Keywords: DNA Genetic Codes, Enhanced Longest Common Subsequence Algorithm (ELCS), Similarity Evaluation, Diagonal Edges, DNA Sequencing, Machine Learning

1. INTRODUCTION

Machine learning, which is generally categorized as artificial intelligence, has demonstrated remarkable ability in a variety of fields. Machine learning has emerged as a crucial tool in genomics to help decipher the complex structure of DNA sequence data.

DNA sequence similarity analysis is a very encouraging approach that entails the examination of genomic sequences to discern similarities, conserved patterns, and plausible functional components. The rapid progress in the fields of genomics and bioinformatics has resulted in an unparalleled increase in the production of biological data, providing invaluable insight on the underlying mechanisms that regulate life. One of the significant obstacles encountered by researchers in this field is to the extraction of valuable knowledge from extensive collections of DNA sequences. In order to decipher the enigmatic information embedded in these molecular blueprints, researchers have progressively relied on machine learning methodologies, effectively using algorithms to identify and analyze patterns, correlations, and anticipatory observations that could otherwise remain concealed.

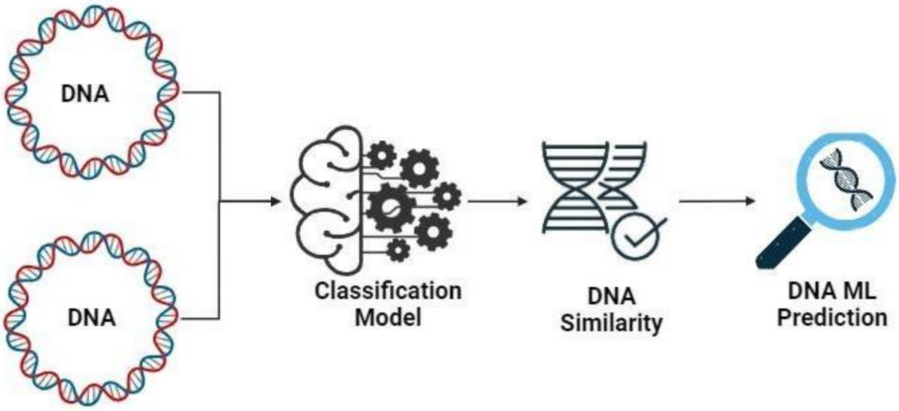


Figure 1.1. DNA Sequence Machine Learning Classification Model

This research article aims to investigate the mutually beneficial association between machine learning and the examination of DNA sequence similarity. This study explores the theoretical foundations of both fields, elucidating the concepts that empower machine learning algorithms to analyze extensive genomic datasets efficiently and accurately. With the exponential growth of genomic data, there is a potential for unlocking new levels of understanding in biology and biomedicine through the combination of machine learning techniques with DNA sequence similarity. Our ultimate objective is to stimulate more investigation and ingenuity, cultivating a more profound recognition of the revolutionary capacity of machine learning in the realm of genomics research. The analysis of DNA sequences is a fundamental task in biological research and has far-reaching implications in various domains, such as evolutionary studies, gene discovery, and genetic disease diagnosis. DNA processing refers to the utilization of biological molecules rather than traditional silicon chips for completing computational tasks. The concept of employing individual particles, including molecules, for computational purposes traces its origins back to 1959, when physicist Richard Feynman of the United States first presented his ideas on nanotechnology. However, the recognition of DNA processing did not occur until 1994, when Leonard Adleman, an American computer scientist, provided a demonstration of how particles may be effectively employed to address computing problems. A calculation might be considered the execution of a calculation, which itself might be characterized as a bit by bit rundown of obvious guidelines that takes a few information, processes it, and produces an outcome. In DNA registering, data is addressed utilizing the four-character hereditary letters in order A - Adenine, G - Guanine, C - Cytosine, and T - Thymine, as opposed to parallel letter set (1 and 0) utilized by customary PCs. A calculation's feedback is thusly addressed (in the least difficult case) by DNA particles with explicit groupings, the directions are done by research facility procedure on the particles, (for example, arranging them as per length or hacking strands containing a specific aftereffect), and the outcome is characterized as some property of the last arrangement of atoms (like the presence or nonappearance of a particular succession).

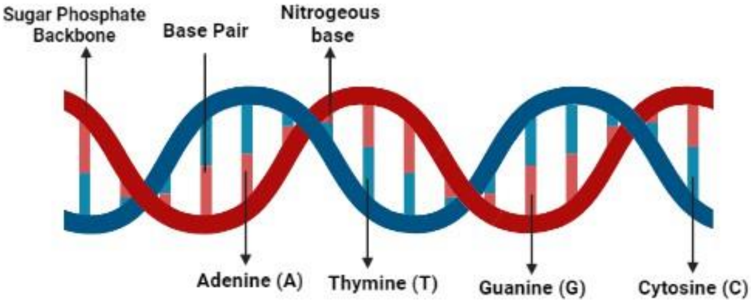


Figure 1.2. double helix structure of DNA

However, in most difficult problems the quantity of potential arrangements develops exponentially with the size of the issue (for instance, the quantity of arrangements could two-fold for each town added). This implies that even somewhat little issues would require unmanageable volumes of DNA (on the request for huge baths) to address every single imaginable response. Adleman's trial was critical on the grounds that it performed limited scope calculations with organic atoms. More significantly, in any case, it opened up the chance of straightforwardly customized biochemical responses. One of the key challenges in this field is to determine the degree of similarity between genetic codes, which aids in understanding evolutionary relationships, identifying functional regions, and uncovering genetic variations associated with diseases. Traditional sequence alignment methods, such as the LCS Algorithm, have been widely employed for comparing DNA sequences and detecting similarities. The algorithm is designed to analyze the sequence that is shared by two provided sequences and has the maximum length. The ease and efficacy of DNA sequence comparison have made it a widely favored option. To overcome these challenges, recent advancements in machine learning have shown promise in various bioinformatics applications. Machine learning models can capture complex patterns and dependencies in DNA sequences, enabling efficient and accurate analysis. In this context, integrating machine learning techniques with the LCS algorithm can enhance its performance by reducing computational requirements and improving alignment accuracy. The design of DNA comprises of two long entwined stands that structure the well-known twofold helix structure as displayed in Figure 1.2. Each strand is worked from a little arrangement of constituent particles called nucleotides. A nucleotide comprises of three sections. The initial two sections are utilized to shape the lace like spine of the DNA strand, and are indistinguishable in all nucleotides.

2. LITERATURE SURVEY

In this paper, the authors used variety of machine learning approaches, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Random Forest classifier, Adaboost, Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), to sequence the DNA on a dataset of human origin. The objective of this analysis is to enhance our comprehension and draw conclusions on the comparative performance of these algorithms [27]. The sequences have been split into substrings of a certain length, where the "k value" determines the length of the substring. This approach is one method for analyzing the sequence [5]. The paper discusses ML algorithms to differentiate DNA sequences based on their nucleotide sequence. [29] Emre Delibas and Ahmet Arslan [1] performed a Dim level surface were made by the qualities doled out to the nucleotides in the DNA groupings. In this article, the Needleman-Wunsch (NW) algorithm was used for sequence alignment and achieved an accuracy of 99.7% with a multilayer perceptron. Aligning sequences means finding sequence similarity between DNA sequences [28]. Closeness estimations were made between these surfaces utilizing histogram-put together surface examinations based with respect to first-arrange measurements. The surface highlights for 3 distinct DNA informational collections of various lengths, and determined the likeness grids. The article introduces a rapid and effective approach for the discovery of DNA sequence similarity. However, it does not explicitly address the utilization of machine learning techniques in the study. In this study, the authors Wang et al. introduced a novel strategy including a 2D k-mer count matrix, which draws inspiration from the CGR technique. They further refine this matrix by evaluating neighboring elements and afterwards evaluate similarities using a mix of pairwise distance (PD) and phylogenetic tree approaches that yield the most accurate results[29]. Authors [2] mentioned examples of the old style LCS issue to identify the portion, variations, and occurrences. In addition, procedure to decrease the size of the subsequent charts. At last, a thorough exploratory assessment utilizing late definite and heuristic most extreme inner circle solvers is introduced. Weiyang Chena, Bo Liao,Weiwei Li [3] proposed a technique to utilizing similitude distance framework can be figured and connections from the measured highlights, It has been shown that the DNA categorization of individuals exhibits the highest level of entropy and the lowest amount of energy. As one progresses from the human species to chimpanzees, orangutans, gorillas, and other distinct species, there is a decrease in entropy and an increase in energy. Authors [4] developed a unique programming calculation that can accurately register a LCAIS between any two successions with rehashed components in O(nm) space, Jiaoyun Yang, Yun Xu, Yi Shang [5] developed a productive equal calculation to tackling LCS issues on GPUs. By changing the information reliance in the score table utilized by unique programming, the calculation empowers more serious level of parallelism and accomplishes a decent speedup on. Costas S. Iliopoulos, M. Sohel Rahman [6] find an answer for Inflexible Fixed Hole LCS to $O(n^3)$. Outstandingly, in each of the above cases, we accept that the two given strings are of equivalent length for example n. However, our outcomes can be effectively reached out to deal with two strings of various length. B.Lavanya, A.Murugan [7] developed and implemented MLCS in a profoundly equal manner, and can be reached out to numerous different information mining applications moreover. In future, tackling even more constant issues in sub-atomic biology is conceivable. Authors [8] proposed a calculation is fundamentally quicker than the best existing successive strategies, arriving at up to 2-3 significant degrees quicker speed on enormous size issues. At last, we present a productive equal execution of the calculation. Assessing the equal calculation on a benchmark set of both irregular and natural successions uncovers a close direct speedup concerning the consecutive calculation.[9] analyzed DNA groupings similitude metric is one of the central issues of bunching. The arrangement free strategy is an extremely well-known method for computing DNA succession closeness. It ordinarily changes over a grouping into a component space considering words' likelihood dissemination as opposed to straightforwardly matches strings. Authors[10] A sophisticated framework has been designed for the purpose of tracking the locations of k-mer throughout the count network. The methodology is implemented across six distinct datasets. The high degree of performance is achieved for two benchmark datasets, including AFproject. Specifically, a 100 percent accuracy is achieved for two datasets, namely 16 S Ribosomal and 18 Eutherian. Furthermore, notable advancements are made in terms of improve performance use when compared such as HEV and HIV-1. In this article, research focuses on the utilization of a neural network method to forecast diseases based on DNA sequence analysis, specifically by considering the GC content within the sequence. Additionally, a medical picture registration technique is employed to track the progression of the predicted diseases. [30].

3. PROPOSED METHODOLOGY

In order to evaluate the similarity of two genetic codes, this study uses the Enhanced Longest Common Subsequences (ELCS) technique inside the Proposed Methodology. The time complexity of this approach is O(n2). This technique uses a new approach to combine edge computation with DNA sequencing, which leads to better results in terms of performance and temporal complexity. This study proposes a method for predicting the similarity assessment of genetic code sequences using a machine learning model. The goal is to use machine learning techniques to precisely determine how similar the two sequences are. Several machine learning methods for classification were compared in this work, including Random Forest, K-Means Clustering, Support Vector Machine, Naive Bayes Classifier, Logistic Regression, Stochastic Process, and XG Boost.

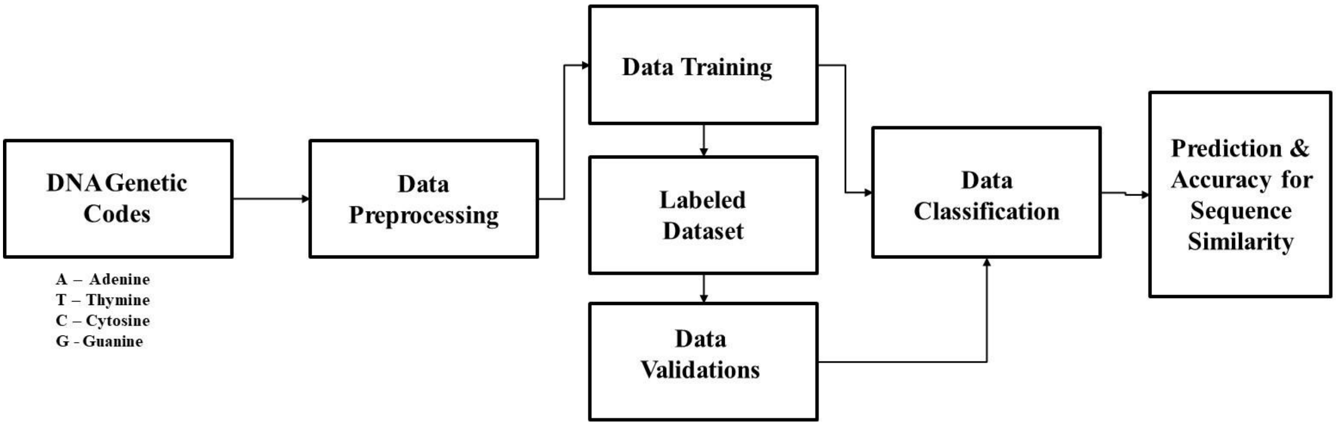


Figure 1.3. Systematic Representation of Machine Learning SVM Classification for Similarity Evaluation of DNA Genetic Codes

In Figure 1.3 states that Systematic Representation of Machine Learning SVM Classification for Similarity Evaluation of DNA Genetic Codes, specifically Support Vector Machine (SVM) classification, can be effectively utilized for similarity evaluation of DNA genetic codes. SVM classification offers a powerful approach to compare and classify DNA sequences based on their genetic similarities. It is a step-by-step overview of the data training process for sequence similarity. This ML Model to perform a variety DNA Genetic Codes to find the Similarity Evaluation, Prediction and Accuracy between DNA. The Dataset DNA genetic codes (A - Adenine, T-Thymine, G-Guanine, C-Cytosine) is labeled by an expert to identify the different types of sequences present in the dataset. After the Preprocessing techniques, machine learning model is trained on the labeled data to learn how to classify the DNA sequences. The performance of the trained model is assessed by evaluating it on a separate validation dataset. The trained model is utilized for the purpose of predicting the degree of similarity between sequences. The evaluation of the model's correctness is conducted by comparing its predictions to the ground truth labels of the validation dataset.

DNA Sequencing and Data Preprocessing

In this Methodology, 10000 DNA Genetic Code Data Samples is used to produces raw data in the form of nucleotide sequences. These sequences can be quite lengthy, so they are typically transformed into fixed-length vectors by encoding the nucleotides (A, C, G, T) into numerical representations or converting them to k-mer representations (substrings of length k). This process converts each DNA sequence into a numerical format suitable for feeding into machine learning classification algorithms.

Feature Extraction

In Feature Extraction, Once the dataset is converted into nucleotide sequence using Data Preprocessing techniques. This process will extract the features from the dataset. In this extraction process of DNA Nucleotide is implemented into Enhanced Longest Common Subsequence Algorithm (ELCS) and improves performance and the time complexity of O(n^2).

Data Labeling

In this methodology, ELCS is deployed in the NCBI DNA Genetic Codes and finds the similarity evaluation of two DNA Genetic codes subsequences. It involves associating each DNA sequence with a specific class or category. For instance, the sequences could be labeled as functional or non-functional process, etc.,

Training the model

In this Training model, Machine Learning Classification Model is developed, NCBI Dataset is tested and trained the model using different classification algorithm. This feature depicts the labeled training data in order to acquire knowledge about the data and associations intermediate and data is extracted with associated class. Support Vector Machine (SVM) algorithm is designed to identify an ideal hyperplane that effectively divides distinct classes within the feature space.

Evaluation Model:

On Comparison of Different Classification Algorithm, Support Vector Machine (SVM) model has been trained and its show good results with the prediction and accuracy of 98%, it is subjected to evaluation using the set of test data. This evaluation aims to determine the model's ability to reliably categorize previously unknown DNA sequences. In Evaluation Model, SVM Algorithm and ELCS is trained with 10000 NCBI Dataset DNA Genetic Codes and it improves the performance in terms of time complexity and similarity evaluation.

Experimental Result

Dna Similarity Evaluation Of Elcs And Machine Learning Classification Algorithm

The Enhanced Longest Common Subsequences (ELCS) Algorithm is employed in the experimental result to examine the similarity evaluation of the DNA genetic codes, and it increases the algorithm's efficiency by reducing its time complexity to O(n2). The machine learning model is integrated with the results of the ELCS Algorithm and similarity evaluation to forecast the corresponding DNA sequence patterns. Several machine learning classification techniques, including Support Vector Machine, Naive Bayes Classifier, Random Forest, K-Means Clustering, Stochastic Process, Logistic Regression, and XG Boost Algorithm, are used to assess the model. The performance of two machine learning techniques, Support Vector Machine (SVM) and XG Boost Algorithm, is compared in this study. The results show that both algorithms offer better accuracy results and predictive power.

Table 1.1. DNA Similarity Evaluation of Enhanced Longest Common Subsequences (ELCS) and Machine Learning Classification Algorithm

Sample 1	**Genetic Codes**	**Sample 2**	**Genetic Codes**	**Matching Sequence**	**% of similarity**
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	CHIMPANZEE	GTCAGATTTG GGGGA TGCTTCTGGCT C------	GTCACGATTTGG GGGAT GCTTCTGGCTC---- --A-	99.8
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	GORILLA	GTCACGATTT GGGGG ATGCTTCTGG CTC ----- A-	GTCAGATTTGGG GGATG CTTCTGGCTC------	99.7
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	ORANGUTAN	GTCACGATTT GGGAG ATGCTTCTGG CTC---- G-	GTCACGATTTGG GGGAT GCTTCTGGCTC---- --A-	91.67
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	BABOON	GTCAGAATTT GGGGG ATGCTTCTGG CTC-----T-	GTCACGATTTGG GGATG CTTCTGGCTC------	88.89
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	MACAQUE	GTCAGAATTT GGGGG ATGCTTCTGG CTC-----T-	GTCAGATTTGGG GGATG CTTCTGGCTC------	88.89
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	VERVET	GTCAGAATTT GGGGG ATGCTTCTGG CTC-----T-	GTCAGATTTGGG GGATG CTTCTGGCTC------	88.89
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	MOUSE-LEMUR	ATCACAG- TTGGGGGATG CCACT GGCCT-----C-	GTCAGATTTGGG GGATG CTTCTGGCTC------	75
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	LEMUR	ATCACAA- TTGGGGG- TGCCACGGTC CT------ C-	TCACGTTGGGGG ATGCC TGGCT---- --	69.44
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC----- -A-	RABBIT	ATCACAATTT GGGGA ACACCACTGG CAT ----- C-	TCACATTGGGGG TGCCG GCT------	69.44

HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC------A-	RAT	GTCACAATTT GGAGG ATGTTACTGG CAT ----- C-	TCACATTTGGGG ACCTG GCT------	77.78
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	MOUSE	GTCACATTTG GGGAT GTTCTGGCT------	GTCACAGTTTGG AGGAT GTTACTGACAT-----C-	72.22
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	HEDGEHOG	GTCAGTTTGA TTTTG GCT------	GTCATAGTTT---- GATTATATGGGC TT ---- C-	58.33
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	DOG	GTCACATTTG GGGGA TCTCTGGCT------	GTCACAATTTGG GGGAT ACTACTGGCAT-----C-	80.56
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	CAT	GTCACGTTTG GGGGA CTCTGGCT------	GTCACAGTTTAG GGGGT ACTACTGGCAT-----C-	72.22
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	HORSE	GTCACATTTG GGTGC CTGGCT------	GTCACAATTTAG GAAGTG CCACTGGCCT-----C-	71.12
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC------A-	COW	GCCTCTCTTT----------- CTGCCCTGCA GGC------	GTCACAGTTTGG AGGAT GTTACTGACAT-----C-	33.33
HUMAN	GTCACGATT TGGG GGATGCTTC TGGC TC A-	ARMADILLO	---------------- TGCTACTAAT AT-----T-	GTCATAGTTT---- GATTATATGGGC TT ---- C-	36.11

To ensure that they are in an appropriate format for use by the ELCS and machine learning algorithm, the DNA similarity evaluation process of the ELCS and machine learning classification algorithm may be preprocessed. The procedure could involve turning the data into a numerical format, cleaning the data, and removing noise. DNA genetic code similarity assessments are produced via the application of the ELCS approach.. The machine learning algorithm is taught using DNA sequences and the matching rules given by the ELCS method. The training procedure facilitates the acquisition of knowledge by the machine learning algorithm, enabling it to effectively categorize DNA sequences into distinct categories. The evaluation of the ELCS and machine learning algorithms is conducted on a separate test dataset that is not used during the training process. This assessment facilitates the comparison of the performance of the two algorithms and enables the identification of the algorithm that exhibits superior performance on the provided dataset.

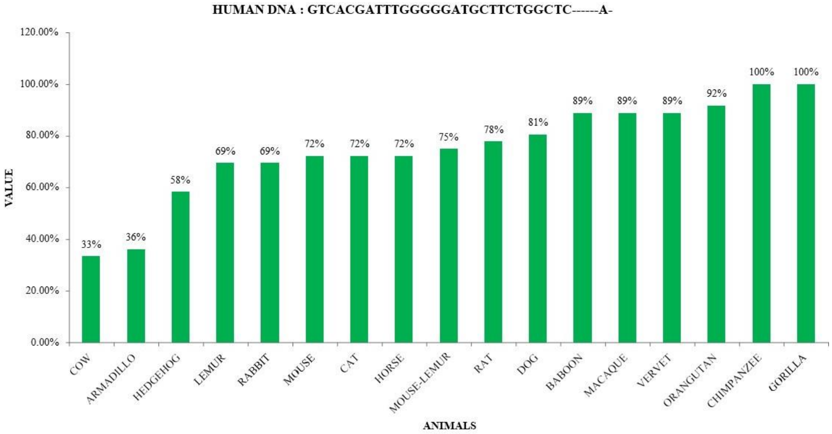


Figure 1.4. Graphical Representation of DNA Sequence Similarity

In the fields of bioinformatics and genomics research, graphical representations are essential resources for learning important lessons about genetic links, motif conservation, and evolutionary trends. Heatmaps, phylogenetic trees, and sequence alignment plots are a few examples of common visualizations used in this area. With the aid of these tools, researchers can more easily and intuitively understand complex DNA data, leading to the identification of conserved regions, evolutionary connections, and functional features present in genomic sequences.

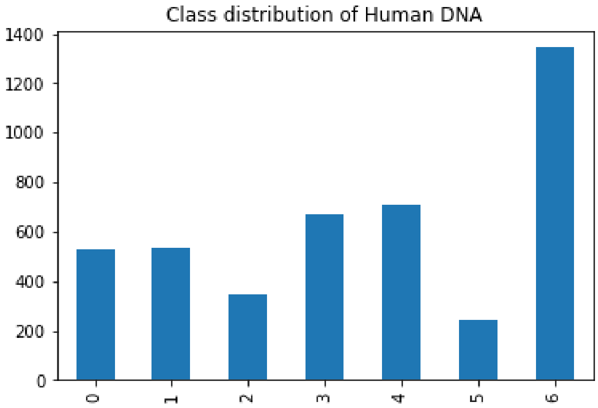


Figure 1.5. Human DNA class distribution

The analysis of class distribution entails the determination of the proportion or frequency of each category within the human genome. The examination of the distribution of human DNA among a population is a crucial undertaking in the process of describing the genetic composition of individuals, as it provides valuable insights into the many functions that different DNA elements fulfill in biological processes and overall well-being.

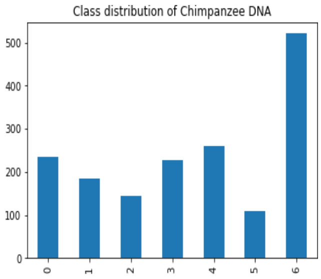


Figure 1.6. Chimpanzee DNA class distribution

The current study is important because it sheds light on the genetic makeup of chimpanzees—our closest living relatives—and helps understand the similarities and differences between the genomes of humans and chimps. It is possible to learn a great deal about evolutionary relationships, conserved elements, and the genetic bases of unique traits or adaptations shared by the two species by comparing the distribution of DNA classes in the two. Studying the spread of chimpanzee DNA in an academic setting advances our overall understanding of primate genetics and the genetic factors that have shaped the evolution of humans.

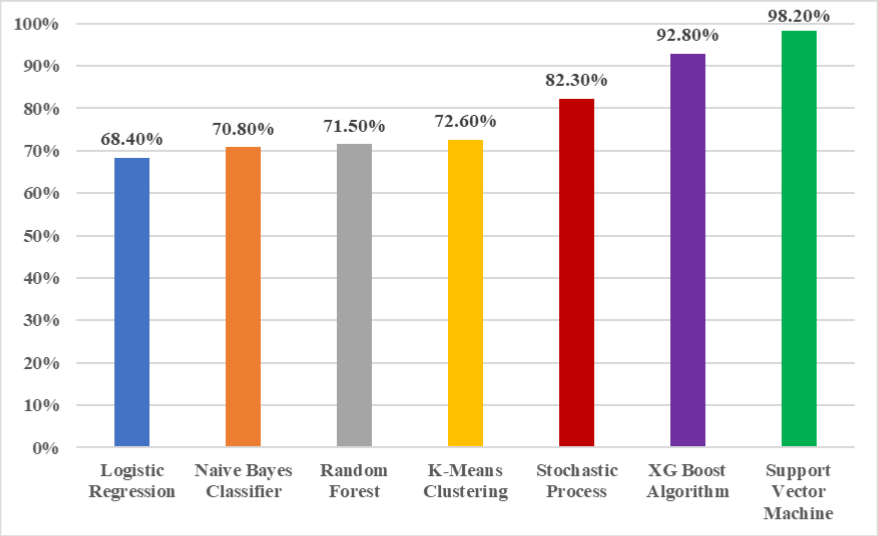


Figure 1.7. Prediction of DNA Similarity Evaluation of Machine Learning Classification Algorithm

In DNA sequences, predicting similarity evaluation , encompasses the advancement and evaluation of computational models that employ machine learning techniques to categorize DNA sequences according to their degree of similarity or dissimilarity.

Analysis of Model Prediction And Accuracy

The predictions made on a DNA sequence test performed on chimpanzees were analyzed using the Pandas crosstab function, which produced the confusion matrix. The following performance metrics are calculated and output by the program: Accuracy, precision, recall, and F1-score are the basic metrics used in the field of machine learning and classification model evaluation. This method included a matrix representation that included a succinct assessment and description of a machine learning technique after it was tested on a dataset. Predicting category labels given input cases is a common use of classification model performance measurement. The amounts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) that the model produced when it was applied to the test data are displayed in the matrix.

Classification Accuracy

Classification accuracy is used to find the exact prediction to the dataset and it is calculated based on the ratio of the given samples.

$$[illegible] = \frac{[illegible] + [illegible]}{[illegible] + [illegible] + [illegible] + [illegible]}$$

$Accuracy = \frac{8+1}{8+1+1+0} = 0.993$

Precision

In this method, the outcome is based on the labelled element such as True positive and it is calculated based on the True Positive and total number of samples.

$$Precision = \frac{TP}{TP + FP}$$

True Positive Counts = 8
False Positive Counts = 0
True Negative Counts = 1
False Negative Counts = 1

$Precision = \frac{8+0}{8+1} = 0.994$

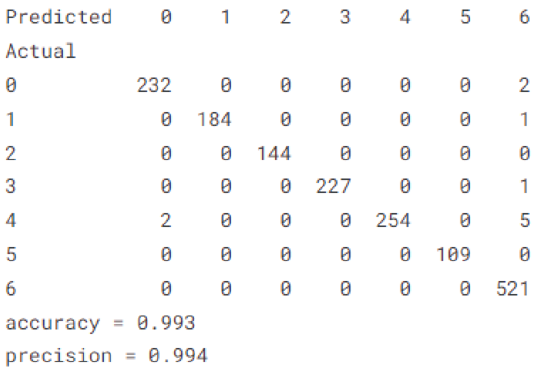
F1-Score

F1-Score is the metric and it used to measure the accuracy model of the given samples. It is combination of the precision and it recall scores of a given model.

$$F1 - Score = \frac{2 * Precision - Recall}{Precision + Recall}$$

$F1 - Score = \frac{2*0.994-0.994}{0.994+0.994} = 0.994$

With a score of 0.993, the model is quite accurate. Furthermore, it performs precisely, as evidenced by a precision value of 0.994. This means that the model correctly predicted 99.3% of the cases, and 99.4% of the positive predictions were correct.



```
Predicted    0    1    2    3    4    5    6
Actual
0          232    0    0    0    0    0    2
1            0  184    0    0    0    0    1
2            0    0  144    0    0    0    0
3            0    0    0  227    0    0    1
4            2    0    0    0  254    0    5
5            0    0    0    0    0  109    0
6            0    0    0    0    0    0  521
accuracy = 0.993
precision = 0.994
```

Figure 1.8. Confusion Matrix for Predictions on Human DNA to Chimpanzee DNA Sequence

4. CONCLUSION

To determine how similar two DNA sequences are, the Enhanced Longest Common Subsequence algorithm is employed. To maximize the ELCS's performance, its computation is accomplished by reducing pointless correlations with earlier research. This tool makes it easier to compare the genetic code sequences of two strings to see how similar they are. Using edge and ELCS computations, the similarity rate between genetic codes is calculated, removing dissimilarity sequences to minimize time complexity. When the program runs inside its time restrictions, the execution process time efficiency exceeds its computing capacity.

5. REFERENCES

1. “An Introduction to Bioinformatics Algorithms”, Neil C.Jones and PavelA.Pevzner, 2004
2. “An Efficient Parallel Algorithm for Longest Common Subsequence Problem on GPUs”, Jiaoyun Yang, Yun Xu and Yi Shang,World Congress Engineering, 2010.
3. “The Complexity of some problems on subsequences and super sequences”, Maier. D, ACM, (25), 1978, 332-336.
4. “Identification of protein motifs using conserved amino acid properties and partitioning techniques”, Wu. T. D and Brutlag.D.L, Proceedings of the 3rd International conference on Intelligent Systems for Molecular Biology, 1995, 402-410.
5. “DISPARE: discriminative pattern refinement for position weight matrices”, Isabelle da Piedade, Man-Hung Eric Tang, and Oliver Elemento, BMC Bioinformatics, 10(388): 2009, 1471-2105.
6. “Discovery of Longest Increasing Subsequences and its Variants Using DNA Operations”, B.Lavanya and A.Murugan, International Journal of Engineering and Technology (IJET),5(2), 2013, 1169-1177.
7. “Mining Longest Common Subsequence and other related patterns using DNA operations”, B.Lavanya and A.Murugan, International Journal of Computer Application, (18), 2012, 38-44.
8. “Discovering sequence motifs of different patterns parallel using DNA operations”, B.Lavanya and A.Murugan,International Journal of Computer Applications, 2011.
9. “A DNA Algorithmic approach to solve GCS Problem”, A.Murugan and B.Lavanya, Journal of Computational Intelligence and Bioinformatics. 3(2), 2010, 239-247.
10. “On the longest common rigid subsequence problem”, Nikhil Bansal, Moshe Lewenstein, Bin Ma, and Kaishong Zhang, Algorithmica, 2010, 270-280.
11. “Comprehensive study on iterative algorithms of multiple sequence alignment”, Hirosawa et al, Computational Applications in Biosciences, 1995, 13-18.
12. “Detecting patterns in protein sequences”, Neuwald.A.F and Green.P, Journal of Molecular Biology, 1994, 698-712.
13. “Sequential Pattern Mining from multidimensional sequence data in parallel”, Mahdi Esmaieli and Mansour Tarafdar, International journal of Computer theory and engineering, 2010, 730-733.
14. “Identification of common molecular subsequences”, Smith. T. F. and Waterman. M. S, Journal of Molecular Biology, 1981 147:195- 197.
15. “A method for fast database search for all k-nucleotide repeats”, Benson. G. and Waterman .M.S, 2nd International conference on Intelligent Systems for Molecular Biology, 1994, 83-98.
16. “Enumerating and ranking discrete motifs”, Neville-Manning. C. G., Sethi. K .S., Wu. D.,andBrutlag. A. D, Proceedings of Intelligent Systems for Molecular Biology, 1997, 202-209.
17. “DNA binding sites: representation and discovery”, Stormo. G, Bioinformatics, 2000, 16:16-23
18. “A generic motif discovery algorithm for sequential data”, Kyle Jensen. L., Mark Styczynski. P., IsidoreRigoutsos, and Gregory Stephanopoulos. V, Bioinformatics, 22(1) 2006, 21-28.
19. “On the complexity of multiple sequence alignment”, Wang. L. and Jiang. T, Journal of Computational Biology, 1994, 337-348.
20. “Analysis of computational tools for motif discovery”, Nan Li and Tompa. M, Algorithms of molecular biology, 2006, 1-8.

21. “Efficient mining of iterative patterns for software specification discovery”, Lo. D and Khoo. S. D. Liu, International Conference.on Knowledge Discovery and Data Mining, 2007, 460-469.
22. “Discovery of frequent episodes in event sequences”, Annila. H. M., Toivonen. H., and Verkamo. A.I, Data Mining and Knowledge Discovery, 1(3), 1997, 259-289.
23. Martinez. M, “An efficient method to find repeats in molecular sequences”. Nucleic Acid Research, 1983, 4629-4634.
24. “Searching for common sequence patterns among distantly related proteins”, Suyama. M., Nishioka. T., and Junichi. O, Protein Engineering, 1995,1075-1080.
25. “Finding sequence motifs in groups of functionally related proteins”, Smith. H. O., Annau. T. M., and Chandrasegaran.S, Proceedings of National Academy (USA), 1990, 826-830.
26. "A polynomial time approximation scheme for the closest substring problem", Bin Ma, LCNS Springer, 2000, 99-107.
27. “DNA Sequencing Using Machine Learning Algorithms“, M.S.Antony Vigil, M. Mirutuhula, Sure Sarvagna, R. Supraja, Gadusunari Priyanka Reddy.
28. “Sequence Alignment Using Machine Learning-Based Needleman–Wunsch Algorithm” Amr Ezz El-Din Rashed, Hanan M. Amer, Mervat El-Seddek, Hossam El-Din Moustafa
29. “A fast and efficient algorithm for DNA sequence similarity identification”, Machbah Uddin, Mohammad Khairul Islam, Md. Rakib Hassan, Farah Jahan, Joong-Hwan Baek
30. “Literature Survey on DNA Sequence by Using Machine Learning Algorithms and Image Registration Technique”, R. Vinodhini1, R. Suganya1, S. Karthiga1, G. Priyanka