International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

Dna Sequence Predictions Using Nlp And Ml

K. Vigneshwar[1], P. Shruthi[2], J. Rahul Naik[3], P. Khaleel Basha[4]
Assistant Professor, Guru Nanak Institute of Technology, CSE Department, Hyderabad[1]
Student,Guru Nanak Institute of Technology, CSE Department, Hyderabad [2,3,4]

Abstract- Deoxyribonucleic acid (DNA) is a biological macromolecule. Its main function is information storage. At present, the advancement of sequencing technology had caused DNA sequence data to grow at an explosive rate, which has also pushed the study of DNA sequences in the wave of big data. Moreover, machine learning is a powerful technique for analyzing largescale data and learns spontaneously to gain knowledge. It has been widely used in DNA sequence data analysis and obtained a lot of research achievements. Firstly, the review introduces the development process of sequencing technology, expounds on the concept of DNA sequence data structure and sequence similarity. Then we analyze the basic process of data mining, summary several major machine learning algorithms like Multinomial NB Classifier & Random Forest, and put forward the challenges faced by machine learning algorithms in the mining of biological sequence data and possible solutions in the future. Then we review four typical applications of machine learning in DNA sequence data: DNA sequence alignment, DNA sequence classification, DNA sequence clustering, and DNA pattern mining. We analyze their corresponding biological application background and significance, and systematically summarized the development and potential problems in the field of DNA sequence data mining using Multinomial NB Classifier & Random Forest. Finally, we summarize the content of the review and look into the future of some research directions for the next step.

Keywords: DNA Sequence Analysis, Machine Learning, DNA Data Mining, Sequencing Technology, Biological Sequence Data, Multinomial Naïve Bayes Classifier, Random Forest, DNA Sequence Alignment, DNA Sequence Classification, DNA Sequence Clustering, DNA Pattern Mining, Big Data Analytics, Bioinformatics, Predictive Modeling, Computational Biology.

I. INTRODUCTION

We live in the era of the genome, advances in science have allowed humans to spy on the mysteries of life. In recent decades, the rapid expansion of biological data is a significant feature of the development of molecular biology, and a massive biological information database has rapidly formed. We must obtain useful knowledge from these huge data, and simultaneously bioinformatics was born. Bioinformatics is an interdisciplinary subject. It comprehensively uses mathematics, life sciences, and computer science to mine biological information in biological data (Chu, 2014), and further guides the relevant researches of biological researchers Specifically, the first step is to obtain information on the protein coding region by analyzing the genomic DNA sequence. Then simulating and predicting the spatial structure of the protein. For complex biological data, on the one hand, it is necessary to solve the problem of storage and management of massive data, and on the one hand, it is necessary to extract effective information from the data on the premise of ensuring that the data reflects the true meaning of biology. Machine learning is an important method to achieve artificial intelligence. It can handle the automatic learning of machines without explicit programming and has been widely used in the field of bioinformatics.

II. LITERATURE REVIEW

Kathleen A. Hill, Lila Kari, and Gurjit S. Randhawa discussed ML-DSP, an alignment-free genome classification method that combines machine learning with digital signal processing for fast and accurate genomic analysis. The study aims to improve genome classification by reducing computational complexity and increasing scalability. The framework converts DNA sequences into numerical representations and applies machine learning techniques for classification. Experimental results showed classification accuracies above 97% and better performance compared to MEGA7 and FFP in terms of speed and accuracy.[1].. Lois Boggess and Liangyou Chen discussed the use of neural network models for genome signature analysis and gene classification. The study aims to improve automatic genomic analysis by applying models such as back-propagation networks, radial basis function networks, self-organizing maps, and committee machines. Experimental results showed average accuracies of 97% for two-way classification and above 83% for four-way classification tasks. The research also discussed methods for preparing training

and testing datasets along with modifications to neural network algorithms for better performance.[2].
Yuhei Kaneshita, Satoshi Asatani, Seiichi Tagawa, Hirohiko Niioka, Takashi Hirano, and Jun Miyake discussed the graphical classification of DNA sequences of HLA alleles using deep learning techniques. The study aims to classify HLA-A DNA alleles by applying stacked autoencoder-based deep learning models. Nucleotide sequence data of 822 bp obtained from the Immuno Polymorphism Database were compressed into two-dimensional representations for visualization. The results showed that the generated plots formed distinct clusters, enabling clear classification and characterization of different HLA alleles.[3].

H. Rowley, S. Baluja, and T. Kanade discussed large-scale machine learning techniques for metagenomics sequence classification. The study aims to improve the binning process in metagenomics by developing fast and accurate classification methods for large genomic datasets. The proposed approach uses compositional methods based on k-mer analysis to assign DNA reads to taxonomic groups. The research highlights that machine learning-based compositional approaches can provide faster solutions with reasonable computational requirements compared to traditional alignment-based methods while maintaining effective classification performance. [4].

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee discussed the classification of DNA microarrays using artificial neural networks and the ABC algorithm. The study aims to improve disease classification by analyzing gene expression data from DNA microarrays. The proposed method uses a swarm intelligence-based feature selection technique to identify important genes, followed by training artificial neural network models such as MLP, RBF, and SVM for classification. Experimental validation on four datasets demonstrated the effectiveness of the approach in accurately classifying disease samples and identifying relevant genes.[5].

III. METHODOLOGY

Genome data analysis faces several challenges in classification, storage, and prediction processes. Biomedical data classification is difficult due to heterogeneous, incomplete, and inconsistent datasets, making data cleaning and preprocessing essential for accurate prediction. Large-scale DNA sequencing data requires reliable storage and proper data integration for efficient access and management. DNA sequence classification is often treated as a binary nonlinear classification problem, while differences in genome lengths require normalization techniques such as up-sampling and down-sampling, which may cause information loss. In metagenomic analysis, the binning step is a major challenge because sequenced reads must be assigned accurately and quickly to taxonomic groups. Support Vector Machine (SVM) learning also faces issues such as convex optimization and detection problems. Additionally, optimization problems in genome analysis are addressed using Genetic Algorithms (GA), which apply evolutionary principles to improve data mining, classification, and model evaluation processes.

Disadvantages of existing system:

- It does not execute very well when the data set has more sound i.e. target classes are overlapping.
- It doesn't perform well when we have large data set because the required training time is higher.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.

Proposed System

DNA sequences contain important information about species, including their characteristics, appearance, and genetic inheritance, making DNA classification essential in modern biology and medicine. DNA sequencing is widely used for identifying cancers, mutations, tumors, and genetic disorders such as Down syndrome. The main goal of DNA sequencing is to determine the order of nucleotides in a DNA segment. By analyzing DNA sequence data using Multinomial Naive Bayes and Random Forest classifiers, high classification accuracy of 98.4% was achieved.

Advantages Proposed System Advantages

- It is easy to implement as you only have to calculate probability.
- Used for both continuous and discrete data.
- It is highly scalable and can easily handle large datasets

System Architecture

The system architecture represents the workflow of DNA sequence data mining and classification. Initially, relevant genomic data is selected and preprocessed to remove inconsistencies and prepare clean target data for analysis. The transformed data is then processed using data mining algorithms such as machine learning classifiers, and the results are evaluated to generate meaningful biological knowledge and accurate predictions.

Modules:

1. Biomedical Data Classification and Prediction – Handling heterogeneous and incomplete biomedical datasets makes accurate classification and prediction difficult, requiring effective data cleaning methods.
2. DNA Sequencing Data Storage – Large DNA sequencing datasets require reliable storage, efficient accessibility, and proper data integration from multiple sources.
3. Binary Nonlinear DNA Classification – DNA sequence classification is treated as a binary classification problem where sequences are grouped based on specific classification rules.
4. DNA Sequence Length Normalization – Variations in genome lengths require normalization techniques such as up-sampling and down-sampling, which may lead to information loss.
5. Binning and Optimization Challenges – Metagenomic binning, SVM optimization issues, and genome analysis challenges require efficient machine learning and Genetic Algorithm techniques for accurate classification and analysis.

IV. IMPLEMENTATION

- **Multinomial NB Theorem:**

The proposed DNA sequence classification system uses the Multinomial Naive Bayes (MNB) algorithm due to its simplicity, speed, and effectiveness in handling large genomic datasets. The model predicts the class of DNA sequences by analyzing nucleotide frequency patterns and calculating probabilities using Bayes' Theorem. To improve classification performance, preprocessing and feature extraction techniques are applied to transform DNA sequences into numerical representations suitable for machine learning analysis. The system evaluates genomic patterns and assigns sequences to the most probable category with high accuracy. Performance metrics such as accuracy, precision, recall, and F1-score are used to validate the model, achieving an overall classification accuracy of 98.4%. The proposed approach supports fast, scalable, and reliable genome analysis for biomedical and bioinformatics applications.

V. EXPERIMENTAL RESULTS

Figure. Home Page

- The image shows a student homepage with the title "DNA Sequence Predictions using NLP and ML" prominently displayed.

Figure. Login Page

- The second image shows the login interface for the "DNA Sequence Predictions" web application.

Figure. Input Page

• This image presents the details page of the DNA Sequence Predictions web application, maintaining the same modern dark-themed interface as the login screen.

Figure. Output of Human Sample DNA

• This is the "output" page of the DNA Sequence Predictions web application.

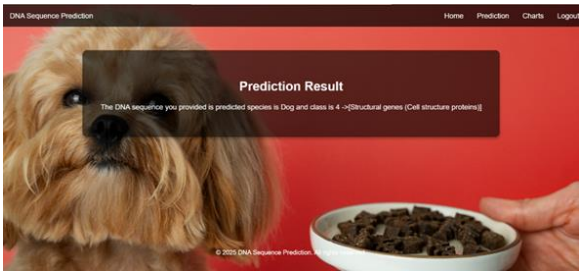

Figure. Result Page

• To get an accurate prediction , please provide key clear and correct DNA data.

VI. CONCLUSION

In recent decades, rapid advancements in hardware and life science technologies have enabled the collection of massive biological data from areas such as genomics, biological imaging, and medical imaging, creating both opportunities and challenges in bioinformatics. Machine learning techniques such as artificial neural networks, deep learning, and reinforcement learning have played an important role in DNA sequence analysis and biological data mining. Increased computing power, faster storage systems, and reduced computational costs have further improved the efficiency of these technologies. However, important challenges still remain, including integrating the biological significance of DNA sequences into the data mining process and handling the continuously growing volume of biological data efficiently. Traditional analysis methods often require high computation time, so distributed and parallel computing techniques are necessary to improve mining efficiency. Additionally, selecting suitable DNA sequence coding methods for specific tasks can improve algorithm performance, reduce training time, and provide more accurate prediction results.

VII. FUTURE ENHANCEMENT

In summary, from the aspects of sequencing technology, DNA sequence data structure, and sequence similarity, this review comprehensively introduces the source and characteristics of DNA sequence data; we briefly summarize the machine learning algorithms and propose biological sequence data Challenges faced by machine learning algorithms in mining and possible solutions in the future. Then, we reviewed four typical applications of machine learning in DNA sequence data: DNA sequence alignment, classification, clustering, and pattern mining, analyzed and discussed their corresponding biological application background and significance, and systematically summarized recent years.

REFERENCES

1. Qingshan Jiang, Dan Wei, Qingda Zhou, "A New Method for Classification in DNA Sequence," in The 6th International Conference on Computer Science & Education, 2011.
2. Yichen Zheng, Ricardo B. R. Azevedo, Dan Graur, "An Evolutionary Classification of Genomic Function," vol. 7, no. 3, p. 4, 2015.
3. Shailendra Singh, Trilok Chand Aseri, Neelam Goel, "An improved method for splice site prediction in DNA sequences using support vector machines," in 3rd International Conference on Recent Trends in Computing, 2015.
4. Karthika Vijayan, Deepa P. Gopinath, Achuthsankar S. Nair, Vrinda V. Nair, "ANN based Classification of Unknown Genome Fragments using Chaos Game Representation," in Second International Conference on Machine Learning and Computing, 2010.
5. Dr P. S. V. Srinivasa Rao, S.S.S.N Usha Devi N, Dr P. Kiran Sree, "CDLGP: A Novel Unsupervised Classifier using Deep Learning for Gene Prediction," in IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, 2017.
6. Steve Wanamaker, Timothy J Close, Stefano Lonardi, Rachid Ounit, "CLARK: Fast and accurate classification of metagenomic and

genomic sequences using discriminative k-mers," vol. 16, p. 13, 2015.
7. Katya Rodriguez, Roberto A. Vazquez, Beatriz A. Garro, "Classification of DNA microarrays using artificial neural networks and ABC algorithm.," vol. 38, p. 13, 2015.
8. Kun Wang, Huixiao Li, Yang Jia, Xiaoqin Wu, Yaning Du, Wei You, "Classification of DNA Sequences Basing on the Dinucleotide Compositions," in 2 nd International Symposium on Computational Intelligence and Design, 2009.
9. Ngoc Giang Nguyen, Favorisen Rosyking Lumbanraja, Mohammad Reza Faisal, Bahriddin Abapihi, Bedy Purnama, Mera Kartika Delimayanti, Mamoru Kubo, Kenji Satou, Dau Phan, "Combined Use of k-Mer Numerical Features and Position-Specific Categorical Features in Fixed-Length DNA Sequence Classification," vol. 10, no. 8, p. 12, 2017.
10. Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, Zhi Xie, Chensi Cao, "Deep Learning and Its Applications in Biomedicine," vol. 16, no. 1, p. 16, 2018.
11. Vu Anh Tran, Duc Luu Ngo, Dau Phan, Favorisen Rosyking Lumbanraja, Mohammad Reza Faisal, Bahriddin Abapihi, Mamoru Kubo, Kenji Satou, Ngoc Giang Nguyen, "DNA Sequence Classification by Convolutional Neural Network," vol. 9, p. 7, 2016.
12. Jason T. L. Wang, Dennis Shasha, Cathy H. Wu, Qicheng Ma, "DNA Sequence Classification via an Expectation Maximization Algorithm and Neural Networks: A Case Study," in IEEE Transactions on systems, man, and Cybernetics—part C: applications and reviews,, 2001.
13. Suprakash Datta, Wendy Ashlock, "Evolved Features for DNA Sequence Classification and Their Fitness Landscapes," in IEEE Transaction on Evolutionary Computation, 2013.
14. Yuhei Kaneshita, Satoshi Asatani, Seiichi Tagawa, Hirohiko Niioka, Takashi Hirano, Jun Miyake, "Graphical classification of DNA sequences of HLA alleles by deep learning," vol. 31, no. 2, p. 4, 2018.
15. Brian Hudson, David Whitley, Martyn Ford, Phil Picton, Hassan Kazemian, Antony Browne, "Knowledge Extraction from Neural Networks," in IEEE International Conference, 2003.