International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1489-1492 www.ijaem.net ISSN: 2395-5252

Detection of gene family with DNA Sequencing Data Machine learning

Athfia Farheen N, Bhavan A J, Bhavani N D, Rahul B, Kavitha H M, Dr. Ravikumar G.K,

4BW17CS010, Dept of CS&E, BGSIT, B.G Nagara
4BW17CS012, Dept of CS&E, BGSIT, B.G Nagara
4BW17CS013, Dept of CS&E, BGSIT, B.G Nagara
4BW17CS051, Dept of CS&E, BGSIT, B.G Nagara
Research Scholar, Dept of R&D, BGSIT, B.G Nagara
Professor, R&D Head, BGSIT, B.G Nagara

Submitted: 25-06-2021 Revised: 06-07-2021 Accepted: 09-07-2021

ABSTRACT–The field of machine learning methods have been widely used in bioinformatics. This aims to develop computer algorithms that improve with experience and to enable computers to assist human in analysis of big and complex data. In genomic research, classifying DNA sequences into existing categories is used to learn the functions of new protein. So identifying and classifying of those genes are important. DNA sequence data frequently are contained into file format called fasta format. Fasta format is single line prefixed by the greater than symbol that contains annotations and another line that contains the sequence. For treating DNA sequences we use k-mer counting.

Keywords- Bioinformatics, DNA Sequence, Fasta format, k-mer counting

I. INTRODUCTION

The DNA (Deoxyribonucleic acid) is an organic moleculewhich carries the genetic information of the organism.The Swiss biochemist Fredrich Miescher first observedDNA in the late 1800s. DNA is like a double helix whichis liable for the co-ordination and functioning of allliving beings and it is inherited from the ancestors to theiroffspring during reproduction. It stores all the importantinformation of the organism and the information for

synthesizing the required proteins. A genome sequence isthe complete list of nucleotides and made up of chemicalbuilding blocks. Nuclotides are made up of phosphategroup, sugar group and nitrogen bases. In our project we mainly focused the four types of nitrogen bases found in nucleotides, they are: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The sequence which is present in these bases determines what biological instructions are present in a strand of DNA. For example, ATCGCT and ATCGTT which instructs for brown eyes and blue eyes. We use these DNA sequences to identify which gene family which they belongs to.

Based on the shared nucleotide or protein sequences genes are categorized into families. In the coding sequence, the positions of exons can be used to identify common ancestry. Knowing the sequence of the protein encoded by a gene can allows to get information on differences among DNA sequences. By training human gene sequence to machine learning model, we predict gene family like G protein coupled receptors, Tyrosine Kinase, Tyrosine phosphates, Synthetase, Synthase, Ion Channel and Transcription factor. We assign id value to each human gene family like, for G protein coupled receptors as 0, for Tyrosine Kinase as 1 and so on. The gene sequence is in text to convert these genes to national language k-mers encoding is used. The human gene sequence is collected and split into testing and training data. We use different Classifier algorithm to train the model and select the best one which give high accuracy rate.

Pre-processing has to be done before using the machine learning model. K-mer encoding is used to convert a long biological sequence and breaks it down into k-mer length. K value is assigned to 6 so that the sequence break in a length of 6 which are called hexamers. Then Count vectorizer and Bag of words describes the occurrence ofwords within a sequence which counts the number of occurrences in hexamers. Natural language processing (NLP) is used for interaction between computer and humans. After pre-processing different Machine learning algorithm are applied and trained and tested, the most accuracy algorithm algorithm is selected.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1489-1492 www.ijaem.net ISSN: 2395-5252

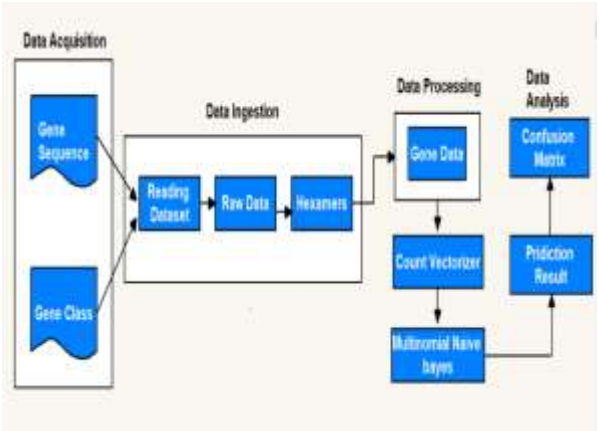
II. OBJECTIVE:

- Creating the "Human Gene Datasets" which consists human gene sequence which is corresponding to respective gene family.
- Converting DNA sequence as a "language" using k-mer counting.
- Developing "multinomial naïve Bayes classifier Model" .
- Training and Testing the model by giving the datasets

III. LITRATURE SURVEY:

1. In the year 2013, Snehal P and Adey Proposed a method to detect cancer using CUDA dissertation with GPU accelerated Pattern Matching Algorithm for DNA Sequences. The objective is to locate the appearance of specific pattern in an array of equal size text. This can be done by parallelization on GPU Using CUDA programming model. This have some limitations such as, it does not give good accuracy and the performance degrades with input errors.
2. In the year 2017, Yang Yang and Katherine Proposed Machine learning for classifying tuberculosis from DNA Sequencing data. For rapid determination of Mycobacterium tuberculosis(MTB) resistance against available tuberculosis drugs for control and management of TB is essential. Use of the best model examined to predict MTB and reduce risk of acquiring multi-drug resistance. This has some limitations, there is low sensitivity for resistance classification as it has single nucleotide polymorphism.
3. In the year 2018, CaioDavii, Andre Pastor, Michel BamshadPropsed to prognosis severe dengue using human genome data and machine learning. To detect dengue fever severity based solely on human genome data. This use only genome markers and can be used to identify individuals at high risk to get dengue phenotype even in uninfected condition. To maintain the policies and security measures the term are still new.
4. In the year 2019 Asad Waqar Malik, proposed pattern matching for DNA sequencing Data Using Multiple Bloom Filters. This gives high accuracy and for efficient solution of pattern matching and storage these bloom filters are used. Evaluation process shows progressive accuracy and time efficiency. For this, the storage space has to be more. Due to large data sets there might be a error.

IV. SYSTEM ARCHITECTURE:



- **Data Acquisition And Data Ingestion**

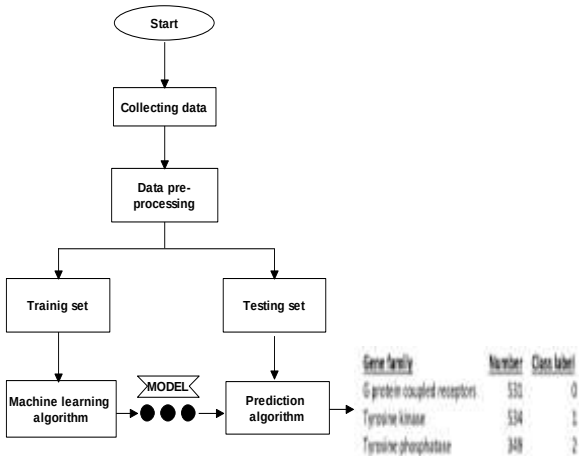
The gene dataset which consists gene sequence and gene class are readed and then they are converted into
Hexamers. To convert genes to Hexamers we use K mer Encoding where k value is set to 6.

- **Data Processing And Data Analysis**

Count vectorizer is used to count number of Hexamers are frequently repeated And then Multinomial Navie Bayes algorithm os applies. And the production results and confusion matrix is calculated.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1489-1492 www.ijaem.net ISSN: 2395-5252

V. DATA FLOW



In the beginning Gene data set is collected, which consists gene data and gene family names. Then it is preprocessed by using k mer encoding which converts gene data to Hexamers. After this the data set is splited to 80% trainingand 20% testing. The data set is trained by Multinomial Navie Bayes algorithm and then respective gene family is predicted.

VI. PROPOSED METHODOLOGY

- By considering a classification model that can predict a gene family based on the human DNA sequence. Human genes are used to predict gene family, by training these genes in a machine learning model. These DNA sequences are evaluated using k-mer counting.
- K-mer refers to all of the subsequences ofall possible length in a sequence. It takes a long biological sequence and breaks it down into k-mer length. For example, consider the sequence "ATGCAT", we use words of length 6 (hexamers) which is "ATGCAT"," TGCATG"," GCATGC"," CATGCA".
- Skit learns Natural language processing tools are used for the k-mer counting. The processing tools covert each gene into string sentences of words, these words are used by the count vectorizer.
- Count vectorizer break downs the sequence into words. Bag of words is a Natural language processing technique of text modelling; this describes the occurrence of words within a sequence. This bag of words is applied to the count vectorizer in NLP.
- Natural language processing used for the interaction between computers and humans. The purpose of Natural language processing is to read, decode and understand human languages.
- Now talking about the algorithms used, in which one of them is the K-means clustering algorithm. The K-means algorithm aims to partition n-observations into k clusters and allocates every data points to the nearest cluster while keeping the observation as small as possible. The accuracy of this algorithm is estimated to be 88%.
- Another one of the algorithms that are used is Random Forest. Random Forest is a bag containing n decision treeswith different parameters and trained under a different subset of data. The accuracy of this algorithm is calculated to be 91%.
- The last algorithm used is the Multinomial Naive Bayes Classifier. This algorithm based on bays theorem. These algorithms are made up of families ofalgorithms, where they all share a common principle that is every pair of features being classified is independent of each other. The accuracy of this algorithm is calculated to be 98%.

VII. APPLICATIONS.

- Genotyping cancer cells and understanding what genes are mis regulated allows physicians to pick the simplest chemotherapy and potentially expose the patient to less toxic treatment since the therapy is ready-made .
- Previously unknown genes could also be identified as contributing to a disease state.
- Single test in a lifetime.
- It requires less time and much less expensive.
- Lifestyle or environment changes that can mediate the effect of genetic predisposition may be identified and then moderated.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1489-1492 www.ijaem.net ISSN: 2395-5252

VIII. CONCLUSION

This project aims at applying several machine learning methods to the construction of classifiers for the detection of promoters in the DNA. In genome research, classifying DNA sequences into existing categories is used to learn the functions of the new protein. Identification and classification of those genes are very important. Newer sequencing methods have drastically cut the cost of sequencing and may eventually allow every person the possibility of personalized genome information. Since we can read the expression of the genes, it gives us the opportunity of advanced medical treatments, but there is certainly more work to be done in generation, understanding, organizing and applying this massive amount of data to human disease.

REFERENCES.

[1]. https://www.sciencedaily.com/releases/2016/11/161128151050.htm
[2]. https://fifarma.org/en/this-is-what-happens-when-a-virus-enters-the-human-body/amp/
[3]. https://www.kaggle.com/thomasnelson/working-with-dna-sequence-data-for-ml
[4]. ML approach for breast cancer detection using DNA sequence recognition. [InternationalJournalofEngineeringandAdvancedTechnology(5thJune2019)]
[5]. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing Data.[InstituteofBiomedicalEngineering,University ofOxfor,2017December12].
[6]. Pattern Matching for DNA Sequencing Data Using Multiple Bloom Filters [Hindawi BioMed Research International Volume2019.]
[7]. Training Knowledge-Based Neural Networks to Recognize Genes in DNA Sequences:Michiel O. Noordewier, Geoffrey G. Towell, Jude W.Shavlik
[8]. A Comparative Study of Machine Learning Methods for Detecting Promoters in Bacterial DNASequences Leonardo G. Tavares, Heitor S. Lopes, and Carlos R. Erig Lima
[9]. ML Approach for Breast Cancer Detection using DNA Sequence Recognition Author'sname:P.Sabitha,KartikGupta,Tejas Sharma,RaviKumarSingh,JugnuKumar Year: June2018
[10]. GPU Accelerated Pattern Matching Algorithm for DNA Sequences to Detect Cancer using CUDA Dissertation, Snehal P. Adey