WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DNA Sequencing with Machine Learning

Pushpalata Verma[1], Shivam Karoria[2], Somesh Mishra[3], Sneha Sharma[4]

[1]*Assistant Professor, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*
[2,3,4]*Student, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*

A B S T R A C T

DNA sequencing is essential to contemporary research. It facilitates the advancement of many fields, including phylogenetics, genetics, and meta-genetics. DNA strands must be extracted and read in order to perform DNA sequencing. Our suggested approach aims to apply an improved prediction model for DNA research and obtain the most precise findings from it. The most popular and well-respected "machine learning models" are those under consideration. Among the suggested models is Naive Bayes. In machine learning, the Naive Bayes approach produced results with a higher accuracy of 98.00 percent.

Keywords: DNA, K-mer encoding, Machine Learning, Prediction Model, Classification.

Introduction

In the era of genomics, the rapid growth of biological data has led to the emergence of bioinformatics, a multidisciplinary field that harnesses mathematics, computer science, and life sciences to extract meaningful knowledge from vast biological databases. This field plays a crucial role in advancing molecular biology by analyzing genomic DNA sequences, modeling protein spatial structures, and facilitating drug design based on protein functions. The volume of biological data has been doubling approximately every 18 months, reaching 162 million biological sequencing data with 150 billion nucleotide bases by February 2013. Two key challenges in bioinformatics are efficiently storing and managing large datasets and extracting relevant information. Machine learning, particularly artificial intelligence, has become a pivotal technique in addressing these challenges, enabling automated learning without explicit programming. Overall, bioinformatics is essential for directing and advancing biological research in the face of the exponential growth of biological data.

DNA

The genome is the complete set of DNA in an organism, and in humans, it consists of 6 billion base pairs arranged into 23 chromosomes. The DNA code is composed of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Despite the uniqueness of individual genomes, more than 99 percent of human DNA is shared among all individuals. The sequence of these bases forms the information necessary for building and maintaining an organism, similar to letters forming words and sentences. DNA bases pair up (A with T, C with G) to create units called base pairs, and these pairs, along with sugar and phosphate molecules, constitute nucleotides arranged in a double helix structure. DNA's ability to replicate itself is crucial during cell division to ensure each new cell inherits an exact copy of the genetic information from the parent cell.

Components of DNA

Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) are the four different types of smaller chemical molecules known as nucleotide bases that make up DNA, which is a linear molecule. The DNA sequence refers to the arrangement of these nucleotides. Genes are sections of DNA that contain genetic information; during reproduction, parents pass these segments of DNA on to their kids.

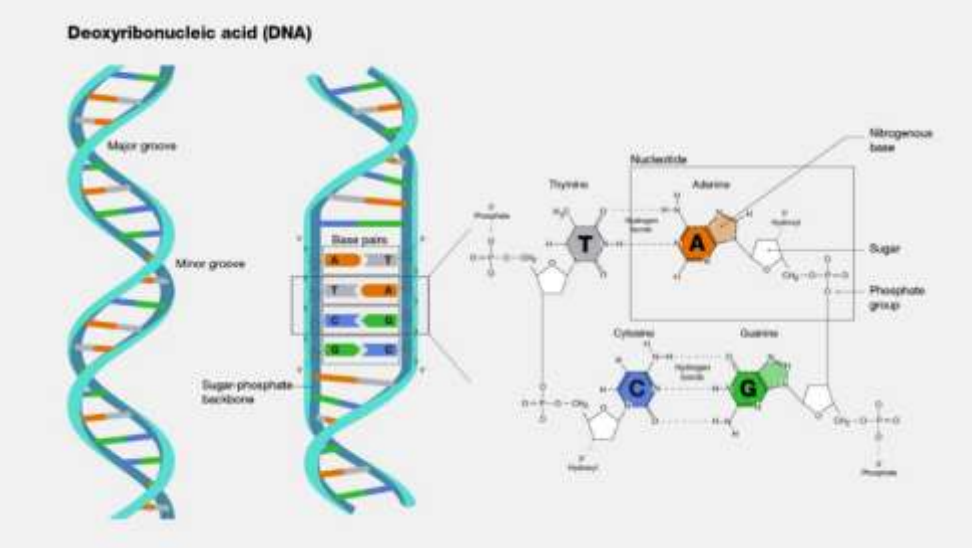


Fig. 1 – diagram of a DNA.

BioPython and Machine Learning

Machine learning is a subset of artificial intelligence that enables software applications to improve their performance on a task over time by learning from data without being explicitly programmed. It involves the use of algorithms that analyze and identify patterns in data, allowing the system to make predictions or decisions.

Biopython, on the other hand, is an open-source collection of Python tools specifically designed for computational biology and bioinformatics. It offers modules and classes to facilitate the manipulation, analysis, and visualization of biological data. Additionally, Biopython provides interfaces to various bioinformatics tools and databases, making it a valuable resource for researchers and developers working in the field of biology.

Making of Prediction Model

DNA data handeling using BioPython

We parse DNA sequence data(fasta) using Bio.SeqIO from BioPython. We can interact directly with the sequence object's characteristics, which includes the length of the sequence and its id. So it produces the sequence id, sequence and length of the sequence.

The next step involves encoding of the DNA sequence. When processing the DNA sequence , it is necessary to covert the string sequence into a numerical value, so as to form a matrix input model training. Gererally speaking , there are three methods for sequence encoding.

- Ordinal encoding DNA sequence.
- One-hot encoding DNA sequence.
- K-mer encoding.

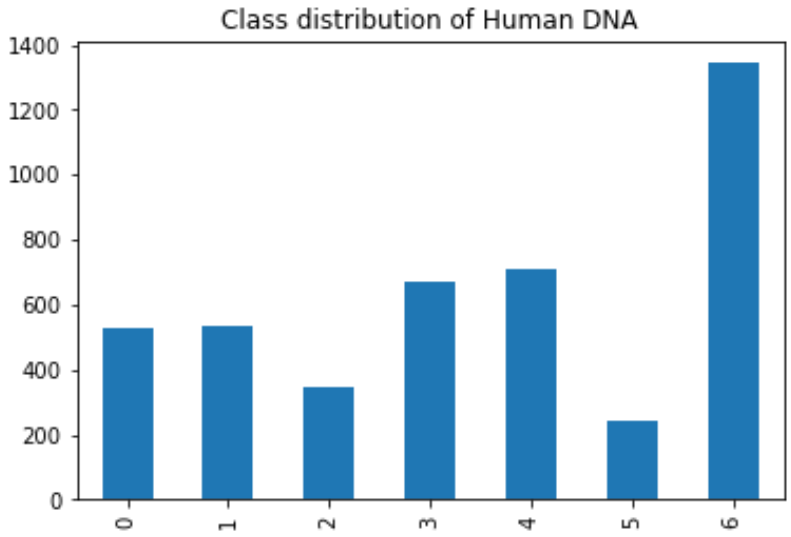
We used here k-mer counting and k-mer function. It returns a list of k-mer "words." We can then join the "words" into a "sentence", then apply our natural language processing methods on the "sentences" as we normally would. We can tune both the word length and the amount of overlap. This allows us to determine how the DNA sequence information and vocabulary size will be important in our application. For example, if we use words of length 6, and there are 4 letters, we have a vocabulary of size 4096 possible words. We can then go on and create a bag-of-words model like we would in NLP.

Then we load our human DNA data and it produces following results as shown below.

Table 1 – Result

	Sequence	class
0	ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCA...	4
1	ATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAG...	4
2	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3

3	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
4	ATGCAACAGCATTTTGAATTTGAATACCAGACCAAAGTGGATGGTG...	3



Here are the definitions for each of the 7 classes and how many there are in the human training data:

Table 2 – Gene family data

Gene family	Number	Class label
G protein coupled receptors	531	0
Tyrosine kinase	534	1
Tyrosine phosphatase	349	2
Synthetase	672	3
Synthase	711	4
Ion channel	240	5
Transcription factor	1343	6

Similarily we will laod chimpanzee and dog DNA data and it will produce similar results. Now we have all our data loaded, the next step is to convert a sequence of characters into k-mer words, default size = 6 (hexamers). The function Kmers_funct() will collect all possible overlapping k-mers of a specified length from any sequence string.

We need to now convert the lists of k-mers for each gene into string sentences of words that can be used to create the Bag of Words model. We will make a target variable y to hold the class labels. Example: y_human . So target variable holds array of class values.

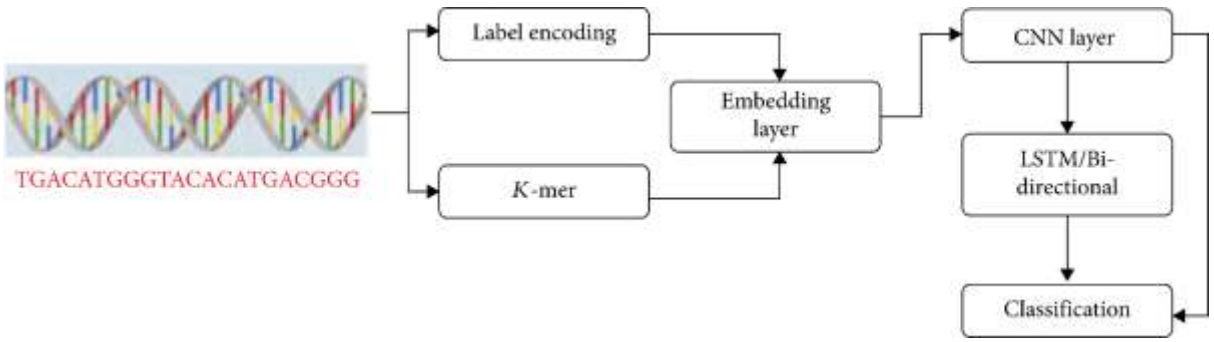
Creating the Bag of Words model using CountVectorizer(). This is equivalent to k-mer counting. The n-gram size of 4 was previously determined by testing. Convert our k-mer words into uniform length numerical vectors that represent counts for every k-mer in the vocabulary.

So, for humans we have 4380 genes converted into uniform length feature vectors of 4-gram k-mer (length 6) counts. For chimp and dog, we have the same number of features with 1682 and 820 genes respectively.

Here we have used the human data to train the model, holding out 20% of the human data to test the model. Then we can challenge the model's generalizability by trying to predict sequence function in other species (the chimpanzee and dog).

Next, train/test split human dataset and build simple multinomial naive Bayes classifier. And our Classification and prediction model is complete.

Flow Diagram



Conclusions

Predicting DNA sequences in different species, such as humans, chimpanzees, and dogs, can yield valuable insights into the genetic similarities and differences among these organisms. Here are some conclusions:

Similar k-mer profiles suggest a higher degree of genetic homology, indicating evolutionary relationships and a shared ancestry.

Conserved k-mers may represent regions associated with essential biological functions, emphasizing the importance of certain genetic sequences in the three species. Understanding these conserved genes can provide insights into fundamental biological processes that have been preserved throughout evolution.

The future scope of DNA sequencing using machine learning holds immense potential for transformative advances in genomics and healthcare. Machine learning (ML) is poised to revolutionize the field by significantly enhancing the analysis of vast genomic datasets. As the volume of genetic information generated through DNA sequencing technologies continues to increase, ML algorithms can efficiently identify intricate patterns, associations, and variations that may elude traditional analytical methods. This capability is crucial for tasks such as variant calling, where ML can distinguish genuine genetic variations from sequencing errors with improved accuracy. Personalized medicine is a key beneficiary of this integration, as ML models can predict disease susceptibility, assess individual risk factors, and optimize drug response predictions based on an individual's genetic makeup. In the realm of cancer genomics, ML contributes to the identification of specific mutations in tumors, guiding targeted therapies and enabling early detection through the analysis of subtle genetic changes. Beyond individual health, ML-driven analyses of population genomics offer insights into population-specific genetic variations and aid in understanding the evolutionary history of different groups. As technology advances, the integration of ML with emerging sequencing technologies, such as single-cell and long-read sequencing, promises a more comprehensive understanding of the genome. Ethical considerations, including privacy and security of genetic data, are also areas where ML can contribute by developing robust privacy-preserving techniques and ethical guidelines. In summary, the future of DNA sequencing using machine learning is characterized by its potential to unravel the complexities of the genome, drive personalized healthcare, and contribute to advancements in various facets of genomics and medicine.

References

Varada Venkata Sai Dileep, Navuduru Rishitha, Rakesh Gummadi, Natarajan.P (September 2022) DNA Sequencing Using Machine Learning and Deep Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE). https://www.ijitee.org/wp-content/uploads/papers/v11i10/J927309111022.pdf

Hemalatha Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, Chandran Venkatesan, C. Suresh Gnana Dhas, "Analysis of DNA Sequence Classification Using CNN and Hybrid Models", Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 1835056, 12 pages, 2021. https://doi.org/10.1155/2021/1835056

Hamed, B.A., Ibrahim, O.A.S. & Abd El-Hafeez, T. Optimizing classification efficiency with machine learning techniques for pattern matching. J Big Data 10, 124 (2023). https://doi.org/10.1186/s40537-023-00804-6

W. Santoso, K. Hulliyah, W. Nurjannah and A. H. Setianingrum, "Systematic Literature Review: Virus Prediction Based on DNA Sequences using Machine Learning and Deep Learning method," 2022 10th International Conference on Cyber and IT Service Management (CITSM), Yogyakarta, Indonesia, 2022, pp. 1-7, doi: 10.1109/CITSM56380.2022.9935921.

Imaizumi, T., Meyer, J., Wakamatsu, M. et al. Clinical parameter-based prediction of DNA methylation classification generates a prediction model of prognosis in patients with juvenile myelomonocytic leukemia. Sci Rep 12, 14753 (2022). https://doi.org/10.1038/s41598-022-18733-4

Juneja, S., Dhankhar, A., Juneja, A., & Bali, S. (2022). An Approach to DNA Sequence Classification Through Machine Learning: DNA Sequencing, K Mer Counting, Thresholding, Sequence Analysis. International Journal of Reliable and Quality E-Healthcare (IJRQEH), 11(2), 1-15. http://doi.org/10.4018/IJRQEH.299963

Hernández, D., Jara, N., Araya, M., Durán, R. E., & Buil-Aranda, C. (2022). PromoterLCNN: A Light CNN-Based Promoter Prediction and Classification Model. Genes, 13(7), 1126. https://doi.org/10.3390/genes13071126

Mohammad H. Alshayeji, Silpa ChandraBhasi Sindhu, Sa'ed Abed (2023). Viral genome prediction from raw human DNA sequence samples by combining natural language processing and machine learning techniques, Expert Systems with Applications. Volume 218, 119641, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.119641.