https://doi.org/10.46335/IJIES.2022.7.8.21 e-ISSN: 2456-3463
Vol. 7, No. 8, 2022, PP. 117-119

International Journal of Innovations in Engineering and Science, www.ijies.net

Optimization and Parallelization of Machine Learning Algorithms for DNA Classification

Sanket Tinkhede[1], Prashant Mishra[2], Tushar Zade[3],Mayuri Askar[4],Prof.Amol Gaikwad[5]

Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India 441110

sankettinkhede0203@gmail.com

***Received on**: 11 June ,2022* ***Revised on**: 02 August ,2022,* ***Published on**: 06August,2022*

***Abstract-** DNA classification is the process of determining to which pre-existential class the sequence belongs. Certain patterns or similarities justify which class a particular sequence belongs in. This classification needs a huge amount of time and manpower to classify if done manually. This paper uses machine learning algorithms to classify DNA sequences. Data obtained from DNA sequences is huge and proves time-consuming even for a machine, so to further reduce the time and obtain accurate classification results, the ML algorithms are optimized and parallelized using appropriate techniques and NLP (natural language processing).*

Keywords- Classification, Optimization, Parallelization, DNA, ML, NLP.

I – INTRODUCTION

The process of DNA sequencing is followed by classifying DNA sequences into various classes, namely 'G Protein Coupled Receptors', 'Synthase', 'Tyrosine Kinase', 'Tyrosine Phosphatase', 'Synthatase', 'Ion channel', and 'Transcription Factor' [14] . This sequence includes lengthy strings consisting of alphabets' A ',' G ','C', and 'T', which represent 'adenine',' guanine ',' cytosine ', and ‘thymine ', respectively [15].

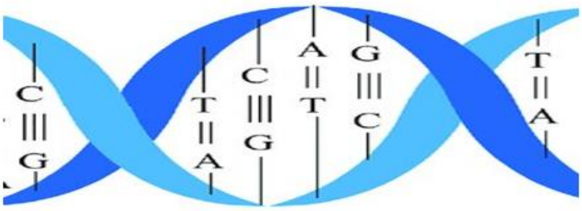


Fig 1- One helix of DNA

These DNA sequences can be as long as millions of alphabetical combinations of ‘AGCT’.

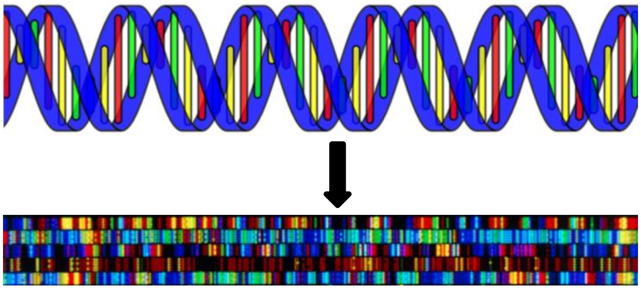
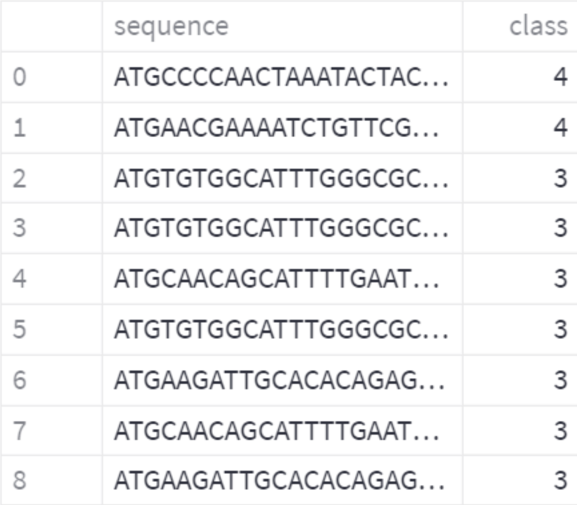


Fig 1.2 Long chain helix structure

Analyzing these sequences and trying to classify them proves to be a tedious task .In this paper, we propose a significantly less time-consuming solution for machine learning algorithms to classify these sequences . ML algorithms can be trained to identify the sequences of different living organisms. This algorithm is not limited to single-species DNA classification. We can have DNA classification of multiple species depending on the data used for training, it can be modified to meet specific needs. The structure of the data needed to train has a certain format[6] i.e., we have to break every sequence into eight words as shown in fig 1.1 ATGCATGC, or it can vary with every helix, and then we have to classify this sequence into different classes. Classification is done using a variety of different algorithms and comparing them to achieve better results[3,4,7]. It comes in handy to choose the best algorithm to achieve the required level of classification. This work is done faster by introducing parallel processing[1], where in all the different algorithms will run at the same time by creating the as individual process and present results even more quickly[2]. These different algorithms are further optimized to yield better results.

II-METHOLOGY

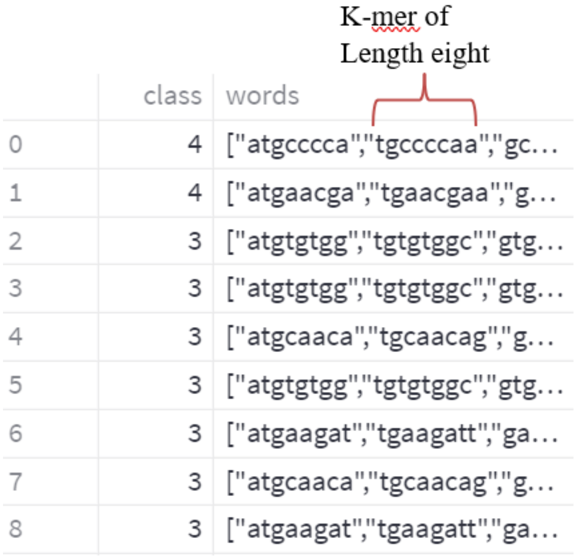
DNA is lengthy combinations of four letters, each representing one base. These strings are uneven in length and change depending on the organism.



	sequence	class
0	ATGCCCCAACTAAATACTAC...	4
1	ATGAACGAAAATCTGTTCG...	4
2	ATGTGTGGCATTTGGGCGC...	3
3	ATGTGTGGCATTTGGGCGC...	3
4	ATGCAACAGCATTTTGAAT...	3
5	ATGTGTGGCATTTGGGCGC...	3
6	ATGAAGATTGCACACAGAG...	3
7	ATGCAACAGCATTTTGAAT...	3
8	ATGAAGATTGCACACAGAG...	3

Fig 2.1 Long DNA sequences

These sequences, acts as data for our ML algorithms,
But they cannot be directly used to train the algorithm. A technique called k-mer [13] counting in NLP is proposed to convert the sequence into trainable data. This method involves taking the long biological sequence and breaking it down into small sequences & Converting sequences into vectors of words of finite length that converts it to trainable data for ML algorithm.


K-mer of
Length eight

	class	words
0	4	["atgcccca","tgccccaa","gc...
1	4	["atgaacga","tgaacgaa","g...
2	3	["atgtgtgg","tgtgtggc","gtg...
3	3	["atgtgtgg","tgtgtggc","gtg...
4	3	["atgcaaca","tgcaacag","g...
5	3	["atgtgtgg","tgtgtggc","gtg...
6	3	["atgaagat","tgaagatt","ga...
7	3	["atgcaaca","tgcaacag","g...
8	3	["atgaagat","tgaagatt","ga...

Fig 2 :- DNA sequences are broken into K-mers of length Eight

These data sets are then further divided into different classes, i.e., 'G Protein Coupled Receptors', 'Synthase', 'Tyrosine Kinase', 'Tyrosine Phosphatase', 'Synthetase', 'Ion channel', and 'Transcription Factor' by grouping all the different data elements present in different classes. This data is then sent to a count vectorizer to create trainable data for ML algorithms.

This data is fed to four different ML algorithms, i.e. SVM, Random Forest, KNN, Decision Tree With the use of ML algorithms, we calculated accuracy, precision, recall, and F1 score, which are the factors that are used to judge the DNA sequences.

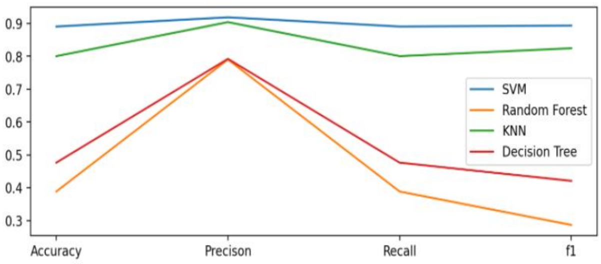


Fig 2.5- Shows the comparison of parameters of different algorithms used.

We prioritized F1 score as it is a better measure to use when we need to seek a balance between precision and recall [5]. So final result is concluded on the basis of F1-score .While testing data with multiple algorithms is a very time-consuming process, so by using parallel computing libraries, i.e., multiprocessing libraries [1], we reduce this time bearing process by creating individual process and running them parallelly rather than working linearly on individual process [10,11,12].This paper presents concludes the best algorithm out of four ML algorithms, i.e. SVM, Random Forest, KNN, Decision Tree to be used for the DNA classification

III - RESULTS AND DISCUSSION

A performance graph is been depicted in figure.3.1. accuracy, precision, recall, and F1 score and of different algorithms used

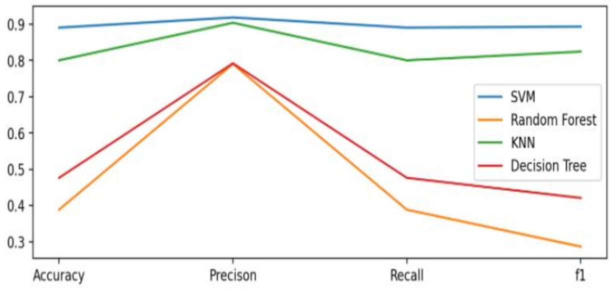
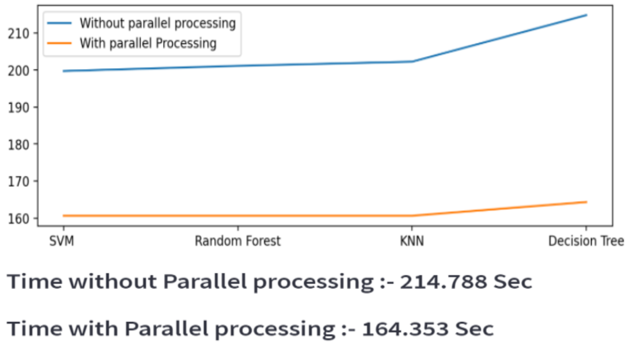


Fig 3.1 Parameter comparison of ML algorithms

```
Best classifier is : SVM with F1 score : 0.893
```

From Figure.3.1, it can be said that SVM has the best performance in classifying DNA sequences compared to the other three algorithms. As the graph suggests, all four factors: accuracy, precision, recall, and F1-Score are the highest in SVM after training it on the provided datasets. Finally, the comparison of results obtained after running all four algorithms in Parallel and Non parallely we get the time comparison as shown in fig 3.1



```
Time without Parallel processing :- 214.788 Sec

Time with Parallel processing :- 164.353 Sec
```

Fig 3.2- Gives the time comparison while parallel processing

From Figure.3.2 it can be clearly seen that by parallel computing we have reduced the time of execution

IV- CONCLUSION

Using four ML algorithms, we have successfully classified DNA sequences. The models are finely optimized for obtaining higher accuracy and better classification of sequences with improved performance. The combined execution time of all four algorithms is significantly reduced by parallel processing.

REFERENCES

[1] *RoboticsTaeHong Kim ,Yoon SeokCha,ByeongChunShin,ByungRaeCha"Survey and Performance Test of Python-based Librariesfor Parallel Processing*".

[2] *Petschow, M., Bientinesi, P.: The Algorithm of Multiple Relatively Robust Representations for Multi-Core Processors. In: PARA 2010: State of the Art in Scientificand Parallel Computing, Python in HPC*

[3] *Issam Hammad, Kamal El-Sankary, and Jason Gu. " A Comparative Study on Machine Learning Algorithms for the Control of a Wall Following Robot." 2019 IEEE International Conference on and Biomimetics (ROBIO). IEEE, 2019*

[4] *Yuvalı, M.; Yaman, B.; Tosun, Ö. Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets. Mathematics 2022, 10, 311. https://doi.org/10.3390/ math10030311*

[5] *S. M. Lim, A. B. M. Sultan, M. N. Sulaiman, A. Mustapha, and K. Y. Leong, "Crossover and mutation operators of genetic algorithms," Int. J. Mach. Learn. Comput.,vol. 7, no. 1, pp. 9–12, 2017, doi: 10.18178/ijmlc.2017.7.1.611.*

[6] *I. V. Kotenko, I. B. Saenko, and A. G. Kushnerevich, "Architecture of the parallel big data processing system forsecurity monitoring of internet of things networks," SPIIRAS Proc., vol. 4, no. 59, pp. 5–30, 2018, doi: 10.15622/sp.59.1*

[7] *J. B. Prajapati and S. K. Patel, "Performance Comparison of Machine Learning Algorithms for Prediction of Students' Social Engagement," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 947-951, doi: 10.1109/ICCMC51019.2021.9418260.*

[8] *E. Tejedor et al., "PyCOMPSs: Parallel computational workflows in Python," Int. J. High Perform. Comput. Appl., vol.31, no. 1, pp. 66–82, 2017, doi: 10.1177/1094342015594678.*

[9] *S. R. M. Zebari and N. O. Yaseen, "Effects of Parallel ProcessingImplementation on Balanced Load-Division Depending on Distributed Memory Systems Client / Server Principles," vol.5, no. 3, 2011*

[10] *.K.Asanovíc et al., "The Landscape of Parallel Computing Research : A View from Berkeley," pp. 1–54, 2006*

[11] *Y.Babuji et al., "Introducing Parsl: A Python Parallel ScriptingLibrary," pp. 1–2, 2017, [Online]. Available:https://doi.org/10.5281/zenodo.891533#.WdOP KS_nvdE.me*

[12] *Y.Babuji et al., "Parsl: Pervasive parallel programming in Python," HPDC 2019- Proc. 28th Int. Symp. High-Performance Parallel Distrib. Comput., pp. 25–36, 2019, doi:10.1145/3307681.3325400.*

[13] *AiminYang,WeiZhang"Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA"*

[14] *fabiocattaneo, germanoguerra, melaniaparisi,marta de marinis, domenicotafuri, mariapiacinelli, and rosarioammendola on " Cell-Surface Receptors Transactivation Mediated by G Protein-Coupled Receptors" Int J Mol Sci. 2014 Nov; 15(11): 19700–19728. Published online 2014 Oct 29. doi: 10.3390/ijms151119700*

[15] *Gunasekaran H, Ramalakshmi K, Rex Macedo Arokiaraj A, Deepa Kanmani S, Venkatesan C, Suresh Gnana Dhas C ."Analysis of DNA Sequence Classification Using CNN and Hybrid Models." Comput Math Methods Med. 2021 Jul 15;2021:1835056. doi: 10.1155/2021/1835056*