Predicting Viral Genome in Human DNA using Deep Learning

Harshal Kallurkar[1], Archana Kadam[2], Pratik Mundokar[3], Sanket Farande, Sahil Bahadure

Abstract*: Classifying the given DNA sequence into discrete normal or abnormal classes has been a longstanding problem in the field of genomics. Previous work includes text-based alignment classification tool of BLAST which has resulted in many genomes being labelled as "unknown". To solve this problem, this paper proposes an approach using convolutional neural network-based (CNN) model, which has two branches, namely the pattern and frequency branches. The former would predict the occurrence of pattern and the latter would predict the frequency of occurrence of the given viral pattern i.e. the DNA sequences. Traditionally CNN's have been used in the area of image processing. But in the past few years, successful experiments have also been conducted in the field of DNA sequencing by using CNN's. The training dataset used, contains DNA sequences obtained from nineteen metagenomic experiments of BLAST (conducted over the period of 2011-2018). This model can be further adapted and can be built as a recommendation model to do further research on sequences which are unlabeled. Detecting and classifying the sequences in this way would eventually help in improving our existing information about the cause of various disease and hence help in faster development of remedies for the same.

Keywords: Convolutional Neural Networks (CNN), Genome, metagenomic contigs, deep learning.

I. INTRODUCTION

To build a model which would accurately classify the human genome as normal or abnormal, detection and then classification of major viruses should be done. For example, there is evidence suggesting that there may be a role of viruses in the development of chronic diseases such as Diabetes and sclerosis. The next generation sequencing technologies provide a tool to obtain the DNA sequences present directly in the clinical samples. As of now, detection of viruses in the human genome is performed by alignment [1], which means that the given virus is matched with an already available database of viral genomes, and would then give the percentage of similarity between them. This method of alignment checking has proven to be not useful in the cases where the DNA sequences might contain several dissimilar viruses which have absolutely no common part with the existing genomes in the available database. So, as a result most of the viruses which would be tested eventually get classified as "unspecified". A different method of classification can be hence used which is, using Deep learning to learn from available examples to classify the viral genomes. In this proposed model, we discuss the implementation of a classifier built using Convolutional neural networks. For the training of model, we have used DNA sequences collected from nineteen metagenomic experiments conducted over the period of nine years by BLAST [1]. BLAST or Basic Local Alignment Searching Tool, is an algorithm which facilitates comparison of nucleotide sequences with a database of DNA sequences. It also provides the features like Nucleotide-Nucleotide comparison, Protein-Protein comparison. This tool, given DNA/protein sequences, returns the most similar DNA/protein sequences from the respective databases. The DNA sequences are derived from samples like skin, blood etc. The dataset contains gene sequence, their corresponding gene id's and the length of each DNA sequence is set as 300 bp (base pairs) and also the class label i.e. whether the given DNA sequence corresponds to a viral genome or not in binary form (0/1). First, we will describe the architecture and flow of working of our model. In the later sections of this paper, we would give the results of performance measures of the model.

II. REVIEW OF LITERATURE

In 2017, Nurul Amerah Kassim and Dr. Afnizanfailzal Abdullah presented an approach through which they used Convolutional Neural Networks are used to classify the DNA sequences of organisms. They used this to specifically identify the genetic marker for liver cancer from Hepatitis-B DNA sequences. [2]

In 2016, Haoyang Zeng, Matthew D. Edwards, Ge Liu and David K. Gifford used CNN to predict the protein-DNA binding. They used a large database of transcription factor binding sites. [3]

In 2016, Giosué Lo Bosco, Riccardo Rizzo, Antonino Fiannaca, Massimo La Rosa and Alfonso Urso presented an approach through which We learnt how deep neural networks can be used in classification of nucleosomes. A nucleosome signifies heritable changes in gene expression & it packs DNA sequences into nucleus of Eukaryotic cells. [4]

In 2016, Michael Leung, Andrew Delong, Babak Alipanahi and Brendan J. Frey presented a paper in which they have demonstrated the importance of analyzing a gene expression and using it further to predict the disease risks well beforehand. They also proposed a computational model for Protein-DNA binding, which can predict cell variables causing various diseases. [5]

In 2017, Sirajul Salekin, Jianqiu (Michelle) Zhang and Yufei Huang proposed an approach in which deep learning was

sed to identify TFBS, which are mainly responsible for controlling the gene expression. [6]

In 2017, Asmaa Abo BakrKamel, Mai S. Mabrouk, Mohamed Waleed Fakhr presented an approach in which Support Vector Machines were used to predict DNA methylated sites, which are a major factor leading to major diseases like cancer, diabetes etc. [7]

III. ALGORITHM DESCRIPTION

We are using the Convolutional Neural network (CNN) [2]-[3] for detecting the abnormality in the human virome sequence. CNN are a class of deep neural networks, used primarily in visual imagery and image classification. A typical CNN consists of an input and an output layer and at the same time many hidden layers. The activation function is usually a RELU layer which is subsequently followed by additional layers such as convolutional layers, fully connected layers etc.

ReLU is the abbreviation for Rectified Linear Unit, which is applied to non-saturating activation function (generally a sigmoid function is preferred as it affects least in terms of accuracy of results).

The architecture of a CNN consists of different layers that transform the input volume into an output volume. The convolutional layer consists of learnable filters. During the forward pass of the CNN, dot product is computed between the entries of the layer and the input.

In the first step of training, two models have been trained (pattern and frequency) independently, output removed from layers and middle layers being used as Frequency & Pattern branches in the final model. In the second step of training, only the parameters in the final layer of architecture are changed i.e. excluding weight and the bias values. In the last step of learning, about two-thousand parameters are changed. Using this approach reduces the amount of overfitting in comparison to other simpler approaches.

IV. ARCHITECTURE & WORKING

The model which we have built uses CNN [2],[4] to detect the presence of viral genome in the DNA sequence. First, the model would get the input as the RAW DNA sequence selected from the 19 metagenomic experiments conducted by BLAST (as shown in fig1.). Preprocessing technique that has been used is One Hot Encoding, in which legal combinations of values are those with a single high bit value (1) and the others as low (0). The preprocessed data is then given as input to two branches namely pattern and frequency branches. The first branch i.e. pattern branch would predict how well the viral genome matches with the DNA sequences. [3],[6] The global max pooling technique has been used which would return only the maximum value for each filter. The second branch i.e. frequency branch returns frequencies of each pattern. In this later, global average pooling would give the average value for each filter. The output of both these models are merged to form a fully-connected layer of CNN [2]. Then, the activation function that is being used is the sigmoid function which would then predict if a viral genome exists in it or not.

The flowchart of working of our model has been shown in fig 1. The architecture of the pattern and frequency branch has also been shown in fig2. and fig3.

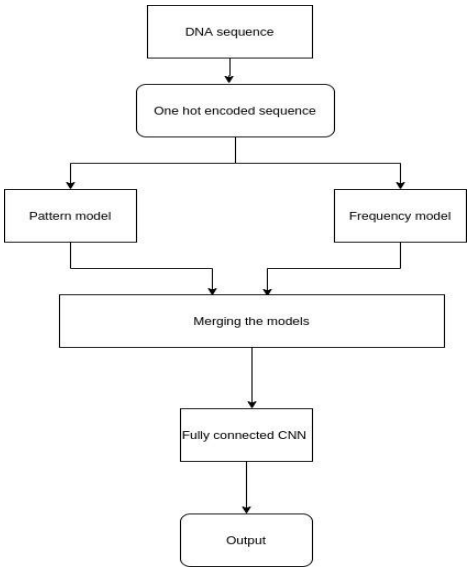


Fig 1. Working and Architecture

The model has been trained on contigs collected from 19 experiments, which were also sequenced from different samples like skin, blood etc. These contigs were then shuffled, again partitioned into training (80%), 10% testing and 10% validation data. The test conducted was only used to evaluate the final performance of the model.

The diagram shown below indicates the working of two of the branches of the model.

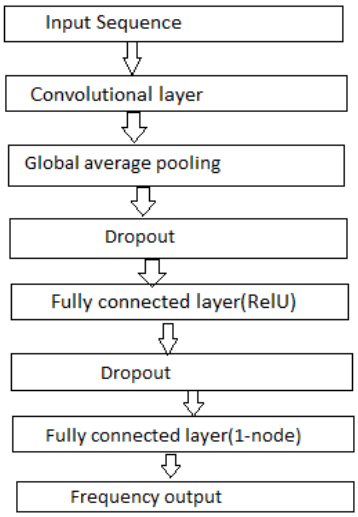


Fig 2. Frequency branch

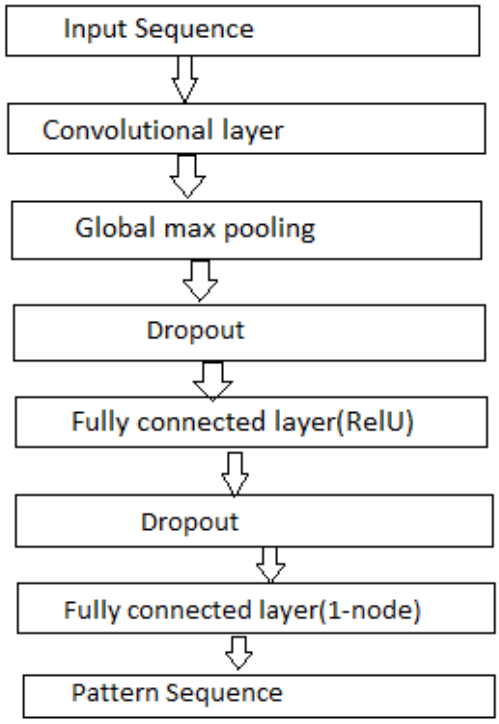


Fig 3. Pattern branch

A short description of working of both branches goes as: after the preprocessing stage, the input sequence being interpreted as 1-D array, is convolved and hence its size is reduced. Through the max/average pooling, respective values are taken from the layers & forwarded to further layers in the network. The dropout value specified will lead to skipping of some of the output values of internal layers.

V. RESULTS OF THE MODEL

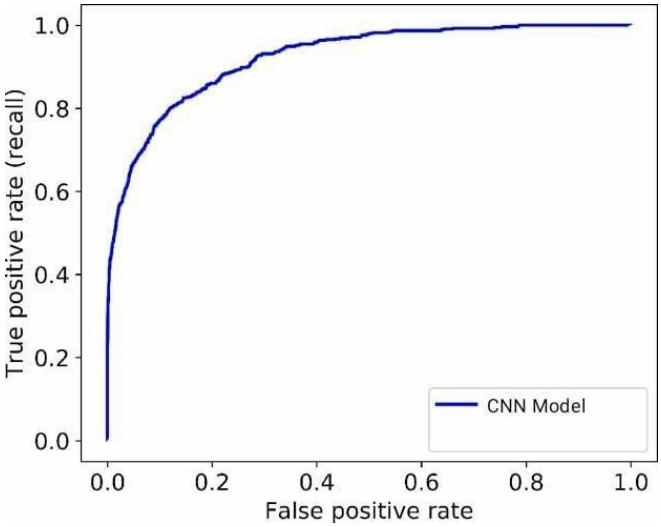
We apply the model to the test set and predict the likelihood of the genome being viral or not. [5] Further, the top N sequences can be analyzed for checking if they are actually viral or not. The output is generated in a log file (merged file), the description of which is as follows: It contains the structure of the model i.e. the number of layers in it, their exact sequence, number of epochs, the AUROC value, the accuracy. In every epoch run, the model is given the all sequences in batch size of 128 which are selected randomly and shuffled. The size of the training data set was 87243 samples.

We also inferred from the output file that Area under the ROC curve for the test set is 0.9448 and for the validation set is 0.9364. We also tested the model on a random sample of 15 DNA sequences as input and got the area under the ROC curve as 0.89285.

The table shown below also shows detailed changes in AUROC values for every epoch, along with loss and accuracy after every run, on validation set.

Epoch No.	AUROC improved from	AUROC improved to	Loss	Accuracy
1	-inf	0.91737	0.0564	0.9849
2	0.91737	0.92519	0.0455	0.9878
3	0.92519	0.92683	0.0415	0.9885
4	0.92683	0.92683	0.0378	0.9894
5	0.92683	0.92867	0.0362	0.9898
6	0.92867	0.93314	0.0321	0.9909
7	0.93314	0.93314	0.0311	0.9910
8	0.93314	0.93639	0.0288	0.9917

The results have also been shown in graphical format which is a plot of True positive rate(recall) against False positive rate. This graph is the final graph of the merged models applied on test data (pattern and frequency branch).



VI. CONCLUSION

In this experiment, the input to the convolution layer is DNA sequence as a 1-D array with one channel per possible nucleotide value. The final output of the model (predicted value) is recorded in a log file containing every detail of Computation. Considering the results mentioned above, it may be concluded that this method be explicitly used as a replacement to an already existing method used by BLAST i.e. sequential text matching. We recommend this system to detect abnormal patterns in DNA sequences so as to cure early. Further the systems can be extended to detect particular disease patterns. Also, accuracy variation can be observed by increasing datasets.

VII. REFERENCES

[1] Basic Local Alignment Searching Tool(Biotechnology), URL: https://en.wikipedia.org/wiki/BLAST_(biotechnology)

[2] Nurul Amerah Kassim and Dr. Afnizanfailzal Abdullah, Classification of DNA Sequences Using Convolutional Neural Network Approach, UTM Computing Proceedings, Innovations in Computing Technology and Applications, Volume 2, Year 2017.

[3] Haoyang Zeng, Matthew D. Edwards, Ge Liu and David K, Convolutional Neural Networks architecture for predicting DNA-protein binding, Oxford University Press, year 2016.

[4] Giosué Lo Bosco, Riccardo Rizzo, Antonino Fiannaca, Massimo La Rosa and Alfonso Urso, A Deep Learning Model for Epigenomic Studies, 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems.

[5] Michael Leung, Andrew Delong, Babak Alipanahi and Brendan J. Frey, Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets, Vol. 104, No. 1, January 2016 | Proceedings of the IEEE.

[6] Sirajul Salekin, Jianqiu (Michelle) Zhang and Yufei Huang, A deep learning model for predicting transcription factor binding location at Single Nucleotide Resolution, 2017 IEEE.

[7] Asmaa Abo BakrKamel, Mai S. Mabrouk, Mohamed Waleed Fakhr, Prediction of DNA Methylation for Every CpG Loci in Chromosome 18 in Embryonic Stem Cells Using Support Vector Machine, 2017 IEEE.

AUTHOR'S PROFILE

Harshal Kallurkar is currently a final year Computer Engineering Student at Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune. His areas of interest are Machine Learning, Algorithms and Computer Networks and wants to pursue Master's in these subjects. He has been a Winter & Summer trainee at Persistent Systems, Pune and has secured a percentile of 97.11 in GATE Computer Science 2020. He is currently working under Prof. Archana Kadam

Archana Kadam is currently a Professor in Computer Engineering Department at Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune. Her areas of interest are Data Science and knowledge Engineering and published many papers in this domain. Her subjects of interest are Data Analytics, Data mining and warehousing, Compilers. She has completed master in Computer Engineering from Pune University. She has done many courses in the field of Machine learning and Data science through many platforms like NPTEL, Coursera etc.

Pratik Mundokar is currently a final year Computer Engineering Student at Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune. His areas of interest are Data Science, Full Stack Development and Core Computer Science Subjects like Operating systems and Computer Architecture. He has completed many courses in the field of Machine learning and Data science through many platforms like NPTEL, Coursera etc. He is currently working under Prof. Archana Kadam.

Sanket Farande is currently a final year Computer Engineering Student at Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune. He has completed the Google Applied CS program conducted at institute level by the Computer Engineering Department. His areas of interest are Algorithms, Databases and Web development in JavaScript. He has completed many certifications by NPTEL including DBMS in which he secured 93% marks and secured a gold medal. He is currently working under Prof. Archana Kadam.

Sahil Bahadure is currently a final year Computer Engineering Student at Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune. His areas of interest are Data Science, Front end Web development in frameworks like React, Angular etc. He also has completed many projects involving Django Framework, Bootstrap, React JS. He is currently working under Prof. Archana Kadam.