Deep Learning Method For Predicting Nucleosome Positioning in Genomes using Genomic Signal Processing Techniques

Akshay Valsaraj
BITS Pilani, K. K. Birla Goa Campus
f20180608@goa.bits-pilani.ac.in

Gourav Saha
BITS Pilani, K. K. Birla Goa Campus
f20180639@goa.bits-pilani.ac.in

Ithihas Madala
BITS Pilani, K. K. Birla Goa Campus
f20180607@goa.bits-pilani.ac.in

Prof. Sumit Biswas
VISTA Lab
Department of Biological Sciences
BITS Pilani, K. K. Birla Goa Campus
sumit@goa.bits-pilani.ac.in

Abstract

Application of Digital Signal Processing techniques (DSP and DIP) to solve Genomics problems initiated the new field Genomics Signal Processing (GSP) which concentrates to encode the Genomics signals based on DSP/DIP framework. In this Genomics era, high throughput DNA sequencing and the use of DNA microarray to simultaneously conduct huge number of experiments has lead to many signal/image processing problems. There is an emergent need to develop signal/image-processing techniques to examine data and determine relationship between genes. In this paper, we will focus on application of DSP and Machine Learning in Biomolecular Sequence Analysis.

1 Introduction

Genomic Signal Processing (GSP) is the processing, analysis and utilization of genomic signals to obtain relevant insights into the gene. This information is then used in the interpretation of particular gene motifs and families further resulting in the ability to classify diseases. DNA's critical role in all life's processes warrants its interpretation from different perspectives and interestingly, signal processing techniques reveal a better understanding. Over the recent years, GSP has captured researchers' attention because it allows quick analysis of genomic data with highly optimised coding for signal processing. Furthermore, this spectral data can easily be transformed into spectogram representations.

The nucleosome is considered the most basic unit of eukaryotic chromatin [6]. The DNA is tightly packaged around the histone proteins under the action of regulatory proteins as well as recognised gene sequences [7]. The DNA packaged in the nucleosomes have regulated accessibility and play an important role in transcriptional control, DNA replication, DNA repair and RNA splicing etc. [1]. We used a dataset containing DNA sequences that either promote or inhibit nucleosome formation to test the accuracy with which our model can differentiate between the two across multiple species.

2 Methods

Datasets

Preprint. Under review.

In this article, we considered the following three species: (i) H.sapiens;(ii) C.elegans; and (iii) D.melanogaster.The data was obtained from [3].According to the paper, H.sapiens genome and its nucleosome map contains a hugeamount of data hence the nucleosome forming sequence samples (positive data) and the linkers or nucleosome inhibiting sequence samples (negative data) were extracted from chromosome 20.

As for the other two species, namely C.elegans and D.melanogaster, the positive and negative data were extracted from their entire genomes. In the datasets thus formed from the three organisms, each of the DNA fragments was assigned with a nucleosome formation score to reflect its propensity to form nucleosome: the higher the score was, the more likely the fragment would be in forming a nucleosome. The DNA fragments with the highest nucleosome formation scores were selected as the nucleosomal sequences, while those with the lowest scores as the linker sequences.

A dataset containing many redundant samples with high similarity would be lack of statistical representativeness. A predictor, if trained and tested by such a biased benchmark dataset, might yield misleading results with overestimated accuracy [2]. To get rid of redundancy and avoid bias, the CD-HIT software (Fu et al., 2012) was used with the cutoff threshold set at 80% to remove those DNA fragments with high sequence similarity (note that the most stringent cutoff threshold for DNA sequences by CD-HIT was 75%).

Species type	Nucleosome forming sequences	Nucleosome inhibiting sequences
H.sapiens	2273	2300
C.elegans	2567	2900
D.mealnogaster	2900	2850

Table 1: Nucleosome forming and inhibiting sequences in the Three datasets used

Converting DNA Sequence to Signal
In order to apply signal processing techniques on gene sequences, there is a need to first convert the gene sequences from the nucleotides to more discrete numerical sequences. The latter can then be treated as digital signals. One of the most common ways used is to assign different numbers to each nucleotide and generating the signal. These mappings can either be fixed (that does not take any biological context) or they might be adaptive depending on adjacent nucleotides or thermodynamics properties. In general, there is no agreement on which representation makes the mos accurate signals. In our case, we use the tetrahedron representation in which each nucleotide corresponds to a vertex of a three-dimensional structure that is characterized by having equal distances between every pair of vertices [8].

Forming Spectrograms
The DNA signal is used to develop spectrogram images for our model. N-point Discrete Fourier Transform (DFT) is first applied to the DNA signal after which Short Time Fourier Transform (STFT) is applied over a sliding window of 9 with an overlap of 6.The hamming function [4]is used for the window to prevent spectral leakage. Once processed, colour spectrograms are defined by super imposing the three signals under the red, green and blue (RGB) values. The tetrahedron coefficients further increase the segregation of the sequences. In the spectrogram, the horizontal axis corresponds to the nucleotide number and the vertical axis refers to the discrete frequency of the DFT measured in the STFT cycles per window size. The spectogram was then resized to $224x224$ so that it could be used as an input for the deep learning model. The tetrahedron coefficients used are :

$ar = 0$, $ag = 0$, $ab = 1$,
$tr = 0.911$, $tg = -0.244$, $tb = -0.333$,
$cr = 0.244$, $cg = 0.911$, $cb = -0.333$,
$gr = -0.817$, $gg = -0.800$, $gb = 0$,

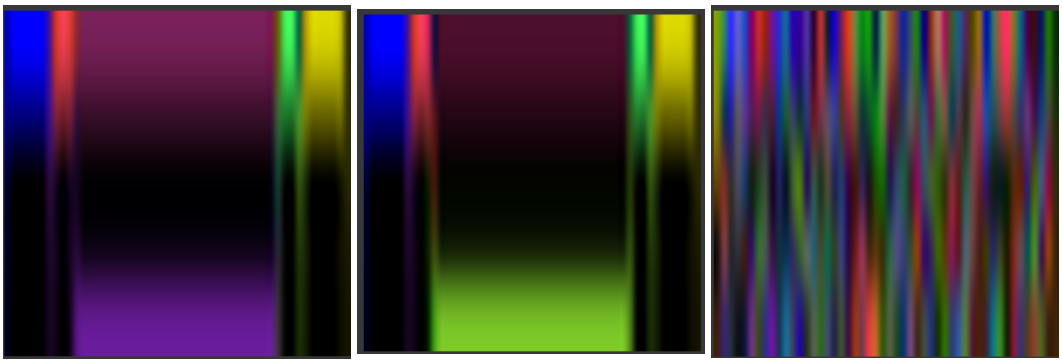

Figure 1: **Left**:Spectrogram of DNA of length 147bp with bases A, T, C, and G , respectively. The nucleotide A is represented by the color blue, T by red, C by green, and G by yellow.The interaction between the various nucleotides is visualized as the superposition of colors representing those nucleotides. This represents the sequence having AT repeat in between, the purple is due to a superposition of red and blue from A and T.
Center : This represents the sequence having CG repeat in between, the greenish yellow is due to a superposition of yellow and green from G and C.
Right: A spectogram of a sequence taken from the dataset.

Model
To train the model we used the densenet architecture [5].DenseNet-121 consists of 121 densely connected convolutional layers with a fully conncted(FC) layer of 1000 units as its final output layer. We removed the final layer and replaced it with a FC layer with two neurons for two class classification.

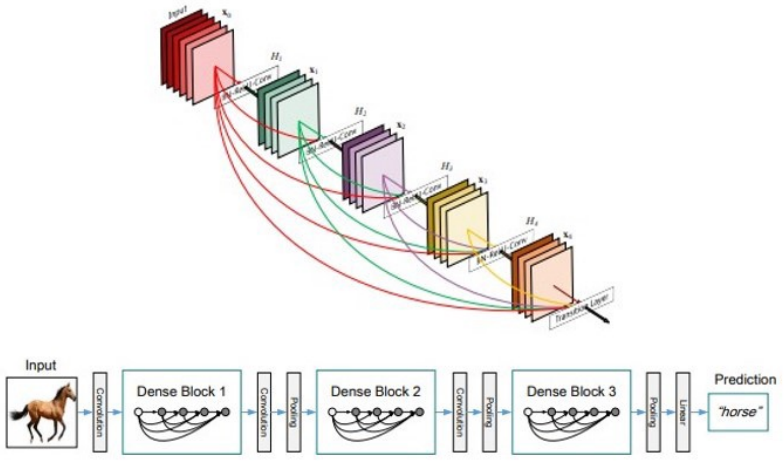

Figure 2: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature-map sizes via convolution and pooling.

3 Results

We present 2 types of results ,For the first type of results we wanted to know how well the pattern of the data could be reflected with different species so we trained on two different species and tested on the third and this was done for all the species. [3] had done a one-vs-All classification results where they trained a model on N-1 sequences and tested it on one of the classes and this was done N times to generate the class for each sequence individually , since this would be very memory intensive we didn't go for this approach.
The second type of results we combined all the 15498 sequences and divided it into 12553 train

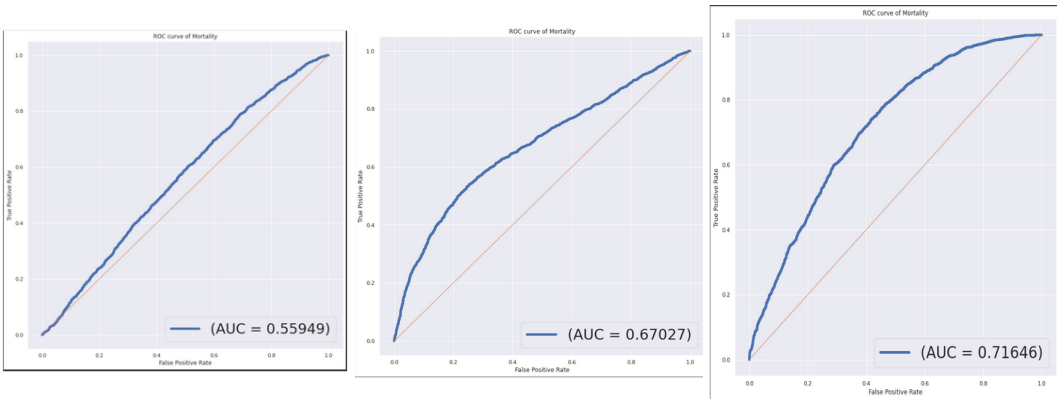

Figure 3: **Left**:ROC curves for C.elegans. **Center** : ROC curves for D.mealnogaster. **Right**: ROC curves for H.sapiens.

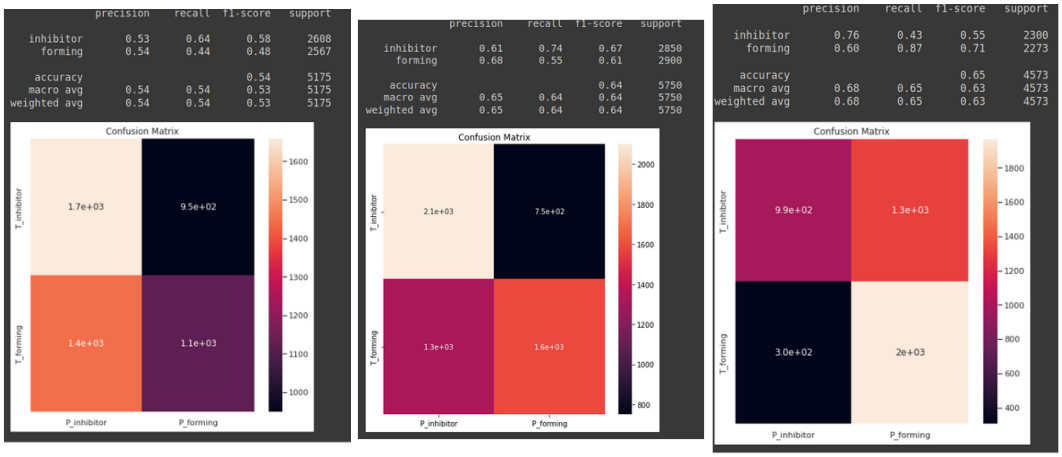

Figure 4: **Left**:Confusion Matrix for C.elegans. **Center** : Confusion Matrix for D.mealnogaster. **Right**:Confusion Matrix for H.sapiens.

sequences, 1395 validation sequences and 1550 test sequenences. In the test sequence 457 sequences belonged to H.sapiens,518 to C.elegans and 575 to D.mealnogaster.
Moreover in the test results we made sure to show the results for each species separately so that we could compare it with the previous results obtained by [3].

Type 1 Results
In this case we trained on two species and tested on the third. For example for H.sapiens we trained on C.elegans and D.mealnogaster and tested on H.sapiens.

Type 2 Results
In this case we trained on 12553 train sequences with 1395 as validation sequences and predicted on 1550 test sequences.

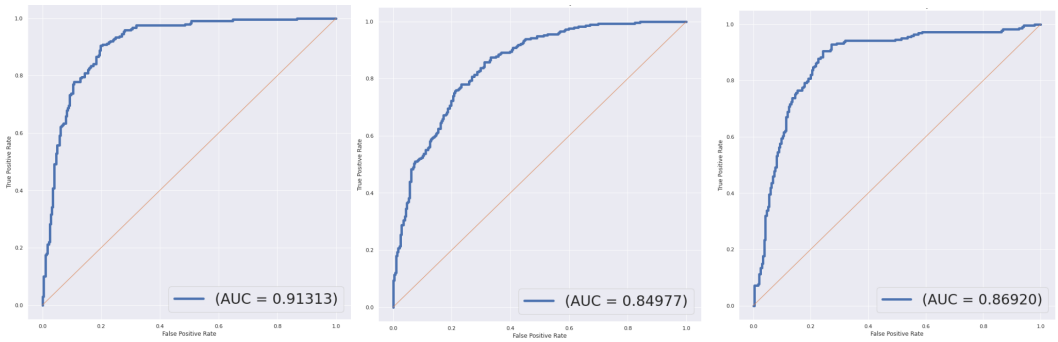

Figure 5: **Left**:ROC curves for C.elegans. **Center** : ROC curves for D.mealnogaster. **Right**: ROC curves for H.sapiens.

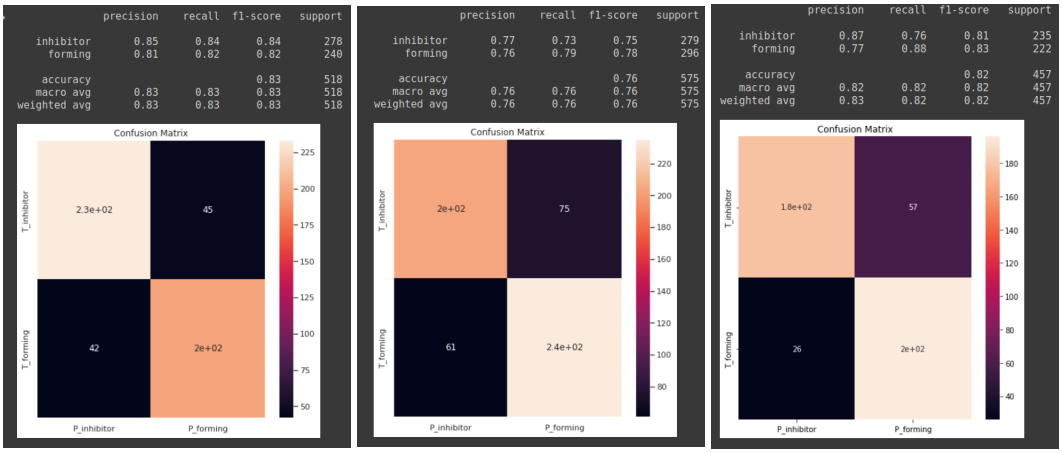

Figure 6: **Left**:Confusion Matrix for C.elegans. **Center** : Confusion Matrix for D.mealnogaster. **Right**: Confusion Matrix for H.sapiens.

4 Conclusion

Species	Optimal parameters			Metrics			
	k	λ	w	Acc (%)	Sn (%)	Sp (%)	MCC
H.sapiens[a]	4	6	0.5	86.27	87.86	84.70	0.73
C.elegans[b]	3	11	0.5	86.90	90.30	83.55	0.74
D.melanogaster[c]	4	7	0.2	79.97	78.31	81.65	0.60

[a]Using the benchmark dataset given in Supplementary Material S1.
[b]Using the benchmark dataset given in Supplementary Material S2.
[c]Using the benchmark dataset given in Supplementary Material S3.

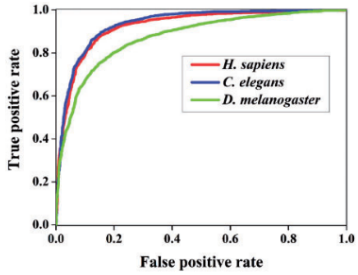

Figure 7: A graphical illustration to show the performance of the iNuc- PseKNC by means of the ROC curves. The areas under the ROC curves, or AUROC, are 0.925, 0.935 and 0.874 for H.sapiens, C.elegans and D.melanogaster,respectively binary

From 7 we can see that we have come close to the results mentioned in the previous paper but we cannot take these as the benchmark as :

1. [3] has used a one vs all cross validation which uses all the data except one whereas was our results are only based on 80% of the sequences which it has been trained on.
2. Compared to previous works we have also presented the results of the model when it is tested on species trained on different species which would prevent similar sequences from the same species providing biases in the data.This could form as the benchmark for future research.

References

[1] N M Berbenetz, C Nislow, and G W Brown. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genet*, 2010;6(9):e1001092: 1001092, 2010-09-02.

[2] K C Chou and H B Shen. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat Protoc*, 2008;3(2):153-162:494, 1038/nprot. 2007.

[3] S H Guo, E Z Deng, and L Q Xu. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, 2014;30(11):1522-1529, 1093/bioinformatics/btu083.

[4] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978. doi: 10.1109/PROC.1978.10837.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[6] T J Richmond and C A Davey. The structure of DNA in the nucleosome core. *Nature*, 2003;423(6936):145-150, 1038/nature01595.

[7] S C Satchwell, H R Drew, and A A Travers. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, 191(4):659–675, October 1986.

[8] B D Silverman and R Linsker. A measure of DNA periodicity. *J. Theor. Biol.*, 118(3):295–300, February 1986.