Protein Classification in Genomics Data with Deep Learning Network

Sthefanie Jofer Gomes Passo
Electrical and Computer Engineering Department
University of Texas at San Antonio
San Antonio, Unided States
Sthefanie.Passo@utsa.edu

Walter Richardson
Department of Mathematics
University of Texas at San Antonio
San Antonio, Unided States
Walter.Richardson@utsa.edu

I. Introduction

Predicting phenotypes from genetic data is also a major area of interest of deep learning [1]. A first step in performing these types of predictions is to specify what genetic variants are present in an individual genome. Genetic variants can improve in different physical characteristics, genetic sickness that people are born with, etc. People would be able to predict sickness and improve in their heath if Machine Learning models learn how to extract characteristics of this genomics data. This problem has been addressed by DeepVariant, which applies a CNN to make variant calls from short read sequencing [2].

Deep learning methods are a class of machine learning techniques capable of identifying highly complex patterns in large data base to genomic data. Similarly, the identification of transcription can be executed with the addition of features from the Encyclopedia of DNA elements [3], [4] (ENCODE) project, as well as transcription-start-site sequencing and RNA-seq signals. The methylation state of DNA, which also influences the expression of genes, has been inferred from three-dimensional genome topology (on the basis of Hi-C) and DNA sequence patterns [5].

Genomicists and general biomedical researchers who seek high-level understanding of how to apply Deep Learning in their data can use this research. This application can be a introduction of machine learning to Computer scientists as well. However, we do not provide a survey of deep learning in the biomedical field, which has been broadly covered in recent reviews [6]–[10]. All the experiments of this paper can be found at https://github.com/SthePasso/DeepLearningAndGenomicData with the comparation of different dimentionality models and techniques models build for interested researchers to study a deep learning network to automatic classification of proteins in the DNA.

II. Method

This section describes the proposed method for the automatic classification of protein in the genomcs data. The "Fig. 1" illustrates the proposed method where the genome data is split in train, validation and test. The train data will be send to the network to train each architecture. The models will evaluate the data and interpret the probabilistic result in a high level of characteristic abstracted. Each of this sections will be described with details in the data base, preprocessing, feature extractions and evaluation metric.

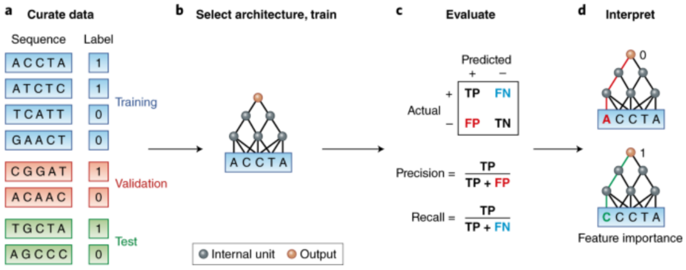

Fig. 1. Deep learning workflow in genomics.

A. Data Base

The Data Base utilized was the Encyclopedia of DNA Elements (ENCODE) project, as well as transcription-start-site sequencing and RNA-seq signals [3], [4], [11]–[13] utilized in previous research [1]. This data base contains 2000 samples, each sample have 50 features of the bases used in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). Each sample is classified with zero (0) or one (1) according is that DNA segment have certain protein that we are searching for. The sequence of this protein is $CGACCGAACTCC$ where for each sample we have a classification if don't have (0) or do have (1) the protein. The "Fig. 3" is clear that the data base is balanced.

This protein sequence could be searched with combinatory analyzes (brute force) but the cost of identifying it in this low level isn't reasonable ("Equ. 1"), once we can analyze the genome in a high level with machine learning and the cost function would be lower considering that for each 500 samples at least 400 be predicted correctly ("Equ. 2"). We also want to do the DNA analysis according to human comportment and methodology that see the data in a high level to figure if it have a protein or not, just like machine learning do with probabilistic models, not with exhausted research looking to combinatory analysis.

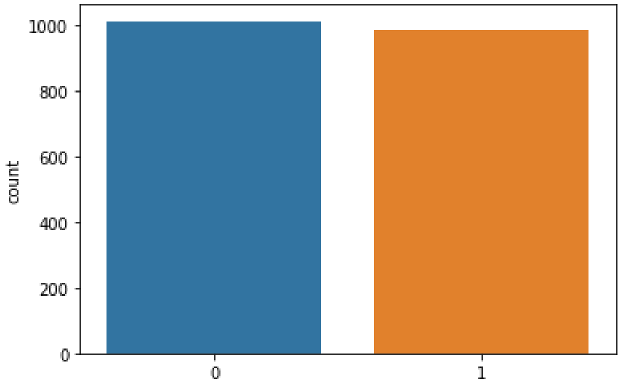

Fig. 2. Genomas that have certain protein or not (CGACCGAACTCC)

$$CombinatoryAnalysis = \frac{50!}{12!(50-12)!} = 12 \times 10^{10} \quad (1)$$

$$CostFunction = Sum[(Actual - Predicted)^2] \times (\frac{1}{Observed})$$

$$= \text{Sum}[(500\text{-}400)^2] \times (\tfrac{1}{50}) = 200(2)$$

B. Preprocessing

The input into a neural network is typically a matrix of real values. In genomics, the input might be a DNA sequence, in which the nucleotides A, C, T and G one hot encoded as [1,0,0,0], [0,1,0,0], [0,0,1,0] and [0,0,0,1]. The data that was (2000, 50, 1) became (2000, 50, 4) after we apply the encoding according to the “Tab. I” and “Tab. II” respectively.

TABLE I
DNA SEQUENCE WITH SHAPE (2000, 50, 1)

DNA Sequence							
C	C	G	...	A	G	T	A

TABLE II
DNA SEQUENCE AFTER APPLYING ONE HOT ENCODING WITH SHAPE (2000, 50, 4)

One hot encoding of Sequence							
0	0	0	...	1	0	0	1
1	1	0	...	0	0	0	0
0	0	1	...	0	1	0	0
0	0	0	...	0	0	1	0

The samples where randomly distributed to the train, validation and test group where fifty (50) percent of the data was to train, twenty five (25) percent to validate and twenty five (25) percent to test [1].

C. Feature Extractions

Many genomic applications, researchers are more interested in the biological mechanisms revealed by the predictive model rather than the prediction accuracy itself [16], [17]. For example, the main motivation for building accurate deep learning models to predict chromatin patterns is a hope to learn new gene-regulation grammar by interpreting the trained model. Although deep learning can achieve state-of-the-art accuracy, it is more challenging to interpret than the more standard statistical models.

In order to understand better what is the math behind the models and how different neural layers work, we did experiments changing the Dimensionality methods and in a second spot changing the Technique methods. In Deep Learning we have layers to do the preprocessing of the data and classification prediction.

Max Pooling, Average Pooling, Global Max Pooling and Global Average Pooling are dimensionality methods related to the preprocessing of the data. Those layers where applied in different models at the same position according to the “Fig. 3”. The Max Pooling was implemented in the state of art [1] where for each 4 features of the genome layer we select the maximum value of them and preserve it reducing the dimension 4 times. Average Pooling select the average value for each 4 layers reducing the dimension 4 times too. Global Max Pooling select the maximum value globally in the layer as well as Global Average Pooling.

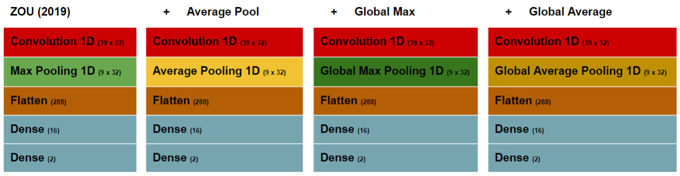

Fig. 3. State of Art with Modified Dimensionalities.

Convolution (CNN), Long short-term memory (LSTM), Convolution Transpose and Dropout are technique related to the prediction of the data. We preserved the models arquiteture of the state of art [1] applying different techniques according to the “Fig. 4”.

CNNs attempts to compress the input, while an upsampling one tries to expand the input. Convolutional neural networks are downsampling by nature, as convolution leaves the output with fewer rows and columns as the input.

LSTMs are explicitly designed to avoid the long-term dependency problem and have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

In a tranposed CNN’s, instead of the input being larger than the output, the output is larger. An easy way to think of it is to picture the input being padded until the corner kernel can just barely reach the corner of the input.

The least technique is Dropout regularization for reducing overfitting and improving the generalization of deep neural networks. It reduce the noise dortin the training when it drop out of the training ramdomly neuronius, in this way the machine learning need to improve itself better do find good weights to do the probabilistic model be accurate even without the neuronius that where droped out of the training.

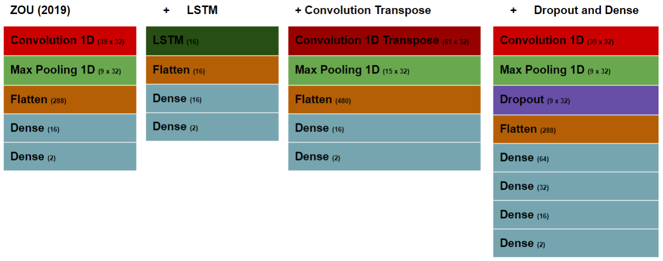

Fig. 4. State of Art with Modified Techniques.

D. Evaluation Metrics

Four (4) main measures usually were considered in the literature for heath care [14], [15] those are the sensitivity (SEN), positive predictive value (PPV), specificity (SPEC) (1 - False Positive Rate (FPR)) and accuracy (Acc) as defined in following:

$$SEN = \frac{TP}{TP + FN} \tag{3}$$

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

$$SPEC = \frac{TN}{TN + FP} \tag{5}$$

$$Acc = \frac{TP + TN}{TN + FP + FP + FN} \tag{6}$$

where TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) indicate the number of heartbeats correctly labeled, number of heartbeats correctly identified as not correspond to the heartbeats, number of heartbeats that incorrectly labeled, and number of heartbeats which were not identified as the heartbeats that they should have been, respectively.

III. Results

The results are presented for two evaluation scenarios of models according to the dimensionality configurations and techniques configurations. "Tab. III" presents a comparison of protein classification results for the proposed methods and the existing algorithms, considering models with different dimensionality configurations. As confirmed by the results, the Max Pooling layer reduce the dimensionality in a better way to the feature extractions and the model infers more accurately if the layer have or not the protein.

When we applied differente techniques configurations the proposed method with Dropout and tree (3) Dense layers can provide a robust solution for the genome data to classify if it have the protein on it as one of the key challenges in dealing with medical data and combinatory analysis. It is shown in "Tab. IV" that our model achieves remarkable outcomes for this data base, the 100% accuracy can happen in small number of labels where it have a linear stability between the data, but when the data is bigger and complex the linear system don't be so assertive, witch is normal.

TABLE III
Comparing the genome classifications models with modified dimensionalities.

Methods	Classe	TP	TN	FP	FN	Acc(%)	se(%)	sp(%)	+p(%)
ZOU [2019]	No	248	239	11	2	98.20	99.21	97.15	97.30
ZOU [2019]	Yes	239	248	2	11	98.20	97.15	99.21	99.17
Average Pooling	No	200	177	59	64	75.40	75.76	75.00	77.22
Average Pooling	Yes	177	200	64	59	75.40	75.0	75.76	73.44
Global Max Pooling	No	183	189	76	52	73.60	75.92	71.37	71.81
Global Max Pooling	Yes	189	183	52	76	73.60	71.37	75.92	75.52
Global Average Pooling	No	183	189	76	52	74.40	77.87	71.32	70.66
Global Average Pooling	Yes	189	183	52	76	74.40	71.32	77.87	78.42

TABLE IV
Comparing the genome classifications models with modified techniques.

Methods	Classe	TP	TN	FP	FN	Acc(%)	se(%)	sp(%)	+p(%)
ZOU [2019]	No	248	239	11	2	98.20	99.21	97.15	97.30
ZOU [2019]	Yes	239	248	2	11	98.20	97.15	99.21	99.17
LSTM	No	255	240	4	1	99.00	99.61	98.36	98.46
LSTM	Yes	240	255	1	4	99.00	98.36	99.61	99.59
Global Max Pooling	No	258	240	1	1	99.60	99.61	99.59	99.61
Global Max Pooling	Yes	240	258	1	1	99.60	99.59	99.61	99.59
Global Average Pooling	No	259	241	0	0	100	100	100	100
Global Average Pooling	Yes	259	241	0	0	100	100	100	100

IV. Conclusion

Deep Learning has demonstrated impressive potential in gemics, and also different mathematics techniques will be better for different kinds of datas. In genomics the combination of Convolution and Dropout got a bigger accuracy than the state of art. The challenge is to verity the comportamente of this data in ordinary differential equations (ODE) and get a predicted result with a high accuracy to replace the Deep Learning network or part of it to my current PhD project. I'm working on a problem of how to combine genome data (or biomedical data), ODE, time-series and Optimal Control. For future research my objectives are to find other genome baselines to similar applicability, test these techniques and other deep learning techniques on them.

References

[1] Zou, J. et al. A primer on deep learning in genomics James. Nature Genetics. Vol 51, page 12–18, www.nature.com/naturegenetics (2019)

[2] Poplin, R. et al. Creating a universal SNP and small indel variant caller with deep neural networks. Preprint at https://www.biorxiv.org/content/early/2018/03/20/092890 (2017)

[3] Liu, F., Li, H., Ren, C., Bo, X. Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Sci. Rep. 6, 28517 (2016).

[4] Li, Y., Shi, W. Wasserman, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics 19, 202 (2018).

[5] Wang, Y. et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci. Rep. 6, 19598 (2016).

[6] Angermueller, C., Pärnamaa, T., Parts, L. Stegle, O. Deep learning for computational biology. Mol. Syst. Biol. 12, 878 (2016).

[7] Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface 15, 20170387 (2018).

[8] Telenti, A., Lippert, C., Chang, P. C. DePristo, M. Deep learning of genomic variation and regulatory network data. Hum. Mol. Genet. 27, R63–R71 (2018).

[9] Yue, T. Wang, H. Deep learning for genomics: a concise overview. Preprint at https://arxiv.org/abs/1802.00810 (2018).

[10] Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. Collins, J. J. Next-generation machine learning for biological networks. Cell 173, 1581–1592 (2018).

[11] Kleftogiannis, D., Kalnis, P. Bajic, V. B. DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res. 43, e6 (2015).

[12] Min, X. et al. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics 18 (Suppl. 13), 478 (2017).

[13] Eser, U. Stirling Churchman, L. FIDDLE: an integrative deep learning framework for functional genomic data inference. Preprint at https://www. biorxiv.org/content/early/2016/10/17/081380 (2016).

[14] Mousavi, S., Afghah, F., Acharya, U. Inter- and Intra-Patient ECG Heartbeat Classification For ArrhythmiaDetection: A Sequence to Sequence Deep Learning Approach.School of Informatics, Computing and Cyber Systems,Northern Arizona University, Flagstaff, USA. 2019.

[15] Passo, S. et al. Classificação de Arritmias com Paradigma Inter e Intra Pacienteutilizando Aprendizagem Profunda. J. Health Inform. Número Especial SBIS - Dezembro: 352-7. 2020.

[16] Hastie, T., Tibshirani, R. Friedman, J. H. The Elements of Statistical Learning Vol. 1 (Springer Science+ Business Media, New York, 2001).

[17] Quang, D. Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 44, e107 (2016).