University of Dundee

Deep Learning in Genomics

J. Divyananda[1,2], S. McKenna[2], V. Prathiba[3], V. Mohan[3] and E. Trucco[2]

1. School of Medicine, University of Dundee, Dundee, UK

2. Computer Vision and Image Processing group (SSE),University of Dundee, Dundee, UK

3. Madras Diabetes Research Foundation, Chennai , India

Introduction

Neural Networks (NN) take inspiration from the neurons of the brain. Neural network is a computational model which is based on the structure and functions of biological neurons. Artificial neural networks consists of layers of interconnected computational units.

Deep Learning (DL) is a subfield of Machine Learning (ML). Genetics is a field of biology which studies how the traits are passed from parents to their offspring, DL can be used to address unique and complicated challenges presented by genetics.

DL techniques: DL tools are rapidly evolving. The main ones used are Convolutional neural network (CNN's) and Recurrent neural network (RNN's). The other methods are deep belief networks, GAN's, fully connected CNN's, Bayesian filters, SVM's were used before deep learning and even now for few ML applications in genomics.

DL address different challenges of genomics like regulatory genomics, protein structure prediction, cellular imaging, and DNA and RNA sequence prediction, splicing of sequence and generating the DNA & RNA sequence. Figure below shows a neuron and neural network
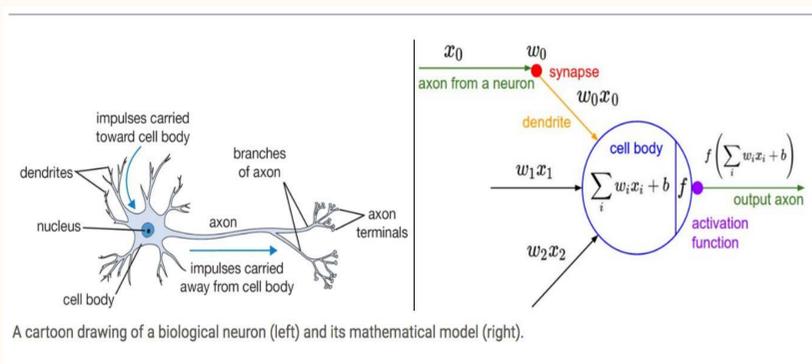

Figure 1 : A biological neuron and it computational model used in an NN [1]

Deep Learning in Genomics

Research on Deep Learning has demonstrated success in various application fields including healthcare and biotechnology [3]. The different applications being hand-writing recognition, robotics, mammography and analysis of molecules in discovery of new drugs [4].

The present applications in genomics found by our review are regulation of gene expression (including epigenetic modifications, interactions between proteins and regulatory sequences, prediction of splicing variants of messenger RNA (mRNA)), genome analysis and SNPs research (SNPs in in coding and non-coding regions of genome, protein structure prediction) and early detection of cancer (research and monitoring of biomarkers)[3]. Many papers discuss the applications of deep learning and different architectures and algorithms have been developed for the same.

The hardware platforms used are titan, GP, nvidia. The number of GPU's used were from two to three. The type data used for analysis is usually a DNA or RNA sequence, ENCODE project datasets which contains nucleotides, mRNA sequences, proteins. Most of the data is publicly available.
The table below shows the summary of DL methods used in genomics:

Deep Learning Packages	Category
Deepmethyl, DeepChrome	Epigenetic modifications
Basset, Deep motif, DanQ, PEDLA, DECRES,FIDDLE, DEEP, DeepBind, DeepCNF	Proteins and regulatory sequences
DeepSlice	Predictions of splice variants of mRNA
DeepSea, Diet Networks, DANN	SNP's in coding and non-coding regions of genome
PconsC2, DNCON2, MUST-CNN	Protein structure prediction
SDAE	Research and monitoring of biomarkers

Table1: DL packages with the different categories they are used for [3].

Deep Learning Model - DeepBind

DeepBind is one of the most common deep learning models, which can be implemented to understand the sequence specifities of DNA- and RNA-binding proteins which are necessary for regulatory processes in biological systems for identifying causal disease variants.

DeepBind model combines 12 terabases of sequence data, which spans over thousands of public PBM, RNAcompete, ChIP-seq and HT-SELEX experiments. DeepBind can also predict deleterious SNVs in promoters, by training a deep neural network to discriminate between high-frequency derived alleles and simulated variants.

DeepBind is based on deep learning, a scalable and modular pattern discovery method, which does not rely on common application-specific heuristics. DeepBind is one of the examples how deep learning algorithms have been developed and implemented to address the challenges of genomics.   [5]
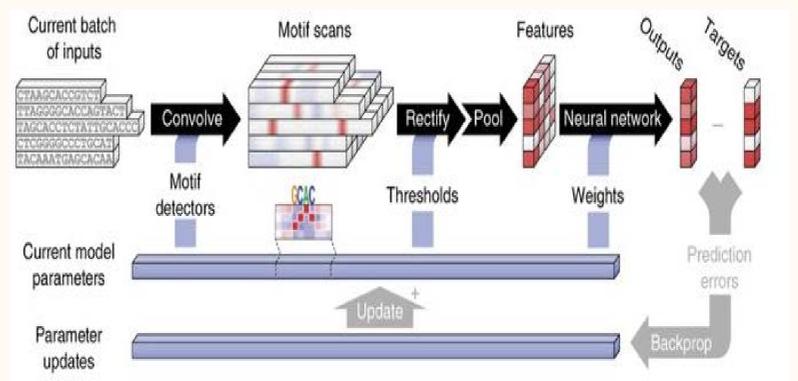

Figure 2: DeepBind architecture [5].

Discussion

The article reviews the current literature on deep learning systems for genomics.

 A set of initial exercises were carried out to acquire familiarity with the literature and techniques of DL, their applications to fields closely related to research in the NIHR project, and DL programming environments and tools.

The review helps us to understand the different DL architectures used to address challenges in genomics. It has provided us with a very good insight into the models, which will enable us to address precision therapeutics for diabetes. The familiarity with the architectures will give us an edge on addressing the project at hand.

References

[1] Charu C. Aggarwal, Neural Networks and Deep Learning, Springer International Publishing, 2018.
[2] Representation learning: a review and new perspectives, Bengio Y, Courville A, Vincent P, IEEE Trans Pattern Anal Mach Intell. 2013 Aug; 35(8):1798-828.
[3] F. Celesti et al., "Big data analytics in genomics: The point on Deep Learning solutions," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 306-309.
[4] Becker, A.S., Marcon, M., Ghafoor, S., Wurnig, M.C., Frauenfelder, T. and Boss, A., 2017. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Investigative radiology, 52(7), pp.434-440.
[5] Alipanahi, B., Delong, A., Weirauch, M.T. and Frey, B.J., 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), p.831.

Acknowledgements