ACTA SCIENTIFIC MEDICAL SCIENCES

Volume 3 Issue 6 June 2019

Mini Review

Deep Learning Genomics

Bhagvan Kommadi*

Department of Architecture, Founder of Architect Corner, India

***Corresponding Author:** Bhagvan Kommadi, Department of Architecture, Founder of Architect Corner, India.

Received: April 15, 2019; **Published:** May 09, 2019

Abstract

In genomics area, the success relies on the interpretability of models. If a model is trained to predict DNA mutations related to disease, it can identify patterns useful for cure of the disease. AI algorithms such as advanced natural language processing algorithms and deep machine learning are playing an important role in Genomics space.

Deep learning is related to a class of machine learning. This class of techniques can identify highly complex patterns in huge datasets. Deep learning has been successful in the area of regulatory genomics, variant calling and pathogenicity scores. Genomics utilizes machine learning to identify data dependencies and find new biological hypotheses. Expressive Machine learning models are required to process huge amounts of genomics data.Using huge data sets, deep learning has changed fields such as NLP and computer vision. Deep learning can help in predicting the impact of genetic variation on gene regulatory mechanisms. The gene regulatory mechanisms which can be used for predicting are DNA accessibility and splicing.

Keywords: Genomics; Deep Learning; DNA

Introduction

Population genetics is related to the study of DNA variation in human beings. For instance, genetics can be used to determine a person's ancestry from his or her DNA sequence. They can help in identifying the mutations which can help populations adapt to a new environment. This area has been evolving for the last 100 years. New DNA sequencing technologies came out because of this evolution.

Classical machine learning techniques involve mathematical theory and equations to detect patterns in a DNA data matrix. For instance, one of the methods can be creating an equation that takes the DNA matrix as input. The equation can outputs a negative number if it appears the population is shrinking over time. It can output a number near zero, if the population size is stable. The output will be a positive value if the population is growing. The classic approach is about developing special feature extractors based on domain specific knowledge and statistical knowledge.

Functional genomic analysis is the area where deep learning has made an impact to date. The availability of data such as DNA, RNA, methylation, chromatin accessibility, histone modifications and chromosome interactions ensure that there is enough training data to predict gene expression, genomic regulation, or variant interpretation. These data groups can help in identifying long noncoding RNAs and splice-site. More data helps in predicting precisely and accurately the genomic features and functions.

Deep Learning in Genomics

Convolutional neural network techniques have feature extractors within the defined approaches. Neural networks can help in feature extraction. However, they cannot give explanations of how the extraction was done. For instance, it is easy to train a neural network to distinguish pictures of tigers from pictures of rabbits. It is tough to explain why a certain picture is a tiger. The algorithms cannot elaborate what distinguishes tigers from rabbits. However classical methods can explain in theory the differences in population genetics. The other problem in the approach is the input matrix size.

Convolutional neural networks is the most successful deep learning modelling technique for image processing. In genomics, CNNs are the basic blocks in computer vision. The technique is related to assimilating a screen of genome sequence as an image. The genome sequence is a fixed length 1D sequence screen with four channels such as A, C, G, T. the genome is made up of 20,000 genes in human beings. The human genome consists of more than 3 billion base pairs of the genetic letters. Sequencing the genome is a critical first step to understanding it. Hence CNNs can analyze single sequence through 1D convolutional kernel. Two class image classification is performed for identifying protein-binding specificity of DNA sequences.

The key feature of CNNs is the adaptive feature extraction when the training process is performed. CNNs are used to discover recurring patterns with small variances. The small variances can be genomic sequence motifs.

Deep learning can be applied to specified sized biological datasets in the order of thousands. The black boxed deep neural networks does not help in complete understanding or transparency of them. Small variations in the input data can have big impacts. These variations can be controlled. The other area deep learning has problems is the size of the input matrix. High-throughput sequencing is related to the sequencing of DNA to occur in a single day. The same process used to take a decade. HTS was prevalent since the 2000s. The latest method named Deep Variant is able to differentiate small mutations from random errors.

Deep Learning genome models consists of mRNA alignments, mappings of DNA repeat elements, gene predictions, gene-expression data and disease-association data. Disease association data is related to the representing the relationships of genes to diseases. Expression tracks help in correlating genetic data with the tissue areas. Expression tracks help in identifying the linkage of a particular gene or sequence with body tissues.

ALDH7A1 gene mutations cause pyridoxine-dependent epilepsy. Amino acid glutamine gets replaced with the amino acid glycine in the antiquitin protein. Antiquitin deficiency builds up α-aminoadipic semialdehyde. This results in the disruption which causes deficiency of vitamin B6 due to the activity of pyridoxine.

Similarly, Breast cancer subtype models are developed using gene expression profiles. These profiles are analysed based on frozen tissue samples. Gene expression profiles analysis identifies transporters, metabolic regulators, immune response regulators, blood vessel development, circulation, and cell-to-cell communication related to the disease. PAM50 signature profile consists of gene expression levels for 50 genes. Based on the profiles, breast cancers are classified into five groups. Five groups are normal-like, luminal A and B, basal-like and Her2- enriched.

Likewise High triglycerides are linked to DNAm variations. These variations are resulted by genetic expressions linked to blood lipid levels, such as high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, triglycerides, and total cholesterol. DNN regression models are built to predict triglyceride concentrations from the blood samples.

33 gene signature helps in differentiating benign tissue controls and localised prostate cancers. Genetic Signature is based on three individual markers of prostate cancer, vets avian erythroblastosis virus E26 oncogene homolog (ERG) expression, prostate specific antigen (PSA) expression and androgen receptor. These gene signatures are used to classify clinical samples and develop deep learning models related to prostate cancer.

Chorn is a inflammatory bowel disease field. The variants from genome expression analysis are in the CARD15, the interleukin 23 receptor and autophagy-related 16-like 1 genes. SNP is a common coding variant which appears for all of the disease risk. In corn case, it is a threonine-to-alanine substitution at amino acid position 300 of the ATG16L1 protein (T300A).

In the case of asthma, genome analysis resulted in total 170 genes as the part of the signature. 57 of these genes were up-regulated. 113 of them were down-regulated in PBMCs. The gene signature has inflammatory and immune response genes, such as NOX5, MALT1, TNFRSF10C, GRK5, CXCL3, RELA, CD40, ABR, RELB, REBB2, PGLYRP1, CD82, RPE, CFTR, KITLG, and IKBKG. Heart stroke related genome analysis identified gene expression related to GATA zinc finger domain contains protein 1. Genome signature had around 250 genes related to chromosome 7q21.

Deep learning is making an impact in the areas such as gene regulation, genome organization, and mutation effects. Deep learning helps to execute four tasks in population genetics such as detecting introgression, estimating historical recombination rates, identifying selective sweeps and estimating demography [1-3].

Conclusion

Predictive toxicology, QSAR, artificial intelligence, data mining, machine learning, pattern recognition, data driven learning, chemo-informatics and Deep learning are the key areas evolving in the genomics applications. Deep learning methods can be compared with machine learning models with fewer parameters. Depending on the type and size of the datasets, deep learning can result in providing benefits or sometimes creating more uncertainty.

Bibliography

1. Deep Learning for Genomics : A Concise Overview.
2. Machine Learning in Genomic Medicine.
3. Deep Learning for predicting disease status using genomic data.

Volume 3 Issue 6 June 2019