International Journal of Smart Education and Urban Society
Volume 10 • Issue 1 • January-March 2019

Machine Learning Techniques for Analysis of Human Genome Data

Neelambika Basavaraj Hiremath, Department of Computer Science and Engineering, J.S.S. Academy of Technical Education Bengaluru, India

Dayananda P., Department of Information Science and Engineering. J.S.S. Academy of Technical Education, Bengaluru, India

ABSTRACT

Human genome data analysis is one of the molecular level information in health informatics, which enables genetic epidemiological analysis of complex data sets. The recent studies of the genomic sequence, a part of genome-wide association studies (GWAS) have led to understand the genetic architecture to identify the area of focus i.e. interactions with single-nucleotide polymorphism (SNP) is linked to causing complex diseases. The study and identification of these interactions and splicing of nucleic acids involves complexity in processing and computation. This article reviews current methods and trends in various machine learning and data mining approaches which are very complex and challenging to model and evaluate the performances.

KEYWORDS

Bioinformatics, Deep Learning Techniques, Gene Expression, Genome, Machine Learning, Neural Network, Omics Data, Single-Nucleotide Polymorphism

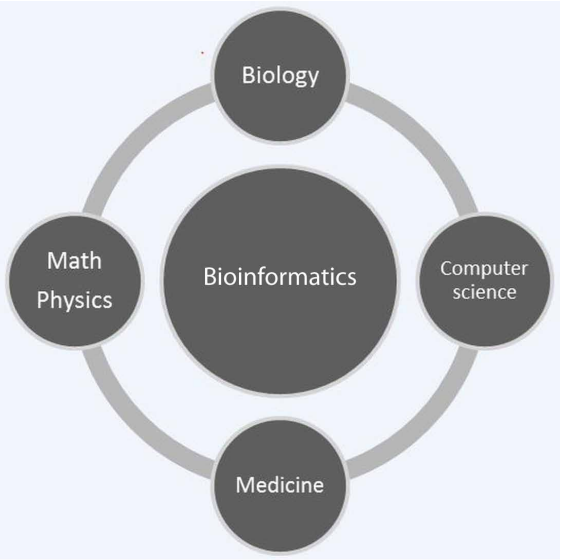
INTRODUCTION

The field of health care domain comprises lots of information and data where it helps to relies goal of diagnosing, treating, helping and healing all patients in need. This domain needs quality of care and research and development (R &D) for new discoveries. The basic goal of Health Informatics is to analyse at all levels of human existence, helping to advance our understanding of medicine and medical practice. The computational models and study real-world medical data with use of biological systems, and to understand the technology for optimizing treatment strategy (Ji, Yan, Li, Hu, & Zhu, 2017) for discovering new drug. According to (Herland et al., 2014) health informatics, is a broader subject where the following studies are covered. Micro level data which deals with molecular level information such as gene expression data which helps clinical predication of diseases of patient. The assessment of gene expression is used to identify histological types of lung cancer disease (Podolsky et al., 2016). The health informatics also covers tissue level, Patient level and Population data for various informational insights.

Bioinformatics research is an important source of health information which revolves around micro level data and focuses on analytical research using molecular data to learn the process of how the human body works. Figure 1 displays the knowledge contribution between another

DOI: 10.4018/IJSEUS.2019010105

Figure 1. Interactions of disciplines contributed to bioinformatics



subject domain. Predictive models can be built by measuring gene expression, splicing, and proteins binding to nucleic acids, which is inclusive of cell variables through the principles of modern biology (Leung et al., 2016). With the growing availability of large-scale data sets, (Olson et al., 2017) mentioned that there are about 165 publicly available datasets were used with machine learning algorithms to fine tune the performance of algorithms, open source packages were used. Advanced computational technique called deep learning architecture which comprises of deep neural networks, recurrent neural networks, convolutional neural networks and emergent architectures were discussed by authors (Min, Lee, & Yoon, 2016). The research community can help users in the advanced age of genomic medicine. Deep learning is used as a computational technique. The study of inheritance and variation of individuals based on DNA (deoxyribonucleic acid) is called genetics. The study of the structure and function of the genome is called genomics. To determine the nucleic acid structure, both bioinformatics and computational techniques are used by the data generated from methods of namely DNA and RNA (ribonucleic acid) sequencing, microarrays, proteomics, and electron microscopy, or optical methods. A genome is an instruction book for building an organism (Leung et al., 2016). The introns and exons are called as alternating regions in a typical gene and they are the most significant valuable information structures. The patterns in the nucleotide sequence (SNP) determines the boundaries between these regions. Disease-causing mutations act by disrupting these patterns. The genomic events which are associated with complex and dynamic aspects of the disease. There are computational models (Sun et al., 2017) built to identify insights on cancer progression.

TYPES OF DATASETS AND TOOLS

The literature survey is being carried out using genome wide association studies (GWAS), which facilitates the genetic variants of individuals associated with disease risk. Various related research papers and literature found in National Centre for Biotechnology Information (NCBI) instituted by National library of Medicine.