Chapter 8
Machine Learning Perspective in Cancer Research

Aman Sharma
IT Department, Jaypee University of Information Technology, Solan, India

Rinkle Rani
https://orcid.org/0000-0002-5970-1026
Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, India

ABSTRACT

Advancement in genome sequencing technology has empowered researchers to think beyond their imagination. Researchers are trying their hard to fight against various genetic diseases like cancer. Artificial intelligence has empowered research in the healthcare sector. Moreover, the availability of opensource healthcare datasets has motivated the researchers to develop applications which can help in early diagnosis and prognosis of diseases. Further, next-generation sequencing (NGS) has helped to look into detailed intricacies of biological systems. It has provided an efficient and cost-effective approach with higher accuracy. The advent of microRNAs also known as small noncoding genes has begun the paradigm shift in oncological research. We are now able to profile expression profiles of RNAs using RNA-seq data. microRNA profiling has helped in uncovering their relationship in various genetic and biological processes. Here in this chapter, the authors present a review of the machine learning perspective in cancer research.

INTRODUCTION

Bioinformatics is playing a vital role in fighting against various stringent diseases such as cancer, diabetics, Alzheimer's, etc.. Cancer is one of the genetic diseases caused due to mutation and variation in genes of the patient's cells. Complexity in tumor microenvironment makes cancer difficult disease from the treatment perspective. Patients with homogeneous cancer type show heterogeneous responses toward the same type of targeted therapies. Clinical trials and the traditional drug discovery process is

DOI: 10.4018/978-1-7998-2742-9.ch008

a time demanding and tedious task. Hence, researchers are trying their hard to design optimal treatment options for such stringent diseases. Availability of huge amount of oncological and pharmacogenomics online data sources have boosted the research in this field. Unlike traditional statistical and computational approaches, bioinformaticians are using artificial intelligence and machine learning to improve the treatment options in genetic diseases.

Cells are the basic building block of all living organisms. There are different types of cells available in the human body such as blood cells, muscle cells, fat cells, etc. Genes are responsible for variation in these cells. Gene helps to carry heredity information and are responsible for various physical and functional processes in the body. Genes are responsible for heterogeneity in genotype and phenotype traits among species. All the information regarding the inheritance of phenotypic traits is carried by genes. Overall if one wants to fight against genetic disease then their root cause i.e. genes need to be studied. Advancement in computational biology and high throughput sequencing is helping to find biomarkers (genes) which are responsible for various diseases.

Further, chip technology in healthcare is considered as the future of the healthcare industry which also provided lab-on-a-chip devices. These chips help in proper diagnosis and prognosis of patients based on their genetic profiles. Various researchers are working in the field of genetics to identify genes which are responsible for the inheritance of diseases. Microarray technology helps to measure the gene expression levels of particular micro-environment. Along with gene expression data, we can collect (genome, transcriptome, and proteome) data such as copy number variations, gene mutation, etc.. Gene expression, drug response data is extensively used in identifying anti-cancer drugs, drug targets, and biomarkers. Some researchers are working to explore various biological pathways corresponding to genetic diseases.

The ratio of the expression level of an individual gene under two variable conditions, obtained by DNA microarray hybridization is called gene expression value. The quantity of mRNA released by gene determines the gene expression value of the individual gene. This quantity may vary based on external stimuli. mRNA helps to carry the information from the genes about protein synthesis. Gene expression data is an asset to various biological research outcomes. It maps the genotype traits to phenotype traits and hence helps to differentiate between various phenotype articulations. It is used to distinguish between different disease phenotypes and identify potential disease biomarkers. Machine learning models take this data as input from genomic assays like m-RNA, DNase-seq, MNase-seq. Active research on various chronic diseases like cancer has exploited its potential and waved various new research outcomes.

Cancer is a complex genetic disease involving various subtypes. There is a need to develop computational approaches which could aid in early diagnosis and prognosis of tumor subtypes. Over the past decade, oncological research has gained serious attention and researchers are trying to personalize treatment therapies for cancer patients (Błaszczyński & Stefanowski, 2015). Apart from biomarker identification researchers are also working for developing computational (in-silico) models/algorithms that can predict disease-specific drug responses, drug synergy, and drug-target interactions.

Many researchers are using supervised and unsupervised machine learning algorithms to solve biological research problems. There are specifically three stages for any supervised machine learning method. First stage deal with the development of a machine learning algorithm that can lead to successful learning. The second stage provides the algorithm with a large amount of data to build a machine learning model. The Model is the generalized summary of rules that came out from the input data set. Third, if any new data point is given to the algorithm, it should predict the label corresponding to it (e.g. classification problems). Whereas, in unsupervised learning, data points are given but no labels