Chapter 11
Application of Deep Learning in Biological Big Data Analysis

Rohit Shukla
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, India

Arvind Kumar Yadav
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, India

Tiratha Raj Singh
https://orcid.org/0000-0003-1109-5626
Department of Biotechnology and Bioinformatics and Centre for Excellence in Healthcare technologies and Informatics (CEHTI), Jaypee University of Information Technology, India

ABSTRACT

The meaningful data extraction from the biological big data or omics data is a remaining challenge in bioinformatics. The deep learning methods, which can be used for the prediction of hidden information from the biological data, are widely used in the industry and academia. The authors have discussed the similarity and differences in the widely utilized models in deep learning studies. They first discussed the basic structure of various models followed by their applications in biological perspective. They have also discussed the suggestions and limitations of deep learning. They expect that this chapter can serve as significant perspective for continuous development of its theory, algorithm, and application in the established bioinformatics domain.

INTRODUCTION

The generation of massive amount of data in this era is a good process in the biological systems and contributes for big data. The big data can be of any type like epigenome, genome, proteome, transcriptome, and metabolome, etc. The big data is defined by its four key characteristics first is volume and others are variety, velocity, and variability. The structure of big data is defined by three types: first is structured, next is unstructured and the last is semi-structured (Mirza et al., 2019). The biological data is very complex as compared to other data because the regulation of one gene or protein depends on the

DOI: 10.4018/979-8-3693-3026-5.ch011

behavior of other genes or regulatory element proteins. The one entity in biological data can regulate many entities and vice versa. In recent years the vast amount of biological data is generated due to the technology advancement towards the high throughput sequencing, medical image processing, genome-wide association studies, gene expression analysis, protein binding motifs and expression studies, pathway and network level analyses, and structural investigations of biological entities etc. These types of data need a complete workflow for the analysis. As earlier described, the biological systems are very complex and not regulated by one entity. Hence in the case of genome-wide association studies, scientists focus on the genetic variants which are associated with the measured phenotypes while only one phenotype is not involved in the disease. It is a very complex process and several elements participate in the disease cascade so by the analysis of one gene or protein or a single type of data, we cannot analyze all the disease spreading factors (C. Xu & Jackson, 2019). Therefore the analysis of all the factors simultaneously can give a better measurement of the disease causing and spreading factors (Zitnik et al., 2019). The other and major challenge regarding the big data is its dimensionality. The big data have high dimensions described by high-resolution data, while in the case of biological data the samples which are collected from the different patients are limited and much less than the number of variables due to its high costs or limited resources like Alzheimer disease patients or replicates of sequencing so they lead to data sparsity, multicollinearity, multiple testing, and overfitting (Altman & Krzywinski, 2018).

The extraction of meaningful information from this complex biological dataset is a challenge in computational biology. Traditionally these types of data were analyzed using various platforms through several statistical and machine learning techniques. Nowadays, the integration of these techniques can handle a large amount of data. With the invention of computation capability in terms of storage and processing, we need to design machine learning-based algorithms which can efficiently extract the meaningful information from the vast amount of data. In that case, deep learning is a part of machine learning and is a very innovative technology that can extract the features on the basis of data characteristics and can also classify the data (Tang, Pan, Yin, & Khateeb, 2019). Machine learning techniques uses a lot of data as a training set for understanding the fundamental patterns, building the models and on the basis of this information, it makes the best fit model for predicting the hidden patterns and information. For Systems biology, proteomics, genomics and several other domains, few well known algorithms like Hidden Markov Model, Random Forest, Bayesian Networks, Support Vector Machine (SVM) and Gaussian Networks were already used (Larranaga et al., 2006). The conventional machine learning performance is heavily based on the feature's identification which is generated from the data. These features are designed by human engineers who have extensive domain expertise in machine learning. The feature selection is a very critical task because we have to select the appropriate features for the appropriate tasks. Sometimes feature selection can be wrong and it can hinder the whole results. To overcome this drawback, deep learning being evolved as a recently emerging and effective technique. By entering "Deep learning in big data analysis" keyword in Pubmed, we got 134 entries as on dated 09/10/2019. Additionally, it is also responsible for the advancements in the various fields where Artificial Intelligence (AI) community has struggled for many years (LeCun, Bengio, & Hinton, 2015). Several advancements in the deep learning capability describe speech recognition and image processing (Chorowski, Bahdanau, Serdyuk, Cho, & Bengio, 2015). The deep learning also has a key role in language translation (Luong, Pham, & Manning, 2015) and natural language processing (NLP) (Kiros et al., 2015), etc.