9

Machine Learning in Bioinformatics

Rahul Yadav[1], Mohit Sharma[2,3] and Nikhil Agrawal[4*]

[1]SRM University, SRM Nagar, Kattankulathur, Tamil Nadu, India
[2]Postgraduate School for Molecular Medicine, Medical University of Warsaw, Warszawa, Poland
[3]Poland Malopolskie Centre of Biotechnology Jagiellonian University, Krakow, Poland
[4]College of Health Sciences, University of KwaZulu-Natal, Westville, Durban, South Africa

Abstract

Human evolution has seen different stages, and at present, we are in the information Age. The revolution to this age started with the advent of the internet. In the present age, data is generated in huge amount in different domains of science. Specially, in biological sciences e.g., genomics, proteomics, molecular modeling etc. The data generated for genomics different from that of molecular modeling. However, the information's can be linked to obtain a better insight into the functionality even at the cellular level. It has become tough to analyze such massive data and conclude in a short period. The present-day scenario is changing with the implementation of Machine Learning methods. Machine Learning provides more in-depth insight into the problems backed up by mathematically models to take a short amount of time in terms of analysis. Machine Learning has been implemented in detection and medication suggestions for cancer patients. In drug discovery, Machine Learning models have been developed to design potential drug molecules. In the present chapter, we have tried to provide an understanding and importance of Machine Learning in the field of bioinformatics and its different domains.

***Keywords*:** Machine learning, bioinformatics, drug discovery, genomics

[*]*Corresponding author*: nikhil.08oct@gmail.com

S. Balamurugan, Anand Krishnan, Dinesh Goyal, Balakumar Chandrasekaran and Boomi Pandi (eds.)
Computation in Bioinformatics: Multidisciplinary Applications, (155–164) © 2021 Scrivener Publishing LLC

9.1 Introduction and Background

Humans have developed their standards of living in due course of evolution with their invention. Throughout our evolution, humans have lived through four primary ages. The first, the "Stone Age" or commonly known as the "Hunter and the Gatherer Age", basically comprised of the humans hunting and gathering food for a living. During the later stages of the Stone Age, humans started to have a more stable life focusing on farming skills as well as other developments like making weapons and utensils during the Mesolithic and the Neolithic phases. The second age is the "Bronze Age" and this age marked to be very important in human evolution. Humans evolved to develop metal tools and weapons as compared to the Stone Age, where the tools and weapons were mostly comprised of stones and wood. The Bronze Age is also known as the Industrial Age. It was during the later stages of this age when humans first started to make industrial machines. This age has not been far behind, and now, we are in the third age of human evolution, the "Computer Age".

The Stone Age remained for a couple of millions of years of age, the Bronze Age for a few thousand years, but the Computer Age has only been a started a few decades ago. Information is being gathered and collected for a long period. However, the past few decades proved to be the most decisive of them all. The first steam-driven computing machine was built by the famous English Mathematician Charles Babbage in the year 1822. Nevertheless, age is not considered as the Computer Age. In the year 1936, Dr. Alan Turing, another mathematician, during World War II made a machine capable of anything computable. Since then, the Computer Age began. Several companies were founded in later years. Most notable are Hewlett-Packard by David Packard and Bill Hewlett and IBM (International Business Machines).

Today's age is the age of the computer, "The Computer Age" also known as the Information Age". The Information Age has seen a significant transformation with the inventions of the modern-day computer. With the advancements during this age, the computers grew in terms of computing capabilities and high information processing speed. However, computers grew smaller in size but became more compact and powerful. Nowadays, we have high-performance computing systems with RAM's of ranging up-to terabytes (TB).

This was enough until recently. With the vast amount of information being gathered each second, the need for computers to be robust and smart became inevitable. Since the past decade, the amount of data that is being generated in each domain of science has drastically engorged.

Artificial Intelligence (AI) is a program, so designed as to solve a complex problem in order to obtain interesting, valuable information. It comprises of two parts. The first part is to find what is the problem and why do we need to solve that problem. The second part is the implementation of particular algorithms for solving the problem and getting the desired output. Various algorithms can be used for solving these problems. The algorithms are backed by mathematical models [1].

The physical limitation of computers led to the innovation of algorithms commonly known as AI. With the amount of data being generated each day, finding necessary information has become like finding a needle out of a haystack. However, digging out information nowadays is becoming comfortable with the use of such programs.

So, what is Artificial Intelligence? Artificial Intelligence or more commonly known as AI is a program so designed which is trained in order to find particular information from an enormous amount of data using less time and space based on some mathematical models. The program is previously trained using a model data set before actually being used. These programs can adapt and act according to the data without any human intervention. An example is Apple's Siri. At the moment, a particular AI program can only be used to perform the task that it was built for. AI itself has two subsets to itself, namely, Machine Learning (ML) and Deep Learning (DL).

A ML program is a somewhat more intelligent artificially intelligent program that can rectify itself with the exposure of data. Although they also need to be trained over a training dataset, yet ML programs are so designed with mathematical models that it can rectify itself with the exposure to data. One of the best examples is IBM's Watson Health. Watson Health is an ML program for cancer diagnostics and treatment. It is used to recommend the treatment of cancer patients by considering the genome, history, and pathology of the patient. The program then recommends a probable treatment with the help of the information previously present.

DL programs are a subset of ML. DL programs are so-called due to the fact that the number of neural networks used in such programs is much more as compared to ML programs. An example of the DL program is the detection of faces from images [4]. Below is an image to provide a brief overview between AI, ML, and DL(Figure 9.1). AI is being used in almost all fields in the present day. Few examples of fields where AI is used are as below.

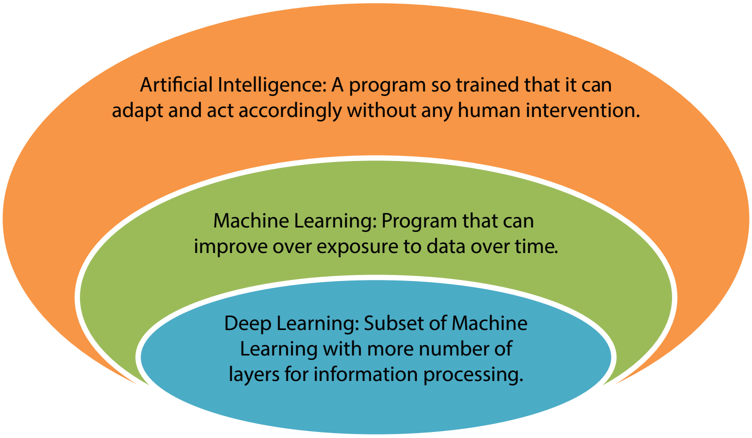
Artificial Intelligence: A program so trained that it can adapt and act accordingly without any human intervention.
Machine Learning: Program that can improve over exposure to data over time.
Deep Learning: Subset of Machine Learning with more number of layers for information processing.

Figure 9.1 Typical example of the correlation between Artificial Intelligence, Machine Learning, and Deep Learning [1–4].

9.1.1 Bioinformatics

In the last few decades, bioinformatics has evolved and is playing a significant role in the fields of biotechnology, drug designing, and related fields. The most notable advancement was the invention of a technique known as DNA sequencing. It was made possible by Dr. Sanger to sequence the DNA of any organism. This technique led us to sequence the human genome comprising of approximately 3.2 billion bases of nucleotides. But with various companies founded during this time, the time has significantly reduced to merely a few days, and the amount of data being produced from the machines is greatly increased to TB. In order to process such data, there is a need for high computing systems as well as better algorithms. Various advancements have also been made in the field of clinical genomics. IBM's Watson is one such example. ML programs are used to identify, diagnose, and provide a probable treatment to the patients [4].

In the field of bioinformatics, ML is also used in the field of computational biology drug discovery is another field, which requires high computational speed as well as a considerable amount of memory in order to perform its task. Support vector machine, a ML algorithm, can be implemented in order to achieve such tasks. SVMs work on regression-based models and are generally supervised based programs. These types of ML programs have gained popularity in the computational biology community. Sometimes, a probabilistic approach based on hidden

Markov models is also used for predictions in the field of computational biology [5].

9.1.2 Text Mining

Another field that uses ML is the field of text mining. The term is first mentioned by Ronen Faldman. Although the actual term coined by Ronen was Knowledge Discovery in Databases, it is a data mining process whose main aim was to extract information that is potentially useful data [6]. ML methods for text mining, retrieval, or extraction of information are also known as natural language processing which uses algorithms, ML methods, and statistics in order to derive important information from a particular text. It uses a method called tokenization, where a document is split into words and all the punctuations are replaced by tabs or other white spaces. These tokenized words form a part of a dictionary. This dictionary is then reduced in size by the use of dimension reduction algorithms where less important words are removed. SVM are used on clean, dimension-reduced, and indexed data [7].

9.1.3 IoT Devices

The best examples in the modern days are the inventions of IoT devices that are small, single units of machines designed in order to perform specific tasks. One such task can be collecting the weather information of a particular geographic location. Hundreds of such devices can be placed at various places within the geographic location. Information such as the air pollution index, CO2 levels, temperature, and moisture can be collected. Companies are combining data produced from hundreds of such devices. This information gathered through these devices collectively becomes vast. Thus, it increases the processing time and needs more memory of the computers [2].

9.2 Machine Learning Applications in Bioinformatics

Bioinformatics is a rapidly emerging field with an ocean of data being generated in each of its domains that needs to be analyzed. This vast amount of data that is being generated has led to the implementation of the ML approach. Pattern recognition is one of the most recognizable aspects of the field of bioinformatics. There are two major categories of data generated in terms of the problem and the approach made toward solving the problem. One is the type of data that is already present and has some information on the approach toward solving the problem is already known. The approach used for types of data is more or less established already. The second type of data is any novel data, any information of which is not known previously. Such type of data requires some novel approach which may also lead to developing entirely new techniques to approach such problems.

Genomics has become one of the most prominent and eminent fields in bioinformatics. In the early phase of bioinformatics, a large amount of data was generated using the microarray. This was because it gained popularity and as an emerging technology within the computational biologists. Microarrays technology allowed us to study the expression of several genes over a panel. The mRNA was first reverse transcribed to DNA which was then attached with a fluorescently labeled dye. On excitation, the fluorescently labeled dyes emitted a particular color that is captured by a sensor. The fluorescent labels were of two colors: green and red. The green color depicts upregulation, and red color depicts downregulation. This is known as the two system hybridization technique. Since it used the property of mRNA expressions, it was found to have applications in the field of medical diagnostics. This method was also used to find the difference between healthy and diseased individuals. The training data for such a problem was the already known information on the expression levels of a particular gene within a normal individual. The ML approach applied to such a problem where a training dataset can be used is known as the supervised learning method. Microarray technology triggered the use of SVM which is based on a supervised learning method. Other applications of ML include the detection of single nucleotide polymorphisms in case of cancer data, prediction of various properties of amino acid residues in a protein, classification, and prediction of cancer. All of this data can be analyzed using artificial neural networks (ANNs).

SVMs are used for the classification of various gene expression data as well as the classification of protein quaternary structures [8].

Computing the phylogenetic distances and plotting of a phylogenetic tree is a daunting task in the field of bioinformatics. The phylogenetic tree represents the closeness or distances between organisms. This is performed by a method called multiple sequence alignment wherein all the sequences comprising of at least more than three organisms are aligned together, and a score is calculated. The sequences are then rearranged and realigned. This process is performed iteratively in order to obtain the best alignment to represent precise distances between the organisms. ML approaches are used in order to perform such tasks implementing various algorithms [9].

9.3 Machine Learning Approaches

ML uses two different approaches, in general, for the processing of data and infers usable information from it. The two different approaches are supervised learning and unsupervised learning. Supervised learning is a ML approach that is trained on similar example data in order to learn how to infer useful information from a dataset. Whereas unsupervised learning is an approach where the ML program is not trained before, yet it tries to infer solutions to the problem [10]. One of the examples of supervised learning in bioinformatics is the use of gene-finding algorithms. A gene finding algorithm is usually trained on a known set of sequences. The transcription start sites and transcription end sites of the genes are already known for these sequences. The splice site information between the genes is also known for the sequences. The programs then predict the genes of any unknown sequences provided to it based on the information derived by the training dataset like the DNA sequence arrangement, the start codons, and stop codons along with the information of 5' and 3' UTR and the introns [10]. When such pieces of information are not available, it comes to the use of unsupervised ML programs. One of the best examples of unsupervised learning methods in bioinformatics would be explained when talking about epigenetics data. Encyclopedia of DNA elements (ENCODE) consortium and the Roadmap Egipegomics Project produce a heterogenous collection of epigenomics data. When working on such a heterogeneous mixture of data to find certain patterns of information, it is best to use unsupervised learning methods. In such cases, the program itself trains on a set of iteration in order to find a particular set of patterns, which it then uses for deriving similar patterns when fed with similar datasets [11]. One of the best examples of such a ML program is a gene-finding tool named GeneMark. GeneMark has both a web-based and a stand-alone version. GeneMark generally has two different variants: one of it uses a supervised learning approach, while the other uses an unsupervised learning approach for *in silico*–based gene finding. The supervised based model is GeneMark.hmm, which uses a type of supervised learning known as the hidden Markov model for gene prediction. Both methods produce quite accurate results. The supervised based model uses a predefined set of information where the start codons and stop codon 5' and 3'

UTR regions of similar data are already known to the program. It then uses the information to find similar patterns in the query dataset and to predict the genes. It is best for the use of already known organisms whose genomic region information is known [11].

Another variant of GeneMark is the GeneMark S which is an abbreviation for GeneMark self-learning and uses an unsupervised based approach for gene finding. The GeneMark S uses two different sets of algorithms for gene finding. First, it uses general GeneMark.hmm, a method with known information for gene finding. It then tries to find patterns from the unknown sequences which, in turn, are used to match with the results obtained from that of GeneMark.hmm with high accuracy of 99% or till the stability of gene finding is saturated at a particular level. If the percentage of accuracy starts to drop, then it stops its iterations [12].

9.4 Conclusion and Closing Remarks

Newer and better algorithms are being developed in order to predict precise results. The ML approach is being implemented in almost all aspects of bioinformatics. Several gene prediction programs make use of ML models in order to predict CDS regions of unknown genes. In next-generation sequencing machine, learning-based methods are used for analyzing data. One such example is the implementation of ML method for analyzing 16S rRNA sequences in QIIME2. The developers of the program have used ML model in order to analyze such data. ML is being used more and more in each field of science. In the field of bioinformatics, it has gained huge popularity due to the vast diversity in the type of data being generated. The vast amount of data being generated in each domain of bioinformatics requires continuously improving ways to analyze with precision. In the real world, the molecules may act in certain ways. In order to reflect the same results, more research is needed to understand the biological function.

References

1. Marr, D., Artificial intelligence—a personal view. *Artif. Intell.*, *9*, 1, 37–48, 1977.
2. Desai, N.S. and Alex, J.S.R., IoT based air pollution monitoring and predictor system on Beagle bone black, in: *2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, IEEE, pp. 367–370, 2017, March.

3. Meidan, Y., Bohadana, M., Shabtai, A., Guarnizo, J.D., Ochoa, M., Tippenhauer, N.O., Elovici, Y., ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis. In: *Proceedings of the symposium on applied computing*, pp. 506–509, 2017 Apr 3.
4. Bini, S.A., Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? *J. Arthroplasty*, *33*, 8, 2358–2361, 2018.
5. Vert, J.P. and Jacob, L., Machine learning for in silico virtual screening and chemical genomics: new strategies. *Comb. Chem. High Throughput Screen.*, 11, 8, 677–685, 2008.
6. Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Zamir, O., Text mining at the term level, in: *European Symposium on Principles of Data Mining and Knowledge Discovery*, 1998, September, Springer, Berlin, Heidelberg, pp. 65–73.
7. Hotho, A., Nürnberger, A., Paaß, G., A brief survey of text mining, in: *Ldv Forum*, 2005, May, vol. 20, No. 1, pp. 19–62.
8. Bhaskar, H., Hoyle, D.C., Singh, S., Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Comput. Biol. Med.*, *36*, 10, 1104–1125, 2006.
9. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Robles, V., Machine learning in bioinformatics. *Brief. Bioinform.*, *7*, 1, 86–112, 2006.
10. Lloyd, S., Mohseni, M., Rebentrost, P.. Quantum algorithms for supervised and unsupervised machine learning. arXiv preprint arXiv:1307.0411, 2013 Jul 1.
11. Libbrecht, M.W. and Noble, W.S., Machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, 16, 6, 321–332, 2015.
12. Besemer, J., Lomsadze, A., Borodovsky, M., GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, *29*, 12, 2607–2618, 2001.