IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 9, September 2023

DOI: 10.17148/IJIREEICE.2023.11907

A Literature Survey on Bioinformatics

Manila M V

Lecturer, Computer Hardware Engg, Government Polytechnic College, Palakkad, India

Abstract: The area of Bioinformatics has arisen from the needs of biologists to utilize and interpret the vast amounts of data that are constantly being gathered in genomics research. The ultimate goal of bioinformatics is to develop in silico models that will complement in vitro and in vivo biological experiments. Bioinformatics encompasses the development of databases to store and retrieve biological data, of algorithms and statistics to analyze and determine relationships in biological data, and of statistical tools to identify, interpret, and mine datasets. Database management, artificial intelligence, data mining, and knowledge representation can provide key solutions to the challenges posed by biological data. AI in bioinformatics provides both basic as well as clinical research with the help of biological sequence matching, proteinprotein interaction and function-structure analysis. This analysis helps in the design and discovery of drugs as well as complex systems. Deep learning (DL) has shown explosive growth in its application to bioinformatics and has demonstrated thrillingly promising power to mine the complex relationship hidden in large-scale biological and biomedical data.

Keywords: Bioinformatics, Data mining, Artificial Intelligence, Deep learning

I. INTRODUCTION

Bioinformatics is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics. The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatics processes in biotic systems. Bioinformatics deals with storing, extracting, organising, analysing and interpreting biological data from huge data banks. It involves the use of computer technologies and statistical methods to manage and analyze a huge volume of biological data about DNA, RNA, and protein sequences, protein structures, gene expression profiles, and protein interactions. A set of software tools are available for molecular sequence analysis. This field is important due to the availability of enormous amounts of public and private biological data and the compelling need to transform biological data into useful information and knowledge.

The two major challenging areas in bioinformatics are data management and knowledge discovery. The information and knowledge from the various disciplines can then be wisely used for applications that cover drug discovery,genome analysis and biological control. The Human Genome Project (HGP) was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. The collection of genes is known as "genome.". Human genome project paved the way for the large scale accumulation of DNA and protein information at various data banks. There are several types of databases available to researchers in the field of biology. They are Primary Databases and Derivative databases. Primary databases contain original biological data. There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide. The primary nucleic acid databases are GenBank, Nucleotide Sequence Database (EMBL-European Molecular Biology Laboratory) and DNA Data Bank of Japan (DDBJ). Derivative databases contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Protein Sequence Databases are Swiss-Prot, UniProt and PIR.

DNA is DeoxyRibonucleic Acid which is a molecule that contains the genetic instructions used in the development and functioning of all living organisms. Proteins are manufactured using the information encoded in DNA. The synthesis of proteins also involve molecules of RNA. DNA is made of chemical building blocks called nucleotides. These are made of three parts a phosphate group, a deoxy ribose sugar and one of four types of nitrogen bases.The four types of nitrogen bases found in nucleotides are adenine (A), thymine (T), guanine (G) and cytosine (C). DNA sequences must be converted into messages that can be used to produce proteins. Each DNA sequence that contains instructions to make a protein is known as a gene. The size of a gene may vary greatly, ranging from about 1,000 bases to 1 million bases in humans. RNA is ribonucleic acid that can read the genetic information carried by DNA. It is similiair to DNA except that it is made up of only one strand, contains the base uracil (U) instead of thymine (T), and contains the sugar ribose instead of deoxyribose.

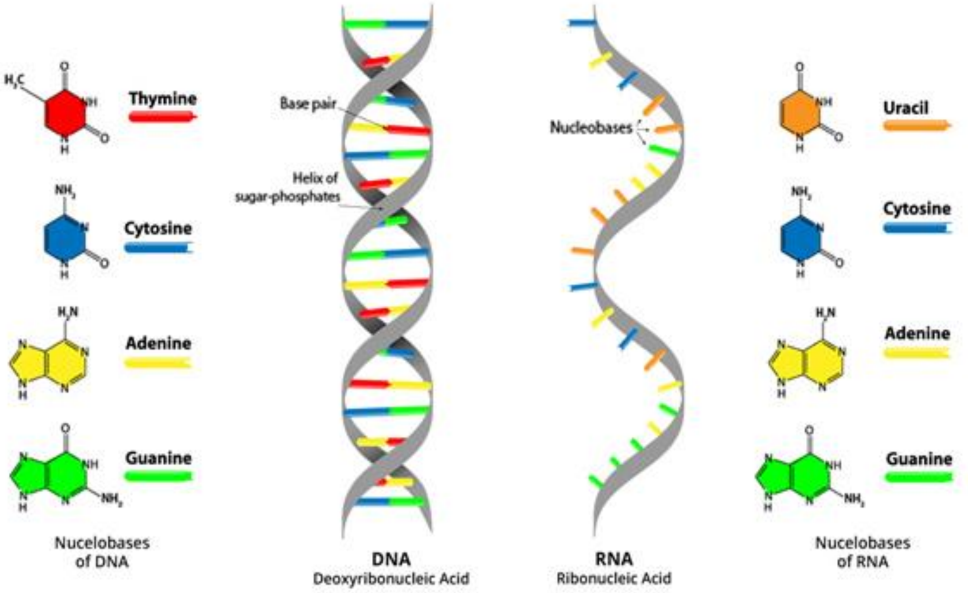

Fig. 1. DNA and RNA

There is a key relationship between DNA, RNA, and the synthesis of proteins, which is often referred to as the central dogma of molecular biology. According to this concept, there is a single direction of flow of genetic information from the DNA, which acts as the information store, through RNA molecules from which the information is translated into proteins. The sequence of bases in the DNA of a gene specifies the sequence of amino acids in a protein chain. The synthesis of protein involves two steps

1) Transcription-The enzymes read the information in a DNA molecule and transcribe it into an intermediary molecule called messenger RNA or mRNA. It is facilitated by RNA polymerase and transcription factors.

2) Translation-The information contained in the mRNA molecule is translated into the amino acids, which are the building blocks of proteins. During translation, the mRNA is ”read” according to the genetic code, which relates the DNA sequence to the amino acid sequence in proteins). Each group of three bases in mRNA constitutes a codon, and each codon specifies a particular amino acid which is the building block of protein.

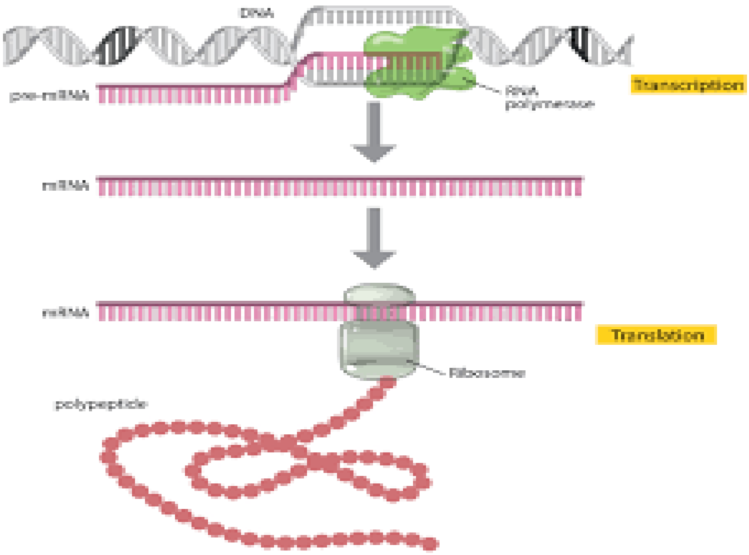

Fig. 2. Central dogma of life

Bioinformatics technology includes a set of technologies like database, data mining, structures, process, modeling, visualization, machine learning, pattern matching, networks, and tools. The existing research in bioinformatics is related to knowledge discovery, sequence analysis, structure analysis, and expression analysis. Sequence analysis is the discovery of functional and structural similarities and differences between multiple biological sequences. This can be done by comparing the new (unknown) sequence with well-studied and annotated (known) sequences. Scientists have found that two similar sequences possess the same functional role, regulatory or biochemical pathway, and protein structure. If two similar sequences are from different organisms, they are said to be homologous sequences. Finding homologous sequences is important in predicting the nature of a protein. This helps greatly in the development of new drugs, and in the performance of phylogenetic analysis. One proposed method for sequence comparison is sequence alignment. It is a procedure for base by base comparison of two (pairwise) or more (multiple) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. To search for an identical character or character patterns, the string matching technique is widely used. Another active research area in the field of sequence analysis is gene prediction. Gene prediction is the process of detecting meaningful signals in uncharacterized DNA sequences. Gene prediction uses homology search to acquire knowledge of the interesting information in DNA.

Structure analysis is the study of proteins and their interactions. Proteins are complex biological molecules composed of a chain of units, called amino acids, in a specific order. They are large molecules required for the structure, function, and regulation of the body’s cells, tissues, and organs. Each protein has unique functions. The structures of proteins are hierarchical and consist of primary, secondary, and tertiary structures. In other words, at the molecular level, proteins can be viewed as 3D structures. The understanding of protein structures and their functions leads to new approaches for diagnosis and treatment of diseases, and the discovery of new drugs. Current research on protein structural analysis involves comparison and prediction of protein structures. Expression analysis includes gene expression analysis and gene clustering. Basically, gene expression analysis is a study that determines the similarities or differences of genes expressed in a particular cell type or tissue. Gene expression, represented by a matrix, can be determined in two ways. First, comparing the expression profiles of genes: if the expression profiles are similar, the genes are co-regulated and functionally related. Second, by comparing the expression profiles of samples, one can consider whether genes are expressed differently. Gene clustering aims to group together genes having similar expression profiles. Genes in a specific group are coregulated and functionally related to each other rather than to genes in different groups. Due to the complexity and gigantic volume of biological data, the traditional computer science techniques and algorithms fail to solve the complex biological problems in the real world. The modern computational approaches called machine learning that can address the limitations of the traditional techniques. Machine learning is an adaptive process that enables computers to learn from experience, learn by example, and learn by analogy. Learning capabilities are essential for automatically improving the performance of a computational system over time on the basis of previous result.

II. LITERATURE SURVEY

The paper [1] provides an introduction to the field of bioinformatics. It provides a basic understanding of cell at molecular level. It also describes the software tools developed for biologists, computer and mathematical cell models and areas of computer science that play an important role in bioinformatics. Genes are contiguous subparts of single stranded DNA that are templates for producing proteins. Genes can appear in either of the DNAs strands. The set of all genes in a given organism is called the genome for that organism. Sequencers are machines capable of reading off a sequence of nucleotides in a strand of DNA in biological samples. The machines are linked to computers that display the DNA sequence being analyzed. A significant difficulty in obtaining an entire genome’s DNA is the fact that the sequences gathered in a wet lab consist of relatively short random segments that have to be reassembled using computer programs. This is referred to as the shotgun method of sequencing. There had been attempts to identify proteins using mass spectroscopy. The technique involves determining genes and obtaining the corresponding proteins in purified form. These are cut into short sequences of amino acids (called peptides) whose molecular weights can be determined by a mass spectrograph. It is then computationally possible to infer the constituents of the peptides yielding those molecular weights. By using existing genomic sequences, one can attempt to reassemble the desired sequence of amino acids. Another powerful tool available in biology is microarrays which allow determining simultaneously the amount of mRNA production of thousands of genes. The microarray approach has been extended to the study of protein expression.The role of bioinformatics is to aid biologists in gathering and processing genomic data to study protein function. Another important role is to aid researchers at pharmaceutical companies in making detailed studies of protein structures to facilitate drug design. The tasks done in bioinformatics include

- Inferring a protein’s shape and function from a given a sequence of amino acids,
- Finding all the genes and proteins in a given genome,
- Determining sites in the protein structure where drug molecules can be attached.

DOI: 10.17148/IJIREEICE.2023.11907

The algorithms used in Bioinformatics can be summarized as follows

1) Comparing Sequences
2) Constructing Evolutionary (Phylogenetic) Trees
3) Detecting Patterns in Sequences
4) Determining 3D Structures from Sequences
5) Inferring Cell Regulation
6) Determining Protein Function and Metabolic Pathways
7) Assembling DNA Fragments 8) Use Script Languages

The paper [2] describes that the major goal in molecular biology is functional genomics which is the study of the relationships among genes in DNA and their function. This function describes the role of a gene product, usually a protein, in reacting with other proteins in a metabolic or signaling pathway.

Biologists and computer scientists may conclude that the ultimate objective of functional genomics is: Given the DNA of an organism, produce a simulator for a cell of that organism. That simulator (or flowchart representing metabolic and signaling pathways) embodies all that it knows about a cell's behavior, allowing in-silico experiments that enable biologists to bypass costly and ethically sensitive in-vitro or in-vivo trials. Biologists deal with essentially four types of data structures

1) Strings-To represent DNA, RNA, and sequences of amino acids
2) Trees-To represent the evolution of various organisms
3) Sets of 3D points and their linkages-To represent protein structures
4) Graphs-To represent metabolic and signaling pathways.

Biological data is often characterized by huge size, the presence of laboratory errors (noise), duplication, and sometimes unreliability. For inferring function from the existing data, a biologist must consider three factors

1) Genes or substrings of DNA capable of generating proteins
2) Protein structures represented in 3D space
3) The roles of these proteins within metabolic and signaling pathways.

The following are some of the most important algorithmic trends in bioinformatics

- Finding similarities among strings (such as proteins of different organisms)
- Detecting certain patterns within strings
- Finding similarities among parts of spatial structures
- Constructing trees (called phylogenetic trees) expressing the evolution of organisms whose DNA or proteins are currently known
- Classifying new data according to previously clustered sets of annotated data
- Reasoning about microarray data and the corresponding behavior of pathways.

The first three can be viewed as instances of pattern matching. It is suggested that both optimization and probabilistic approaches are necessary for developing biology-oriented pattern-matching algorithms. In the 1970s, a dynamic programming technique was devised to match two strings, taking into account the costs of insertions, deletions, and substitutions which is called global pairwise alignment. This technique was subsequently extended to consider local alignments. But dynamic programming is time consuming and therefore cannot be applied in a practical way to strings with hundreds of thousands of symbols. A remarkable bioinformatics development from the 1990s is a pattern-matching approach called BLAST, or the Basic Local Alignment Search Tool, that mimics the behavior of the dynamic programming approach and efficiently yields good results.

BLAST is the most frequently used tool for searching sequences in genomic databases. Another widely used and effective technique is multiple alignment, which helps align several sequences of symbols, so identical symbols are properly lined up vertically, with gaps allowed within symbols. The sequences may represent variants of the same proteins in various species. The goal is to find conserved parts of the proteins that are unchanged during evolution. Finding conserved parts of proteins also provides hints about a protein's possible function. Methods for multiple alignments are based on dynamic programming techniques developed for pairwise alignment. The Solutions to various problems in Bioinformatics can be explored through approaches from machine learning, neural networks, genetic algorithms, and clustering.

The paper [3] reviews the machine learning methods used for bioinformatics. It presents modelling methods, such as supervised classification, clustering and probabilistic graphical models for knowledge discovery, as well as deterministic and stochastic heuristics for optimization. Applications in genomics, proteomics, systems biology, evolution and text mining are also mentioned. The exponential growth of the amount of biological data available raises two problems i)efficient information storage and management and ii) the extraction of useful information from these data. The second problem is one of the main challenges in computational biology, which requires the development of tools and methods capable of transforming all these heterogeneous data into biological knowledge about the underlying mechanism. These tools and methods should allow us to go beyond a mere description of the data and provide knowledge in the form of testable models. There are several biological domains where machine learning techniques are applied for knowledge extraction from data.

Genomics is one of the most important domains in bioinformatics. The number of sequences available is increasing exponentially. These data need to be processed in order to obtain useful information. The first step is to extract the location and structure of the genes from genome sequences. Sequence information is also used for gene function and RNA secondary structure prediction. If the genes contain the information, proteins are the workers that transform this information into life. Proteins play a very important role in the life process, and their three-dimensional (3D) structure is a key feature in their functionality. In the proteomic domain, the main application of computational methods is protein structure prediction. Proteins are very complex macromolecules with thousands of atoms and bounds. Hence, the number of possible structures is huge. This makes protein structure prediction a very complicated combinatorial problem where optimization techniques are required. In proteomics, machine learning techniques are applied for protein function prediction. Another interesting application of computational methods in biology is the management of complex experimental data. Microarray essays are the best known domain where this kind of data is collected. Complex experimental data raise two different problems. First, the data need to be pre-processed, i.e. modified to be suitably used by machine learning algorithms. Second, the analysis of the data. In the case of microarray data, the most typical applications are expression pattern identification, classification and genetic network induction. Systems biology is another domain where biology and machine learning work together. It is very complex to model the life processes that take place inside the cell. Thus, computational techniques are extremely helpful when modelling biological networks, especially genetic networks, signal transduction networks and metabolic pathways. Evolution and, especially phylogenetic tree reconstruction also take advantage of machine learning techniques. Phylogenetic trees are schematic representations of organisms' evolution. They were constructed according to different features (morphological features, metabolic features, etc.) but with the great amount of genome sequences available, phylogenetic tree construction algorithms are based on the comparison between different genomes . This comparison is made by means of multiple sequence alignment, where optimization techniques are very useful. A side effect of the application of computational techniques to the increasing amount of data is an increase in available publications. This provides a new source of valuable information, where text mining techniques are required for the knowledge extraction. Thus, text mining is becoming more and more interesting in computational biology, and it is being applied in functional annotation, cellular location prediction and protein interaction analysis. In addition to all these applications, computational techniques are used to solve other problems, such as efficient primer design for PCR, biological image analysis and back translation of proteins. Machine learning is programming computers to optimize a performance criterion by using example data or past experience. The optimized criterion can be the accuracy provided by a predictive model in a modelling problem, and the value of a fitness or evaluation function in an optimization problem. In a modelling problem, the 'learning' term refers to running a computer program to induce a model by using training data or past experience. Machine learning uses statistical theory wh jhen building computational models since the objective is to make inferences from a sample. The two main steps in this process are to induce the model by processing the huge amount of data and to represent the model and making inferences efficiently. It must be noticed that the efficiency of the learning and inference algorithms, as well as their space and time complexity and their transparency and interpretability, can be as important as their predictive accuracy. Optimization problems can be posed as the task of finding an optimal solution in a space of multiple possible solutions.

It is mentioned in the paper [7] that even though bioinformatics has made great advancements with conventional machine learning, Deep learning is producing more promising results. Machine learning provided more viable solutions with the capability to improve through experience and data. Although machine learning can extract patterns from data, there are limitations in raw data processing, which is highly dependent on hand-designed features. To advance from handdesigned to data-driven features, representation learning, Deep learning can be used. Representation learning can discover effective features as well as their mappings from data for given tasks. Furthermore, deep learning can learn complex features by combining simpler features learned from data. In other words, with artificial neural networks of multiple nonlinear layers, referred to as deep learning architectures, hierarchical representations of data can be discovered with increasing levels of abstraction. The successes of deep learning are built on a foundation of significant algorithmic details and generally can

DOI: 10.17148/IJIREEICE.2023.11907

be understood in two parts: construction and training of deep learning architectures. Deep learning architectures are basically artificial neural networks of multiple nonlinear layers and several types have been proposed according to input data characteristics and research objectives. Deep learning architectures are classified into four groups: deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and emergent architectures(DST-NNs, MD-RNNs, and CAEs). The goal of training deep learning architectures is optimization of the weight parameters in each layer, which gradually combines simpler features into complex features so that the most suitable hierarchical representations can be learned from data. A single cycle of the optimization process is organized as follows. First, given a training dataset, the forward pass sequentially computes the output in each layer and propagates the function signals forward through the network. In the final output layer, an objective loss function measures error between the inferenced outputs and the given labels. To minimize the training error, the backward pass uses the chain rule to backpropagate error signals and compute gradients with respect to all weights throughout the neural network. Finally, the weight parameters are updated using optimization algorithms based on stochastic gradient descent. Another core element in the training of deep learning architectures is regularization, which refers to strategies intended to avoid overfitting and thus achieve good generalization performance. DNNs have been widely applied in protein structure prediction research. Since complete prediction in three-dimensional space is complex and challenging, several studies have used simpler approaches, such as predicting the secondary structure or torsion angles of protein. Few studies have used CNNs to solve problems involving biological sequences, specifically gene expression regulation problems. RNNs are expected to be an appropriate deep learning architecture because biological sequences have variable lengths, and their sequential information has great importance. Several studies have applied RNNs to protein structure prediction, gene expression regulation, and protein classification. Emergent architectures have been used in protein structure prediction research.

The paper [9] reviewed some selected modern and principled DL methodologies, some of which have recently been applied to bioinformatics, while others have not yet been applied. This perspective may shed new light on the foreseeable future applications of modern DL methods in bioinformatics.Current trend in principled DL are Attention mechanism, Reinforcement learning, Few-shot learning, Deep generative models, Meta learning, Symbolic reasoning empowered DL. Attention mechanisms -were first proposed to conduct machine based translation tasks that can alleviate the problems faced by RNNs when applied to bioinformatics problems, thus expanding their domain of applications in bioinformatics. The self-attention layer can translate the original representation of an input sequence into another representation of the sequence. For each position in the sequence, the other positions in the input sequence try to better characterize that position for capturing the semantic meaning of the sequence and interactions between different sequential positions. Attention mechanisms can potentially be used in a wide range of biosequence analysis problems, such as RNA sequence analysis and prediction, protein structure and function prediction from amino acid sequences. Reinforcement learning - considers what actions to take, given the current state of the partial solution to maximize the cumulative reward. After each action, the state can change. Observations about the set of change-of-state become guiding information for future actions. This type of reinforcement learning has recently been incorporated into the DL paradigm, referred to as deep reinforcement learning. Note that a key distinguishing feature is that users do not have to predefine all the states, and a model can be trained in an end-to-end manner, which has become an increasingly active research field with numerous algorithms being developed. Reinforcement learning can be applied in collective cell migration , DNA fragment assembly , and characterizing cell movement. DNA fragment assembly is a technique that aims to reconstruct the original DNA sequence from a large number of fragments by determining the order in which the fragments have to be assembled back into the original DNA molecule. The reinforcement learning model shows less computational complexity and unnecessary external supervision in the learning process compared with the genetic algorithm and supervised approach. Few-shot learning- Although there is a large amount of data in the bioinformatics field, data scarcity still occurs in biology and biomedicine. Few-shot learning is designed to handle these cases. Few-shot learning trains an ML model with a very small quantity of data. In extreme cases, there is only one training sample for one class, referred to as one-shot learning Similarly, zero-shot learning when a class has no training sample. Using few-shot learning algorithms, a model can be trained with reasonable performance on some difficult problems by utilizing only the existing limited data. Few-shot learning is suitable for many problems in bioinformatics that have limited data, such as protein function prediction and drug discovery. Deep generative models- Deep generative models, such as variational autoencoders (VAEs) are powerful networks for information derivation using unsupervised learning, which has achieved remarkable success in recent years. Generally, it is impossible to model the exact distributions of any property of such datasets, those methods are designed to model an approximate distribution that is as similar to the true distribution as possible, implicitly or explicitly. Deep generative models can be applied to problems related to drug design, protein structure design, 3D compound design, protein loop modelling, and DNA design. Meta learning also known as 'learn-to-learn', attempts to produce such models, which can quickly learn a new task with a few training samples based on models trained for related tasks. A good meta learning model should generalize to a new task even if the task has never been encountered during the training time.

DOI: 10.17148/IJIREEICE.2023.11907

The key idea is that when training a model is finished, the model needs to be exposed to a new task during the testing phase, several steps of fine-tuning are performed, and then the model's performance on the new task is checked. thus meta learning outputs an ML model that can learn quickly. Meta learning can be used in B-cell conformational epitope prediction in continuously evolving viruses, which is useful for vaccine design. Symbolic reasoning empowered DL-In the bioinformatics field, symbolic reasoning is applied and evaluated on structured biological knowledge, which can be used for data integration, retrieval, and federated queries in the knowledge graph . This method combines symbolic methods, in particular, knowledge representation using symbolic logic and automated reasoning, with neural networks that encode for related information within knowledge graphs, and these embeddings can be applied to predict the edges in the knowledge graph, such as drug target relations.

III. CONCLUSION

In the era of big data, transformation of biomedical big data into valuable knowledge has been one of the most important challenges in bioinformatics. To extract knowledge from big data in bioinformatics, machine learning has been a widely used and successful methodology. Machine learning algorithms use training data to uncover underlying patterns, build models, and make predictions based on the best fit model. Deep learning (DL) has shown explosive growth in its application to bioinformatics and has demonstrated thrillingly promising power to mine the complex relationship hidden in large-scale biological and biomedical data. The modern DL technologies may shed new light on the foreseeable future applications of modern DL methods in bioinformatics.

REFERENCES

[1] J. Cohen, "Bioinformatics: An Introduction for Computer Scientists," ACM Computing Surveys, 36(2), 122-158, 2004.

[2] Jacques Cohen, Computer Science and Bioinformatics, Communications of the ACM, March 2005/Vol. 48 No.3, Pages 73-78

[3] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu ˜ Galdiano, Inaki Inza, Jos ˜ e A. Lozano, Rub ´ en Arma ´ nanzas, Guzm ˜ an´ Santafe, Aritz P ´ erez, Victor Robles, ´ Machine learning in bioinformatics, Briefings in Bioinformatics, Volume 7, Issue 1, March 2006, Pages 86–112,

[4] Jones, N. and Pevzner, P. An Introduction to Bioinformatics Algorithms.MIT Press, Cambridge, MA, 2004.

[5] Krane, D. and Raymer, M. Fundamental Concepts of BioInformatics.Addison Wesley-Benjamin Cummings, Boston, 2003

[6] Yi-Ping Phoebe Chen, Bioinformatics Technologies

[7] Seonwoo Min, Byunghan Lee, and Sungroh Yoon Deep Learning in Bioinformatics

[8] LeCun Y, Bengio Y, Hinton G. Deep learning Nature 2015;521(7553):436-44

[9] Haoyang Li, Shuye Tian, Yu Li, Qiming Fang, Renbo Tan, Yijie Pan, Chao Huang, Ying Xu, Xin Gao, Modern deep learning in bioinformatics, Journal of Molecular Cell Biology, Volume 12, Issue 11, November 2020, Pages 823–8