Survey on Usage of Machine Learning Techniques in Different Biological Domains.

Divya K S [1], Dr M A Dorairangaswamy [2], Jain Stoble B [3]

[1]Research Scholar, Computer Science and Engineering Dept., ASIET, Kalady, India
[2]Professor, Computer Science and Engineering Dept., ASIET, Kalady, India
[1]Assistant Professor, Computer Science and Engineering Dept., ASIET, Kalady, India

Abstract *- Nowadays, one of the most challenging problems in computational biology is to transform the raw data into knowledge. Different Machine learning techniques can be used to carry out this transformation. There are several biological realm where machine learning techniques are applied for knowledge extraction from raw data. We can categorize these domains into genomics, proteomics, microarrays, systems biology, evolution and text mining. This paper gives a brief overview of different biological domains where machine learning techniques can be applied.*

Keywords - *Bioinformatics, Genomics, Proteomics, Microarrays, Systems biology, Evolution, Text mining.*

I. INTRODUCTION

Computational science, a part of science including the utilization of computer and computer science to the understanding and modelling of the structures and processes of life. It involves the utilization of computational strategies (e.g., algorithms) for the portrayal and reproduction of organic frameworks, as well as for the interpretation of experimental data, often on a very large scale Preceding the rise of machine learning algorithms, bioinformatics calculations must be expressly modified by hand which, for issues, for example, protein structure forecast, demonstrates very troublesome.[2] Machine learning techniques allow the algorithm to make use of automatic feature learning which suggest that based on the dataset alone, the algorithm can analyse how to select multiple features of the input data into a more efficient set of features from which to conduct more learning. This multi-layered approach to learning patterns in the input data allows such systems to make very unpredictable forecasts when prepared on enormous datasets. In recent years, the size and number of available biological datasets have increased, empowering bioinformatics researchers to make use of these machine learning systems.

II. BIOLOGICAL DOMAINS

Following sections explain various biological domains in which machine learning can be applied

A. Genomics

Genomics is an interdisciplinary field of biology concentrating on the evolution, structure, function, mapping, and editing of genomes. A genome is the complete set of DNA of an organism, including all of its genes. Genomics is one of the most important domains in bioinformatics [1].With Genomic sequences we can perform gene prediction, Splice site prediction, Motif Identification etc.

In computational biology, gene prediction or gene finding refers to the way of recognizing the regions of genomic DNA that encode genes. This incorporates protein-coding genes as well as RNA genes, but may also incorporates identification of other elements such as regulatory regions. Gene finding is one of the first and most significant steps in understanding the genome of a species once it has been sequenced. The coding region of a gene, also known as the CDS (from coding sequence) is the portion of a gene's DNA or RNA that codes for protein. The region usually begins at the end by a start codon and ends at the 3' end with a stop codon.

Machine learning Techniques play main role in Splice site prediction .RNA splicing is a procedure that removes the intervening, non-coding sequences of genes (introns) from pre-mRNA and connect the protein-coding sequences (exons) together in order to allow translation of mRNA into a protein. Correct identification of splice sites in DNA sequences plays one of the main roles in gene structural prediction in eukaryotes.

Motifs are termed as short, usually fixed length, sequence patterns that may specify important structural or functional features in DNA and protein sequences such as transcription binding sites, splice sites, active sites, or interaction interfaces [3]. Motifs are candidates for functionally important sites .Presence of a motif may be used as a base of protein classification.[2] Supervised learning based on deep learning neural networks for enhancer motif prediction has become popular recently.

B. Proteomics

Proteomics involves the applications of technologies for the identification and quantification of overall proteins present content of a cell, tissue or an organism. Proteomics based studies are used in different research settings such as detection of various bio markers, candidates for vaccine production, understanding pathogenicity mechanisms, alteration of expression patterns in response to different signals and interpretation of functional protein pathways in various diseases..[4]

Proteomics involves protein structure prediction, protein Function prediction and Proteomics annotations. Protein structure prediction is one of the most important goal of bioinformatics and theoretical chemistry. It has significance in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes).Protein function prediction play an important role in the development of new drugs, better crops, and even the development of synthetic biochemical such as biofuels.

C. Microarrays

A DNA microarray (also usually called silicon chip or biochip) may be a assortment of microscopic DNA spots connected to a solid surface. Machine learning methods can be used in micro array data analytics, micro array data pre-processing and micro array image analysis.

Microarray could be a rapidly growing technology employed in biological processes. There are several uses of existing microarrays within the space of cancer, diabetic and genetic diagnoses, gene and drug discovery in molecular biology, etc. Microarrays aid computation of many thousands of genes at the same time. Pre-processing is a crucial opening within the analysis of microarray information, to correct for effects arising from imperfections within the technology instead of real biological variations. Microarray image analysis is that the method of extraction and decoding sequence data [5].

D. Text Mining

Text mining tools are progressively more accessible to biologists and computational biologists and these will usually be applied to answer scientific queries together with alternative bioinformatics tools. Text mining applications for bioinformatics include subcellular localization prediction such as Sherloc and Epiloc [8, 9] and protein clustering such as TXTGate [10]. Text mining tools can be used for annotating biological databases in the same fashion other bioinformatics tools are used. [6].Some of the most investigated applications of text mining in Bio-informatics are Information Retrieval (IR), devoted to obtain relevant information from a collection of information resources and a user query, Document Classification (DC), which assigns one or more classes or categories to a document, Named Entity Recognition/Normalization (NER/NEN), devoted to extract named entity from unstructured or semi-structured machine-readable documents, Summarization (SUM), which synthesizes input text covering all contents of analysed documents[7].

E. Systems Biology

Systems biology includes the study of systems of biological elements, which can be molecules, cells, organisms or entire species. Systems Biology deals with knowledge and models at many alternative scales, from individual molecules through to complete organisms. Systems Biology deals with information and models at many alternative scales, from individual molecules through to complete organisms. Computational systems biology addresses queries basic to our understanding of life and progress here can result in sensible innovations in medicine, drug discovery and engineering. It aims to develop and use economical algorithms; information structures visual image and communication tools with the aim of computer modelling of biological systems. Many systems biology approaches involve mathematical and computational modelling, the progress, maintenance, and dissemination of tools for systems biology is in itself a greatest challenge. Examples of this include development of data repositories, data standards and software tools for simulation, analysis and visualization of system components such as biochemical networks. Another example are applications of high-throughput molecular identification technologies which regularly need subtle processing and analysis, and generally involve parts of signal processing and applied mathematics analysis. Because the ensuing quantitative measurements are transferred to formal mathematical models for the aim of modelling, the endeavour becomes perhaps more systems biology and fewer bioinformatics [11].

F. Evolution

The study of biological process relationships among protein sequences was one among the primary applications of bioinformatics. The relationship between bioinformatics and the study of protein evolution was further strengthened with the advent of large-scale sequencing projects. The growing number of sequences stored in the databases, including those of complete genomes, provided a completely new dimension to the study of protein evolution: that of the evolution of complete proteomes. New bioinformatics tools were developed that allowed the comparison of complete genomes, the efficient detection of orthology relationships and the reconstruction of the evolution of complete proteomes. Sophisticated tools for the comparison of protein sequences and the reconstruction of phylogenetic trees have allowed a better understanding of protein evolution at the molecular level.[12].

III. CONCLUSIONS

Bioinformatics and machine learning are developing as interdisciplinary science machine learning approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theory of life's organization at the molecular level. Machine learning and bioinformatics are fast growing research area today. It is necessary to look at what are the important analysis problems in bioinformatics and develop new machine learning strategies for scalable and effective analysis.

REFERENCES

[1] Machine learning in bioinformatics Pedro Larran‹aga,BorjaCalvo, Roberto Santana,ConchaBielza, Josu-Galdiano,In‹akiInza, Jose¤ A.Lozano, Rube¤nArman‹anzas,Guzma¤nSantafe¤, AritzPe¤rezand Victor Robles

[2] https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect08_Bind_Motifs.pdf

[3] Motif Discovery in Protein SequencesBy Salma Aouled El Haj Mohamed, MouradElloumi and Julie D. Thompson-Submitted: April 13th 2016Reviewed: August 30th 2016Published: December 14th 2016 .DOI: 10.5772/65441

[4] Bilal Aslam, MadihaBasit, Muhammad AtifNisar, MohsinKhurshid, Muhammad HidayatRasool, Proteomics: Technologies and Their Applications, Journal of Chromatographic Science, Volume 55, Issue 2, 1 February 2017, Pages 182–196, https://doi.org/10.1093/chromsci/bmw167

[5] Microarray Image Segmentation Using Clustering Methods VolkanUslan and ĐhsanÖmürBucak,Department of Computer Engineering,Fatih University, 34500,B.Çekmece,Đstanbul, Turkey.vuslan@fatih.edu.trand ibucak@fatih.edu.tr

[6] Rodriguez-Esteban, Raul. “Biomedical text mining and its applications.” PLoS computational biology vol. 5,12 (): e1000597. doi:10.1371/journal.pcbi.1000597

[7] Text Mining Basics in Bioinformatics Carmen De Maioa , Giuseppe Fenzab , Vincenzo Loiab , MimmoParentebaDipartimento di Ingegneriadell’InformazioneedElettrica e MatematicaApplicata, University of Salerno, 84084 Fisciano (SA), Italy bDipartimento di ScienzeAziendali - Management & Innovation Systems, University of Salerno, 84084 Fisciano (SA), Italy

[8] Shatkay H, Höglund A, Brady S, Blum T, Dönnes P, et al. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics. 2007;23:1410–1417.Available: http://www-bs.informatik.uni- tuebingen.de/Services/SherLoc2/ [PubMed] [Google Scholar]

[9] Brady S, Shatkay H. EpiLoc: a (working) text-based system for predicting protein subcellular location. Pac SympBiocomput. 2008:604–615. Available: http://epiloc.cs.queensu.ca/ [PubMed] [Google Scholar]

[10] Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, et al. TXTGate: profiling gene groups with text-based information. Genome Biol. 2004;5:R43. Available: http://tomcat.esat.kuleuven.be/txtgate/[PMC free article] [PubMed] [Google Scholar]

[11]]Likić VA, McConville MJ, Lithgow T, Bacic A. Systems biology: the next frontier for bioinformatics. Adv Bioinformatics.2010;2010:268925. doi:10.1155/2010/268925

[12] Gabaldón T. Evolution of proteins and proteomes: a phylogenetic approach. EvolBioinform Online. 2007;1:51–61. Published 2007 Feb 24.