ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

Overview of Bioinformatics Tools and Implementation of Clustal omega Tool

[1]P.Sudhasini, [2]Dr. B. Ashadevi

[1] Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India
[2] Assistant Professor, Department of Computer Science, M.V. Muthaiah Govt. Arts College for women, Dindugul, Tamilnadu, India
[1]sudhabaskee@gmail.com,[2] asharajish2005@gmail.com

Abstract*— The processing, storage, and analysis of biological data are all part of bioinformatics. Building databases to store experimental results is one example of this. anticipating how proteins fold and modeling the interactions between all of the chemical processes that take place within a cell. Bioinformatics methods may be used for genome comparison, gene annotation, and gene expression analysis. This knowledge may be used to develop novel biotechnology applications such as gene therapy, personalized medicine, and the development of new drugs. A bioinformatics typically entails the following actions: develop a computer model by combining statistical information with biological data, address a problem with computational modeling and examine and evaluate a computational algorithm. In order to determine the pairwise identity matrix for the mushroom 5.8s rRNA sequences, we will examine fundamental bioinformatics techniques and the implementation of the clustal omega tool in this study.

Index Terms*— Bioinformatics, Clustal Omega, Mushroom, Protein, rRNA, Sequences

I. INTRODUCTION

An interdisciplinary discipline of research called bioinformatics creates techniques and software tools for comprehending biological data, particularly when the data sets are huge and complicated and are accessible by the user (the biologist), who could be a layperson without computer technology experience. Bioinformatic tools are computer programs made for the purpose of performing sequence or structural analysis as well as retrieving relevant data from the vast amount of molecular biology and biological databases. Given the widespread nature of the scientific research community, these software tools must be made online accessible. With the help of bioinformatics, we can manage and make sense of the enormous volumes of data involved [1]. Bioinformatics entails the handling, archiving, and analysis of biological data. Examples of this may be: building databases to hold experimental results, Modelling how all of the chemical activities in a cell interact with one another and predicting how proteins fold. Gene annotation, gene expression analysis, and genome comparison may all be done using bioinformatics techniques. novel biotechnology applications, such gene therapy, customized medicine, and the creation of novel medications, can be created using this knowledge. Typically, a bioinformatics involves the following actions are Compile statistical information from biological data to create a computer model, Deal with a computational modelling issue and Analyze and assess a computational algorithm. [2]. In this paper we will look into basic bioinformatics tools and the implementation of clustal omega tool for the mushroom 5.8s rRNA sequences to know the pairwise identity matrix.

II. DATABASES FOR PROTEIN ANALYSIS (AMINO ACID SEQUENCE DATABASES)

Swiss-Prot: In order to provide a high level of annotations—such as descriptions of a protein's function, the structure of its domains, post-translational modifications, variants, etc.—as well as a low level of redundancy and a high level of database integration, the curated protein sequence database SWISS-PROT was created. http://www.ebi.ac.uk/swissprot/access.html), Swiss-Prot in ExPASy (http://us.expasy.org/sprot/) [3].

TrEMBL: Automatic TRanslations of European Molecular Biology Laboratory nucleotide sequences are known as TrEMBL. It is a library of protein sequences made up of unreviewed computer translations annotated with fresh DNA sequences found in nucleotide sequence databases. The EMBL Nucleotide Sequence Database contains all of the coding sequences received from (EMBL-BANK) that have not yet been annotated in Swiss-Prot. External link: TrEMBL (http://www.ebi.ac.uk/trembl/) [3].

PIR: The Protein Information Resource (PIR) is a comprehensive, open-access resource for protein informatics that aids in the advancement of genomic and proteomic study and knowledge. PIR manages the Protein Sequence collection (PSD), a collection of approximately 283,000 annotated protein sequences that spans the whole taxonomic spectrum. There are four sub-bases in the Protein Information Resource, each with a lower degree of annotation. External link: PIR (http://pir.georgetown.edu/) [3].

ENZYME: A wide range of characteristics and activities are covered by enzyme functional databases, including occurrence, the kinetics of processes that are mediated by

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

enzymes, structure, and metabolic activity. It establishes a connection between the Swiss-Prot sequences and the whole enzyme activity categorization.. External link: ENZYME (http://us.expasy.org/enzyme/) [3].

PROSITE: In order to identify protein families and domains, PROSITE is a library of motif descriptors with annotations. Patterns or profiles, which are obtained from numerous alignments of homologous sequences, are the motif descriptors utilized in PROSITE. Information on families, domains, and secondary structures of proteins is included.. External link: PROSITE (http://us.expasy.org/prosite/) [3].

INTERPRO: By grouping proteins into families, predicting domains, and identifying key locations, InterPro enables functional analysis of proteins. InterPro employs prediction models, referred to as signatures, supplied by several databases (referred to as member databases) that make up the InterPro collaboration to categorize proteins in this way. In order to provide a comprehensive integrated database and diagnostic tool, we merge protein signatures from several member databases into a single searchable resource. It combines data from several secondary structure databases, including PROSITE, and provides references to further databases and in-depth information. External link: INTERPRO (http://www.ebi.ac.uk/interpro/index.html) [3].

PDB: A database for the three-dimensional structural information of big biological entities like proteins and nucleic acids is called the Protein Data Bank (PDB). The information, which is often gathered by X-ray crystallography, NMR spectroscopy, or increasingly, cryo-electron microscopy and provided by biologists and biochemists from across the world, is publicly available on the Internet via the websites of its member organizations.. External link: PDB (http://www.rcsb.org/pdb/) [3].

III. MAJOR CATEGORIES OF BIOINFORMATICS TOOLS

On certain uses, there are both pre-made items and things that may be specially made. Both visualization tools and data-mining software are available for analyzing and retrieving information from proteomic databases and genomic sequence databases, respectively. These fall under several categories, including homology and similarity tools, tools for protein functional analysis, tools for sequence analysis, and other tools

A. Homology and Similarity Tools

Sequences that have diverged from a common ancestor are said to be homologous. Thus, whereas the homology of two sequences might be either true or incorrect, the degree of similarity between them can be quantified. This collection of tools may be used to find connections between innovative query sequences with illustrative structure and function and database sequences with known structure and function.

B. Protein Function Analysis:

The secondary (or derived) protein databases that provide details on protein motifs, signatures, and domains may be compared to your protein sequence using the tools in this category of applications. It is possible to approximatively determine the biochemical role of your query protein based on very significant hits against these several pattern databases.

C. Structural Analysis:

With the help of this collection of tools, we may do further, more thorough analyses on the query sequence, such as compositional bias analysis, evolutionary analysis, and the detection of mutations, hydropathy areas, CpG islands, and mutations. These and other biological characteristics can be used as hints in the investigation of the precise function of your sequence. The Table 1. mentioned about the basic bioinformatic tools.

TABLE 1. BIOINFORMATICS TOOLS [4]

Tools	Description
BLAST	It is a search tool, used for DNA or protein sequence search based on identity[4].
HMMER	Homologous protein sequences may be searched from the respective databases using this tool[4].
Clustal Omega	This application is capable of doing multiple sequence alignments[4].
Sequerome	This Application is used for Used for sequence profiling[4].
ProtParam	Used to predict the physico-chemical properties of proteins[4].
JIGSAW	To find genes, and to predict the splicing sites in the selected DNA sequences[4].
novoSNP	Used to find the single nucleotide variation in the DNA sequence[4].
ORF Finder	This tool may be used to discover Open Reading Frames (ORF) in putative genes[4].
PPP	A tool for predicting the promoter sequences that are present upstream of a gene in eukaryotes[4].
Virtual Foorprint	This tool may be used to analyze promoter regions with several regulator patterns as well as the entire bacterial genome (with one regular pattern) [4].
WebGeSTer	This collection of transcription terminator sequences is used to forecast the locations where genes will be terminated during transcription [4].

IV. IMPLEMENTATION OF CLUSTAL OMEGA TOOL

Multiple sequence alignment is frequently done using the program Clustal Omega.The benchmarks, which include a recently disclosed approach based on secondary structure prediction, are based on comparisons or predictions of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

protein structures. The precision of protein alignments is great when compared to other programs, and Clustal Omega is generally quick enough to create very large alignments [5]. The TABLE 2. Represents data collection from NCBI online database of mushroom 5.8s rRNA sequences around tamilnadu[6][7].

TABLE 2. DATA COLLECTION [6][7]

ID	DETAILS OF STRAIN
KY491659.1	Fulvifomes fastuosus strain LDCMY43 [6]
KY491658.1	Phellinus sp. strain LDCMY23 [6]
KY471289.1	Ganoderma sp. strain LDCMY12 [6]
KY471288.1	Phellinus sp. strain LDCMY45 [6]
KY471287.1	Inonotus rickii strain LDCMY52 [6]
KY471286.1	Phellinus sp. strain LDCMY 24 [6]
KY111254.1	Coriolopsis caperata strain LDCMY42 [6]
KY111253.1	Ganoderma wiiroense strain LDCMY11 [6]
KY111252.1	Fomitopsis ostreiformis strain LDCMY21 [6]
KY111251.1	Ganoderma sp. strain LDCMY16 [6]
KY111250.1	Ganoderma sp. strain LDCMY41 [6]
KY111249.1	Phellinus badius strain LDCMY36 [6]
KX957805.1	Phellinus sp. strain LDCMY34 [6]
KX957804.1	Phellinus badius strain LDCMY31 [6]
KX957803.1	Phellinus sp. strain LDCMY29 [6]
KX957802.1	Phellinus sp. strain LDCMY28 [6]
KX957801.1	Phellinus badius strain LDCMY27 [6]
KX957800.1	Ganoderma sp. strain LDCMY05 [6]
KX957799.1	Ganoderma resinaceum strain LDCMY01 [6]
KX957798.1	Fulvifomes fastuosus strain LDCMY39 [6]
KY009873.1	Ganoderma wiiroense strain LDCMY19 [6]
KY009872.1	Ganoderma sp. strain LDCMY14 [6]
KY009871.1	Ganoderma sp. strain LDCMY22 [6]
KY009870.1	Ganoderma sp. strain LDCMY18 [6]
KY009869.1	Ganoderma wiiroense strain LDCMY17 [7]
KY009868.1	Trametes elegans strain LDCMY37 [7]
KY009867.1	Ganoderma wiiroense strain LDCMY08 [7]
KY009866.1	Ganoderma sp. strain LDCMY04 [7]
KY009865.1	Ganoderma sp. strain LDCMY06 [7]
KY009864.1	Ganoderma wiiroense strain LDCMY02 [7]

The FIGURE 1. Represents the online clustal omega tool to perform multiple sequence alignment using the above 30 mushroom sequences. In which we have to choose the sequence either RNA, DNA or protein as input data. here we have DNA sequences so we have to choose the DNA option in the dropdown box.

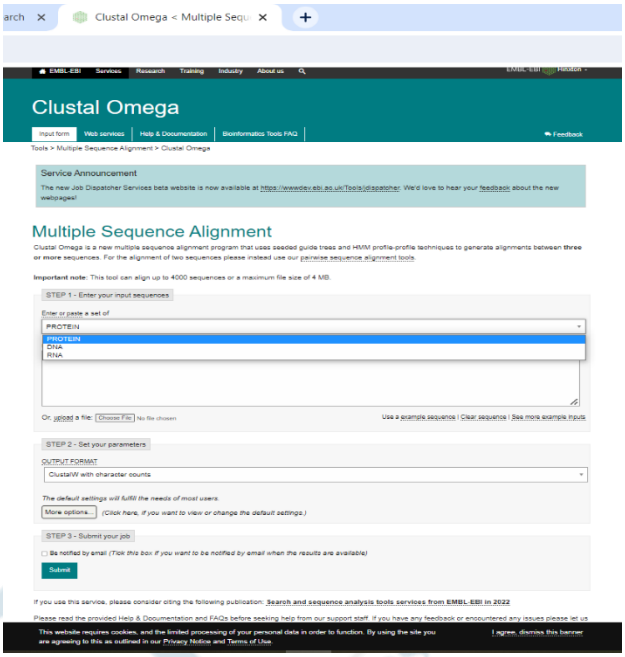

FIGURE 1. Clustal omega Tool

Then select the way of process which in mentioned in the FIGURE 2. Here we have to choose the clustalw and charater count option to denote the process of Clustal multi alignment process in which employ an alignment engine rather than dynamic programming and profile alignment to align the profiles of hidden Markov models (HMMs). A profile HMM created from previous alignments can contain additional information that Clustal Omega can read. For instance, if a user already has a globin alignment and wants to utilize it in conjunction with the sequence input file to align a collection of globin sequences, they can do so. This HMM is described as a "external profile" in this context, and its application as "external profile alignment" (EPA). Each sequence in the input set is aligned to the external profile throughout the EPA process. Then, position by position, pseudocount data from the external profile is transferred to the input sequence. This would ideally be used to massively curated alignments of specific proteins or domains of interest, such those utilized in metagenomics investigations [8][9].

Finally choose the submit button, it will process our input by validataing the sequences and gaps everything. Untill then **FIGURE 3.** Screen will be shown for the user. Still its been using as online tool so that user can access this information anywhere anytime at any place. The result have been stored for 7days for the further references after that it will be vanished we have to do the process again to get back the result.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

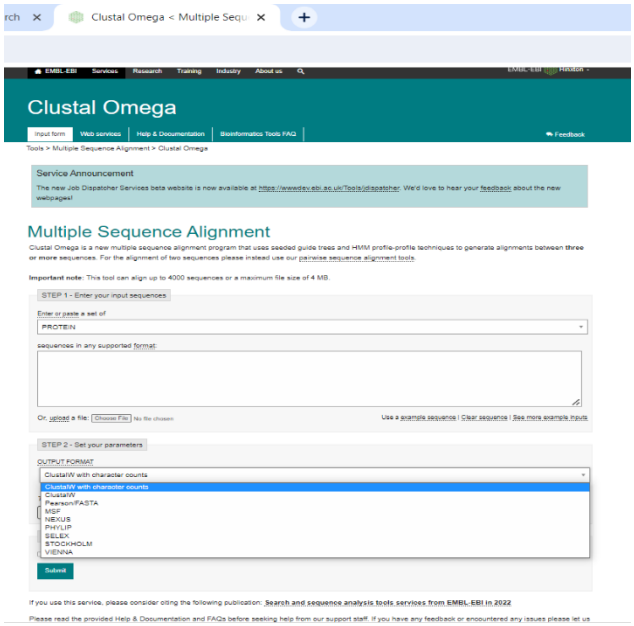

FIGURE 2. CLUSTAL OMEGA TOOL

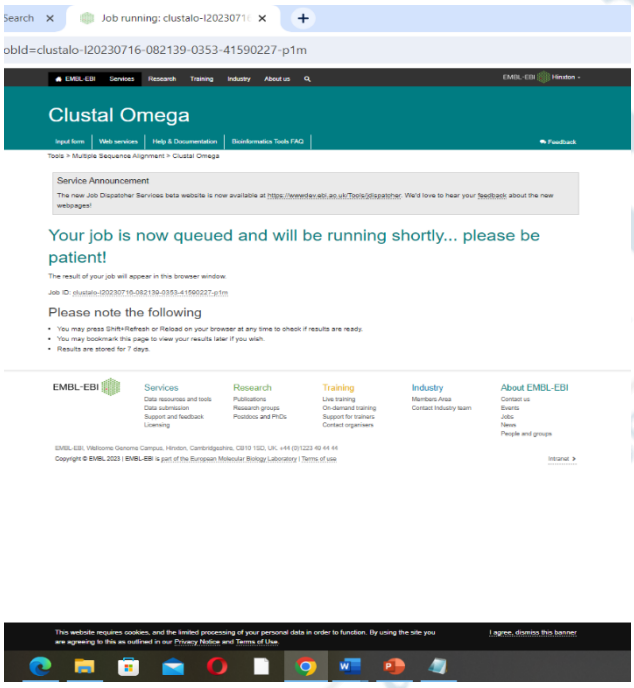

FIGURE 3. CLUSTAL OMEGA PROCESS 3

The FIGURE 4. Represents the entire result summary of our given input like, basic alignment of entire input data, guide tree, phylogentic tree, pairwise identity matrix.

The FIGURE 5. Shown the representation of phylogentic tree in which the entire 30 mushroom sequences was arrangened as family tree by the way of hierarchical order. According the result of similarity between each other which is mentioned in the Figure 6. The pair wise identity matrix given 30*30 matrix format similarity score for evey sequences, totally 900 data points invloved in the sheet[10].

So according to this PIM value of each other the phylogentic tree was delivered, here in Figure 5. It shown the order of gene sequence alignemnt in which we can identify the similarity and order of family which the sequnces belongs to like ancestors of the sequences.

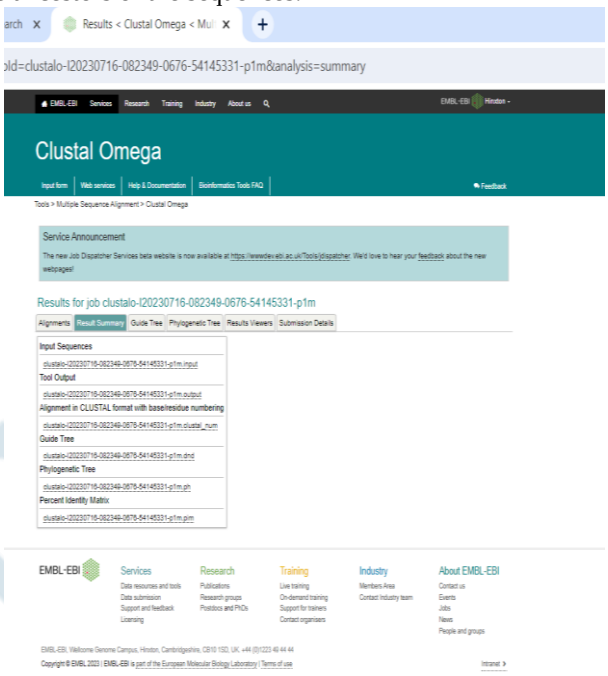

FIGURE 4. CLUSTAL OMEGA RESULT SUMMARY

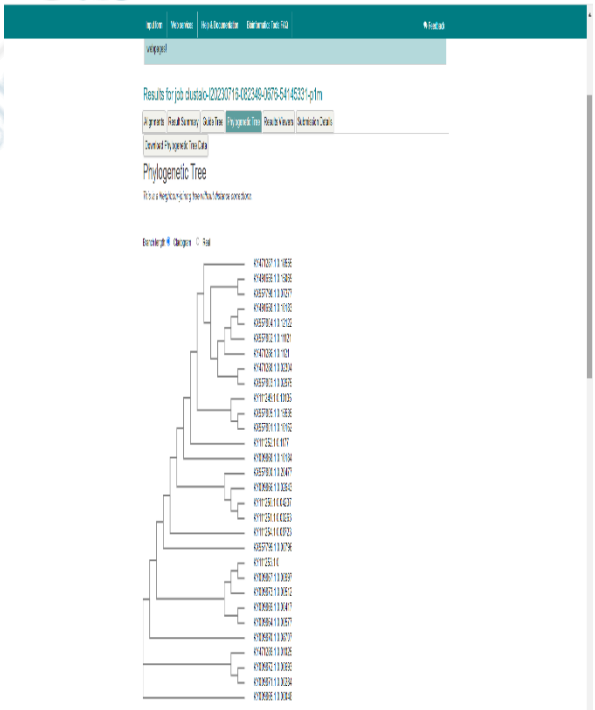

FIGURE 5. CLUSTAL OMEGA PHYLOGENTIC TREE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

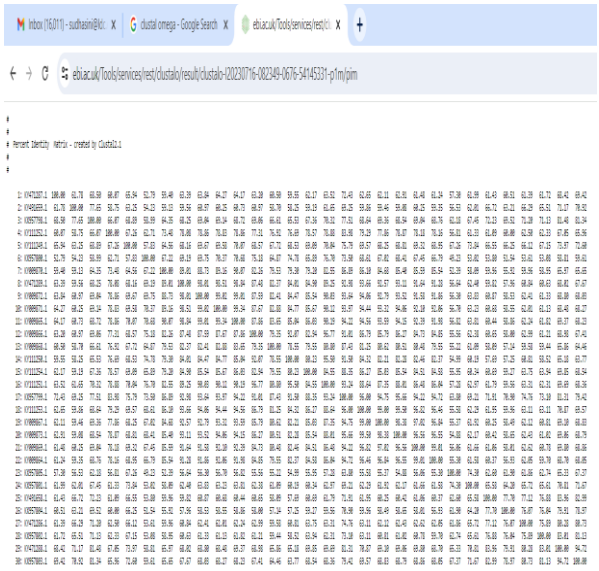

FIGURE 6. CLUSTAL OMEGA PIM MATRIX

V. CONCLUSION

In particular, when the data sets are large and complex and are accessible by the user (the biologist), who may be a layperson without computer technology experience, bioinformatics, an interdisciplinary field of research, develops methods and software tools for understanding biological data. The large quantity of molecular biology and biological databases may be accessed using bioinformatic tools, which are computer programs designed to perform structural or sequence analysis and to retrieve pertinent data. These software tools must be made available online because of how widely used the scientific research community is. Bioinformatics can help us manage and make sense of the massive amounts of data involved. In addition to performing the implementation of the cluster omega tool for the mushroom 5.8s rRNA sequences to determine the pairwise identity matrix, the article elaborated on basic bioinformatics tools that are useful for large data sequences. This information was used to fine-tune the phylogenetic tree to identify the ancestors of each sequence.

REFERENCES

[1] Bindel, https://omicstutorials.com/bioinformatics-tools-softwares-programmes/

[2] Can T, “Introduction to bioinformatics. Methods”, Mol Biol. 2014;1107:51-71. doi: 10.1007/978-1-62703-748-8_4. PMID: 24272431, 2014

[3] Edna María Hernández-Domínguez, Laura Sofía Castillo-Ortega, Yarely García-Esquivel, Virginia Mandujano-González, Gerardo Díaz-Godínez and Jorge Álvarez-Cervantes, “Bioinformatics as a Tool for the Structural and Evolutionary Analysis of Proteins”, Computational Biology and Chemistry, ISBN978-1-78985-691-0, DOI: 10.5772/intechopen.89594, 2020

[4] Mehmood MA, Sehar U, Ahmad N (2014) Use of Bioinformatics Tools in Different Spheres of Life Sciences. J Data Mining Genomics Proteomics 5: 158. doi:10.4172/2153-0602.1000158

[5] Fabian Sievers, Desmond G. Higgins, “Clustal Omega for making accurate alignments of many protein sequences”, Tools for Protein Science, Wiley online library, vol 27, issue 1, Pages 135-145, https://doi.org/10.1002/pro.3290, 2018

[6] P.Sudhasini, Dr.B.Ashadevi, “ Pairwise Sequence Alignment Similarity Score Prediction on Mushroom Biological data “ , International Journal of Advanced Science and Technology Vol. 29, No. 4s, (2020), pp. 1844-1867 , 2020

[7] Sudhasini P, Ashadevi B. Clustering Mushroom 5.8s rRNA Sequences using k-means Algorithm with Predicted k Value. In: 5th international conference on intelligent computing and control systems, IEEE. 2021. Available from: https://doi.org/10.1109/ICICCS51141.2021.9432167

[8] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li , Rodrigo Lopez , Hamish McWilliam , Michael Remmert , Johannes So¨ding , Julie D Thompson and Desmond G Higgins, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega“, Molecular Systems Biology 7; Article number 539; doi:10.1038/msb.2011.75, 2011

[9] Sievers, F., Higgins, D.G. (2014). Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In: Russell, D. (eds) Multiple Sequence Alignment Methods. Methods in Molecular Biology, vol 1079. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-646-7_6, 2013

[10] Masoodi F, Quasim M, Bukhari S, Dixit S, Alam S. Applications of Machine Learning and Deep Learning on Biological Data. 1st ed. and others, editor;Auerbach Publications. Taylor & Francis. CRC press. 2023. Available from: https://www.routledge.com/Advances-in-Computational-Collective intelligence/book-series/ACCICRC