International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 5 27 - 36

Sequential Pattern Mining Aide to Bio-Informatics

Ratna Giri Devavarapu
Asst Professor, Dept of Information Technology,
Sagi Ramakrishnam Raju Engineering College,
Chinamiram, Bhimavaram-534 204, A.P., India.
e-mail: drsrkrit *@gmail.com*

Dr. G.Murali
Professor, Dept of Computer Science and Engineering,
KKR &KSR Institute of Technology,
Guntur, India.
e-mail: drgudipatimurali@gmail.com

Abstract— Practical Bio-Informatic is the study of all vicinities of development, testing and novel appliances for statistical and computational techniques for prototype and study of all types of scientific data, in addition to further areas of Information Technology and Sciences. Bio-Informatics is a novel approach to conceptualize the natural science in provisions of molecules and apply Informatic methods is derived from computer science and applied mathematics regulation, for instance, info to be grateful for and systematize them in order to relate with these molecules, on a large scale for use in future research studies. Bio-Informatics is the study of upcoming appliances in natural science, chemistry, pharmaceuticals, medicine, and agriculture and various additional fields of research and development. Many pharmaceutical manufacturing companies are attracted in mining sequential patterns from the databases. Sequential Pattern Mining is doing good technique of data mining, which recognizes the temporal relationship between different drugs and it can help in estimating the treatment course for patients. These studies give an improvement in the sympathetic of the loom of Sequential Pattern Mining and Bio-Informatics play a part to a vital role in a biomedical study in the storage of patient's case reports which is useful in providing treatment to other patients.

Keywords- *Bio-Informatics, Chemistry, Pharmaceuticals, Medicine, Agriculture, Sequential Pattern Mining etc.*

*****

I. INTRODUCTION

Bio-Informatics is a distinction of study by various persons in different ways to gain the knowledge from computer analysis of biological data. It is a quickly growing division of natural science inter-disciplinary by using various methods and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has a lot of appliances in the different vicinity of natural science and medicine. It can describe the make use of computer machines to handle the natural science information. Practical Bioinformatics is a long-standing problem on protein folding and there have been a lot of experimental, theoretical and computational revisions so far efforts by many types of research have exposed mysterious folding mechanisms of proteins(1). This statement can be understood as the make use of computer machines to characterize the molecular gears of income organisms. Bio-Informatics is the meadow of science emerged from various other disciplines combination like natural science, Computer Science and Information Technology and forms a solitary discipline.

Bio-Informatics take part in three most critical roles that they are initially, improvement of new-fangled algorithms and figures to evaluate associations among the members of huge datasets. Secondly, the investigation and understanding of different kinds of data together with nucleotide and amino acid series, protein areas and protein buildings, and lastly, the development and execution tools that allow us to well-organized and accessing the different types of information professionally. Practical bioinformatics includes method development for storage, retrieval, and analysis of data. The present research fields like Computational biology, Pharmacokinetics, Cheminformatics, Structural genomics, Comparative genomics, Biophysics, Biomedical informatics, Mathematical biology, Genomics, Proteomics, Pharmacogenomics, Computational Chemistry, Functional genomics, Pharmacoinformatics, Agro-informatics etc., are some of the fields related to Bioinformatics era. Throughout the years, the accumulation of biological data/information has been quickly expanding because of the advancements and enhancements to existing technologies and different facilities. A good example of this would be the Human Genome Project which was founded in 1990 by the United State Department of Energy and the US National Institutes of Health and was lastly finished in 2003. To increases, the high volume data, in the end, that leads to use of calculation and arithmetical methods to organize, store, maintain, examine and take the biological results(2).

Bio-informatics to work out the biology engage the employ of various methods, which includes informatics, statistics, applied mathematics, computer science, artificial intelligence, chemistry, and biochemistry, etc., to solve biological evils typically on the molecular level. The most ordinary evils that arise in model-based biological processes at the molecular level and making any conclusion from data collected. Bio-Informatics that gives a solution in which involves following steps: initially collect the statistics from biological data

obtained, it builds a computational model to solve computational prototype problems and lastly to test and evaluate the computational algorithms (3).

Bioinformatics Scope: Most recent daily biological research to use the clinical and medical laboratories are producing bulky datasets using various methods and analyzing such data is not an easy task and requires other various computational techniques. Thus, bioinformatics has a crucial role in modern-day research and development as it involves various other disciplines like statistics, information technology, etc. Areas of research field including medical genomics and genetic disorders are increasing and personalized medicines are mostly depended on various bioinformatics methodologies.

Bio-Informatics endeavors to understand the biological issues and predict behavior through logic and mathematics which is made through computing experimental data into consent graphics and attempts to extrapolate algorithms and recurrences. In large-scale studies in bioinformatics, the main purpose is to determine how much life is computable. As bioinformatics developed mainly the wide appliances of high-throughput progression methods and more transcriptional units with no or small protein-coding discover (4). Due to the interdisciplinary nature of bio-informatics, multiple bioinformatics techniques have become popular in different fields and in consequence, these disciplines, in turn, have a prominent influence on future of bioinformatics development(5). Bioinformatics examination of forensic DNA has important inference for the organization of forensic evidence for identifying criminals and the integration of crime databases with public health and population genetics databases (6). The important Bio-Informatics has become is a part of clinical laboratories which are involved in generating, maintaining, analyzing, and interpreting data acquired from molecular genetics testing(7). The fastest growth of data from genomics and proteomics studies worldwide caused bioinformatics to gain importance and became famous and because of its multidisciplinary nature created its much demand by various biologists and in computing(8).

Our Aim of Bioinformatics helps in interpreting the relationship between living things and analyzes and solves biological problems by combining basic biological concepts with computer technologies and foundational theories. In short, Bio-Informatics is an MIS (Management Information System) for molecular biology and has several realistic appliances.

The aims of practical bio-informatics are:

1. Bio-Informatics arranges the data in a manner that supports researchers to use and obtain existing information and to submit new entries as they are produced by everyday researchers.
2. It will help in designing new tools and resources that support in the scrutiny and supervision of data in large scales obtained from huge industries and pharmaceutical companies.
3. Bioinformatics help uses the available data to scrutinize and construe results in nature in an essential way.
4. To help researchers in the Pharmaceutical industries and research laboratories in understanding the protein structures which in turn helps in designing new drugs for therapeutic purpose.

B. Bio-Informatics Tasks: Knowledge acquisitions of the biological and medical labs uses the various approaches that produce exceedingly in large data sets in which cannot be analyzed by hand - for example, sequencing human genomes which are very large in length and managing of it by hand is impossible. Thus contemporary biological and medical research and development cannot be done without the help of bio-informatics which gives the use of computer knowledge and mathematical modeling. A modern scholar and researchers effort in the area of genetic diseases and medical genomics is quickly rising and the future of personalized medicine depends on bioinformatics approaches to compare previously sequenced data obtainable in different databases. The errands of Bio-Informatics involve the examination of sequential information that is previously congregated and stored. This involves the following activities:

1. Identify the genes in the DNA series as of different organisms and comparing them with already sequenced organisms to analyze genetic similarity among them.
2. Classifying families of related sequences and the development of novel models.
3. Align the analogous series and generate the Phylogenetic trees to scrutinize evolutionary relations.
4. Finding all the genes and proteins of a genome from a given sequence of amino-acids.

In this regard, an active predicting sites in the protein structures to attach drug molecules helping in drug designing.

C. Bioinformatics Applications: Bio-Informatics is a significance of computer skills to get the information that accumulates in certain kind of biological databases. To understand the influence of Practical bioinformatics that provides central, internationally easy to get to databases that enable scientists to submit, search and analyses information and that is newly obtained, and also compare them with previous information available in databases. Bio-Informatics is a characterized as the administration, obtaining, control, and the introduction of complex biological information or biological data sets, and clinical informatics is the application of information management in the healthcare sector to encourage safe, efficient, effective, personalized, and responsive care. Computational contents manipulate different types of raw data such that researchers can review and analyze

for clinical coverage stored in various databases (9). It offers analysis of various kinds of data using software applications intended for data study and comparisons and provides apparatus for modeling, visualizing, exploring the interpreting data.

It is our major aspiration is to convert a huge number of intricate the data into useful information and knowledge. Computational biology has its appliances in many areas, prominently in the field of science. It helps in providing practical tools to explore Proteins and DNA in a number of further ways which in previous years are very difficult or not possible. Interrelating structures and functions of various sequences and checking similarity among them is hard to define and for the purpose, Bio-computing provides a recognition technique which eases the task. Bioinformatics tool (RasMol) has important application in predicting the 3D structure of the protein from linear amino acids sequence data. RasMol is one of the software of bioinformatics which is used to visualize macromolecular structure. The problem is to make simpler of sympathetic intricate genomes by analyzing, simple organisms and then applying the same ideologies to additional problematical ones. This would result in identifying potential drug target by scrutiny homologies of essential microbial proteins. To study the emerging genomic series data and the human genome project is a familiar sight attainment for bioinformatics. Any individual from clinicians to information technologists with access to the internet can liberally obtain the composition of biological molecules such as protein sequences and nucleic acid sequences stored in different databases using bioinformatics gears. Large-scale industries and organizations for example pharmaceutical companies make use of Bio-informaticians uphold and fulfill the need for bioinformatics tools in the organization (10) Bioinformatics is useful in designing drugs. In primary research indication of similarities that have been employed to investigate disease resemblance and in line to identify new uses for existing drugs and it was recognized that symptom-based similarities alone are inadequate to foresee that new curative uses for existing drugs (11).

D. Bioinformatics Challenges: The major Bio-informatics confronts are mainly related to the everyday flood of raw data obtaining through research studies, aggregate meaning information, evolving knowledge arising from the study of the genome and its manifestation. Informatics has assisted to launch the molecular biology into genome epoch. The utilization of informatics in organizing, managing, and analyzing genomic data (the genetic material of an organism) has become an important part of biology and medical research. A new IT related discipline bioinformatics combines computing, mathematics, and biology to meet the many computational challenges in modern molecular biology and medical research that solves various problems that arise in biomedical researchers. Bioinformatics is mainly based on two: Data management and knowledge discovery and rely on effectively adopting techniques developed in IT for biological data, with IT scientists playing an essential role in developing various software ultimately helping environment and mankind. There is a solid inspiration in the utilization of machine learning strategies in knowledge discovery and data mining to create models of biological inference. The historical backdrop of the connection between machine learning and science is viewed as long and complex.

The nonstop increment in the volume and detail of information is obtained by data scientists and biological researcher, for example, the rise in online networking, Internet of Things (IoT), and interactive media has delivered a mind-boggling stream of information in either organized or unstructured organization. Tending to enormous information is a challenging and time-consuming task that requires a huge computational framework to guarantee fruitful information handling and examination. However, big data is changing in healthcare, science, engineering, finance, business, and finally the society. The development of new technologies and methods in storing of data and data mining techniques allow for the safeguarding the continuously growing amounts of data which was described by a change in the nature of data held by different organizations in various fields (12).

The huge volumes of data generated by the genome sequencing projects and making use of it is a big challenge for scientists of molecular biology as it is confusing and hard to understand. Bioinformatics simplifies complex data using different software's benefitting researchers in various fields across the globe. Bioinformatics introduced use of computers in molecular biology studies/experiments. The prime challenge faced by Bioinformatics is to handle diverse types of data, to provide a well understanding of the functions of the genes. As part of this, the data must be annotated, filtered and visualized in a better way. Bioinformatics found useful in providing new tools for prediction of the structure of proteins sequences which help in studying various functions of the body of single or multi-cellular organisms which in turn helps in designing of new drugs. A well-versed Bioinformaticians get opportunities in a variety of fields like Genomic study projects, Proteomics study, Drug discovery, gene therapy and many more which is not limited to biological science.

E. Bioinformatics Career Opportunities: Bioinformatics is an emerging scientific field which is growing at a good pace as new researchers and gathering of important data is increasing simultaneously. Bio-Informatic majorly employs web-based programs and databases to access the wealth of data to answer questions relevant to the data obtained by comparing it with already gathered information available on the web in different databases. There are lots career opportunities available for different stream students involved in bioinformatics studies;

which includes Database Programmer for storing of excessive data obtained from daily routine research laboratories, Software Developer for designing new software for prediction of protein 3D structure, etc.

II. BIOINFORMATICS AND GENOMICS

Genomics: Genomics is the development and application of molecular mapping, sequencing, characterization, computation, and analysis of entire genomes of organisms and whole sets of gene products. According to Xintao Wu, a computer scientist at the University of Arkansas, protecting genetic privacy has become a significant and challenging topic as the era of personal genomics is quickly approaching. The analysis of whole genome gives insight into the global organization, expression, regulation, and evolution of the hereditary materials. Biologists from the very long time have been focusing on understanding the natural biological processes and mechanisms that occur in the human body. Numerous fundamental questions and problems are unanswered and are waiting for the solution, for example, how the genome regulates the full cellular metabolic network and how unicellular organisms evolve into multi-cellular organisms through various changes in the environment. In the same way, one of the major questions for the conventional biological researchers is that how the phenotypic characteristics of a particular cell or organism are controlled by its genome. Synthetic biology adapts a reverse approach and practices an entirely new direction to address this problem. Synthetic biology involves modifying or de novo synthesize an organism's genome in order to enable designed biological features (13). Genomics has three subfields, they are:

1. Structural genomics which deals with the genetic mapping, physical mapping and sequencing of entire genomes.
2. Functional genomics is the global study of the structure, expression patterns, interactions and regulation of the RNA and Protein encoded by the genome.

Comparative genomics facilitates the comparison of entire genomics of different species with the goal of enhancing the understanding of the functions of each genome including evolutionary relationships. A living organism is characterized by its fundamental set of genetic information called genome. A genome is composed of one or more DNA molecules each organized as a chromosome, for example, in humans in each cell contains 23 pairs of chromosomes. The DNA has all information encoded in it for a proper functioning of the cell. Any modification in the cell genome leads to apoptosis or leads to the formation of a tumor. The primary role of nucleic acids is to carry the encoding of the primary structure of proteins. The Protein sequence is derived from the DNA using various bioinformatics technologies. The Protein's structure determines the protein function. Different sequences of bases in DNA specify different sequences of bases in RNA, and the sequence of bases in RNA specifies the sequences of amino acids in proteins. DNA is a linear double helical structure as determined by Watson and Crick. The double helix is composed of two intertwined chains called nucleotides. Each nucleotide consists of a Phosphate group, a Deoxyribose sugar molecule and one of four different Nitrogenous bases Adenine (A), Guanine (G), Cytosine(C) and Thymine (T). RNA is the other major nucleic acid and it is single-stranded, unlike DNA which is double-stranded. It contains ribose instead of Deoxyribose. There are three types of RNAs in the cells for use in protein synthesis. They are Messenger RNA (mRNA), Ribosomal RNA (mRNA) and Transfer RNA (tRNA). mRNA acts as a template for protein synthesis. The rRNA and tRNA form a part of the protein synthesizing machinery. An interesting development in experimental biology is the use of RNA- i (RNA-interference).

III. BIOINFORMATICS AND PROTEOMICS

A. Proteomics: The total complement of proteins present in a cell or cell type is known as its proteome and the study of such large data sets defines the field of Proteomics. The term of proteomics was invented in parallel to genomics. Important experimental techniques in proteomics include protein microarrays which allow the detection of the relative levels of a large number of proteins present in a cell. This is the cataloging and analysis of Proteins to determine the expression times of a protein and the interactions of various Proteins. Proteomics represents the genome at work and it is a dynamic process. The appearance or non-appearance of proteins and immediate estimation of the relatively large quantity of proteins can help to comprehend cellular processes and might be valuable for recognizing drug targets and diagnostic/prognostic markers. Proteomics is one of the most active fields of biological research, as it has its wide range of applications in biological field research studies (14).

Drugs that are designed to potentially target specific proteins which control various protein-related human diseases. In the analysis done in 2007, it is estimated that most number of drug targets to be in hundreds. However, the number of druggable proteins given in Drug Bank database Web site (http://www.drugbank.ca/) is substantially greater according. The disclosure of medications in light of in silico docking of the inhibitors in models of the 3-dimensional (3D) structures of the protein targets has proven to be of significant value in the processes of rational drug design as well as drug screening and has majorly donated to conserving resources in the area of drug discovery. Possibly by using this information will help in considerable saving the cost and time required for designing and release of drug or medication (15).

Proteomics can be divided into two types:

1. Expression Proteomics is the study of global changes in Protein expression.

2. Cell map Proteomics is the systematic study of Protein-Protein Interactions.

The protein expression levels changes throughout the life of an organism. Proteomics attempts to catalog and characterize these Proteins and compare variations in their expression levels in healthy and diseased tissues.

B. Basic Local Alignment Search Tool (BLAST): Basic Local Alignment Search Tool (BLAST) is a sequence similarity search program that is available on the web or as a stand-alone computer tool to compare a user's query to a database of sequences which is previously obtained by different researcher across the globe. Various forms of BLAST involve in comparing nucleotide or protein sequences that are stored in protein or nucleotide databases.

BLAST is a heuristic process or method that finds short matches between two sequences and tries to start alignments by comparing them from these hot spots. BLAST requires or it decreases search space or a number of comparisons it makes which in turn increases its speed of the alignment. BLAST involved in performing alignments, not only this but it also provides statistics of the data about the alignment which is called expected value or false positive rate (16). BLAST is computer algorithm software that is freely available for use online at the National Center for Biotechnology Information (NCBI) website, as well as many other sites and can be downloaded. Most of the biological and clinical researchers or pharmacists use BLAST which speedily aligns and compare a query of protein or DNA sequences with that of the sequences present in databases which makes it valuable tools in genomic and proteomic research. Prior to using BLAST, computer science principles have been not applied in biological studies, before BLAST, other dynamic programs use dynamic algorithm programs such as Needleman-Wunsch and Smith-Waterman algorithms. But researchers used to face certain challenges using these dynamic algorithm programs like these programs requires long processing time an use or parallel computer processors or supercomputers(17). Recent years studies involve the application of computational techniques to biological data and that play a major part of science, specifically in the analysis of protein and DNA sequences for various therapeutic purposes. Examining the similarities between nucleotide or protein sequences is a method by which bioinformatics provides understanding and solving of various biological problems. Commonly used bioinformatics tools to compare or align these DNA or protein sequences are BLAST and FASTA which is used worldwide. These tools compare the pair of nucleotide or protein sequences based on their local similarities which help in studying evolution in various living organisms (18).

C.Fixed Local Alignment Search Tool (FASTA): There are two software packages available on the web for similarity searching which is one of the most powerful strategies for distinguishing newly obtained sequences from biological research studies. The FASTA programs give an inclusive set of fast similarity searching tools, which is very much identical to those provided by the BLAST package, as well as it provides programs for slower, optimal, local and global similarity searches for more accurate results for searching with short peptides and oligonucleotides.STA programs can also develop BLAST-like alignment and tabular output, for smoothing the incorporation into existing analysis pipelines, and can analyze small, representative databases, which finally report results for a larger set of sequences, using links from the smaller dataset. The FASTA programs work with various types of database formats, which include MySQL and PostgreSQL databases. Different researchers widely across the world run FASTA program through web servers, like for example similarity searching tools at European Bioinformatics Institute (http://www.ebi.ac.uk/Tools/sss). This FASTA package can also be downloaded and run on various computer operating systems like UNIX, MacOS terminal, or Windows console (19). This FASTA program finds deoxyribonucleic acid (DNA) databases and protein databases with statistically important similarities among the sequences. This FASTA package compares protein and DNA similar to BLAST package and in addition, it also compares short peptides and oligonucleotide and gives results on computers after searching and analysis.

Both FASTA and BLAST packages attempt to identify homologous proteins or DNA sequences in the obtained data by comparing them with already stored data. Compared to FASTA package, BLAST is faster, but FASTA is more flexible which provides rigorous (SSEARCH, LALIGN, GGSEARCH and GLSEARCH) and heuristic (FASTA, FASTX/Y, TFASTX/Y and FASTS/M/F) algorithms, also huge range of scoring matrices and different approaches for estimating statistical importance (20).

IV. BIOINFORMATICS DATABASE SYSTEMS

Present-day biological databases not only include important information, but also complicated query facilities and bioinformatics data analysis tools. Bioinformatics Database Systems provides a thorough examination through the world of Bioinformatics Database Systems. The innovative bioinformatics research most popular and storing places now available. Bioinformatics is Knowledge acquisition and Experimental study which also includes popular primary genetic and protein sequence databases, structure and pathway databases, microarray databases, phylogenetic databases, and boutique databases. These database systems also explore the data quality and information combination issues in the management of bioinformatics databases, including data quality problems that have been obtained, and strong efforts in the data cleaning field. Biological data integration problems

are likewise canvassed inside and out, and Bioinformatics Database Systems shows how information coordination can make new archives to address the requirements of the biological societies. It additionally displays regular information combination models utilized in current bioinformatics databases. Bioinformatics Database Systems covers biological data mining and biological data processing approaches using cloud-based technologies. The growing usage of Information Technology in Biological sciences paved the way to a large number of websites, databases, tools and software's available on the World Wide Web for open access. In addition, an enormous amount of literature is also available for ready reference. The internet has changed the way in which the data in a Central Data Warehouse is shared by the researchers across the globe.

Biological Databases: A biological Database is a huge collection of persistent data supported by software meant for the update, retrieve and maintain data. There are different types of databases depending on the nature and type of data stored in the database. The data in a biological database may be of type sequences or structures 2D gel or 3D structure images. In the case of Protein sequence analysis Primary, composite and secondary databases are needed. These databases store different levels of protein sequence information.

A. Primary Databases: The growing demands for the sequence information during the 1980s, a lot of primary database projects were taken up and resulted in the creation of nucleic acid and Protein sequence databases. Some of the significant DNA sequence databases are DDBJ (Japan), EMBL (Europe) and Gene Bank (USA). These databases exchange data on regular basis to ensure consistency of data. The early 1960s witnessed the development of the Protein Sequence database at the National Biomedical Research Foundation (NBRF). Currently, this database is split into four distinct sections designated as PIR1 thru PIR4. They differ in terms of the quality of data and the level of annotation. There are few protein sequence databases of great significance like MIPS, SWISS-PROT, etc. A major problem faced by primary databases is a rapid increase of data and storing them in the database creates the majority of problems. Such problems can be reduced by developing other protein sequence databases that help search more successful.

B.Secondary Databases: Secondary Databases uses other databases as their source of important information which can be primary databases. The mostly secondary database contains pattern data. Widely used and the popular secondary database includes PROSITE, profiles, Pam, etc. Similarly, tertiary database arises from the data stored in secondary databases. A bibliographic database is slightly different from these databases as it stores biological and other fields of data published in the form of article, books, and abstracts. Scientific writers and researchers extensively used bibliographic databases for their exposure to new research. One of the bibliographic databases extensively used in the world includes PubMed.

V. BIOINFORMATICS TOOLS

An earlier generation of Bioinformatics used tools and applications with a text-based interface. BLAST is a most popular tool that is widely used by biologists. This is an algorithm for searching large databases of Protein or DNA sequences. The NCBI provides a web-based implementation that searches the massive sequences and annotated data. Programming languages like Perl and Python are used to interface with biological databases and parse output from programs written in routine languages like C, C++ etc., to implement bioinformatics algorithms. Bio-Informatic meta-search engines like sequence profiling tools are available to find relevant information from several databases. For various bioinformatics applications, SOAP-based interfaces have been developed which authorize one program of the computer running any part of the world to use data resources on another computer located in any other part of the world. Bioinformatics research and web services along with the open source bioinformatics the large obtainability of collections lead to the following generation bioinformatics tools called integrated bioinformatics platform. These tools range from a web-based interface to an extensible bioinformatics workflow development environment. There is some open and free source of software packages available online developed by bioinformatics programmers and distribute these tools as they produce. As a result of this most recent and up-to-date tools are available to the researchers.

A. Sequence Alignment: From the perspective of the biologist's sequence Comparison is motivated by the fact that all living organisms are related by evolution. The genes of organisms that is closer to each other show signs of similarity at DNA level. Alignment of sequences of DNA, RNA or Proteins is a type of arrangement of primary sequences to detect similarity among them for any structural, functional or any evolutionary relationships in the sequences of different species. Sequence alignment may be considered as the identification of residue-residue correspondence. The Sequence alignment is a basic tool of bioinformatics. By using sequence alignment one can measure the similarity between two or more sequences, determine the residue-residue correspondences, observe patterns of conservation and variability and infer evolutionary relationships. By performing these operations one can search the databanks for related sequences. Two sequences from two different organisms can be said to be homologous if the sequences are similar and have a common ancestor sequence. Alignment of protein sequences

is a task to search and identify evolutionary or structurally related targets in multiple sequences of amino acids. However aligning protein sequences problem has been studying from decades, most of the studies stated that there is a notable increase or growth in multiple or pairwise alignment tools which improve accuracy in aligning and easy for researchers to compare various new sequences with the stored important sequences (21).

The alignment of sequences is used to identify the regions of similarity. Sequence alignment is the process of lining up of sequences to achieve the maximal level of identity (Cohen). In biology, the nucleotide or amino acid sequences may be compared.

There are two important types of sequence alignment namely pairwise alignments and multiple sequence alignments

There is a massive growth in the number of sequences available for comparison. Improvements in the sequence alignment algorithms and use of information technology made the alignment simple, faster and easy.

B. Pairwise Alignment: The comparing of two sequences to look for possible alignment of characters between sequences using the Dot matrix analysis method. This method is also used for finding direct or inverted repeats in Protein and DNA sequences and for predicting regions in RNA that is self - complementary and have the potential of forming secondary structure.

Dynamic programming is a computational method used for pairwise alignment of two Protein or nucleotide sequences. It provides a very good alignment between sequences. This method compares every pair of characters in the two sequences and generates an alignment. This method provides a reliable computational method for aligning DNA and Protein sequences. This algorithm can be used for both local and global types of alignments. This method generates a matrix of numbers that represent all possible alignments between the sequences. Using a distance scoring scheme, dynamic programming method can also be used for an alignment that highlights evolutionary changes.

The word or k- tuple method is used by FASTA and BLAST algorithms. These methods align two sequences, first by searching for identical short stretches of sequences called words or k – tuples and then by joining these words into an alignment by the dynamic programming method. The FASTA and BLAST algorithms are heuristics. FASTA program achieves a high level of sensitivity for similarity searching at high speed. BLAST works under the assumption that high scoring alignments are likely to contain short stretches of identical or near identical letters. FASTA and BLAST are essentially local similarity search methods that concentrate on finding short identical matches that contribute to a total match.

C. Multiple Sequence Alignment: Multiple sequence alignment (MSA) is an extension of the pairwise alignment. Multiple sequence alignment uses two or more sequences. The goal of multiple sequence alignment is to generate a concise, information-rich summary of sequence data in order to make decisions on the relatedness of sequences to a gene family. Multiple sequences of a set of sequences provide information about the most similar regions in the set. These regions in Proteins represent conserved functional or structural domains. There are many methods to carry out multiple sequence alignment such as Profiles, Blocks, Fingerprints, ClustalWetc. ClustalW performs a global sequence alignment. It performs pairwise alignments of all of the sequences and uses the alignment scores to produce a Phylogenetic tree. The quantitative measures of sequence similarity and difference are measured using One is the Hamming distance between two strings of equal length is defined as the number of mismatching characters. The other one is Levenshtein or edit distance between two strings of equal length is defined as the minimal number of edit operations required to change one string into the other.

D. Phylogenetic Analyses: A phylogenetic analysis is intended for the evaluation of sequence relatedness. The field of phylogenetics is meant for working out the relationships among species, populations, individuals or genes. The basic principle is that the origin of similarity is common ancestry. The similarity is the measurement of resemblance or difference. Similarity can be observed from the data collected and involves no historical hypotheses. Phylogeny is a description of biological relationships, usually expressed as a tree. Phylogeny states a topology of the relationships based on classification according to the similarity of one or more sets of characters or on a model of evolutionary processes. In many cases, phylogenetic relations based on different characters are consistent. In Computer Science, a tree is a kind of graph. In phylogenetic trees, the length of the edge signifies some measure of the dissimilarity between two species. Phylogenetics use sequence alignments in the construction and interpretation of Phylogenetic trees.

IV. HOMOLOGY, ORTHOLOGY, AND PARALOGY

Homology is a similarity between characters that is due to their common ancestry. In genetics, homology is measured by comparing protein or DNA sequences and genes that share a high sequence identity. Sequence homology indicates a common function. Shared ancestry may be evolutionary or developmental. In evolutionary ancestry, the structures are evolved from common ancestry, whereas developmental ancestry, the structures are derived from the same tissue in embryonal development. Homology is different from analogy.

More recently, sequencing facilities can create a lot of quality and protein sequences in a brief timeframe, consequently

leading to various complete genomes of organisms that are available today in databases for more in detailed comparative studies. The first step in comparative genomics study is the finding of homologous genes and more specifically orthologous genes. Homologous genes called homologs are obtained from a gene in the last common ancestor. Researcher Fitch in 1970 divided homologs into two categories that are orthologous and paralogous genes. Orthologous genes (orthologs) are homologs that are developed by speciation event in their last common ancestor. Speciation event is the lineage splitting event that creates a separation of species in two or more types. Similarly, paralogous genes (paralogs) are homologs developed by duplication of genes in their last common ancestor.

The finding of orthologs became more important and plays a vital role as it orthologs tends to share similar functions and helps in understanding similarities. Orthologs are of great interest in similarity studies as they can be used in transferring functional annotations which includes protein-protein interaction to newly designed/sequences genomes (22).In various biological and informatics studies, it is very common to search orthology relationship dependent on sequence similarities for the passing of important information from organisms to describe genes that are newly sequences which in turn solves various biological problems. Finding orthologs is of great use and it is a relevant issue in molecular biological studies in the understanding structure of protein and nucleotide sequence and evolutionary conclusions (23).

In Bioinformatics homology among proteins is concluded on the basis of sequence similarity. If the gene sequences of two or more genes are highly similar, it is likely that they are homologous. Sequence similarity may be because of a number of factors like transcription factor and similar by chance. This type of similarity is not homologous. Events like gene fusion may cause partial homology. There is two type of homology of sequences: Orthologous and Paralogous. Homologous sequences are said to be Orthologous if they are separated by speciation. Orthologous genes are such genes that came from a single ancestor with divergent copies in the resulting species; these genes are more or less similar to each other as these are derived from common species. These orthologous genes provide useful information in the taxonomy. Paralogous sequences are obtained by duplication of the gene in homologous sequence. If a gene copied twice or duplicated in any other location within the genome then such copies are said to be paralogous. A set of sequences that are paralogous are called paralogs of each other. Paralogs typically have a similar function. Paralogous sequences provide useful insight into the way genomes evolve. Paralogous sequences often belong to the same species. The same genome is always analogous using multiple homologs. But this does not mean that paralogs will always be restricted to the same genome as evolution progresses. Orthology is derived from gene speciation while paralogy proceeds from duplication of the gene (24).

V. SEQUENTIAL PATTERN MINING

Bioinformatics and sequence mining's are the application and development of data mining techniques to resolve problems by comprehending biological data. Sequence analysis is the greatest primitive operation in sequence mining techniques. Particular studying sequential patterns which are appropriate and different from modern sequence study. Developing retrieved sequences similarity and distance between dissimilar protein sequences can be analyzed (27). Sequence pattern mining is one of the approaches to data mining which is mostly used to recognize different patterns of ordered events (28). This system is useful in finding of sequential patterns that take place in huge databases. This method identifies subsequence mostly in repeated in short intervals and are in sequence as patterns from various sequence databases (29). In the real world, a massive amount of data is needed to be collected continuously and stored in the databases. Many industries mostly Biotech and Pharma are becoming interested in mining sequential patterns from these databases for developing new drugs in fighting various diseases. Most of the methods earlier designed methods use Apriori methods but such methods face challenges when mining data from huge databases. In sequential mining pattern growth methods are highly effective and efficient (30).

In sequential pattern mining, the frequent sequential pattern data obtained through mining has equal significances. Every sequence constitutes a different transaction and the use of each one is different from each other in various fields (31). Sequential pattern mining has so many applications in the field of science for medical sequences such as DNA research analysis and in various other fields like customer shopping behavior identification, etc. In Biotech and Pharma field sequential pattern mining plays a key role in identifying symptoms of patients suffering from different diseases and observing their patterns obtained from various data stored by different researchers across the world and such data can provide useful information in designing medicine for the diagnosis (32).In following generation sequencing, two types of errors may follow experimental and computational. The digital post-processing of sequenced samples, and are the main subject from Computational errors are curtailed. Post-processing involves procedures for instance quality-scoring, aligning, assembling, variant calling, genotyping and error-correction of the data (33). Various sequential patterns are hidden in huge databases and it is a major challenge in sequential pattern mining. Sequential pattern mining has broad applications in various fields different kinds of data available in databases are in the time-related format, for instance, customer purchase behavior stored in the database, a sequence can be used in developing marketing strategies (34). Time

interval sequential pattern mining is presented to mine time interval data between various successive items. Time interval sequential patterns disclose time order of different items but also time interval between two or more successive items (35).

VI. CONCLUSION

This study briefly discusses sequential data mining applications in various domains of studies like bioinformatics, biotechnology, healthcare, web usage mining, etc. Sequential pattern mining methods have been found to be applicable in a large number of domains. Sequential pattern mining methods have been used to analyze data and identify subsequence patterns. Sequence pattern mining helps in tracking down of frequent sub-sequences as patterns from huge databases. Such patterns have been used in scientific experiments and in the treatment of various diseases and mostly in the analysis of DNA sequences, etc. This paper helps in understanding various patterns from extracted data through different data mining tools and using different algorithms. Further study on sequential pattern mining is highly recommended as the growth of the data is increasing every day with new researchers.

REFERENCES

[1]. Kikuchi T (2018). Recent Topics in Protein Folding. J Proteomics Bioinform 11: 075-078. doi: 10.4172/jpb.1000469.

[2]. Gerard G Dumancas, IndraAdrianto, Ghalib Bello and Mikhail Dozmorov(2017). Current Developments in Machine Learning Techniques in Biological Data Mining. Bioinformatics and Biology Insights; 11: 1–4.

[3]. Can T(2014). Introduction to bioinformatics. Methods Mol Biol. 2014; 1107: 51-71.

[4]. Xue Liu, LiliHao, Dayong Li, Lihuang Zhu, SongnianHu(2015).Long Non-coding RNAs and Their Biological Roles in Plants. Genomics, Proteomics and BioinformaticsVolume 13, Issue 3, 1 June 2015, Pages 137-147.

[5]. Perez-Iratxeta C[1], Andrade-Navarro MA, Wren JD(2007). Evolving research trends in bioinformatics. Brief Bioinform. 2007 Mar; 8(2): 88-95.

[6]. Brusic V.(2007). The growth of bioinformatics.Brief Bioinform. Mar; 8(2): 69-70.

[7]. Oliver GR, Hart SN, Klee EW(2015). Bioinformatics for clinical next generation sequencing.Clin Chem. Jan; 61(1): 124-35.

[8]. Ojo OO, OmabeM(2011). Incorporating bioinformatics into biological science education in Nigeria: prospects and challenges. Infect Genet Evol. 2011 Jun; 11(4): 784-7.

[9]. Michael R Clay1 and Kevin E Fisher(2017). Bioinformatics Education in Pathology Training: Current Scope and Future Direction. Cancer Informatics 16: 1–6.

[10]. ArdeshirBayat(2002). Bioinformatics. BMJ. Apr 27; 324(7344): 1018–1022.

[11]. Hameed et al(2018). BMC Bioinformatics 19:129.

[12]. Ibrahim AbakerTargioHashem, IbrarYaqoob, NorBadrulAnuar, SalimahMokhtar, AbdullahGani, SameeUllah Khan. The rise of “big data” on cloud computing: Review and open research issues. Information Systems 47 (2015): 98–115.

[13]. Yueqiang Wang, YueShen, Ying Gu, Shida Zhu, Ye Yin (2018) Genome Writing: Current Progress and Related Applications. Genomics, Proteomics & Bioinformatics. In Press.

[14]. Manuel Mauricio Goez, Maria Constanza Torres-Madroñero, Sarah Röthlisberger, Edilson Delgado-Trejos (2018) Preprocessing of 2-Dimensional Gel Electrophoresis Images Applied to Proteomic Analysis: A Review. Genomics, Proteomics & Bioinformatics. In Press.

[15]. Larissa Catharina, Carlyle Ribeiro Lima, Alexander Franca, Ana Carolina Ramos Guimarães, Marcelo Alves-Ferreira,

[16]. Pierre Tuffery, Philippe Derreumaux, Nicolas Carels(2017). A Computational Methodology to Overcome the Challenges Associated With the Search for Specific Enzyme Targets to Develop Drugs Against Leishmania major.

[17]. Jian Pei (2002) Pattern-growth methods for frequent pattern mining. PhD Thesis. SIMON FRASER UNIVERSITY June 13, 2002.

[18]. Lobo, I. (2008) Basic Local Alignment Search Tool (BLAST). Nature Education 1(1):215.

[19]. Eric S Donkor, Nicholas T K D Dayie, Theophilus K Adiku(2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). Journal of Bioinformatics and Sequence Analysis; 6(1): 1-6.

[20]. William R. Pearson(2016). Finding Protein and Nucleotide Similarities with FASTA. Curr. Protoc. Bioinformatics. 2016 Mar 24; 53: 3.9.1–3.925.

[21]. William R Pearson(2014). FASTA Search Programs. eLS.

[22]. Chuong B. Do and KazutakaKatoh (2008) Protein Multiple Sequence Alignmen. *Methods in Molecular Biology*, 484: Functional Proteomics: Methods and Protocols Edited by: J. D. Thompson et al., DOI: 10.1007/978-1-59745-398-1.

[23]. SoheilJahangiri-Tazehkand, LimsoonWong, ChangizEslahchi. OrthoGNC: A Software for Accurate Identification of Orthologs Based on Gene Neighborhood Conservation. Genomics, Proteomics & Bioinformatics. In Press.Luca Ambrosino1 and 24 . Maria Luisa Chiusano(2017).Transcriptologs: A Transcriptome-Based Approach to Predict Orthology Relationships. Bioinformatics and Biology Insights; 1–8.

[24]. Roy A Jensen(2001). Orthologs and paralogs - we need to get it right. Genome Biol. 2001; 2(8): interactions1002.1–interactions1002.3.

[25]. Khan NT et al (2018).Data Mining – Basics of Bioinformatics. Transcriptomics 6: 142. doi:10.4172/2329-8936.1000142.

[26]. Donovan E (2018) Cloud Based Electronic Health Record Applications are Essential to Expeditionary Patient Care. J ComputSciSystBiol 11: 154-163. doi:10.4172/jcsb.1000265.

[27]. Devarapu R, Murali G, Thota H (2018) Identification of Protein Biomarkers for Diabetic Retinopathy using Sequence Mining Techniques. J Proteomics Bioinform 11: 094-098. doi: 10.4172/jpb.1000472

[28]. Wright AP et al(2014). The use of sequential pattern mining to predict next prescribed medications. J Biomed Inform ,http://dx.doi.org/10.1016/j.jbi.2014.09.003.

[29]. S.Vijayarani, S.Deepa (2013) Sequential Pattern Mining – A Study. International Conference on Research Trends in Computer Technologies (ICRTCT - 2013).

[30]. Jian Ye, Scott McGinnis, and Thomas L. Madden(2006). BLAST: improvements for better sequence analysis. Nucleic Acids Res. Jul 1; 34(Web Server issue): W6–W9.

[31]. Peng Huang (2013) Improved algorithm based on Sequential pattern mining of big data set. DOI: 10.1109/ICSESS.2016.7883028.

[32]. ChetnaKaushal, Harpreet Singh (2015) Comparative Study of Recent Sequential Pattern Mining Algorithms on Web 34 . 34 . Clickstream Data. 2015 IEEE Power, Communication and Information Technology Conference (PCITC) Shiksha 'O' Anusandhan University, Bhubaneswar, India.

[33]. Abnizova I, teBoekhorst R, Orlov Y (2017) Computational Errors and Biases in Short Read Next Generation Sequencing. J Proteomics Bioinform 10: 1-17. doi: 10.4172/jpb.1000420.

[34]. Masseglia, Florent&Teisseire, Maguelonne&Poncelet, Pascal. (2005). Sequential pattern mining: A survey on issues and approaches. Encyclopedia of Data Warehousing and Mining. 10.4018/978-1-59140-557-3.ch193.

[35]. Ya-Han Hu, Fan Wu, Chieh-I Yang (2010) Mining multi-level time-interval sequential patterns in sequence databases. The 2nd International Conference on Software Engineering and Data Mining.