Journal of The Gujarat Research Society

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

A Study on Bioinformatics and Its Applications

Dr Mohd Irfan

SOBAS, Sanskriti University, Mathura, Uttar Pradesh, India

Email Id- irfan.sobas@sanskriti.edu.in

ABSTRACT: ***The genetic foundation of phenotypes has made significant progress in recent years thanks to technological developments, genomics has shifted the paradigm of biological issues on a whole genome-wide scale (genome-wide), unveiling an avalanche of data and bringing up a slew of new options. On the other hand, the massive volume of data created highlights the problems that must be solved in terms of biological data storage and processing. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. Bioinformatics and computational biology have attempted to address these issues in this setting. This study explores methods, new methodologies, and technologies that might give biological significance to data collected, as well as bioinformatics and its usage in the analysis of biological data. In the future, bioinformatics can help in order to extract the accurate data associated with the genome.***

KEYWORDS: ***Data analysis, Databases, Genomics, Systems biology.***

1. INTRODUCTION

The rapid growth of the Internet, as well as its usage in storing and accessing enormous volumes of biological data and the explosion of genetic data, has generated an unprecedented demand for scientists with a thorough understanding of biological sciences and computational methodologies. Bioinformatics has shown to be a powerful instrument for advancing research and development in the multibillion-dollar sector of biotechnology. It is critical for lowering the cost and time required to produce new goods such as medicines, vaccines, plants with specialized characteristics and pest and disease resistance, novel protein molecules, biological materials, diagnostics, and so on. This is significant for the biotech sector since, as a result of the WTO agreement, both goods and methods are now covered by intellectual property rights (IPR). Integration of various databases has become necessary as complete genome sequences, data from microarrays, proteomics, and particular data at the taxonomic level have been available. It can only be done with the help of advanced bioinformatics software. The organization of raw data into appropriate databases and the development of software tools are important issues in which India may play a role. Because of its huge skilled personnel with strengths in Chemistry, Physics, Mathematics, Computer Science, Software Development, Health Care, and Biological Sciences, India has an edge over other nations in this regard.

1.1. Scope of bioinformatics:

To gather, store, organize, archive, analyses, and display biological data, bioinformatics technologies are necessary. Different areas of bioinformatics include genomics and genome analysis, protein informatics, microarray analysis, structural or functional proteomics, lead development, protein marker development, target identification and validation, molecular modelling, Chemo-informatics, new drugs and discovery, analogue based drug design, and traditional drug design. Agricultural science, food preservation, chemical industry, detergents, insecticides, paints, and cosmetic industry are all possible uses. Manpower is required to create algorithms utilizing pure mathematics, statistics, numerical analysis, optimization methods,

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

and artificial intelligence software in programming languages such as C++, ORACLE, Java/Biojava, PERL/BioPERL, and XML/BioXML. Designing databases, managing Gene Banks, Protein Banks, and Banks for Biochemical Reactions; analyzing Biochemical and 3D structural databases; creating mirror sites and project oriented links; studying genomics, proteomics, functional proteomics, motifs, profiles, phylogenesis, human SNP (single nucleotide polymorphism) databases as well as maps, protein modelling and analysis are all tasks that require skilled personnel[1].

1.2.Genetic Informatics:

The sequence of nucleotide bases [adenine (A), guanine (G)] (purines) or [thymine (T) and cytosine (C)] (pyrimidines) of the DNA molecule stores complete information (genetic information) in the living organism. DNA is a lengthy polymer (106 to 107 nm) that is folded into a bundle of strings inside the cell nucleus (chromosomes). Glycosidic linkages bind the bases to the five-membered methylated sugar (deoxyribose) phosphate backbone, passing through the sugar's C1' atom to N9 of purines or N1 of pyrimidines. The DNA molecule is typically made up of two strands that are twisted antiparallel to one other, like a right-handed screw (the orientation of C5' to C3' in two strands is opposite). On the two-strands, complementary bases (A-T) and (G-C) face each other. A pair of hydrogen bonds link them together (Watson and Crick, 1953). Van der Waal's forces stack them on top of each other. This information is 'conserved' during cell division through the process of 'replication,' in which a 'complementary strand' is created over the current or 'template' strand[2].

1.3.Four levels of protein structure:

The proteins are arranged in a hierarchical order. Only an amino acid sequence makes up the ‘primary' structure. Stretches of amino acid chains can form different ‘secondary structures' (-helix, -sheet, c-coil, and t-turn) characterized by a hydrogen-bonded network between the carboxylic as well as amino group of the backbone atoms of the amino acids, depending on the physio-chemical properties of the constituent amino acids. Energy drives the secondary structure, which is stabilized by H-bonds. As a result of interactions between side chains, backbone, and environment, amino acid chains fold into a particular 3-D shape (the protein ‘monomer’s 'tertiary' structure). A large number of these protein "monomers" can clump together to form a "quaternary structure" of the proteins. Specific ‘active' or ‘catalytic' sites are found in both monomeric and multimeric protein/enzyme complexes.

The interaction of a specific ‘activator' or ‘inhibitor' (effector) with the characteristic ‘architecture' and ‘chemical properties of the residues' at the active site can regulate the majority of the body's critical functions. It's simple to see why a minor change in a protein's "active site residue" or a "functional group" of a "effector" may drastically alter an enzyme's activity, but many changes go unreported or only influence the "kinetics" of select processes. The change in ‘free energy' of contact between the two molecules determines whether an enzyme substrate or a protein interacts with another.

1.4. DNA sequence data:

Genes are specific regions of DNA that, when disrupted, cause visible changes in protein quantity, composition, characteristics, or illness. There are, however, a few hundred genes that code for RNAs but do not produce proteins. Similarly, there are sequences that have a direct impact on RNA (transcription) and protein (translation), despite the fact that they do not encode any protein. DNA sequences are used to encode genes. Coding DNA, or exons, are used to

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

make RNA. Introns, or intervening sequences, are spliced out. Amino acid sequences are translated from mature RNA. A fraction of mRNA (shown by striped boxes) is not translated and does not encode any protein. *Whole genome sequencing projects:*

A genome can be sequenced in one of two methods. During the late 1980s and 1990s, a group of researchers from the Human Genome Project (HGP), led first by Nobel Laureate James Watson and then by Francis Collins, Director, National Institute of Health (NIH) Washington, D.C., devised techniques that became known as the BAC-to-BAC approach. The technique is ever evolving and changing. It moves slowly yet steadily. Craig Venter, GNN president, invented the 'SHOT GUN' technology, also known as the 'whole genome sequencing method,' in 1996 while working at the Institute for Genomic Research (TIGR). It brings speed into the picture, enabling researchers to do the job in months to a year[3].

1.5.Protein sequence or structural data:

Proteomics is a branch of research that analyses a cell's complete protein composition at the same time. Understanding how the genome determines phenotypes begins with knowing when and at what amounts genes are expressed. While the quantity of mRNA controls the amount of protein in a cell, it is exposed to post-translational changes that are not detectable by hybridization. To determine the quantity of protein in a cell, a variety of experimental techniques are necessary. Although these approaches are outside the focus of this chapter, they must be listed here in order to analyses their data quality and issues. Molecular biologists have traditionally employed gels to separate various components based on their masses. Different components would move at different speeds across a gel matrix. Each point on a gel map indicates a distinct protein. The procedure is similar to that of analyzing 'DNA Microarrays.' Each spot's resolution and location are then assessed. After that, the site is located, and the sequencing information is utilised to link the location to the gene sequence. The bound protein can be sequenced directly, or the spot can be removed and examined using mass spectrometry techniques such as Electro Spray Ionization-Mass Spectrometry (ESI-MS) and Matrix Assisted Laser Desorption Mass Spectrometry (MALDMS)[4].

1.6. Biological databases:

Digital chain molecules like as DNA and proteins have the ability to be cast in the form of digital symbols such as: Adenine, Guanine, Thymine, as well as Cytosine are nucleotides, while amino acids such as Tyrosine (Y), Glycine (G), Histidine (H), Lys (K), and others are amino acids. As a result, experimentally determined sequences may be produced with absolute confidence in theory. There is no lower limit of uncertainty when it comes to measuring efficiency. The nucleotide sequence in genomic DNA and the related amino acid sequence can be disclosed entirely if we have adequate economical resources[5].

1.7.Sequence comparison methods:

1.7.1. Scoring matrices: Dot matrix, PAM and BLOSSUM:

One can compare two sequences at a time (pair wise sequence alignment) or even a large number of sequences at once (multiple sequence alignment). The primary issue is how to compare two nucleotides or amino acids, as well as how to deal with the "gaps." A 'dot matrix' showing the presence or absence of a specific residue, or scoring matrices, are used in pair wise alignment. PAM (Percentage of Accepted Mutations) matrices based on Dayhoff's mutational data as well as BLOSUM (Block Substitution Matrix) matrices based on observed amino acid substitutions are both common scoring matrices. The BLOSUM matrices are based on a huge

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

data set of about 2000 conserved amino acid patterns known as BLOCKS. These BLOCKS were discovered in a protein sequence database that had over 500 groupings of related proteins.

1.8. Application of Bioinformatics:

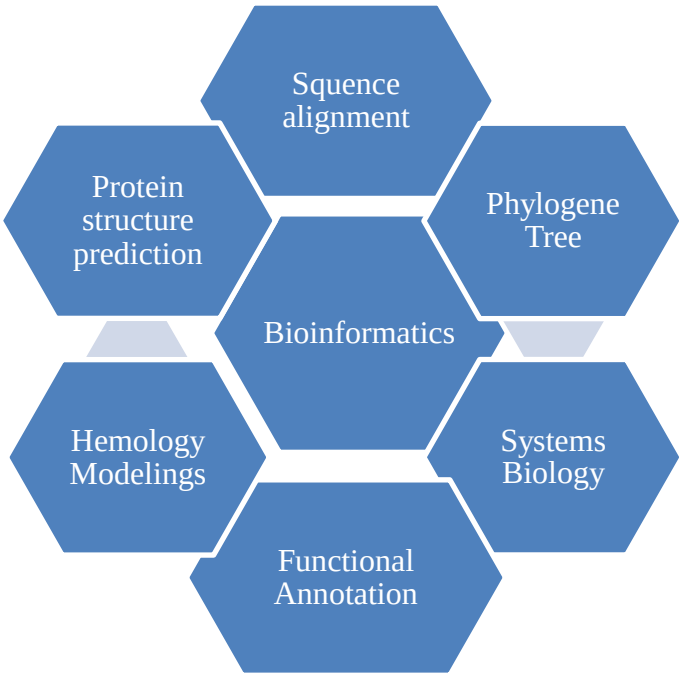

Figure 1: Illustrate diagram showing some major application of bioinformatics.

1.9 Application of bioinformatics to drug designing or chemo-informatics:

Diabetes, cancer, hypertension, and infectious illnesses caused by parasite infections are all serious health risks. Many illnesses are still incurable due to fast resistance to available medicines and significant antigenic diversity. This demands the adoption of cutting-edge drug development approaches. The basic idea of 'rational drug design' is fairly straightforward. It is based on the 'Lock & Key' theory for enzymatic activity proposed by E. Fisher and Paul Ehrlich. Every medication, according to this theory, interacts with a specific 'target' molecule. An enzyme, receptor, circulating messenger, storage location, ion channel, or membrane attached molecule might all be drug targets. It might also be a molecule of DNA or RNA. Drug design may be divided into two categories: analogue based as well as target structure based or rational drug design[6].

1.10 Cheminformatics:

The availability of 3D- parameters in chemical databases is the primary impediment to the widespread use of 3D- QSAR. This is the field of chemoinformatics. It is based on physics, chemistry, coordinate systems and transformations, building molecules as well as polymers, conformational studies on small molecule, quantum chemical calculation on small molecules, molecular mechanics (MM) and molecular dynamics (MD) calculations, or conformational studies on small molecules. There are several programming programs available for this purpose. It's one of bioinformatics' most significant fields.

1.11 Target structure based drug designing:

Journal of The Gujarat Research Society

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

One of the most significant uses in modern drug design is the creation of a novel strong and selective ligand for a certain protein or receptor. The area is also known as "structure-based," "rational," or "de novo" drug design since it makes use of data from the target molecule's 3D structure. In general, there are three fundamentally distinct ways to build a new molecule. The first step is to do a "active site analysis." These approaches are used to figure out which atoms or functional groups are optimal for interacting with the "active site. The third and most frequent method is known as "sticking together parts" (growers and builders). These components might be fragments (functional groups, rings, and so on) or a single atom. The process is referred to as "connecting methods and de novo design.

The Institute for Systems Biology describes systems biology as a holistic approach to understanding the complexity of a system in which "the whole is greater than the sum of its parts." This is an interdisciplinary science that aims to build new technologies and discover new possibilities. a new dimension of data to create new discoveries and ideas, resulting in an innovation cycle (See Figure 2).

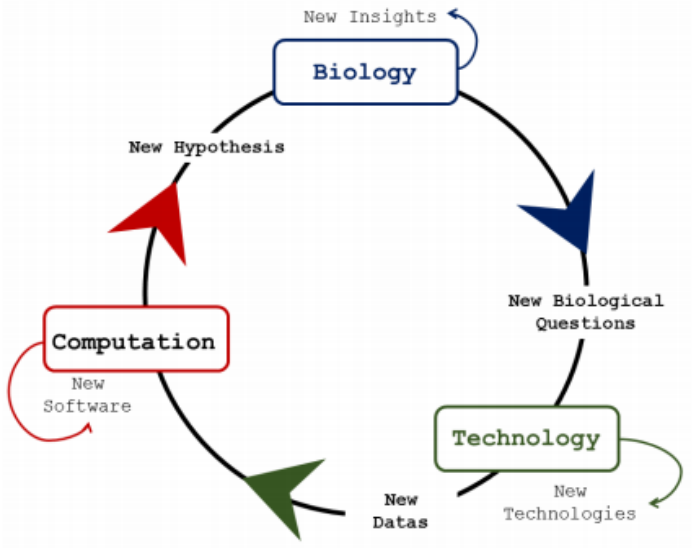

Figure 2: Bioinformatics as an Interdisciplinary Discipline, Systems Biology is Fascinating. Adapted from the Institute for Systems Biology[7]**.**

2. LITERATURE REVIEW

Seung et al. studied about Today's plant research relies heavily on bioinformatics. As the volume of data rises dramatically, so does the demand for data management, visualization, integration, analysis, modelling, and prediction tools or methodologies. At the same time, many biology researchers are inexperienced with bioinformatics methodologies, tools, including databases, which may result in missed opportunities or misunderstanding of data. The major ideas, methodologies, software packages, and databases used in bioinformatics are described in this overview, with an emphasis on those important to plant research. They also go through the basics of biological sequence analysis, transcriptome analysis, computational proteomics, computational metabolomics, bio-ontologies, including biological databases. Finally, they look at a few new bioinformatics research areas[8].

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

Liu et al. discussed about One of the most significant approaches in data mining is discriminative pattern mining. This difficult devour is identifying a group of patterns that appear in disproportionately large numbers in data sets with varied class designations. Such patterns are extremely useful for detecting group differences and building classifiers. In biological data analysis, discriminative pattern mining approaches have shown to be quite useful. Phosphorylation motif finding, differentially expressed gene identification, discriminative genotype pattern recognition, and other bioinformatics applications are common. They give an overview of discriminative pattern mining and the related successful approaches in this article, and then show how they may be used to bioinformatics issues. Finally, they address possible obstacles and future work for this job in general[9].

Anthony et al. studied about Pacific Biosciences' single-molecule, real-time sequencing has larger read lengths than second-generation sequencing, making it ideal for unresolved issues in genome, transcriptome, and epigenetics research. PacBio sequencing can fill gaps in existing reference assemblies and assess structural variation (SV) in personal genomes utilizing highly contiguous denovoassemblies. Due to its capacity to sequence full-length transcripts or fragments of substantial lengths, ac Bio transcriptome sequencing is useful for the identification of gene isoforms and allows the reliable discovery of new genes and novel isoforms of recognized genes. Furthermore, PacBio's sequencing technology offers data that may be used to directly detect base changes like as methylation. Hybrid sequencing techniques are often more cost-effective and scalable, especially for small facilities, than PacBio Sequencing alone. Pac Bio sequencing has made available a wealth of information previously unavailable through SGS alone[10].

DISCUSSION

aim to highlight some of the recent achievements in bioinformatics in the basic fields of sequencing, gene expression, protein, including metabolite studies, databases, and ontologies, as well as present limits and developing areas in these areas, in this review. Data and database integration, automated knowledge extraction, robust inference of phenotype from genotype, and training or retraining of students and experienced researchers in bioinformatics are all outstanding challenges in bioinformatics today. Each scientist will invest more time on the computer as well as the Internet generating and describing data and experiments, analyzing the data and finding other people's data relevant for comparison, finding existing research in the areas and relating it to his or her results into the current body of knowledge, and publishing his or her findings to the world.

CONCLUSION

Advances in data collection and processing skills, as well as the interpretation of outcomes, have pointed to a bright future. However, rapid advancement in all fields of research has resulted in the creation of novel analytical techniques. While we continue to learn more about how the body functions, we should shift our focus from molecular to systemic methods, which has the potential to revolutionize our knowledge of how complex biological systems are regulated. Data integration, on the other hand, is not the end. It's the start of fresh findings and theories, resulting in a feedback loop. Furthermore, significant health advancements will be made, such as the application of genomic technology in gene therapy and customized medicine. This possibility emphasizes the necessity for scientists who are experts in a variety of fields, as well as the effectiveness of interdisciplinary research teams, in which the complementarity of diverse talents will allow for significant scientific advancements.

ISSN: 0374-8588
Volume 22 Issue 1, January 2020

REFERENCES:

[1] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, no. 6, pp. 315–327, 2010, doi: 10.1016/j.ygeno.2010.03.001.

[2] R. Li *et al.*, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Res.*, vol. 20, no. 2, pp. 265–272, 2010, doi: 10.1101/gr.097261.109.

[3] X. Zhou, L. Ren, Q. Meng, Y. Li, Y. Yu, and J. Yu, "The next-generation sequencing technology and application," *Protein Cell*, vol. 1, no. 6, pp. 520–536, 2010, doi: 10.1007/s13238-010-0065-3.

[4] S. Batzoglou *et al.*, "Batzoglou_arachne_2002," pp. 1–13, 2001, doi: 10.1101/gr.208902.7.

[5] A. M. C. Brown and N. S. Willetts, "A physical and genetic map of the IncN plasmid R46," *Plasmid*, vol. 5, no. 2, pp. 188–201, 1981, doi: 10.1016/0147-619X(81)90020-2.

[6] G. Valle, "TIGR Assembler: A new Tool for Assembling Large Shotgun Projects," *Genomics*, vol. 1, no. 1, pp. 9–19, 2012.

[7] W. J. S. Diniz and F. Canduri, "Bioinformatics: An overview and its applications," *Genet. Mol. Res.*, vol. 16, no. 1, pp. 1–21, 2017, doi: 10.4238/gmr16019645.

[8] S. Y. Rhee, J. Dickerson, and D. Xu, "Bioinformatics and its applications in plant biology," *Annu. Rev. Plant Biol.*, vol. 57, pp. 335–360, 2006, doi: 10.1146/annurev.arplant.56.032604.144103.

[9] X. Liu, J. Wu, F. Gu, J. Wang, and Z. He, "Discriminative pattern mining and its applications in bioinformatics," *Brief. Bioinform.*, vol. 16, no. 5, pp. 884–900, 2014, doi: 10.1093/bib/bbu042.

[10] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications," *Genomics, Proteomics Bioinforma.*, vol. 13, no. 5, pp. 278–289, 2015, doi: 10.1016/j.gpb.2015.08.002.