

Chapter 1

Introduction to basics of bioinformatics

Rajesh Kumar Pathak¹, Dev Bukhsh Singh² and Rahul Singh³

¹*School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India,* ²*Department of Biotechnology, Institute of Biosciences and Biotechnology, Chhatrapati Shahu Ji Maharaj University, Kanpur, India,* ³*Department of Basic and Translational Sciences, School of Dental Medicine, University of Pennsylvania, Philadelphia, PA, United States*

1.1 Introduction

Researchers are now constantly making efforts to explore the function of the biological system. Efforts are only at a stage of unprecedented development and growth, expressed in the amount of data produced from each experiment (Avashthi et al., 2014; Kumar, Pathak, Gupta, Gaur, & Pandey, 2015; Mount, 2001). Via bioinformatics, these huge datasets from the experiments are turned into usable information (Mount, 2001; Wang, Zaki, Toivonen, & Shasha, 2005). Bioinformatics is recognized as the science of the 21st century and has tremendous potential for decoding complex biological systems via analysis and integration of multiomics data. It uses information technology (IT) to allow various types of biological data to be analyzed, linked, and manipulated to understand new biological insights (Avashthi et al., 2020; Kumar et al., 2015; Pathak, Taj, Pandey, Arora, & Kumar, 2013).

In other words, bioinformatics is a data management and manipulation method for molecular biology, biochemistry, the health sector, environmental biology, and agriculture, addressing the storage of data sets, data mining and processing, structural and functional annotations of gene and protein, system modeling, and drugs' discovery (Avashthi et al., 2018; Jayaram & Priyanka, 2010; Verma, Pathak, Kasana, & Kumar, 2017). It is used to predict the structure and function of a newly examined protein and protein sequences to create a cluster of similar family sequences and construct phylogenetic trees for the study of evolutionary relationships (Jayaram & Priyanka, 2010; Mount, 2001; Wang et al., 2005). Bioinformatics has a very important role to play in agriculture-dependent countries, where it can be used to boost nutritional content, increase agricultural produce yields, and implant resistance to many biotic and abiotic stresses (Jayaram & Priyanka, 2010; Meidanis, 2003; Pathak, Giri, Taj, & Kumar, 2013).

For the agricultural science sector, plant and animal genome sequencing should have an enormous yield. In both the integration and analysis of genomics, transcriptomics, and other high-throughput sequencing results, bioinformatics plays a vital role, with great potential in redesigning to boost productivity. To understand the function and interaction of many genes, there has been a paradigm change from the single-gene approach (i.e., gene-by-gene approach) (Kumar et al., 2015). This change has resulted from the discovery that cross-talking of several biomolecules acting in an interdependent manner and results in any biological answer. As a consequence, several high-throughput technologies have been developed that offer insight into all the molecules involved in a process (Kumar et al., 2015). Study in the field of genomics has accelerated the process. There is, however, a large difference in the expression of a trait between the genotype and the phenotype (Kumar et al., 2015; Pathak & Singh, 2020b). Studies are performed at various levels to fill this gap: the whole system, organism, biochemical, gene, and protein levels. All these fields have contributed to the collection of vast amounts of biological knowledge due to unprecedented research efforts. Bioinformatics, which culminates in biology and computational technology, aims to develop novel strategies for wide-scale analysis of biological system (Pathak & Singh, 2020a).

Bioinformatics techniques, such as simulation, docking, protein–protein interaction, and analysis of next-generation sequencing (NGS) data, may be used to investigate or modify the sequence for better fitting of essential genes for a particular function and to study the function of these genes or proteins at the system level. It was then possible to use this specified genetic, genomic, and proteomic information to grow resistant, nutritionally improved, and profitable crops and also discover therapeutic drugs (Agnihotry, Pathak, Srivastav, Shukla, & Gautam, 2020; Singh & Pathak, 2020). Some important tools and databases along with their application and link of availability are highlighted in Tables 1.1 and 1.2.

TABLE 1.1 List of important and popular database resources for bioinformatics.

S. no.	Database	Application	Availability	References
1.	National Centre for Biotechnology Information (NCBI)	It offers access to biomedical and genomic information to boost research activity.	https://www.ncbi.nlm.nih.gov/	Benson, Boguski, Lipman, and Ostell (1990)
2.	GenBank	It is a nucleotide database available at NCBI. It is used for the retrieval of nucleotide sequences.	https://www.ncbi.nlm.nih.gov/genbank/	Benson et al. (2012)
3.	European Nucleotide Archive	It is a nucleotide database available at EBI-EMBL. It provides a detailed record of nucleotide sequencing data, covering raw sequencing data, information about sequence assembly, and functional annotation.	https://www.ebi.ac.uk/ena/browser/home	Leinonen, Sugawara, and Shumway (2010)
4.	DDBJ	It is a nucleotide database available at NIG, Japan. It is used for the retrieval of nucleotide sequences.	https://www.ddbj.nig.ac.jp/index-e.html	Tateno et al. (2002)
5.	Protein Information Resource (PIR)	PIR is an integrated database to support research and scientific studies in genomics, proteomics, and systems biology.	https://proteininformationresource.org/	Wu et al. (2003)
6.	UniProt	UniProt aims to provide a complete, high-quality, and freely available protein sequence and functional information resource to the scientific community.	https://www.uniprot.org/	Apweiler et al. (2004)
7.	Pfam	It is a broad set of protein families database used for domain analysis.	https://pfam.xfam.org/	Bateman et al. (2002)
8.	CATH (Class, Architecture, Topology, and Homology)	The CATH database offers information on the evolutionary relationships of protein domains as a free, publicly accessible online database resource.	https://www.cathdb.info/	Orengo et al. (1997)
9.	SCOP	The purpose of the SCOP database is to provide a detailed and systematic explanation of the structural and evolutionary relationships between proteins deposited in the RCSB Protein Data Bank.	http://scop.mrc-lmb.cam.ac.uk/	Murzin, Brenner, Hubbard, and Chothia (1995)
10.	Protein Data Bank (PDB)	It is a structural database that contains the 3D structure of macromolecules. It is crucial for research in the area of structural bioinformatics, drug discovery and protein structure prediction, etc.	https://www.rcsb.org/	Berman et al. (2000)
11.	PubChem	It is the largest database of chemical information. Here, we search chemical molecules and retrieve their structure for molecular docking/virtual screening.	https://pubchem.ncbi.nlm.nih.gov/	Bolton, Wang, Thiessen, and Bryant (2008)
12.	ZINC	It is a database that contains commercially available molecules for computational screening. We can search and retrieve analogs for any molecule based on similarity. It plays a key role in the discovery of lead molecules for drug development.	https://zinc.docking.org/	Sterling and Irwin (2015)

(Continued)

TABLE 1.1 (Continued)

S. no.	Database	Application	Availability	References
13.	GEO	GEO is a database for functional genomics. It contains gene expression/microarray data. Here, we can download gene expression profiles submitted by the scientific community throughout the world for further investigation as per need.	https://www.ncbi.nlm.nih.gov/geo/	Edgar, Domrachev, and Lash (2002)
14.	Sequence Read Archive	It is the largest publicly available repository of high-throughput sequencing data. Here, we can download NGS data submitted by the scientific community throughout the world for further investigation as per need.	https://www.ncbi.nlm.nih.gov/sra	Leinonen et al. (2010)
15.	The Arabidopsis Information Resource	It is a database of the model plant <i>Arabidopsis thaliana</i> provides genetic and genomic information to the scientific community.	https://www.arabidopsis.org/	Rhee et al. (2003)
16.	Rice Genome Annotation Project	It provides sequence and annotation data for the rice genome.	http://rice.plantbiology.msu.edu/	Kawahara et al. (2013)
17.	Gramene	It is a curated, open-source, integrated database platform for research in the area of comparative functional genomics of crop plant species.	https://www.gramene.org/	Ware et al. (2002)
18.	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Database of gene/genome sequencing information for the understanding function of biological system	https://www.genome.jp/kegg/	Kanehisa and Goto (2000)
19.	Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	It is a database resource for known and predicted protein–protein interactions derived from experimental, computational methods, and text mining.	https://string-db.org/	Szklarczyk et al. (2019)
20.	BioModel	It is a repository of mathematical models. It provides a wide variety of current physiologically and pharmaceutically applicable mechanistic models based on literature in standard file formats.	https://www.ebi.ac.uk/biomodels/	Le Novere et al. (2006)

1.1.1 Concept behind bioinformatics, in silico biology, and computational biology

There are many definitions available for bioinformatics in different books and the Internet. We can say that “Decoding problems arising in the field of biological sciences *via* computation” is known as bioinformatics. It handles biological data obtained through several experimental techniques for their documentation in the form of databases for further availability to the scientific community. This dataset is analyzed for dissecting the complexity of biological systems, and acquired knowledge accelerates research activity and helps scientists to get fruitful results and novel insight. The term “in silico biology” is concerned with bioinformatics. Generally, we can say that performing any work using a computer for biology is known as in silico biology, for example, retrieval of biological sequences from databases.

Computational biology is different from bioinformatics and in silico biology. Here, we are talking about the development of algorithms and theoretical methods for solving biological problems. These developed algorithms and methods are further utilized in the design and development of biological software or tools that help in analyzing biological data using the computer to create knowledge, that is, bioinformatics. Therefore these disciplines are linked to each other and use several disciplines of biology and natural sciences along with the computer, IT, mathematics, and statistics. That is why it is called interdisciplinary science.

TABLE 1.2 A list of commonly used bioinformatics software.

S. no.	Software/tool	Application	Availability	References
1.	Basic Local Alignment Search Tool (BLAST)	It is a very common tool in bioinformatics used for similarity searching and identification of homologs and paralogous sequences.	https://blast.ncbi.nlm.nih.gov/Blast.cgi	Altschul, Gish, Miller, Myers, and Lipman (1990)
2.	CLUSTAL	It is a widely used program for multiple sequence alignment. Many versions of the Clustal program are available for sequence analysis.	http://www.clustal.org/omega/	Chenna et al. (2003) and Sievers et al. (2011)
3.	Molecular Evolutionary Genetics Analysis (MEGA)	It is a software program for performing statistical analysis of molecular evolution and for building phylogenetic trees.	https://www.megasoftware.net/	Kumar, Tamura, and Nei (1994)
4.	Modeller	It is a well-known program for modeling protein 3D structure based on sequence information.	https://salilab.org/modeller/	Eswar et al. (2006)
5.	PyMOL	PyMOL is a comprehensive software package for the visualization and animation of 3D structures.	https://pymol.org/2/	DeLano (2002)
6.	Swiss PDB Viewer	Swiss PDB Viewer is a bioinformatics program that offers a user-friendly interface for the simultaneous study of many proteins. It is used for structural alignments and comparison, calculation of H-bonds, angles, and distances between atoms.	https://spdbv.vital-it.ch/	Guex and Peitsch (1997)
7.	Chimera	It is an interactive visualization and analysis software for molecular structures and related data, including maps of density, trajectories, and alignments of sequences. It is possible to create high-quality images and animations.	https://www.cgl.ucsf.edu/chimera/	Pettersen et al. (2004)
8.	MarvinSketch	It is used to draw, edit, import, and export chemical structures, and also for the conversion of structural file formats.	https://chemaxon.com/products/marvin	Bunin, Siesel, Morales, and Bajorath (2007)
9.	AutoDock	It is a computer program used for the identification of lead compounds through molecular docking.	http://autodock.scripps.edu/	Goodsell, Morris, and Olson (1996)
10.	Gromacs	It is a molecular dynamics package developed specifically for protein, lipid, and nucleic acid simulations.	http://www.gromacs.org/	Van Der Spoel et al. (2005)
11.	Velvet	It is a de novo based genome assembly program developed especially for short-read sequencing data.	https://www.ebi.ac.uk/~zerbino/velvet/	Zerbino and Birney (2008)
12.	Trinity	It is a transcriptome/RNA-seq data assembly program generated by the Illumina NGS platform based on the de novo algorithm.	https://github.com/trinityrnaseq/trinityrnaseq/wiki	Grabherr et al. (2011)
13.	FastQC	It provides a quick way to check the quality control of raw sequencing data generated from NGS platforms.	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	Andrews (2010)
14.	Trimmomatic	It is a widely used flexible read trimming tool for NGS data coming from Illumina.	http://www.usadellab.org/cms/?page=trimmomatic	Bolger, Lohse, and Usadel (2014)
15.	Cutadapt	It detects and removes adapter sequences, poly-A tails, primer sequences, and other regions from sequencing reads.	https://cutadapt.readthedocs.io/en/stable/	Martin (2011)
16.	DIAMOND	It is a sequence aligner for protein or translated DNA searches, developed for big sequence data analysis. It is very fast as compare to BLAST.	http://www.diamondsearch.org/index.php	Buchfink, Xie, and Huson (2015)

(Continued)

TABLE 1.2 (Continued)

S. no.	Software/tool	Application	Availability	References
17.	Blast2GO	Used for gene ontology, and annotation of genomic data.	https://www.blast2go.com/	Conesa et al. (2005)
18.	EffectorP	It is a tool used for the prediction of fungal effector proteins. It is trained in plant-pathogenic fungi to differentiate secreted proteins from secreted effectors.	http://effectorp.csiro.au/	Sperschneider et al. (2016)
19.	CellDesigner	It is a computational systems biology program used for pathway modeling and simulation analysis.	http://celldesigner.org/	Funahashi, Morohashi, Kitano, and Tanimura (2003)
20.	Cytoscape	It is a widely used and well-known software platform in the field of complex network science and network biology for visualization and analysis. Different types of Cytoscape Apps are available for different types of analysis.	https://cytoscape.org/	Shannon et al. (2003)

NGS, Next-generation sequencing.

1.1.2 Scientific discipline and support systems for bioinformatics

Bioinformatics is a central discipline that is linked with several disciplines of science. These disciplines of science and technology provide infrastructure and interdisciplinary nature to bioinformatics. In the sciences discipline, several traditional and advanced subjects are associated with bioinformatics, such as plant sciences, animal sciences, molecular biology, genetics, and evolutionary biology, pharmaceuticals, mathematical, and statistical sciences, and omics. Besides, in the support system, bioinformatics is closely associated with computer science, IT, and computational resources because they work as a backbone for bioinformatics. Therefore these scientific disciplines and support systems are combined to build a new discipline called bioinformatics for decoding intricate problems linked to biological systems via innovative manner with fast processing due to the involvement of computer science and IT (Fig. 1.1).

1.1.3 Needs of bioinformatics

Bioinformatics is a need of time because due to advances in several omics platforms have produced big data in biology related to genomes, transcriptomes, genotyping by sequencing, proteome, metabolome, etc. The management and analysis of these high-throughput data need bioinformatics. In the era of omics and availability of organism and discipline/subject-specific data, that is, genomic data, proteomic data, etc., there is a need to develop specific databases for their documentation. This will further accelerate the research and help scientists to integrate these available data with novel discoveries. Besides, these datasets will help in improving the accuracy of available tools and enable scientists to train data for the development of new tools via machine and deep learning approaches.

Thus in view of the above facts and contiguous process of experimental data generation needs bioinformatics professionals to manage and analyze these big data of biology to boosting research and development in a smart way for mining useful information. This will differentially help society.

1.2 Historical background of bioinformatics

The term “bioinformatics” was primarily used from the late 1980s onward to refer to computational analysis of genome data and described more broadly as the “study of informatic processes in biotic systems” (Hogeweg, 2011). Paulien Hogeweg and Ben Hesper first used the term “bioinformatics” at the beginning of the 1970s, and the term was published by him in 1978 (Hogeweg, 1978, 2011; Hogeweg & Hesper, 1978).

More than 50 years ago, bioinformatics took place when desktop computers and deoxyribonucleic acid (DNA) sequencing were a dream. Not much was known about deoxyribonucleic acid (DNA) in the early 1950s. At that moment, its role as the carrier molecule of genetic information was still controversial (Gauthier, Vincent, Charette, &

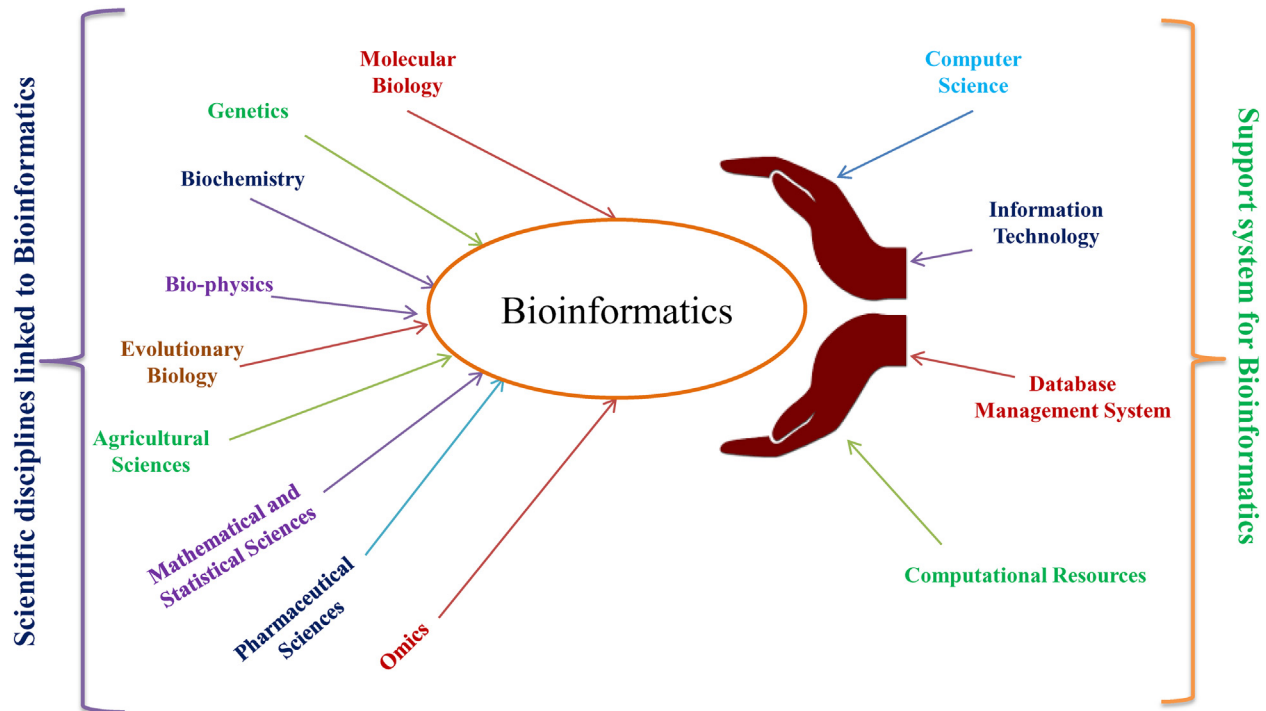


FIGURE 1.1 Scientific disciplines associated with bioinformatics and their support systems.

Derome, 2019). Avery, Colin, and McCarty (1944) demonstrated that the absorption of pure DNA from a virulent bacterial strain could give a nonvirulent strain virulence, but their findings were not immediately recognized by the scientific community. Many assumed that proteins were worked as genetic information carriers. Hershey and Chase confirmed the role of DNA as a genetic information encoding molecule in 1952 when they proved beyond a reasonable doubt that it was DNA, not protein, which was absorbed and transmitted by bacterial cells infected by a bacteriophage (Gauthier et al., 2019; Hershey & Chase, 1952). So, the information about the function of DNA is known but its structure was not determined. In 1953 Watson, Crick, and Franklin finally resolved the double-helix structure of DNA. Despite this major discovery, it would take 13 more years before the genetic code was deciphered and 25 more years before the first methods for DNA sequencing were available. Consequently, in DNA analysis, the use of bioinformatics lagged almost two decades behind the analysis of proteins whose chemical nature was already better understood than that of DNA (Maxam & Gilbert, 1977; Nirenberg & Leder, 1964; Sanger, Nicklen, & Coulson, 1977; Tamm, Shapiro, & Lipshitz, 1953; Watson & Crick, 1953). Therefore the analysis of protein was the starting point in bioinformatics (Gauthier et al., 2019).

The major progress in solving protein structure through crystallography during the late 1950s, the first sequence of a protein, insulin, was published. Furthermore, many advancements in the determination of protein sequence and their structure were reported. Margaret Dayhoff (1925–1983), the first bioinformatician, was an American physical chemist, known for her significant contribution and application of computational methods in the field of biochemistry and protein sciences. Therefore she is known as the mother and father of bioinformatics (Jaskolski, Dauter, & Wlodawer, 2014; Moody, 2004; Sanger & Thompson, 1953).

The first dynamic programming algorithm for pairwise alignments of protein sequence was developed by Needleman and Wunsch in 1978 and the first multiple sequence alignment (MSA) algorithms emerged in the 1980s. The software CLUSTAL was introduced in 1988 for MSA. Then, the concept of a mathematical framework for amino acid substitutions was introduced with the development of a point accepted mutation matrix by Dayhoff. In 1970–80 the paradigm has been shifted from protein to DNA analysis. Between 1980 and 1990 the parallel advances in biology and computer science occurred. In the establishment of the National Centre for Biotechnology Information (NCBI) database in 1988 and at NIH and launched of Human Genome Project in 1990, bioinformatics gains huge importance due to its application in bioinformatics for the management and analysis of biological data, which lead to the development of genomics and structural bioinformatics between 1990 and 2000. In 2000–10 the concepts of high-throughput

bioinformatics were introduced and gain a lot of attention by the scientific community in this field and their importance in biology (Dayhoff, Schwartz, & Orcutt, 1978; Gauthier et al., 2019; Murata, Richardson, & Sussman, 1985; Needleman & Wunsch, 1970; Sievers & Higgins, 2014).

Yoshizumi Ishino first defined the signature repeat spacer architecture of CRISPR arrays in 1987 (Ishino, Shinagawa, Makino, Amemura, & Nakata, 1987). Francisco Mojica later revealed, backed by bioinformatics analyses, that CRISPR arrays were present not only in *Escherichia coli* but also in most archaeal and several bacterial genomes (Mojica, Díez-Villaseñor, Soria, & Juez, 2000). Subsequent studies of bioinformatics showed bacteriophageal spacer matches, leading to the right conclusion that CRISPR-Cas systems function as an acquired immune system. Furthermore, another bioinformatics analysis on spacer matches fruitfully investigated that CRISPR-Cas systems will mostly target DNA rather than RNA (Mojica, García-Martínez, & Soria, 2005). It was an excellent breakthrough in the area of genome editing (Alkhnabashi, Meier, Mitrofanov, Backofen, & Voß, 2020). Bioinformatics has a very big history from 1950 to yet, where many tools and databases were introduced, they have tremendous potential for solving complex biological problems and data management. Therefore it is not possible to cover the complete history of bioinformatics in a single chapter.

Nowadays, due to advances in high-throughput technology, in terms of data mining and management, the advent of “Big Data” has raised new problems, calling for more computer science skills in the area. Biological Big Data has (and continues to have) significant consequences for the predictive capacity and reproducibility of bioinformatics outcomes coupled with an ever-growing number of bioinformatics tools. Many universities have included bioinformatics in their course curriculum to address this issue. The ever-increasing complementarities between computer science and biology have led to recent subdisciplines, such as synthetic biology, systems biology, and whole-cell modeling (Gauthier et al., 2019).

1.3 Aim of bioinformatics

As we know that bioinformatics is a need of time in the era of omics, you do not think about research without bioinformatics or it is necessary for finding new information. The aim of bioinformatics-related projects is generally

- Documentation of useful information about medicinal plants available in literature/public domains in the form of a database.
- Documentation of available information about plants and animal genetic resources.
- Development of useful organism/species-based databases for the documentation of omics data.
- Improvement of the contents and usefulness of already developed/available databases.
- Development of better and fast graphical user interface (GUI)-based tools for data integration and analysis.
- Improvement of available tool/software for biologist/wet lab scientists/noncomputer background scientists for their effortless use.
- Development of platform-independent software's for the computational analysis of biological data.

1.4 The recent development in the field of bioinformatics

To understand the code of life that is DNA, bioinformatics has developed and contributed to the growth of research projects associated with high-throughput DNA sequencing and other fields of biology. Its ultimate aim is to visualize the wealth of biological knowledge hidden in sequences, structures, literature, and other big data of biology. In the era of omics, the use of computational tools becomes necessary for the investigation of useful information from large sets of biological data. The analysis of these data required good bioinformatics skills and most biologists are not familiar with command-line tools, bioinformatics analysis, and interpretation. Therefore much collaboration has been established among experimental biologists and bioinformatics groups to make their data meaningful. Now in recent years, several GUI-based software platforms have been developed that can easily understandable for biologists to integrate and analyze high-throughput omics data.

Bioinformatics has evolved into a full stream of multiple opportunities for researchers from different areas, such as medicine, biotechnology, plant breeding, chemistry, statistics, mathematics, computer sciences, and IT. The main motto of this field is to turn data into knowledge through computation. With the advancement of many other significant research areas, such as systems biology, synthetic biology, nanotechnology, and pharmacogenomics, it is still growing. Due to advances in technology and data accumulation at an unprecedented rate, bioinformaticians are expected to be in

constant demand to decode the complexity of data for addressing the biological problem (Moussa, Vannier, Yennamalli, & Singh, 2016; Pathak, Baunthiyal, Pandey, & Kumar, 2020).

1.5 Challenges in bioinformatics

Different disciplines of science and technology are associated with bioinformatics. This association is also a challenge to understand the subject due to its interdisciplinary nature for the peoples in different backgrounds, for example, people with biology background generally facing problem with computer and people from computer science background facing some problem with the concept of biology. Therefore bioinformatics gained importance in recent years as compulsory courses for many postgraduate programs run by universities in different disciplines of science and technology to trained students in this area. The major challenges in the area of bioinformatics faced by scientists from biology background are:

- Design of small drug-like molecule,
- Development of a quantitative model for signal transduction pathway,
- Accurate prediction of protein secondary and tertiary structure using amino acid sequence,
- Understanding the evolution of protein,
- Understanding the speciation event,
- Understanding gene ontology,

Besides, there are many challenges faced by computer scientists working in the area of bioinformatics, such as:

- Management of big data in biology,
- Information management,
- Improve the accuracy of software, and
- Programmability.

Therefore it is an urgent need to develop manpower in this emerging field, they know about biology and computer along with associated disciplines/subjects for solving complex problems in biology (Meidanis, 2003; Moussa et al., 2016). This will ultimately accelerate the outcome of the research projects but developing interest in computer science/programming to the biology students and in biology to the computer science/mathematics students is also a common challenge right now. Due to dependency on the computer for every work, it may be solved by the near future.

1.6 Application of bioinformatics

Bioinformatics has great potential to solve complex problems in biology and associated disciplines. With the start of the human genome project, it becomes an essential tool for scientists and playing a key role in the area of life sciences, chemical sciences, physical sciences, agricultural sciences, and medical sciences (Fig. 1.2). Some important applications of bioinformatics are discussed in the following sections.

1.6.1 Sequence analysis

Sequence analysis is one of the major applications of bioinformatics with the development of the Basic Local Alignment Search Tool (BLAST) program in 1990 and has become popular. The area of sequence analysis is very broad; here, we analyze the nucleotide or protein sequence of any organism for several purposes. Here, we can analyze single or multiple sequences to find out similarity and identity among them via several sequence alignment tools, such as BLAST and FASTA Clustal. It is also used in the annotation of newly discovered sequences, find out conserved regions, and other regulatory regions among them. Besides, the prediction of the physicochemical properties of sequences also comes under the sequence analysis (Chinchole, Pathak, Singh, & Kumar, 2017; Sidhu, Bhangu, Pathak, Yadav, & Chhuneja, 2020).

1.6.2 Phylogenetic analysis

Phylogenetic analysis is another important research area in bioinformatics. Here, we can visualize the evolutionary event and construct a relationship among organisms or sequences. It is extensively used in evolutionary biology for the objective of determining the evolutionary event via multiple sequence alignment followed by tree construction. It also supports the identification of key regions within sequences and plays an important role in vaccine and

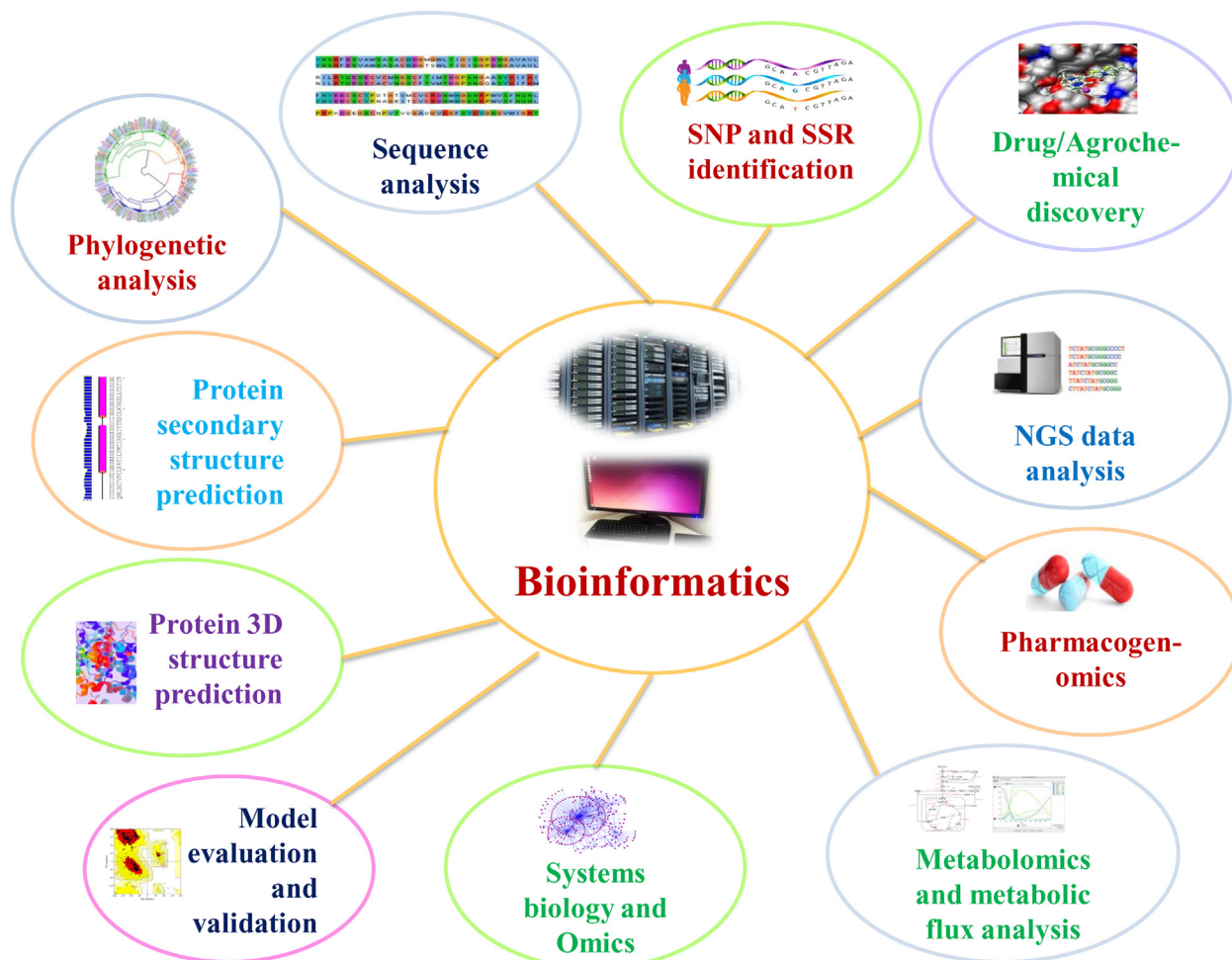


FIGURE 1.2 Application of bioinformatics in different areas of biological, agricultural, and medical sciences.

drug designing programs etc. (Allaby & Woodwark, 2004; Chinchole et al., 2017; Gupta et al., 2018; Mount, 2001).

1.6.3 Prediction of protein secondary structure

Right now, three computational methods, that is, Chau Fasman, GOR, and PHD, are available for protein secondary structure prediction. As it is a key area of the structure in the field of structural bioinformatics, various computational algorithms have been developed and used. Artificial Neural Network and Hidden Markov Model are the commonly used methods for the different available software. The secondary structure prediction methods are generally used to determine the number of amino acid residues involved in the formation of different types of secondary structures, that is, which or how many amino acid residues are involved in the formation of the helix, strand, coil, and turn. It provides support in the building of the 3D structure of a protein based on secondary structure information when a suitable structural homolog is not available in the protein data bank (Pathak, Singh, Sagar, Baunthiyal, & Kumar, 2020; Sidhu et al., 2020).

1.6.4 Protein 3D structure prediction

Prediction of protein 3D structure from sequence information is a challenging task in bioinformatics for the quality of the model. Right now three computational approaches are used, that is, homology or comparative modeling, threading or fold-recognition, and ab initio. It plays a crucial role when the experimentally determined structure of the target

protein is not available in the PDB database. It is used to determine the structure of newly discovered proteins or other proteins their structures are not solved by NMR or X-ray crystallography for ligand small molecule screening for identification of lead molecules for drugs/agrochemicals development. Besides, modeled structures are utilized for predicting protein–protein interactions, structural comparison, alignment, etc., for new insight via computation (Gupta, Gaur, Pathak, Gupta, & Kumar, 2014; Gupta et al., 2018; Singh & Pathak, 2020).

1.6.5 Evaluation and validation of predicted protein model

Various computational methods are developed by bioinformatics scientists for the quality analysis and validation of the predicted protein 3D model. The most popular program used for evaluation and model validation is Swiss PDB viewer, structural analysis and verification server (SAVES), Rampage, and PROCHECK. We can evaluate and refined our predicted model with the help of the above tools for further investigation (Mamgain, Sharma, Pathak, & Baunthiyal, 2015; Sidhu et al., 2020).

1.6.6 Discovery and designing of small molecules leading to drugs/agrochemical development

Bioinformatics plays a vital role in the discovery of lead molecules. Techniques, such as molecular docking, virtual screening, and molecular dynamics simulation, are widely used for this purpose (Sagar et al., 2014). With the help of molecular docking and virtual screening, we can identify novel molecules for the treatment and prevention of diseases in humans, animals, and plants (Pathak et al., 2016). Furthermore, it can be validated in terms of conformational behavior and stability during target–ligand interaction with respect to time using a molecular dynamics simulation study (Rana, Pathak, Shukla, & Baunthiyal, 2020). This study helps in the identification of lead compound quickly with the help of computational tools and reduced the time and experimental cost, the fast drug discovery process for downstream validation (Mamgain, Dhiman, Pathak, & Baunthiyal, 2018; Pathak & Sagar, 2012; Pathak, Baunthiyal, Taj, & Kumar, 2014; Pathak, Baunthiyal, Shukla, et al., 2017; Rai, Pathak, Singh, Bhatt, & Baunthiyal, 2021; Singh, Gupta, & Pathak, 2020).

1.6.7 Next-generation sequencing data analysis

Due to progress and recent development in several omics platforms and various sequencing technologies, a big amount of data has been generated every day that need bioinformatics for their analysis and management. In recent years a bunch of tools and databases has been developed for the management, integration, and analysis of such big data. These tools are useful in the identification of differentially expressed genes via analysis of microarray and RNAseq/transcriptome data, assembly and annotation of genome and transcriptome, SNP discovery, genome-wide association study, identification and characterization of gene families, the discovery of new genes, etc. The obtained information from big data analysis will help in different research programs associated with plant and animal breeding, marker development, drug target identification, and drug discovery (Magi et al., 2010; Pathak, Baunthiyal, et al., 2020; Pereira, Oliveira, & Sousa, 2020).

1.6.8 SNP and SSR identification

Progress in NGS and bioinformatics has enabled the large-scale discovery of SNP and SSR markers. It has revolutionized research related to genomics and assists in the molecular breeding program, and it will also help in analyzing genetic diversity and population structure, designing genetic maps of high density, and providing genotypes for genome-wide association research.

1.6.9 Pharmacogenomics

In the postgenomic era, pharmacogenomics is recognized as one of the keys and highly demanded research areas to study the response of a drug based on genetic variants for the development of personalized medicine. As we know that some drugs are working on a particular population and or group of peoples but no desired response of these drugs was reported in other populations or countries due to minute change in genomic regions. So that the concept of pharmacogenomics has been introduced for the development of a person or population or area-specific medicine to reduce the risk of side effect and increase their potency. Therefore bioinformatics tools are rapidly utilized for the analysis of

high-throughput genomics data for the detection of how genes alter the response of a drug; this investigation will lead to the development of personalized medicine (Aneesh, Sekhar, Jose, Chandran, & Zachariah, 2009; Takahashi, Luzum, Nicol, & Jacobson, 2020).

1.6.10 Metabolomics and metabolic flux analysis

A quantitative and qualitative understanding of metabolic rates or fluxes over a metabolic network and in particular cellular compartments offers insights into metabolic regulation and helps to understand the contribution of metabolic changes to pathology. Bioinformatics facilitates researchers to analyze metabolomics data and develop a metabolic model for flux balance analysis. This will help in the metabolic engineering program for various purposes.

1.6.11 Systems biology and omics data integration

Nowadays systems biology is one of the essential areas in bioinformatics and has outstanding contributions in decoding the complexity of biological systems via modeling, simulation, and network analysis. It helps us to integrate different omics data, that is, genomics, proteomics, metabolomics, transcriptomics, and other omics, for the development of a model that will describe the behavior of biological systems in different conditions holistically with respect to time (Pathak, Baunthiyal, Pandey, Pandey, & Kumar, 2017).

In past years scientists were working on a single gene or protein, but in recent years the paradigm has been a shift from a reductionist approach to a holistic approach due to the augmentation of omics technology and systems biology. So, with the help of systems biology, we can predict the behavior of whole systems and identify key components involved in different biological processes through network modeling and analysis. The systems-based approaches are highly useful in investigations of key genes/proteins, drug targets, modeling of pharmacokinetic and pharmacodynamic activity of drug molecules at particular concentration/amount with respect to time computationally before going to the experimental studies for evaluating the drug-likeness of molecules. Besides, it is useful in the modeling and simulation of host–pathogen interaction to identify key components involved in disease progression as well as resistance for the development of suitable disease resistance strategy and their implementation (Kokane, Pathak, Singh, & Kumar, 2018; Kumar et al., 2018; Pathak, Baunthiyal, Pandey, Kumar, 2018; Pathak, Gupta, Shukla, Baunthiyal, 2018).

1.7 Future perspective

Bioinformatics got a lot of attention from the time of the human genome project and has become an essential tool for research and development right now. Without the use of bioinformatics, we cannot think about the novel discovery in the 21st century. In the current era of omics, many leading research groups around the world imagine that bioinformatics is the future of next-generation science and technology because everywhere most of the work has been conducted by computer software for speed and accuracy. Therefore in the future, it becomes an essential part of every research lab. Near the future, lots of bioinformaticians are required for handling big data in biology for their management and mining of key information for further research, which will directly decrease experimental work and cost. As we know that the world population is increasing rapidly, more and more food is required to meet the need of peoples. In the future, the development of high-yielding crop varieties with nutritional value and disease-resistant crop varieties is needed, and it is possible by the use of bioinformatics to dissect the complexity of crop plants systems for better productivity. However, it will play a major role in the fundamental and advance research in the area of health, biomedical, and veterinary sciences for the discovery of medicines, therapeutics, and management of diseases for a better future for mankind.

1.8 Conclusion

In recent years bioinformatics has become an essential subject and hot research topic that is associated with several discipline and methodology. The present chapter highlights the power of bioinformatics and its approaches for solving a complex biological problem, which leads to research and development. Besides, it presents several important database resources and tools along with their uses and availability. In the future, rapid advancement in the generation of big data through omics and other advancements in different areas are expected in which bioinformatics provide a broad range of applications for their management, analysis, and novel discovery. Furthermore, this chapter provides career opportunities to students and researchers in the area of computational and integrated science, that is, bioinformatics.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Agnihotry, S., Pathak, R. K., Srivastav, A., Shukla, P. K., & Gautam, B. (2020). *Molecular docking and structure-based drug design. Computer-aided drug design* (pp. 115–131). Singapore: Springer.
- Alkhnabshi, O. S., Meier, T., Mitrofanov, A., Backofen, R., & Voß, B. (2020). CRISPR-Cas bioinformatics. *Methods (San Diego, Calif.)*, *172*, 3–11.
- Allaby, R. G., & Woodwark, M. (2004). Phylogenetics in the bioinformatics culture of understanding. *Comparative and Functional Genomics*, *5*(2), 128–146.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data.
- Aneesh, T. P., Sekhar, S., Jose, A., Chandran, L., & Zachariah, S. M. (2009). Pharmacogenomics: The right drug to the right person. *Journal of Clinical Medicine Research*, *1*(4), 191.
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Martin, M. J. (2004). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *32*(1), D115–D119.
- Avashthi, H., Gautam, B., Jain, P. A., Tiwari, A., Pathak, R. K., Srivastava, A., . . . Kumar, A. (2014). In silico identification of MAPK3/6 substrates in WRKY, bZIP, MYB, MYB-related, NAC and AP-2 transcription factor family in Arabidopsis thaliana. *International Journal of Computational Bioinformatics and In Silico Modeling*, *3*, 454–459.
- Avashthi, H., Pathak, R. K., Gaur, V. S., Singh, S., Gupta, V. K., Ramekte, P. W., & Kumar, A. (2020). Comparative analysis of ROS scavenging gene families in finger millet, rice, sorghum and foxtail millet revealed potential targets for antioxidant activity and drought tolerance improvement. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*, 33.
- Avashthi, H., Pathak, R. K., Pandey, D., Arora, S., Mishra, A. K., Gupta, V. K., Kumar, A. (2018). Transcriptome-wide identification of genes involved in Ascorbate–Glutathione cycle (Halliwell–Asada pathway) and related pathway for elucidating its role in antioxidative potential in finger millet (*Eleusine coracana* (L.)). *3 Biotech*, *8*(12), 499.
- Avery, Oswald T., Colin, M. MacLeod, & McCarty, Maclyn (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, *79*(2), 137–158.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Ewinger, L., Eddy, S. R., Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Research*, *30*(1), 276–280.
- Benson, D., Boguski, M., Lipman, D. J., & Ostell, J. (1990). The national center for biotechnology information. *Genomics*, *6*(2), 389–391.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120.
- Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). *PubChem: Integrated platform of small molecules and biological activities*, . *Annual reports in computational chemistry* (4, pp. 217–241). Elsevier.
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.
- Bunin, B. A., Siesel, B., Morales, G. A., & Bajorath, J. (2007). *Cheminformatics theory* (pp. 1–49). Netherlands: Springer.
- Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, *31*(13), 3497–3500.
- Chinchole, M., Pathak, R. K., Singh, U. M., & Kumar, A. (2017). Molecular characterization of EcCIPK24 gene of finger millet (*Eleusine coracana*) for investigating its regulatory role in calcium transport. *3 Biotech*, *7*(4), 267.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–3676.
- Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). *Chapter 22: A model of evolutionary change in proteins. Atlas of protein sequence and structure*. Washington, DC: National Biomedical Research Foundation.
- DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, *40*(1), 82–92.
- Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210.
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Sali, A. (2006). Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics*, *15*(1), 5–6.
- Funahashi, A., Morohashi, M., Kitano, H., & Tanimura, N. (2003). CellDesigner: A process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, *1*(5), 159–162.

- Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6), 1981–1996.
- Goodsell, D. S., Morris, G. M., & Olson, A. J. (1996). Automated docking of flexible ligands: Applications of AutoDock. *Journal of Molecular Recognition*, 9(1), 1–5.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Chen, Z. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7), 644.
- Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis*, 18(15), 2714–2723.
- Gupta, A. K., Gaur, V. S., Pathak, R. K., Gupta, S., & Kumar, A. (2014). Dof1 transcription factor interacts with only specific regions of the promoters driving the expression of genes involved in carbon and nitrogen metabolism. *International Journal of Computational Bioinformatics and In Silico Modeling*, 3(4), 412–422.
- Gupta, S., Pathak, R. K., Gupta, S. M., Gaur, V. S., Singh, N. K., & Kumar, A. (2018). Identification and molecular characterization of Dof transcription factor gene family preferentially expressed in developing spikes of *Eleusine coracana* L. *3 Biotech*, 8(2), 82.
- Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36, 39–56.
- Hogeweg, P., & Hesper, B. (1978). Interactive instruction on population interactions. *Computers in Biology and Medicine*, 8(4), 319–327.
- Hogeweg, P. (1978). Simulating the growth of cellular forms. *Simulation*, 31(3), 90–96.
- Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology*, 7(3), e1002021.
- Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., & Nakata, A. (1987). Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of Bacteriology*, 169(12), 5429–5433.
- Jaskolski, M., Dauter, Z., & Wlodawer, A. (2014). A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *The FEBS Journal*, 281, 3985–4009.
- Jayaram, B., & Priyanka, D. (2010). *Bioinformatics for better tomorrow. Supercomputing facility for bioinformatics and computational biology*. New Delhi: Indian Institute of Technology.
- Meidanis, J. (2003). Current challenges in bioinformatic. In M. A. Nascimento, E. S. de Moura, & A. L. Oliveira (Eds.), *SPIRE 2003, LNCS 2857* (pp. 16–27). .
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.
- Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Childs, K. L. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6(1), 4.
- Kokane, S. B., Pathak, R. K., Singh, M., & Kumar, A. (2018). The role of tripartite interaction of calcium sensors and transporters in the accumulation of calcium in finger millet grain. *Biologia Plantarum*, 1–10.
- Kumar, A., Pathak, R. K., Gayen, A., Gupta, S., Singh, M., Lata, C., Gupta, S. M. (2018). Systems biology of seeds: Decoding the secret of biochemical seed factories for nutritional security. *3 Biotech*, 8(11), 460.
- Kumar, A., Pathak, R. K., Gupta, S. M., Gaur, V. S., & Pandey, D. (2015). Systems biology for smart crops and agricultural innovation: Filling the gaps between genotype and phenotype for complex traits linked with robust agricultural productivity and sustainability. *OMICS: A Journal of Integrative Biology*, 19(10), 581–601.
- Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: Molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics (Oxford, England)*, 10(2), 189–191.
- Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Snoep, J. L. (2006). BioModels database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(1), D689–D691.
- Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Hoad, G. (2010). The European nucleotide archive. *Nucleic Acids Research*, 39(1), D28–D31.
- Leinonen, R., Sugawara, H., & Shumway, M. (2010). International Nucleotide Sequence Database Collaboration, the sequence read archive. *Nucleic Acids Research*, 39(1), D19–D21.
- Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., & Brandi, M. L. (2010). Bioinformatics for next generation sequencing data. *Genes*, 1(2), 294–307.
- Mamgain, S., Dhiman, S., Pathak, R. K., & Baunthiyal, M. (2018). *In silico* identification of agriculturally important molecule (s) for defense induction against bacterial blight disease in soybean (*Glycine max*). *Plant Omics*, 11(2), 98.
- Mamgain, S., Sharma, P., Pathak, R. K., & Baunthiyal, M. (2015). Computer aided screening of natural compounds targeting the E6 protein of HPV using molecular docking. *Bioinformation*, 11(5), 236.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17(1), 10–12.
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of the Sciences of United States of America*, 74, 560–564.
- Mojica, F. J., Díez-Villaseñor, C., Soria, E., & Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology*, 36(1), 244–246.
- Mojica, F. J., García-Martínez, J., & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, 60(2), 174–182.
- Moody, G. (2004). *Digital code of life: How bioinformatics is revolutionizing science, medicine, and business*. London: Wiley.
- Mount, D. W. (2001). *Bioinformatics—sequence and genome analysis* (Vol. 156, pp. 75–85). New York: CSHL.

- Moussa, A., Vannier, B., Yennamalli, R.M., & Singh, T.R. (2016). Recent trends and developments in bioinformatics: Challenges and opportunities.
- Murata, M., Richardson, J. S., & Sussman, J. L. (1985). Simultaneous comparison of three protein sequences. *Proceedings of the National Academy of the Sciences of United States of America*, 82, 3073–3077.
- Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), 536–540.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443–453.
- Nirenberg, M., & Leder, P. (1964). RNA code words and protein synthesis. The effect of trinucleotides upon the binding of sRNA to ribosomes. *Science (New York, N.Y.)*, 145, 1399–1407.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH—A hierarchic classification of protein domain structures. *Structure (London, England: 1993)*, 5(8), 1093–1109.
- Pathak, R. K., & Sagar, M. (2012). Green tea: A natural source of drug for liver cancer. *International Journal of Medical Sciences*, 5(1-2), 7–9.
- Pathak, R. K., Baunthiyal, M., Pandey, D., & Kumar, A. (2018). Augmentation of crop productivity through interventions of omics technologies in India: Challenges and opportunities. *3 Biotech*, 8(11), 454.
- Pathak, R. K., Baunthiyal, M., Pandey, D., & Kumar, A. (2020). Computational analysis of microarray data of *Arabidopsis thaliana* challenged with *Alternaria brassicicola* for identification of key genes in *Brassica*. *Journal of Genetic Engineering & Biotechnology*, 18(1), 1–20.
- Pathak, R. K., Baunthiyal, M., Pandey, N., Pandey, D., & Kumar, A. (2017). Modeling of the jasmonate signaling pathway in *Arabidopsis thaliana* with respect to pathophysiology of *Alternaria* blight in *Brassica*. *Scientific Reports*, 7(1), 16790.
- Pathak, R. K., Baunthiyal, M., Shukla, R., Pandey, D., Taj, G., & Kumar, A. (2017). *In silico* identification of mimicking molecules as defense inducers triggering jasmonic acid mediated immunity against *Alternaria* blight disease in *Brassica species*. *Frontiers in Plant Science*, 8, 609.
- Pathak, R. K., Baunthiyal, M., Taj, G., & Kumar, A. (2014). Virtual screening of natural inhibitors to the predicted HBx protein structure of hepatitis B Virus using molecular docking for identification of potential lead molecules for liver cancer. *Bioinformation*, 10(7), 428.
- Pathak, R. K., Giri, P., Taj, G., & Kumar, A. (2013). Molecular modeling and docking approach to predict the potential interacting partners involved in various biological processes of MAPK with downstream WRKY transcription factor family in *Arabidopsis thaliana*. *International Journal of Computational Bioinformatics and In Silico Modeling*, 2, 262–268.
- Pathak, R. K., Gupta, A., Shukla, R., & Baunthiyal, M. (2018). Identification of new drug-like compounds from millets as Xanthine oxidoreductase inhibitors for treatment of hyperuricemia: A molecular docking and simulation study. *Computational Biology & Chemistry*, 32–41.
- Pathak, R. K., & Singh, D. B. (2020a). *Integrated omics for dissecting host–pathogen interaction: Challenges and opportunities. Recent trends in 'computational omics: Concepts and methodology'*. United States: Nova Science Publisher.
- Pathak, R. K., & Singh, D. B. (2020b). *Systems biology approaches for food and health. Advances in agri-food biotechnology* (pp. 409–426). Reading, MA: Springer.
- Pathak, R. K., Singh, D. B., Sagar, M., Baunthiyal, M., & Kumar, A. (2020). *Computational approaches in drug discovery and design. Computer-aided drug design* (pp. 1–21). Singapore: Springer.
- Pathak, R. K., Taj, G., Pandey, D., Arora, S., & Kumar, A. (2013). Modeling of the MAPK machinery activation in response to various abiotic and biotic stresses in plants by a system biology approach. *Bioinformation*, 9(9), 443.
- Pathak, R. K., Taj, G., Pandey, D., Kasana, V. K., Baunthiyal, M., & Kumar, A. (2016). Molecular modeling and docking studies of phytoalexin (s) with pathogenic protein (s) as molecular targets for designing the derivatives with anti-fungal action on 'Alternaria' spp. of 'Brassica'. *Plant Omics*, 9(3), 172.
- Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of Clinical Medicine*, 9(1), 132.
- Petersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612.
- Rai, S. K., Pathak, R. K., Singh, D. B., Bhatt, A., & Baunthiyal, M. (2021). Chemo-informatics guided study of natural inhibitors targeting rho GTPase: A lead for treatment of glaucoma. *In Silico Pharmacology*, 9(1), 1–11.
- Rana, G., Pathak, R. K., Shukla, R., & Baunthiyal, M. (2020). *In silico* identification of mimicking molecule(s) triggering Von Willebrand Factor (VWF) in Human: A molecular drug target for regulating coagulation pathway. *Journal of Biomolecular Structure & Dynamics*, 38(1), 124–136.
- Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Miller, N. (2003). The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, 31(1), 224–228.
- Sagar, M., Pathak, R. K., Pandey, R. K., Singh, D. B., Pandey, N., & Gupta, M. K. (2014). Binding affinity analysis and ADMET prediction of epigallocatechin gallate (EGCG) derivatives for AP-1 protein: A drug target for liver cancer. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 3, 66. Available from <https://doi.org/10.1007/s13721-014-0066-x>.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of the Sciences of United States of America*, 74, 5463–5467.
- Sanger, F., & Thompson, E. O. P. (1953). The amino-acid sequence in the glycol chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 53, 366–374.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.

- Sidhu, K. S., Bhangu, S. K., Pathak, R. K., Yadav, I. S., & Chhuneja, P. (2020). Identification of natural lead compounds for leaf rust of wheat: A molecular docking and simulation study. *Journal of Proteins and Proteomics*. Available from <https://doi.org/10.1007/s42485-020-00048-5>.
- Sievers, F., Higgins, D. G., & Clustal, Omega (2014). Accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, 1079, 105–116.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Thompson, J. D. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), 539.
- Singh, D. B., Gupta, M. K., & Pathak, R. K. (2020). Natural products in cancer chemoprevention and chemotherapy. *Frontiers in Natural Product Chemistry*, 6, 3–34.
- Singh, D. B., & Pathak, R. K. (2020). Computational approaches in drug designing and its application. In N. Gupta, & V. Gupta (Eds.), *Experimental protocols in biotechnology* (pp. 95–117). Springer Nature.
- Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Taylor, J. M. (2016). EffectorP: Predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, 210(2), 743–761.
- Sterling, T., & Irwin, J. J. (2015). ZINC 15—Ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11), 2324–2337.
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Jensen, L. J. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1), D607–D613.
- Takahashi, T., Luzum, J. A., Nicol, M. R., & Jacobson, P. A. (2020). Pharmacogenomics of COVID-19 therapies. *npj Genomic Medicine*, 5(1), 1–7.
- Tamm, C., Shapiro, H. S., Lipshitz, R., et al. (1953). Distribution density of nucleotides within a desoxyribonucleic acid chain. *The Journal of Biological Chemistry*, 203, 673–688.
- Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., & Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Research*, 30(1), 27–30.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16), 1701–1718.
- Verma, S., Pathak, R. K., Kasana, V., & Kumar, A. (2017). Binding affinity analysis of cinnamanilide and α -aminophosphonic acid derivatives for acetohydroxyacid synthase through molecular docking. *International Journal of Agriculture, Environment and Biotechnology*, 10(3), 271.
- Wang, J. T., Zaki, M. J., Toivonen, H. T., & Shasha, D. (2005). *Introduction to data mining in bioinformatics*. Data mining in bioinformatics (pp. 3–8). New York: Springer.
- Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., McCouch, S. (2002). Gramene: A resource for comparative grass genomics. *Nucleic Acids Research*, 30(1), 103–105.
- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171, 737–738.
- Wu, C. H., Yeh, L. S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Vinayaka, C. R. (2003). The protein information resource. *Nucleic Acids Research*, 31(1), 345–347.
- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829.