Databases and Bioinformatics Tools for Data Mining

6

Pallabi Pati, Sushil Kumar Rathore, and Manoj Kumar Gupta

Abstract

Data, information, and knowledge play an interesting role in human life. Huge repositories of data generated because of the recent development of technologies demand the development of novel tools and techniques that can retrieve more important information. Data mining is a kind of knowledge discovery technique that extracts useful information from heterogeneous biological data by employing various machine learning, artificial intelligent systems, and decision-making techniques. Thus, in this chapter, the authors attempted to understand how data mining approaches have revolutionized biological research. The topic of data mining is discussed in brief, including the application it has in bioinformatics. This chapter also illustrates some of the emerging problems and opportunities in data mining in bioinformatics by utilizing this analogy.

Keywords

Data Mining · Bioinformatics · Databases · KDD

P. Pati (✉)
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_6

Abbreviation

ANN	Artificial neural network
BLAST	Basic local alignment search tool
DBMS	Database management system
DDBJ	DNA databank of Japan
DNA	Deoxy Ribonucleic acid
EMBL	European molecular biology laboratory
ESTs	Expressed sequencing tags
GOLD	Genomes Online Database
INE	Integrated Rice Genome Explorer
INSDC	International nucleotide sequence database collaboration
IRGSP	International Rice Genome Sequencing Project
KDD	Knowledge discovery in database
KNN	K Nearest Neighbor
NCBI	National center for biotechnology information
PDB	Protein databank
RGP	Rice Genome Research Program
RNA	Ribonucleic acid
TBP	TATA box binding protein
TIGR	The Institute for Genomic Research
UCEs	Upstream control elements
VEP	Variant Effect Predictors

6.1 Introduction

The digital revolution has made it easy to record, process, store, distribute, and share digitized information. With major developments in computing and related technology and their ever-expanding use in various walks of life, vast quantities of data of various features continue to be gathered and processed in databases. If the amount of data in the world doubles every 20 months, it is possible that the size and number of databases will rise at a similar rate. Thus, it is really a challenge to discover information from this huge amount of data. Data mining is an effort to make sense of the abundance of knowledge embedded in this large data volume [1]. Data mining uses various techniques, models, or algorithms to analyze the vast amount of data stored in the databases. In data mining, a pattern or rule is discovered that helps in establishing a hidden relationship between variables. The key objective is to manipulate the computer's data processing ability with the capacity of humans to interpret patterns [2].

In 1989, Piatetsky-Shapiro coined the phrase "knowledge discovery in database" (KDD). Application of data mining and KDD is inevitable due to the huge size of data collected from different sources and difficulty in handling and analyzing these data manually. KDD can be seen as an inclusive method of extracting useful information from information, while data mining could be defined as the core of KDD, which involves algorithms that discover unknown patterns of data, construct models, and discovery [3]. In the last two decades, the development of various data mining techniques in various fields such as artificial intelligence, machine learning, soft computing, and statistics has led researchers to create and apply new data mining methodologies. Data mining generally works on information stored in the database, which may be interrelated or relevant and inconsistent and irrelevant sometimes. Thus, it requires an application that allows the administrator of that data to manage it in order to exploit and control the necessary data. Maintenance and manipulation of the database is known as database management systems (DBMS) [4]. For instance, “Integrated Rice Genome Explorer” (INE) is a database that helps us to integrate genetic as well as information regarding physical mapping with the genome sequences generated by the collaboration with the “International Rice Genome Sequencing Project” (IRGSP). These Databases contain a various tools to analyze and compare the genomic databases of rice and maize. This system is also very much helpful in the development of different kinds of grass crop species. This kind of comparative genomics analysis helps us gather updated knowledge regarding total structural and functional data of various basic plant genome, which will lead to a better era in the field of biological research by focusing on bioinformatics tools techniques. Considering the above information, in the current chapter, the authors tried to highlight the basics concept of databases and their standards and the benefits of DBMS. It then describes the principle of data mining and how data mining processes are useful in biological research.

6.2 The Databases Concept

Databases are having a significant influence on the increasing usage of computers. It is certainly fair to say that databases are in use in most areas where computing is required, including business, engineering, medicine, law, education, and library [4]. The word database is used so often that to define it, and we must define the concept of a database. A database is a list of many related data. By data, we mean factual information that can be recorded and that has implicit meaning. Consider the names, addresses, and phone numbers of your friends. You may have recorded this data on a personal computer and stored it in a database using a database management application, like Microsoft EXCEL or ACCESS. This is a set of information with a predetermined context and therefore is a data set. The previous concept of the database was very generalized; for instance, consider the list of terms this page includes as similar data. Therefore, this page may serve as a database. However, nowadays, the term database is generally used more specifically. A database should have an implicit set of properties [4].

A database can reflect few parts of the natural world, called the mini world, or the universe of discourse (UoD). The database reflects changes within the mini world. A database is a systematically organized data that holds some meaning which cannot be altered or deleted by the database administration. A random collection of data, even if very large, cannot be called a database. A database is created, designed, and populated with data for a specific purpose. It has both an intended group of users and some preconceived applications in which they may be interested. Thus, a database has a "source" from which details are collected (collection of records, memories, etc.), "any degree of contact" with events that "happening in the real world" (i.e., events that are true), and its information "used by an audience that is actively involved" (i.e., its users who frequently view, change, and erase the information contained in the database). For the purpose of regulatory compliance, a database can be of any size and complexity. On the one hand, it is likely some database may be limited (say, no more than a few hundred records). For example, a collection of names and addresses with a basic layout like: "John Smith" and "Kolkata". On the other hand, there may be on the order of half a million books in large libraries, in different categories such as by author's last name, subject, and book title. Each book could be arranged in alphabetical order within its respective category [4].

A database may be developed and managed either by anyone manually (by a physician), or by computer. Because the library card catalog is a manually edited database, it is an example of a database where someone can make mistakes. A computerized database may be established and sustained either by the services of the application program(s) designed for that purpose or by the services of a DBMS. Here error is minimal in comparison to a manual database. A DBMS allows users to build and retain databases that help us to keep track of items and arrange them logically for easier access to knowledge. As a comprehensive software system, the DBMS is a general-purpose system that facilitates the processes of defining, constructing, and manipulating the database for multiple purposes [4].

The concept of a database includes defining the types of data, its composition, and the rules for which data can be applied to the database. To build a database, first data is stored in some storage medium controlled by the DBMS, and then an interface is written to access this data. A database may be manipulated by making series of queries against a database to obtain specific data, updating the database to reflect changes in the mini-world, and generating reports from the data. In order to define a database, a programmer must first determine the data type for the data to be stored, then the configuration of the data, then what is the output data type of the data will be in the data structure, and finally, the configuration for that output data type is taken into consideration. The process of building the database is the process of storing the data itself on some controlled medium, like a hard drive or flash drive. In order to exploit a database, it needs to be queried to acquire requested data, which then needs to be modified to represent improvements within the mini world, which needs to be queried again to obtain more knowledge about the mini world and eventually needs to be used to generate a report or other kind of result data [4]. A general-purpose DBMS is not necessary use to implement a computerized database. We might compose our series of programs to build and preserve the database, in essence, developing our special-purpose DBMS applications (software that handles databases). Regardless of whether we use a general-purpose database system or not, we typically have to recruit a significant amount of software to manipulate the database [4].

6.3 Advantages of DBMS

DBMS allows end-users to build, view, edit, and erase data. It is a layer that links programs and data. Compared to the File-Based Data Management System, DBMS is a superior management application (https://www.tutorialspoint.com/). Few important advantages of DBMS are reducing data redundancy, sharing of data, data integrity, data security, privacy, backup and recovery, and data consistency. The file-based data storage systems spanned several files, each located in several separate places in a system and sometimes residing in another device in several locations. Due to this redundancy, multiple copies of the same file can lead to data redundancy. In a database, we can set a password in an encrypted way and preventing anyone from accessing our database so that they cannot alter our database's setup. As a result of this, there is no chance of encountering duplicate data. Users of a database may also share the data among themselves. However, there are multiple kinds of authorization to access the data. As with multiple layers of security, the data can only be disclosed depending on the class of authorizations. Remote users, who are working together, can also access the database at the same time, and they can also share the data they are looking at within the database.

Data integrity ensures data reliability and consistency. Data integrity is very important because multiple databases are stored on a single database server. All these databases contain information that is either visible to a lot of people or is connected to a lot of people. In order to ensure that the data used are accurate and consistent, it is essential for the data to be verified on multiple sources and it get exploited by different predefined users. Data security is a critical principle of database management. Only approved users should be given entry to the framework, and there should be a username and password for each authorized user. Unauthorized users are not permitted to access the database under any circumstances because it is highly illegal and against the security policies.

According to the laws, the privacy law in a database ensures that the approved users can only read, change, or erase the database's data. The information about the database is given and, once it is given, can be seen only by the user with the necessary authority. For example, some sites (like Facebook) require only one account and a certain password to create an account for a particular user. Others have their own username and password. A database management framework automatically also takes control of replication and recovery. The DBMS will remember all necessary changes and informs the user when it is time to back up their data. As well, in case of a system error or crash, it restores the database to the same state it was before the error occurred. The software configuration also guarantees that the anomalies cannot exist and that data redundancy is not a problem. To make some claims, all data will appear consistently across the database, and all users viewing the database will agree on all of it. It is not just an effective storage structure, but the versatility is still fantastic. If any improvements are made to it, they automatically go to all the users, and there is no data discrepancy.

6.4 Knowledge Discovery and Data Mining

The conventional way of converting data into information focused on doing manual review and evaluation by a domain specialist in order to identify valuable trends in data for decision support. For instance, early the work of Reeder & Feller employing methods was crucial in diagnosing and treating fever [5]. In 1996, this process was described by different steps starting from data selection, preprocessing, data transforming, data mining, and interpretation [6]. Data selection involves the previous know-how of and target of the application. The selection of a dataset or a subset of variables is made through ranking via selection technique [1, 7]. Data pattern processing is needed to enhance the actual data quality for mining. This also enhances the productivity of mining by decreasing the data processing time. Data preprocessing requires data cleaning, data transformation, data integration, compact representation, data reduction or data compression, etc. Data cleaning comprises operations like normalization, noise elimination, and missing data handling, redundancy reduction, etc. Real-world data is frequently erroneous, incomplete, and contradictory, likely due to technical errors or defects in the implementation of the system. It is important to clean up such low-quality data before data mining. Data integration plays a significant role. This operation involves the integration of various heterogeneous datasets created from various sources. Reduction and projection of data involve identifying useful features to represent the data (depending on the objective of the task) and using methods of reduction of dimensionality, discretization of features, and extraction (or transformation) of features. The application of data compression principles can help in data reduction, which has potential in the future to grow, especially in the field of multimedia dataset knowledge discovery. Data mining mainly involves classification, regression, clustering, description, image retrieval, the discovery of association rules and functional dependencies, rule extraction, etc. Interpretation deals with the deduction of patterns found and the possible visualization patterns of extracted information. To classify the genuinely interesting or useful patterns for the user, one may evaluate the extracted patterns automatically or semi-automatically. Using discovered knowledge, we integrating earlier generated knowledge into the performance system and taking knowledge-based acts [1, 7].

Thus, data mining is basically a subset of Knowledge Discovery. While the original notion was "Knowledge Discovery in Databases" (KDD), nowadays, in order to emphasize that Data Mining is an essential part of the knowledge discovery method, the current most common notion is "Knowledge Discovery and Data Mining" (KDD) (Fig. 6.1). It is important to note that KDD (knowledge discovery and data mining) is not simply a process but also encompasses the complete value-added chain from the data's extremely physical side to the very human side of knowledge—i.e., the latter characterized from a cognitive point of view: knowledge

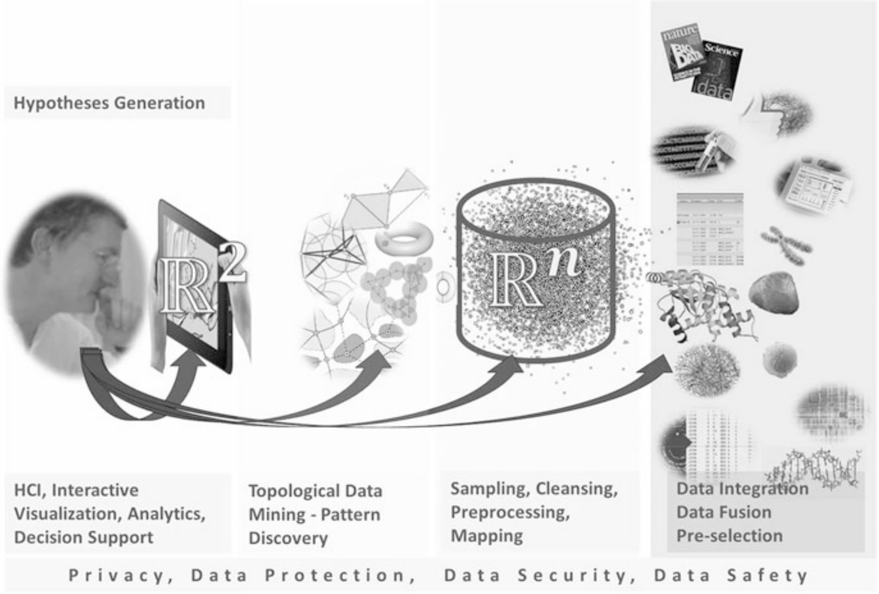
Hypotheses Generation
$\mathbb{R}^2$
$\mathbb{R}^n$
HCI, Interactive Visualization, Analytics, Decision Support
Topological Data Mining - Pattern Discovery
Sampling, Cleansing, Preprocessing, Mapping
Data Integration Data Fusion Pre-selection
Privacy, Data Protection, Data Security, Data Safety

Fig. 6.1 The general knowledge discovery process that is widely employed in the life sciences (Adapted from [7]

as a set of expectations [1, 7]. Recently, Holzinger describes the novel technique that extends the original definition by Fayyad and the team [6] by having an actual human make the decisions. As core theories of human-computer interaction, HCI & KDD, together with a novel approach, aims to bring all two together into this research project to advance knowledge in a specific context [7]. The core principle of HCI-KDD is to allow end-users to identify and classify previously unknown and potentially valuable and accessible knowledge. It is defined as the process of identifying new data patterns from unstructured data. The goal, in this case, is a visualization of data that was previously unseen. There is a specialist in the framework with clear domain expertise. By allowing them to interactively explore their data sets, they may be able to recognize, interpret, and appreciate valuable details, acquire new, and previously unknown information [7].

6.4.1 Data Mining

Initially, the data mining technique was widely used in economics. But nowadays, several data mining techniques are also used in the field of agricultural research. It helps to improve the prediction of the yield of the specific variety of cultivation that will be very much beneficial for the plant. As we all know the agricultural production largely depends on climatic condition, soil type, irrigation method, cultivation strategies, data mining will help us to predict the best dependencies or fitting model, which in turn may help us in increasing the grain yield. Typically, a data mining algorithm comprises three major components, namely, the model, preference criterion, and search algorithm [1]. A model includes parameters that must be calculated using a specific representational type or tool from the data for the chosen function. The preference criterion is the preference, based on the data given, for one model or set of parameters over another. The criterion is generally some sort of the model's goodness-of-fit function to the data, perhaps modified by a smooth term to prevent overfitting or to generate a model with too many degrees of freedom to be limited by the data given. The search algorithm is an algorithm for finding particular models or patterns and parameters, provided the data, model(s), and criterion of preference. The model-preference-search components are typically instantiated by a particular data mining algorithm [1]. Tasks for data mining are divided into two key categories: predictive and descriptive. Six key functions of data mining are described classifying, regression, clustering, modeling of dependencies, variance detection, and summarizing [6]. Classification, regression, and anomaly detection are categorized under the predictive category, whereas clustering, dependency modeling is categorized under the descriptive category. Predictive model forecasts use certain variables in the dataset to predict unknown values of other related variables while descriptive model classifies patterns or relationship and utilizes human-understandable pattern and trends in data [3].

Classification is part of the classical methodology of data mining that is based on machine learning. In a database, it finds mutual properties among a set of objects and categorizes them according to the classification model into diverse groups. Its primary objective is to scrutinize the training data and construct an accurate definition or model for each class using the data function. Statistical techniques such as decision trees, neural networks, and statistics are used in this process [3]. Regression describes the relationship between dependent and independent variables. Prediction is reached by endorsing regressions. Statistical regression is a mathematical model that relates the values of the dependent variable to the values of the other predictor or independent variable. The predicted variable in regression could be a continuous variable. Real-valued prediction variables in regression are mapped from elements of a learning function. Some of the widely used regression techniques include statistical regression, Neural Network, Support Vector Machine regression. More complex methods could also be used to predict future values, such as logistic regression, decision trees, or neural networks, and these techniques could also be combined to achieve better results [3].

Clustering is a technique of data mining that groups physical or abstract objects into related object classes. Clustering is a technique of dividing data sets (records/tuples/objects/samples) into multiple groups (clusters) based on predetermined similarities. The main objective of clustering is to find affinity-based groups (clusters) of objects so that there is a great resemblance to each other within individual clusters, while clusters are sufficiently diverse from each other. Clustering is a type of unsupervised learning in machine learning terminology [3]. Dependency Modeling (or association rule mining) is one of the finest recognized data mining techniques and is categorized under an unsupervised data mining technique that seeks to identify links or relationships between items or records belonging to a large dataset and identifies significant variables dependencies [3]. Anomaly detection is synonymous with the uncovering of the most significant changes or aberrations from standard behavior [3]. Although not part of data mining techniques, the summary is the result of these techniques and deals with the determination of a compact representation for a subset of data synonymously referred to as generalization or description [3]. Sequential Pattern is used over a business cycle to determine sequential patterns or associations or periodic events/trends between variable data fields [3].

6.4.2 Data Mining Architecture

The architecture of data mining can be mainly classified as below:

Knowledge Base It acts as the start of the entire process of data mining. It serves as a guide for looking for the resulting trends or evaluating their interestingness. This form of information can involve hierarchies of concepts that organize attributes or values into different abstraction stages.

Data Mining Engine It forms the main element of the mining framework, consisting of all the modules required to perform data mining tasks, such as characterization, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern Evaluation Module This module is generally correlated with measures of interest. In order to stay focused on looking for interesting trends, it persistently interacts with the data mining engine. Many times, depending on the data mining method used, it uses thresholds to sieve out discovered patterns or can use the pattern evaluation module incorporated with the mining module.

User Interface The module acts as a link between users and the framework for data mining. It makes it simple and effective for users to communicate with the system without thinking about the convolutions behind the operation.

Data Sources (Www, Data Warehouse, Archive, Other Repositories) These are the actual data sources, and for efficient data mining, a huge amount of historical data is needed. In databases or data centers, companies usually store data. Often the data warehouse includes more than one database or text file, or spreadsheet. Another big source of data is www.

Server Database or Data Warehouse Includes concrete data that are set to be retrieved. Its main duty is to retrieve data at the request of users.

Other Processes Data must be cleaned and merged before it is passed on to the data warehouse server, as data are obtained from different sources and are in different formats such that it cannot be used directly for mining processes. The data need to be cleaned, integrated, and it is only important to pick and move on the secure data to the data warehouse server. Numerous techniques for cleaning, integration, and selection may be needed for the operation [3].

6.5 Databases for Biological Data Mining

In recent decades, a huge number of genome-scale experimental data sets have been made available. Thus, for storing and analyzing, several biology databases have appeared online. These databases can be classified according to data form, data processing techniques, data coverage scope, and database accessibility. These databases contain a wide range of data ranging from the genome of model and nonmodel plants (https://asia.ensembl.org/index.html) to protein information (https://www.rcsb.org/) (Table 6.1).

6.5.1 Databases for Genes, Genomes, and Variations

While breeding has been effective, the method of choice for farmers remained traditional, e.g., for studies contrasting two genes, A and B, in a test plant. With the aid of genomics and new sequencing techniques, scientists can study the underlying genetic makeup of plants, and these findings are helping us figure out how plant breeding contributes to the development of desired traits. Even though it is still in its infancy, Next-Generation Sequencing (NGS) technologies are permitting the mass sequencing of genomes and transcriptomes to produce a vast amount of genomic information. By means of bioinformatics technologies, the NGS data analysis, as evidenced by the huge collections of markers, allows the new genes and sequences discovery and the location and arrangement of their occurrence on the genome. By studying the gene expression level of a variety of breeds, breeders get an understanding of the molecular basis of complex traits. Genome-wide association studies, or GWAS, include TILLING and Eco-mutation in genome sequencing technologies, which can make it possible to scan mutant as well as germplasm collections for allelic variants in target genes [8]. It is very useful to re-sequence an organism’s genome more than once in order to find markers that can be used in high-throughput genotyping platforms like SNPs and SSRs, or the construction of a genetic map. These tools and resources make it much easier to study genetic diversity, which is important for maintaining germplasm, enhancing, and application. Also, they can be used to help identify some of the genes in those regions of the genome that might also be involved with the disease, and they can be used to find markers linked to those genes as well. New markers for quantifiable characteristics based on DNA, such as the ones mentioned above, are employed for marker-assisted selection, including breeding by design, marker-assisted backcross selection, and genome selection. Thus, advances in genomics provide breeders with novel tools

Table 6.1 List of few important Biological databases

Database Name	Description	Species	Link
BAR (bio-analytical resource for plant biology)	Provides a user-friendly interface for the exploration of gene expression data	Several plant species including *O. sativa*	http://bar.utoronto.ca
CoP database	Microarray data based integrated database for co- expressed genes and biological processes in plants	*Arabidopsis thaliana*, *Vitis vinifera Glycine max*, *Oryza sativa*, *Populus trichocarpa*, *Hordeum vulgare*, *Triticum aestivum*, and *Zea mays*.	http://webs2.kazusa.or.jp/kagiana/cop0911
CSRDB (cereal small RNA database)	Consists of large maize and rice datasets smRNA sequences provided by high performance pyrosequencing	*O. sativa* and maize	http://sundarlab.ucdavis.edu/smrnas
CyVerse (former iPlant collaborative)	Provides a strong computing platform allowing massive datasets and complex research to be discovered using the data	Plants, animals, and microbes	http://www.cyverse.org
DDBJ updated on daily bases	A repository of nucleotide sequence data	*O. sativa* and several organisms species	http://www.ddbj.nig.ac.jp
Diurnal	An internet-based site to keep track of diurnal and circadian genome wide expression profiles from results of model plants	Plant species	http://diurnal.mocklerlab.org
DroughtDB	Manually curated genes associated with stress response to drought	*O. sativa ssp. japonica cv. Nipponbare Zea mays*, *Arabidopsis thaliana*, *Sorghum bicolor*, *Hordeum vulgare*, *Brachypodium distachyon*, *Solanum lycopersicum*, *Secale cereale, and Aegilops tauschii*.	http://pgsb.helmholtz-muenchen.de/droughtdb
EMBL	Comprehensive compilation and annotation of nucleotide sequences	*O. sativa* and several organisms species	http://www.ebi.ac.uk/about

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
	from available public sources		
Ensembl plants	Provides numerous genomic data sets and analysis and visualization tools for several plant species in the genome browser	*O. sativa* and other organism species	http://plants.ensembl.org/index.html
ExPath	It offers data on metabolic pathways inferred from transcriptomic data based on microarrays, gene annotation, and orthologous genes	*Oryza sativa, Arabidopsis thaliana, and Zea mays*	http://expath.itps.ncku.edu.tw
FamNet	Enables the user to retrieve data from one or more plant species linked to preserved structural-functional domains within proteins	Arabidopsis, *Oryza sativa*, *Medicago truncatula*, *Populus tremula*, *Hordeum vulgare*, *Glycine max*, *Nicotiana tabacum*, and *Triticum* spp	http://www.gene2function.de/famnet.html
Galaxy	A software framework that allows experimentalists to conduct complex large-scale research with only a web browser without informatics or programming skills		http://galaxyproject.org
GenBank updated on daily basis	NIH genetic sequence database, a repository of publicly available DNA sequences	*O. sativa* and other organism species	http://www.ncbi.nlm.nih.gov
Genevestigator	Provides powerful tools to explore gene expression across a wide variety of biological contexts	Arabidopsis, *Oryza sativa*, *Medicago truncatula*, *Populus tremula*, *Glycine max*, and *Triticum* spp	https://genevestigator.com/gv
Gramene	An open data resource for comparative functional genomics in cereals and other plant species	O. sativa and other plant species	http://www.gramene.org
GRASSIUS (grass regulatory information services)	It consists of a series of databases relating to the regulation and interaction of gene expression in grasses	*Zea mays, Oryza sativa, Saccharum* spp., and *sorghum bicolor*	www.grassius.org

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
	with agronomic features. Includes transcription factors, promoters, transcription and co-regulators Factor-clones ORF		
GreenPhylDB	The database having catalog of gene families from various green plants	*O. sativa* and other plant species	http://www.greenphyl.org/cgi-bin/index.cgi
IsomiR Bank	Integrated resource that contains the sequence and expression of isomiRs	*Arabidopsis thaliana*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, *Oryza sativa*, *Drosophila melanogaster*, *Zea mays,* and *Solanum lycopersicum*.	http://mcg.ustc.edu.cn/bsc/isomir
Mercator pipeline	Functional annotation of plant "omics" data	Arabidopsis, Chlamydomonas, rice	http://mapman.gabipd.org/web/guest/app/Mercator
MoChA ("molecular characteristics database for allergens")	Database of allergenic proteins acquired by bioinformatics methods or proof of binding to IgE. It has obtained accurate experimental genome, transcriptome, proteome data, and molecular properties	2000 organisms	http://lilab.life.sjtu.edu.cn:8080/mocha/main-7.9-2.html
MPIC ("mitochondrial protein import components") database	Searchable details on plant and nonplant mitochondria protein import equipment	*O. sativa* and 23 other organism species	http://www.plantenergy.uwa.edu.au/applications/mpic
NIASGBdb ("National Institute of Agrobiological sciences planttfdb database")	A database having information on simple sequence repeat (SSR) polymorphisms in plant genomes	*O. sativa* and other plant species	http://www.gene.affrc.go.jp/databases_en.php
OryGenesDB	A rice reverse genetics database, created with flanking sequence tags of different mutagens and data on functional genomics	*O. sativa ssp. indica and japonica*, and two other plant species	http://orygenesdb.cirad.fr/index.html

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
PDB (protein data Bank)	Worldwide archive of structural data of biological macromolecules	*O. sativa* and other organism species	http://www.rcsb.org/pdb
Phytozome	An annotated plant genome and gene familial data comparison center. Provides an overview of each plant gene's evolutionary history at the level of sequence, gene structure, gene family, and genome organization [70]	*O. sativa* and 64 other plant and algae species	http://www.phytozome.net
PLANEX (PLAnt co-expression) database	Have publicly available GeneChip data received from the gene expression omnibus	*Arabidopsis thaliana, Hordeum vulgare, Glycine max, Vitis vinifera, Oryza sativa, Triticum aestivum, Solanum lycopersicum,* and *Zea mays*	http://planex.plantbioinformatics.org
PlantAPA (alternative polyadenylation)	A internet based server for query, visualization, and analysis of poly (A) sites in plants, helping in profiling various cleavage sites and quantify expression pattern of poly(A) sites across different conditions	*Oryza sativa, Chlamydomonas reinhardtii, Medicago truncatula*, and *Arabidopsis thaliana*	http://bmi.xmu.edu.cn/plantapa
PlantArrayNet	Information on co-expressed genes using microarray-transcriptomic data	Rice, Arabidopsis, and *Brassica rapa*	http://arraynet.mju.ac.kr/arraynet
PlantDHS (plant DNase I hypersensitive site database)	Incorporate histone modification, transcription factor binding sites, RNA sequencing, genomic sequence, and nucleosome positioning/occupancy	*Arabidopsis thaliana, Oryza sativa, and Brachypodium distachyon*	http://plantdhs.org
PlantGDB	A database of sequence data from different plant species	*O. sativa* and other plant species	http://www.plantgdb.org

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
Plant homolog database	A database comprised of plant homologous genes	16 plant sp. Including 10 *Oryza* species	http://phd.big.ac.cn
Plant MPSS (massively parallel signature sequencing) databases	Information on the expression status of genes, and potential unique transcripts (antisense transcripts, alternative splice isoforms, and regulatory intergenic transcripts)	Grape. Arabidopsis, *Magnaporthe grisea,* and rice	http://mpss.udel.edu
Plant-PrAS (plant protein annotation suite) database	Database of properties related to physicochemical and structural information, and unique functional region in plant proteomes	*Arabidopsis thaliana, Glycine max, Populus trichocarpa, Oryza sativa, Physcomitrella patens,* and *Cyanidioschyzon merolae*	http://plant-pras.riken.jp
Planteome	For plant and species-specific crop ontologies, a resource for popular reference ontologies. It also provides ontology-based rice gene annotation, QTLs, phenotypes, and germplasms	*Oryza and plant species*	http://www.planteome.org
PlantRNA	Assembles transfer RNA (tRNA) gene sequences obtained from fully annotated plant nuclear, plastid, and mitochondrial genomes	Five flowering plants (*Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, *Populus trichocarpa*, and *Brachypodium distachyon*), a moss (*Physcomitrella patens*), two green algae (*Ostreococcus tauri and Chlamydomonas reinhardtii*), a pennate diatom (*Phaeodactylum tricornutum*), one glaucophyte (*Cyanophora paradoxa*), and one brown alga	http://plantrna.ibmp.cnrs.fr

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
		(*Ectocarpus siliculosus*).	
PLEXdb	A single resource of gene expression for plants and plant pathogens. It is a phenotype genotype, hypothesis building knowledge warehouse, leveraging highly parallel expression data to associated genetic, physical, and pathway data with seamless portals	*Oryza, Vitis*, maize, *Fusarium graminearum*, *Arabidopsis*, soybean/*Phytopthora*/soybean cyst nematode, *Brachypodium*, cotton, poplar, Citrus, tomato, and *Medicago*	http://www.plexdb.org
PlnTFDB (plant transcription factor database)	A web interface to navigate various plant species' broad sets of transcription factors. Information is given for each family, including protein sequences, coding regions, genomic sequences, expressed sequence tags (ESTs), domain architecture, and scientific literature	*O. sativa ssp. indica* and *japonica* and other plant species	http://plntfdb.bio.uni-potsdam.de/v3.0
PmiRKB (plant miRNA Knowledge Base)	Information available for four major functional modules-"SNPs", "Pri-miRNAs", "MiR—Tar", and "self-reg"	*21 O. sativa* and Arabidopsis	http://bis.zju.edu.cn/pmirkb
PMRD (plant MicroRNA database)	A plant miRNA data repository containing associated sequence information, secondary structure, target genes, miRNA expression profiles, and their mapping to the browser of the species-specific genome	*O. sativa* and 120 plant species	http://bioinformatics.cau.edu.cn/PMRD
PO (plant ontology)	Robust and flexible controlled vocabulary that accurately represents the biology	*O. sativa* and other plant species	www.plantontology.org

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
	of plant structures and stages of development		
PODC (plant omics data center)	A repository for expression data of annotated gene and omics data analysis tools	*Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, *Nicotiana tabacum*, *Oryza sativa*, spreading earthmoss (*Physcomitrella patens*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), sorghum (*Sorghum bicolor*), grape (*Vitis vinifera*), corn (*Zea mays*)	http://bioinf.mind.meiji.ac.jp/podc
POGs (putative orthologous groups 2) database	A database that combines data from, Arabidopsis, rice and maize into "putative orthologous groups" (POGs) and permits comparisons among orthologs and extrapolation of annotations among species.	*Arabidopsis thaliana*, *Oryza sativa,* and *Zea mays*	http://pogs.uoregon.edu
Ppdb (plant promoter database)	Information available on Y patches, regulatory element groups (REGs), transcription start sites (TSSs), and core promoter structure (TATA boxes, initiators, GA and CA elements)	*Arabidopsis thaliana,* poplar*, Physcomitrella patens,* and *Oryza sativa.*	http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi
STIFDB2 (stress responsive transcription factor database)	Group of responsive genes for biotic and abiotic stress with options to detect possible transcription factor binding sites in their promoters. The data have been characterized by an integrated biocuration and genomic data mining approach	*O. sativa ssp. japonica* and *Indica* and Arabidopsis	http://caps.ncbs.res.in/stifdb2

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
UniProtKB	A central annotated protein resource consisting of two sections: UniProtKB/Swiss-Prot for annotated manual entries and UniProtKB/TrEMBL for annotated computer entries	*O. sativa* and other organism species	http://www.uniprot.org

and methodologies that permit a great leap forward in plant breeding, including the genetic dissection and breeding for complex traits and super domestication of crops.

The algorithms and methods used to store and process genomic data created by various technical platforms will rely on what kind of data is being used and what outcome is predicted. The 3000 genome project (http://iric.irri.org/resources/3000-genomes-project) and 1001 Arabidopsis genomes (http://1001genomes.org/) are good examples of why a genetic interface is needed to help breeders with the information they need. Once the information is obtained, results are made available to the breeders [9]. Attracted by the fame and glory of the invention of the global web page, a common and often successful approach to providing the information is through a web page that can be easily browsed. Several extensive bioinformatics resources exist to help scientists study plant and human genetics, like GenBank (http://www.ncbi.nlm.nih.gov/genbank/), the European Bioinformatics Institute (http://www.ebi.ac.uk/), and the Swiss-Prot database (http://expasy.org/sprot/). These above databases are dedicated to storing knowledge for all organisms, although many other more basic databases based on species of importance to the breeders still exist, including the Gramene (http://www.gramene.org/), Sgn (http://solgenomics.net/), Phytozome (http://www.phytozome.net/), which contain information that may have more specific usage for breeding programs. For instance, the "MSU Rice Genome Annotation Project" (http://rice.plantbiology.msu.edu/), the International Rice Genome Sequencing Project (IRGSP) [10], RAPdb [11], and the Oryza Genome Evaluation project [12] are primarily providing assembly, annotation, and related information of rice genome. These genomes are provided by constructing a built-in web resource for rice, including a rice species-specific genome explorer, whole-genome alignment, synteny, genetic and physical maps with genes, gene trees, ESTs and QTL positions, genetic diversity data including SNPs, and advises them on their genome sequence [13].

6.5.2 Databases for Gene Expression Datasets

With the invention of the microarray in the 1980s, it became possible to measure the abundance of all transcripts at the genomic scale. This is now known as the transcriptome. To this date, several gene expression data from such experiments have been stored in public repositories, like the EBI ArrayExpress (AE; https://www.ebi.ac.uk/arrayexpress/) and NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/), after the implementation of the "Minimum Information About a Microarray Experiment" (MIAME) standard [14]. Unlike the International Nucleotide Sequence Databases (http://www.insdc.org/), these two databases, namely AE and GEO, for gene expression have not been sharing data with each other. There have been several instances where AE had started importing GEO data in the past but have recently stopped doing so. Though we still have access to all archive data of AE in GEO, all the new data are not available for us anymore [14]. Thus, at present, researchers operating on a specific topic would need to scan both of the databases since these databases have been independently maintained. Besides that, the DNA DataBank of Japan (DDBJ) recently started another repository for the investigation of gene expression called Genomic Expression Archive (GEA; https://www.ddbj.nig.ac.jp/gea/). The GEA is a repository of functional genomics data such as genotyping SNP arrays, epigenetics, and gene expression. Genomic or DNA microarray data and sequence-based data are acceptable in the MAGE-TAB format, in strict compliance with MIAME and MINSEQE guidelines, respectively [15]. As a consequence, there is a need for the integration of these public gene expression databases. Recently, Bonon, therefore, developed an index of public gene expression databases called All Of the gene Expression (AOE). Thus, he used a database of all gene expressions (All Of the gene Expression) to get a clearer idea of the average amount of genes in a community of employees (AOE). The aim of AOE is to compile and bring together all of the gene expression results and make them all searchable. He has been maintaining the AOE website for 5 years, and it has been helpful for pursuing functional genomics studies [14].

6.5.3 Database for Gene-Interactomes, Pathways, and Ontologies

A gene interaction network is the collection of genes, each linked by an edge indicating a functional relationship between these genes. These edges are called interactions because the two genes are assumed to have either a physical connection with their gene products, e.g., proteins, or one of the genes changes or influences the function of another gene of interest [16]. "The functional products of genes, e.g., proteins, work together to achieve a particular task, and they often physically associate with each other to function or to form a more complex structure. These interactions can be long-lasting, such as forming protein complexes, or brief, when proteins modify each other such as the phosphorylation of a target protein by a protein kinase. Since these interactions are important to carry out most biological processes, knowledge about interacting proteins is crucial for understanding these biological functions, which can be easily done via studying networks of these interactions" [16].

There are other, more complicated genetic variations. So, when all of these gene variants work together, the resulting influence does not manifest itself in just one gene alone. Moreover, it does not manifest itself at all. At high throughput, we can also measure the gene combinations to help further understand this disease. There are two general categories of such interactions: synthetic lethal (Synths) interactions and suppressor (Syns) interactions. The effect is lethal as a result of two nonessential genes combining to form lethal effects, and suppressive effects occur when a lethal variance within one gene "cancels out" or is "negated" by that of another gene. Much more research needs to be done on how drugs act in the human body. This way, we can understand how they work and use this knowledge to prevent or treat diseases [17, 18].

With the involvement of high-throughput methodologies like co-immunoprecipitation followed by mass spectrometry, yeast two-hybrid (Y2H), or tandem affinity purification, studies have been performed, which help to classify physical protein-protein associations for a wide variety of species. The fundamental genetic mechanism of drug action was mapped, and its influence on molecular pathways important in many biological systems, both in humans and for organisms [16]. The expansion of the number of proteins and how they "interact" has continued over the last decade. This has led to the creation of public databases that can be shared among scientists. As in predicting enzyme-specific interactions, computational techniques are used to forecast protein-protein interactions. At some point, the use of genomic data will help us understand the complicated relationship between protein pairs [19, 20] or help us predict novel interactions we have not yet experienced [21].

All the interactions performed in the lab are recorded so that they can be made accessible to the public at large. Scholars continue to be able to utilize various databases as multiple organisms' protein-protein interactions connect these new organisms with many other organisms. A database first gathered their samples from multiple sources. Nevertheless, biomolecular interaction databases like the International Molecular Exchange Consortium now allow researchers to compare protein-protein interactions from a wide taxonomic spectrum of species. Further, these databases agree to create publicly accessible datasets in standardized formats such as MITAB or PSI-MI XML 2.5. Currently, the databases recorded in MPIDB (http://www.jcvi.org/mpidb), DIP (http://dip.doe-mbi.ucla.edu), IntAct (http://www.ebi.ac.uk/intact), MINT (http://mint.bio.uniroma2.it/mint), Pact (http://mips.gsf.de/genre/proj/mpact), MatrixDB (http://matrixdb.ibcp.fr), BioGRID (http://www.thebiogrid.org), InnateDB (http://www.innatedb.com), and BIND (http://www.blueprint.org) are actively generating relevantly large numbers of relevant documents, and provide these through the "Proteomics Standard Initiative Common Query Interface" (PSICQUIC) service. This database can contain and hold interactions all intended towards a specific organism, such as the BioGrid (https://thebiogrid.org/) database, or it can contain and hold interactions specifically targeted at a specific biological domain, such as the MatrixDB (http://matrixdb.univ-lyon1.fr/) database. However, regardless of the file format that is used, this data is available in a standards-compliant, tab-delimited, and XML format. Presently, these databases share a lot of the same documents. We still have a lot of work to do to track the complicated relationships between patients and interactions. However, once the new data entry pipelines for each healthcare system are set in place, accurate reporting across an organization will be easier to track [16].

Understanding and visualizing the networks of these connections are also essential to researchers. Recent advances have been made in the types of software. There is software for different platforms, like Cytoscape (https://cytoscape.org/), Osprey [22], Pajek [23], etc. This software can display the network by employing a graph layout algorithm and will display the network layout attributes as nodes and are visual representations in each node (e.g., protein images, coloring). Moreover, through the usage of a number of various items, such as a plug-in and a filter, it analyzes these interactions and aids in the incorporation of external data sources like gene ontology [24, 25].

6.5.4 Databases for Gene Ontology

The Gene Ontology (GO) resource is the most straightforward and commonly utilized method available in terms of identifying the roles of genes. In GO, all functional knowledge is arranged as well as represented in a form amenable to computational analysis, which is essential for modern biological research. The GO database is organized using a formal ontology by specifying groups of genes and the connections between them. GO words (such as "GO:00086467", "GO:00093381", and "GO:00093385") contain meanings that are sometimes stated as "equivalence axioms" (axioms saying that two terms are identical if they are both closely connected to the same things), since they can be computationally inferred utilizing rational reasoning. The GO framework has been carefully built over the span of 20 years by a small team of ontology developers; it is continually changing in reaction to recent scientific findings and consistently refined to reflect the most current state of biological understanding. The members of the ontology creation team include specialists on biological knowledge representation. They read the literature to validate the correctness of the representation and involve biocurators (those who study and curate biological knowledge) to collaborate alongside them to establish this representation of biological knowledge [26]. The Planteome seems to be a current database (www.planteome.org) which has provided a common collection of ontologies for use in genome, expression, and phenomic projects. Additionally, it provides ontology-based annotations for approximately 85 plant species, including a number of *O.sativa* subspecies *indica* and *japonica* rice as well as wild *Oryza* species.

6.5.5 Databases for Pathway

The word "pathway" is poorly described and may arguably, be used to characterize any sequence of action between biomolecules until a specific product is created. The Reactome project [27] is online, open access, database of biological pathways that has been curated. The reactions are ordered hierarchically, with a series of single reactions required at the lowest level and a succession of interconnected pathways at the top-level [28]. The data in Reactome are obtained from scientific literature, with information being collected by researchers, editors, reviewers, and curators. Subsequently, Reactome contains links to other databases such as Ensembl, UniProt, and KEGG [29].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database linking genetic, biochemical, as well as phenotypic details from various sources [30]. It provides knowledge regarding chromosomes and metabolic processes. The proteins and enzymes that belong to such pathways, along with information regarding genetic, molecular, and environmental mechanisms, diseases, and drug mechanisms. Many connections are given to various databases, including the UniProt database and the NCBI Entrez Gene database. Unfortunately, access to KEGG was discontinued in 2011, and we no longer have the ability to access KEGG via FTP. Instead, they only have the feature to access KEGG through their API (application program interfaces). It also inhibits system developers' ability to incorporate the KEGG pathway in the software.

WikiPathways is an open access initiative that is different from the other pathway databases [31]. The tool is part of MediaWiki and enables anyone to contribute to and manage biochemical pathways on the Wikipedia website [32]. WikiPathways encompasses several various signaling pathways involved in multiple biochemical processes over several organisms. WikiPathways is now a novel approach to preserving and processing vast amounts of genetic information in response to the public's desire to ensure and organize the data, thus ensuring its ultimate performance.

Ultimately, MetaCyc [33] is a massive systematic repository of pathways and enzymes among all aspects of existence, with the majority of evidence derived from current literature asserts that it is the most exhaustive set of metabolic pathways accessible. MetaCyc is purported to be the biggest array of curated metabolic pathways. No pathway database can ever be accurate, and lesser of around 10% of predicted genes or proteins can be mapped to a given pathway or reaction in some environmental samples [34]. Thus, it is common to use multiple complex databases and algorithms to get the conclusions that best fit the available data. Gene ontology and pathway enrichment analysis have been discussed in detail in Chap. 12.

6.6 Bioinformatics Tools in Data Mining

To date, several bioinformatics tools have also been developed which are used in various bioinformatics analysis, including sequence alignment (Chap. 7), gene identification and structure annotation (Chap. 8), phylogenetic analysis (Chap. 9), RNA structure prediction (Chap. 10), structural proteomics (Chap. 11), and gene ontology & pathway enrichment analysis (Chap. 12), high-throughput sequencing technologies (Chap. 13), DNA–Protein Interaction Analysis (Chap. 15), RNA–Protein Interaction Analysis (Chap. 16), SNP identification and discovery (Chap. 17), microsatellite markers discovery (Chap. 18), genome-wide association study (Chap. 19), expression profiling and discovery of microRNA (Chap. 20), identifying long noncoding RNA (Chap. 21), metagenomics (Chap. 23), and single-cell RNA sequencing (Chap. 25). Detailed information about each tool and its utility is described in detail in each chapter later.

6.7 Conclusion and Future Perspective

In conclusion, biological databases are life science knowledge collections collected from experimental observations, written literature, technologies for high-throughput experiments, and quantitative analysis. They provide information from study areas such as genomics, proteomics, metabolomics, microarray genes, and phylogenetics. While numerous databases and online resources for protein bioinformatics have been established to assemble and store numerous biological details, there are challenges as well as opportunities to build Next-Generation databases, including resources that facilitate the integration of data, generation of data-driven hypotheses, as well as exploration of biological information [35]. Effective storage and handling of vast quantities of data is the first obstacle that machine biologists would meet. Huge parallel disk technologies (file systems that are distributed, clustered, or parallel) were investigated, in addition to stronger hardware support. Lustre (http://lustre.opensfs.org) and “Hadoop Distributed File System (HDFS)” (http://hadoop.apache.org) are the best examples.

The collection and handling of information is only one side of the same coin. The goal of high-throughput omics studies is to translate clinical data into expertise in biomedical science and healthcare systems. We need accessible computing facilities and an effective data processing system to achieve precision medicine and improved therapies. Cloud computing appears like an inexpensive option for large-scale data processing relative to conventional HPC cluster computing. Bioinformatics research is also altering how the analysis is carried out by hosting cloud-based data storage with massive amounts of high-throughput data. Code is instead going to the data instead of transferring data to the application code. In addition, the performance of converting data into information often involves modern and powerful machine learning and data mining techniques, and analytical architectures. Apache Spark (http://spark.apache.org), for large-scale lightning-quick in-memory clustering computation, is a newly developed fast and general-purpose computing engine. It enabled a wide variety of higher-level software, along with data collection and organization, GraphX for graph processing, MLlib for machine learning, Spark SQL for SQL, and Spark Streaming for apps with scalable streaming. In Big Data analysis, the most difficult challenge is to cope with the data’s variability, variety, and uncertainty and to find a better way to incorporate them. Along with analyzing the versatility of NoSQL technology, the implementation of ontology and Semantic Web technology is another exciting area. Ontology plays a perfect function in solving the issues of the variability of data sources as a systematic, a precise description of a commonly accepted conceptualization of a topic of concern. The rapid growth and acceptance of ontologies have helped the scientific community use structured ontologies to annotate and incorporate biological and biomedical data and automate the discovery and design of web resources and workflows for bioinformatics. Linked Data infrastructure offers a means to publish and interconnect organized knowledge on the internet. Bio2RDF [36] and the EBI RDF platform [37] are active Linked Data ventures in the area of bioinformatics. Through identifying a series of basic conventions to construct RDF(s) compliant Linked Data from a diverse collection of heterogeneously structured resources derived from multiple network providers, they utilize Semantic Web technology to develop and provide the largest network of Linked Data for Life Sciences. The task of Linked Data integration is to create software that can ingest such data, extract, and display meaningful biological information in a user-friendly manner.

Sensitive site design that makes the web page appear nice on all platforms is becoming more relevant with the pervasiveness of mobile devices (tablets and phones). Protein bioinformatics databases of the next decade can provide consumers with an optimized viewing and interaction interface through a wide variety of devices utilizing technologies such as Bootstrap (http://www.getbootstrap.com), JQuery (https://www.jquery.com), and Dojo Toolkit (https://dojotoolkit.org), etc. The creation of NoSQL technology and a high-performance index and search framework such as Lucene/Solr (http://lucene.apache.org) for rapid information retrieval has also been motivated by the need for pace, particularly for web-based applications.

Conflict of Interest None.

Additional Information Fig. 6.1 (CC0 1.0) [7] have been reused under Creative Commons Attribution licenses.

References

1. Mitra S, Acharya T. Data mining: multimedia, soft computing, and bioinformatics. 1st ed. Hoboken: Wiley-Interscience; 2003. 424 p.
2. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011. 740 p.
3. Mittal S, Zaman M. A review of data mining literature. Int J Comput Sci Inform Sec. 2016;14 (11):437.
4. Ramez E, Shamkant N. Fundamentals of database system. London: 7th ed., Pearson Education; 2017. 1272 p.
5. Reeder MM. Reeder and Felson's Gamuts in radiology: comprehensive lists of roentgen differential diagnosis. New York: Springer Science & Business Media; 2013. 691 p.
6. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AIMag. 1996 Mar 15;17(3):37.
7. Holzinger A, Dehmer M, Jurisica I. Knowledge discovery and interactive data Mining in Bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics. 2014 May 16;15(6):I1.
8. Pérez-de-Castro AM, Vilanova S, Cañizares J, Pascual L, Blanca JM, Díez MJ, et al. Application of genomic tools in plant breeding. Curr Genomics. 2012 May;13(3):179–95.

9. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008 Mar;24(3):142–9.
10. Sasaki T, Burr B. International Rice genome sequencing project: the effort to completely sequence the rice genome. Curr Opin Plant Biol. 2000 Apr;3(2):138–41.
11. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. 2013 Feb;54(2):e6.
12. Song S, Tian D, Zhang Z, Hu S, Yu J. Rice genomics: over the past two decades and into the future. Genomics Proteomics Bioinformatics. 2018 Dec 1;16(6):397–404.
13. Garg P, Jaiswal P. Databases and bioinformatics tools for rice research. Curr Plant Biol. 2016 Nov 1;7–8:39–52.
14. Bono H. All of gene expression (AOE): an integrated index for public gene expression databases. PLoS One. 2020 Jan 24;15(1):e0227076.
15. Kodama Y, Mashima J, Kosuge T, Ogasawara O. DDBJ update: the genomic expression archive (GEA) for functional genomics data. Nucleic Acids Res. 2019 Jan 8;47(D1):D69–73.
16. Bebek G. Identifying gene interaction networks. Methods Mol Biol. 2012;850:483–94.
17. Avery L, Wasserman S. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet. 1992 Sep;8(9):312–6.
18. Dolma S, Lessnick SL, Hahn WC, Stockwell BR. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. Cancer Cell. 2003 Mar;3(3):285–96.
19. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 1999 Mar 16;96(6):2896–901.
20. Goh C-S, Cohen FE. Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol. 2002 Nov 15;324(1):177–92.
21. Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC Bioinformatics. 2007 Sep 13;8:335.
22. Breitkreutz B-J, Stark C, Tyers M. Osprey: a network visualization system. Genome Biol. 2003;4(3):R22.
23. Mrvar A, Batagelj V. Analysis and visualization of large networks with program package Pajek. Complex Adap Syst Model. 2016 Apr 6;4(1):6.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
25. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005 Aug 15;21(16):3448–9.
26. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330–8.
27. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2016 Jan 4;44(D1):D481–7.
28. Haw R, Stein L. Using the Reactome database. Curr Protoc Bioinformatics. 2012;38 (1):8.7.1–8.7.23.
29. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front Genet [Internet]. 2017;8. [cited 2020 Dec 26]; Available from: https://www.frontiersin.org/articles/10.3389/fgene.2017.00023/full#B23.
30. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucl Acids Res. 2008 Jan 1;36(suppl_1):D480–4.
31. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012 Jan 1;40(D1): D1301–7.
32. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008 Jul 22;6(7):e184.

33. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016 Jan 4;44(D1):D471–80.
34. Wallace RJ, Rooke JA, McKain N, Duthie C-A, Hyslop JJ, Ross DW, et al. The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics. 2015 Oct 23;16(1):839.
35. Chen C, Huang H, Wu CH. Protein bioinformatics databases and resources. Methods Mol Biol. 2017;1558:3–39.
36. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008 Oct;41(5):706–16.
37. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, et al. The EBI RDF platform: linked open data for the life sciences. Bioinformatics. 2014 May 1;30(9):1338–9.