

# Systems biology and big data analytics

Rohit Shukla<sup>1,\*</sup>, Arvind Kumar Yadav<sup>1,\*</sup>, William O. Sote<sup>2</sup>, Moacyr Comar Junior<sup>2</sup> and Tiratha Raj Singh<sup>1,3</sup>

<sup>1</sup>Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Solan, India, <sup>2</sup>Campus Centro-Oeste, Department of Biochemistry, Federal University of Sao Joao Del Rei, Minas Gerais, Brazil, <sup>3</sup>Centre of Excellence in Healthcare Technologies and Informatics (CHETI), Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Solan, India

## 25.1 Introduction

There are two types of approaches in science, first is the reductionist approach and the second is a holistic approach. The whole living biological systems can be characterized by network modeling and this approach is called a holistic approach (Ideker, Galitski, & Hood, 2001; Kitano, 2002). This can be done by combining all the genes, proteins, metabolites, etc., and can be predicted the effects of whole system. In the reductionist approach, the one entity (gene or protein or metabolite) can be characterized using the wet lab methods and their effect can be observed alone. Hence, in the era of big data, the holistic approach is more useful to characterize the perturbation of one gene/protein/drug in a whole biological system. From here the concept of systems biology originated, which deals with the holistic approach and can characterize the real effects of one gene or protein against the whole system. Systems biology is used to reveal the system-level information of an organism and the aim of systems biology can be summarized in the following points:

1. The understating of the system's all components based on its molecular level.
2. To make the system that can predict the future growth of a system in the normal environment?
3. To make the system that can predict the response of an organism due to the external stimulus?
4. The ability to predict the systems-level changes in the gene knockout, based on whole system.

The systems biology approach is a very powerful tool to reveal the role of previously undescribed components in the biological system and can decipher the multivariate relationship between different organisms (Aderem et al., 2011; Gupta, Singh, Shukla, & Misra, 2013). In systems biology, we can analyze any system ranging from an ecosystem, a single organism, tissue level, single-cell dynamics, or a group of biological entities or molecules. The biological molecules that make a comprehensive system are very diverse in the terms of their structure, function, subcellular localization, etc. Hence, designing a comprehensive system for the analysis is very critical. The major thing is that how one can design the system and its components and can construct a network and can derive meaningful information to make the final prediction and ultimately solve their hypothesis-driven research purpose is still a research problem and scientists are seeking these answers through their research at the global level (Ideker et al., 2001; Kitano, 2002).

## 25.2 Big data in general and in the context of biology

The analytics of big data is potentially revolutionary. Various leading companies, such as Facebook, Amazon, Uber, and Netflix, are revolutionarily changed their success due to the big data analytics only. The collection, creation, and analysis of the generated big data by any industry or academic institute can give at least some of the benefits, such as increase revenue by increasing the effectiveness and performances and reduce the cost of the products and processes due to less manpower. In this row, the biomedical and pharmaceutical companies and academic and scientific institutions are not resistant to this process and these are also facing data-driven challenges, such as storage, development, and extracting meaningful information.

---

\*Rohit Shukla and Arvind Kumar Yadav contributed equally to write this chapter.

The big data are described by the data scientists as four Vs (volume, velocity, veracity, and variety). Volume, that is, first V, represents the data at scale, and that data obtained from various resources with a lot of data points. In the case of biomedical data, there are next-generation techniques that are widely used to produce the high-throughput data of a sample or multiple samples, tissues, or from the whole cell. The one-time sequencing of one biological sample is not enough and cannot provide a good result hence the multiple samples with three sequencings of each sample generates a large volume of data. A large amount of biological data can be produced due to the advent of single-cell sequencing, although thousands of cells can be analyzed for each tissue or patient at a lesser depth (Wang & Song, 2017; Wu, Wang, & Wu, 2017).

The data that are in Petabyte to Exabyte scale can be stored in the big data domains. The big data domains are the specific server that is specially designed for storing a large amount of data. One billion gigabytes is equal to one exabyte and now day's most portable devices are storing the data on the gigabyte scale. The generated volumes of the data are much higher as compared to the storage volume; hence, due to this limitation, the global sum up of the data can be stored in the high-performance servers. This high amount of data is called intermediate data because it contains a lot of redundant information and the best data can be selected by quality control and data reduction processes. From the last few years, historical growth in biological data is recorded, which is almost double in each 7-year since the first illumina genome sequencing was done in 2008, and this growth is even much faster than Moore's law-based predicted growth.

Due to this faster growth in genomics, this can be compared with physics, astronomy, and social media fields, such as Facebook, YouTube, and Twitter. Various academic and research institutes started to sequence the genome of thousands of normal and disease persons, such as the Saudi Human Genome Program (Project Team, 2015) and Genomics England project (Genomics England, 2017b) (Table 25.1). The UK genome project will sequence 100,000 human genomes, which will result in 20 petabytes of data. The sequencing of several other organisms is also carrying out, which will generate a lot of data in the coming years. In the field of agriculture, thousand to millions of vegetable varieties are sequenced. A thousand varieties of rice are also sequenced in recent years (Zhu, 2012). Personalized medicine is a field that is solely based on individual genomes; hence, it is targeted to sequence the genome or exome of at least a significant portion of the world population. Thus it will largely increase the growth of biological big data (Stephens et al., 2015).

The second "V" called velocity indicates the infrastructure speed of the stored data. It will ensure the fast transferring of the data from one place to another place. The large amounts of data need large remedies for data manipulation. The four-wheeled method (truck delivery) is the best, secure, and fastest way to transfer the Exabyte scale data to this time (<https://aws.amazon.com/snowmobile/>). But in the current internet speed, if we want to upload 100 petabytes of data, it will take approximately 30 years. Although the volume and velocity are not a major issue for biological data since we can store and move the data inside the walls of the same institution until this time. In biology, the data are not always transferred through the internet. The data can be transferred from the service provider to the client through the physical storage also if it is very big.

The veracity and variety represent the other two "Vs," which are more critical than the first two "Vs." The data uncertainty is represented by the veracity. In the genomic sequence data, the biases are intrinsic and can occur due to the experimental batch effects, error rates, and different statistical models applied. The fourth and last "V" refers to the variety of the data. Variety is a natural thing in the case of biological data. The biological data comes from different domains in different forms hence it is very heterogeneous. Big data often means distinct signals and detection systems from the same source in this regard. Heterogeneity in the case of biological data is also an obstacle. Due to this integration of different sources, we can predict some novel information about unassumed results from this data.

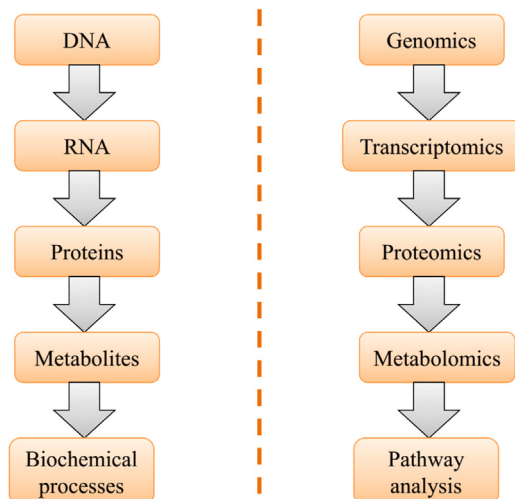
For reproducible research by using biological data: the collection, integration, availability, and organization of data are the bigger issues (Iqbal, Wallach, Khoury, Schully, & Ioannidis, 2016). If scientists can spread this data for the researchers, it will be a great contribution to science. Hence, developing such a project that involves a lot of data and that is useful for researchers is also beneficial to science, but it is a crucial task. Several diverse projects are going on, which generates a lot of data and that data is available in the public domains. We are summarizing a few of these here in Table 25.1.

### 25.3 Types of data in systems biology

Systems biology can deal with the heterogeneous type of data, which are generated by various kinds of experimental techniques from several resources. The detailed central dogma of biological data is described in Fig. 25.1. Some of the popular data types are described below in detail.

**TABLE 25.1** The table describes several International big data resources.

Project	Description	Reference
Saudi Genome Program	This project aims to identify the genomes of all the Saudi populations and analyze those using bioinformatics techniques.	<a href="#">Project Team (2015)</a>
Project baseline	This project aims to collect comprehensive health data and then their use for the prevention of disease.	<a href="#">Maxmen (2017)</a>
Genomics England	In this project, 70,000 people will be participating and 100,000 genomes will be sequenced to see the genetic variability.	<a href="#">Genomics England (2017a, 2017b)</a>
iPOP	The Integrated Personal Omics Profiling aims to provide the layout foundation for personalized medicine.	<a href="#">Li et al. (2017)</a>
PMI	This project aims to participate engage 1 million or more US participants to increase our understanding of the individual's lifestyle, genetic architecture, and environmental factors for treating the disease. The PMI Cohort Program is a landmark longitudinal research effort.	<a href="#">Collins and Varmus (2015)</a>
NextGen-Jane	( <i>Company</i> ) This project aims is to develop a method to detect that what the body of individuals is telling. They can reveal the important individual health information in their home with privacy.	<a href="#">Erickson et al. (2014)</a>
PheKB	This project aims to detect the phenotypes from electronic medical records.	<a href="#">Kirby et al. (2016)</a>
PGP	The Personal Genome Project, launched in 2005, is a global vision and alliance of projects committed to the development of public data on genomes, health, and traits.	<a href="#">Ball et al. (2012)</a>
PheWAS	This project (Phenome-wide association studies) aims to identify many phenotypes than the single nucleotide polymorphisms or other attributes.	<a href="#">Denny et al. (2010)</a>
eMERGE	This project is funded by the National Human Genome Research Institute. The aim of this project is to provide high-throughput genetic research by the combining of DNA biorepositories with electronic medical record systems for genomic medicine.	<a href="#">Lemke et al. (2010)</a>
FAIR	An educational resource that includes data and metadata standards, several databases, and data policies that is interrelated.	<a href="#">Wilkinson et al. (2016)</a>
Human disease network	It is a comprehensive resource that provides the human disease network where each node corresponds to the specific disease and the size of the nodes indicates the genes, which are associated with this disease. If the disease has common genes so these nodes (disease) are connected.	<a href="#">Goh et al. (2007)</a>
B2DK	Big Data to Knowledge is a trans-NIH initiative to promote research and development with innovative methods and technologies to optimize and accelerate the incorporation into biomedical research of big data and data science.	<a href="#">Bourne et al. (2015)</a>

**FIGURE 25.1** The first panel describes the natural biological process while the second panel describes the respective data types, which are used in systems biology analysis.

### 25.3.1 Biological sequences

The major biological information center called DNA is a double helix molecule that is made up of four nucleotide bases (adenine, thymine, guanine, and cytosine) in the form of two complementary strands from 5' to 3' and vice versa. DNA stores the information in the form of nucleotide in a sequence. Therefore it is required to determine the accurate sequences. A gene that is the smallest unit of DNA codes a particular protein. The proteins are made up of amino acids and these are essential molecules to maintain the function of an organism. The continuous three nucleotides of a gene called a codon and these codons code amino acid and this series of codons form a polypeptide chain by coding the series of amino acids. For example, the MTNWFLPSSL is a 10-amino-acid sequence and coded by the 30-character nucleotide sequence (ATGACCAATTGGTTTCTACCCCTTCCTTG).

So it proves that one codon codes one corresponding amino acid. The one amino acid can be coded by multiple codons and it gives the redundancy of the genetic codes. The start and stop codons are also found, which can start and stop the transcription process. There were differences in the use of codon among the various species and these connections can be formed between the use of codon and the biological characteristics of an organism (Kanaya, Yamada, Kinouchi, Kudo, & Ikemura, 2001; Xu et al., 2013). DNA sequencing techniques were introduced in the 1970s and different sequencing techniques have been developed since then. Due to the advancement in sequencing technology, the scientist can sequence the whole genome, transcriptome, parts of chromosomes or one full chromosome, one gene, or the whole genome. These emerged technologies save the time and cost of sequencing (Hall, 2007; Tucker, Marra, & Friedman, 2009).

### 25.3.2 Molecular structure

The prediction of the three-dimensional structure of DNA, RNA, and protein is very crucial to understand the actual function of the omics molecule. The sequence data are increasing day by day through the high-throughput sequencing technologies while the structure prediction is limited. The DNA is tightly packed inside the cell in a protein–DNA structure called chromatin. The histones are the primary protein component of the chromatin. The packaging of DNA allows us to control the binding of transcription factors (TFs) to the DNA and regulates the damage of DNA. The proteins can bind with other proteins and form the biologically active complex. Similar to DNA, RNA can also interact with protein or with each other to perform diverse biological functions. Therefore it is very important to understand the binding sites of these molecules to explore the binding pattern and for the identification of the binding pattern; a proper 3D structure is required, which can reveal the interacting residues and their binding sites. There are several experimental methods, such as X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy, which can correctly predict the 3D structure of the biological macromolecules. Nowadays there are several computational methods also available, which can accurately predict the 3D structure of the protein (Emanuelsson, Nielsen, Brunak, & von Heijne, 2000; Zhang, 2008).

The methods, which can predict the DNA and RNA structures, are also available (Reinharz, Major, & Waldispühl, 2012). Several servers, such as HADDOCK (Dominguez, Boelens, & Bonvin, 2003), Pathdock, SymmDock (Schneidman-Duhovny, Inbar, Nussinov, & Wolfson, 2005), and Hex (Macindoe, Mavridis, Venkatraman, Devignes, & Ritchie, 2010), are available to perform the molecular docking between protein and other biological macromolecules. Several computational tools are available, which can be used for the visualization and analysis of these complexes, such as UCSF Chimera (Pettersen et al., 2004) and Pymol, and all these are available free of cost for researchers.

### 25.3.3 Gene expression

The gene expression represents the information present in the gene. This is the initial step that produces the mRNAs (protein-coding RNAs) and other functional RNAs, such as rRNA or tRNA. Gene expression is a very basic process and appears in all organisms, which build the macromolecular machinery for an organism. Although all the cells contain the same gene, their expression is different in various tissues. In specific conditions, some genes are expressed in particular places while some are not expressed. The latest technologies, such as microarray, RNAseq, and ChipSeq, are used to capture the RNA expression level in different cells (Lashkari et al., 1997; Torres, Metta, Ottenwälder, & Schlötterer, 2008). If many copies of the mRNAs of a particular gene are present in the cell so it represents the highly expressed genes. This expression is measured by taking the samples from the disease and normal tissues or in different environmental conditions or during different growing stages, etc. The knowledge on gene expression is expressed in the form of a matrix where rows represent the genes and the experimental state is represented by columns.

### 25.3.4 Binding sites and domains

Various important cellular processes, such as DNA packing, RNA transcription, DNA recombination, DNA replication, and DNA repair mechanism, are initiated and controlled by the binding of a specific protein to a specific region of DNA. The specific position is called motif (pattern) and these are represented by the position weight matrix (PWM) (Ben-Gal et al., 2005). The PWM is also known as the position-specific scoring matrix or position-specific weight matrix), which provides any given fixed length substring with a weighted match. A protein called DNA-binding domain has a specific fold and pattern, which recognizes a specific part of DNA and can bind with the single or double-stranded DNA (Rudnick & Bruinsma, 1999). Hence, the protein domains or motifs and the specific part of DNA are crucial to recognizing the binding between DNA and protein. Therefore predicting and understanding the type of interaction could be beneficial to increase or decrease the affinity of these interactions.

### 25.3.5 Protein–protein interaction

In the cells, various proteins play various types of role, such as the component of cellular machinery (such as ribosomes), gene expression regulator, and enzyme catalysts. Some proteins are synthesized and perform the role in the same cell compartment while various other proteins moved from that compartment and play the key role in other compartments of the cell. The protein can work as a monomer or the multiple same chains can binds and form the homodimer, and then, it can perform the biological functions. Sometimes various proteins bind with each other and perform cellular functions. RNA polymerase is a classic example of protein complexes. The proteins are bound together with various forces and the PPI data represents the binary relationship between one protein to another protein and is widely used in systems biology to decipher the relationship between protein and protein.

### 25.3.6 Mass spectroscopy

Mass spectroscopy is a technique by which the spectra (singular spectrum) are generated to measure the molecular masses. These spectra can define the elemental composition of the given sample and can elucidate the chemical structure of the given sample, such as a peptide, metabolites, and chemical compounds. The mass spectroscopy used charged molecules or molecule fragments to measure their mass-to-charge ratio by ionizing chemical compounds (Bogges, 2001). The 2-dimensional and 3-dimensional arrays are used to represent the 2D (molecular weight vs magnitude) and 3D (molecular weight vs. magnitude vs. time) mass spectroscopy data.

### 25.3.7 Metabolic pathways

The metabolism is an important process of the cell that produces the building material and generates the energy by using several enzymes in a sequential process. All organisms receive the macronutrients from the environment for their required growth and proper functioning. After that these foods are processed by thousands of reactions. In cells, chemical reactions always occur that transfers the ions and electrons and breaks the chemical molecules. These processes or reactions are called metabolic pathways. These reactions can be considered as a biological network where nodes and edges are considered as metabolites and reactions between them. The edges can be connected with the multiple enzymes and one node can also connect with many other nodes in a biological network. The metabolic reactions follow the law of physics and chemistry; hence, if a scientist wants to make a real model of these metabolic reactions, then the scientist has to consider many physicochemical constraints (Palsson, 2006).

In summary, the biological data can be divided into four categories: first is the sequence data, second is the 3D structure data, third is the multivariate data, and the last is the network data. However, in the current scenario, the biological data are continuously increasing and require more computational power, algorithms, and tools for analysis. These data can be converted from one type to another type according to the user's need for the analysis.

## 25.4 Biological big data resources

A lot of resources are available that give the genomic and transcriptomic information about the different genes, disease and normal tissue sample, single-cell sequencing data, genetic variations, such as copy number variation, frame shift mutations, and single nucleotide polymorphism. We are describing a few global resources here in tabular form in the coming sections.

**TABLE 25.2** Genomics and transcriptomics resources.

Resource	Description	References
1000 Genome Project	The genome sequencing of 2504 individuals is done and then the mutations are predicted and stored in the database.	<a href="#">Auton et al. (2015)</a>
MGI-MPO	A comprehensive resource for associations between mouse phenotypes and gene knockouts.	<a href="#">Blake et al. (2014)</a>
ArrayExpress	It has the data produced from the DNA microarray and high-throughput sequencing techniques.	<a href="#">Kolesnikov et al. (2015)</a>
ENCODE	A complete genome map of transcription factor binding sites.	<a href="#">Rosenbloom et al. (2013)</a>
dbSNP	The database provides comprehensive information about genetic variations.	<a href="#">Sherry et al. (2001)</a>
Roadmap epigenomics	A complete genome map of Human contains putative target genes and other information.	<a href="#">Chadwick (2012)</a>
Gene expression omnibus	A comprehensive repository of the microarray and other high-throughput functional genomics data.	<a href="#">Barrett et al. (2013)</a>
Sequence read archives	It provides the raw reads generated from high-throughput sequencing techniques.	<a href="#">Leinonen, Sugawara, and Shumway (2011)</a>
GTEx	Across tissues and the population, genetic aberrations, and gene expression are profiled.	<a href="#">GTEx Consortium (2013)</a>
RNA-seq Atlas	The RNA sequencing experimental data from various human tissue samples.	<a href="#">Krupp et al. (2012)</a>
Cancer target discovery and development (CTD2)	Fitness of tested cell lines against pharmacological and genetic disturbances.	<a href="#">Basu et al. (2013)</a>
Online Mendelian inheritance in man	Database of human genes associated with diseases and disorders, and their genetic phenotypes.	<a href="#">Amberger, Bocchini, Scott, and Hamosh (2009); Amberger, Bocchini, and Hamosh, (2011)</a>
TCGA	Genomic aberrations, gene expression, DNA methylation, and expression of miRNA, tumor-sampled proteomics, clinical results.	<a href="#">Muzny et al. (2012)</a>
The gene ontology	The functional annotations of the genes to predict the role across species.	<a href="#">Ashburner et al. (2000)</a>

### 25.4.1 Genomics and transcriptomics resources

The genomics and transcriptomics are the major big data in biology and it plays a major role in systems biology. Hence, we have summarized the resources of genomics and transcriptomics data in [Table 25.2](#).

### 25.4.2 Proteomics resources

The proteomics data can tell us about the functional proteins and several other kinds of information. It is a key constituent of the systems biology analysis. We have included reference protein databases and 3D structure databases in this chapter. The resources are summarized in [Table 25.3](#).

### 25.4.3 Cellular metabolome

The metabolomics data are the key components in biological pathways and these are widely used in network generation. Hence, we have summarized the metabolome resources in [Table 25.4](#).

### 25.4.4 Protein–protein interaction databases

The protein–protein interactions are the core of the systems biology approach. These data are widely used in biological network construction; hence, we have summarized the protein–protein interaction database in [Table 25.5](#).

**TABLE 25.3** Some useful proteomics resources.

Resource	Description	References
Human Protein Reference Database	The comprehensive resource about protein–protein interaction, posttranslational modifications, and enzyme–substrate relationships.	Keshava Prasad et al. (2009)
IntAct	Protein–protein interaction database.	Hermjakob et al. (2004)
PROSITE	It provides a comprehensive resource about the protein families, domains, functional sites, associated patterns, and their association profile.	Hulo et al. (2008)
Protein Data Bank	It is the largest and widely used repository of the three-dimensional structure of biological macromolecules.	Berman et al. (2000)
UniProt	A comprehensive secondary database of protein sequence and functions.	UniProt Consortium et al. (2017)
The Human Protein Atlas	The proteomic database is generated from immunohistochemistry and immunocytochemistry experiments.	Thul and Lindskog (2018)

**TABLE 25.4** List of cellular metabolome databases.

Resource	Description	References
METLIN	The compound's description generated from mass spectroscopy and all metabolites with the detailed information can be found in this resource.	Guijas et al. (2018)
Small Molecule Pathway Database	Database of interactions of small molecules, including metabolic processes, metabolic disorders, signaling of metabolites, and pathways of drug action.	Frolkis et al. (2010)
Human Metabolome Database	The comprehensive description of the small molecule metabolites, which provides the structure, their description, disease associations, gene sequence data, and pathway information.	Wishart et al. (2007)

**TABLE 25.5** List of protein–protein interaction databases.

Resource	Description	Reference
WikiPathways	Curated biological pathway database.	Kelder et al. (2012)
BioCarta	Collections of curated cell signaling pathways.	
Reactome	A curated cell signaling pathways database.	Croft et al. (2014)
Biological Repository for Interaction Datasets (BioGRID)	A dataset of protein–protein interactions repository.	Chatr-aryamontri et al. (2013)
CORUM	Immunoprecipitation protein complexes, accompanied by mass-spectrometry.	Ruepp et al. (2008)
MINT	A detailed molecular interaction database.	Licata et al. (2012)
IntAct	A detailed molecular interaction database.	Kerrien et al. (2012)
STRING	A comprehensive protein–protein interaction network generation database.	Franceschini et al. (2013)
PhosphoSitePlus	Predict the phosphosites in a given kinase.	Hornbeck et al. (2012)
NetworKIN	Predict the phosphosites in a given kinase.	Linding et al. (2008)

**TABLE 25.6** Some important chemical compounds and drug information-related databases.

Resource	Description	Reference
ChEMBL	The best resource to find bioactive compounds against various receptors with the literature annotation.	Gaulton et al. (2012)
FAERS	Reports of side effects experienced by patients following drug treatment.	Sakaeda, Tamon, Kadoyama, and Okuno (2013)
PharmGKB-OFFSIDES	Side effects of drugs after bias correction and filtering obtained from FAERS studies.	Tatonetti, Ye, Daneshjou, and Altman (2012)
Clinicaltrials.gov	A comprehensive resource to find the detail of clinical trials.	
DrugBank	A comprehensive resource of drug and their targets, which are approved worldwide.	Wishart et al. (2008)
SIDER	Side effects of the marketed drugs.	Kuhn, Campillos, Letunic, Jensen, and Bork (2010)

**TABLE 25.7** Some other databases were used in the systems biology analysis.

Resource	Description	References
ExPASy	This is metadata that has various databases and several software tools, which can explore genomics, proteomics, structure analysis, systems biology, evolutionary biology, population genetics, transcriptomics, and medicinal chemistry.	Artimo et al. (2012)
Kyoto Encyclopedia of Genes and Genomes	A comprehensive database of pathways related to human diseases.	Kanehisa et al. (2014)
National Center for Biotechnology Information	It is a collection of various databases that can provide different types of data, such as genetic variation, and high-throughput sequencing data.	
Broad Institute Cancer Cell Line Encyclopedia	Provides a comprehensive resource about the cancer cell lines.	Barretina et al. (2012)
Cancer Genome Project Genomics of Drug Sensitivity in Cancer	Profiles of genomic aberrations, gene expression signature, and other cancer-related information.	Garnett et al. (2012)

### 25.4.5 Drug and chemical compound databases

We have also summarized the drugs and chemical compound databases in [Table 25.6](#). These databases describe the  $IC_{50}$ , drug–target interaction, and lots of other information.

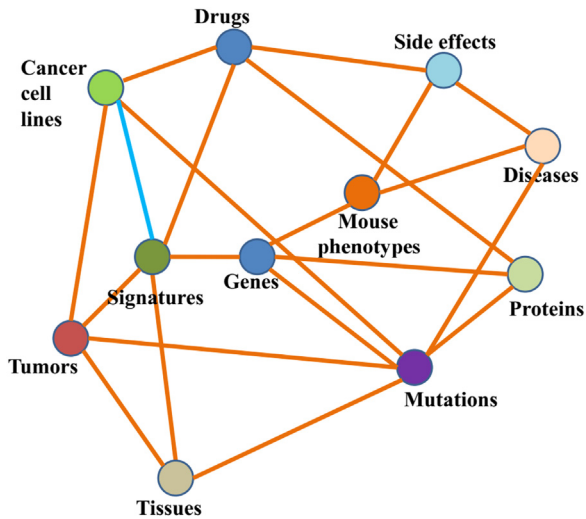
### 25.4.6 Different other databases

These are other very popular databases that serve as important resources for systems biology analysis ([Table 25.7](#)).

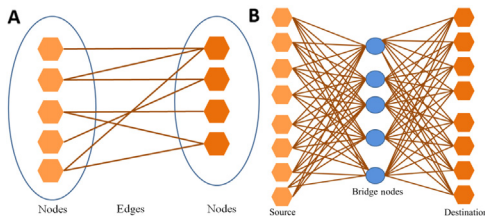
## 25.5 Network generation and its analysis from various sources of data

For the integration of data from various resources, which are described above, the single-node type networks, gene-set libraries, and bipartite graphs can be used, which are relatively simple data structures in network biology ([Fig. 25.2](#)).

In all these networks, the single-node-type network ([Fig. 25.3A](#)) is a very simple and widely used data structure for the gene–gene and protein–protein interaction network data abstraction. Various other types of networks, such as patient–patient, disease–disease, phenotype–phenotype, drug–drug, side effect–side effect, and cell line–cell line, can also be constructed based on the correlation between these or some other similarity measures, which can infer the relationship between entities in the network. The connection or edges between nodes in single node-type networks can



**FIGURE 25.2** The figure describes the complex biological mechanisms, which are interrelated with each other, and makes a complex network.

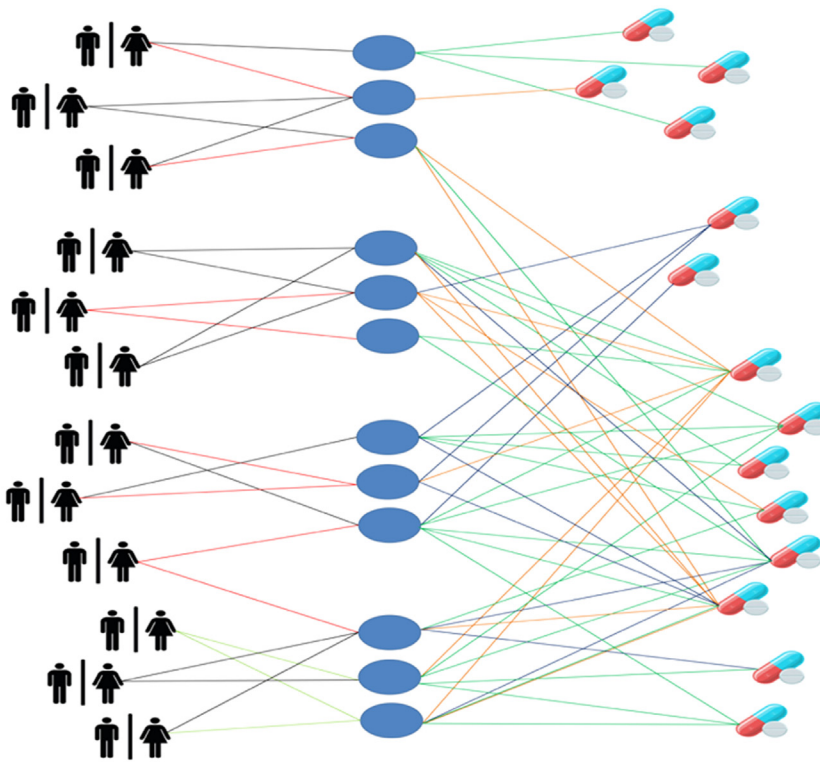


**FIGURE 25.3** Single node type and bipartite graph. (A) Figure represents the simple single node type networks where nodes are connected with the edges. (B) Figure represents the bipartite graph in which more than two nodes are involved and connected with the edges.

be weighted. Like, if an edge is connected by two proteins so it represents that both proteins have the same binding site or these proteins can show the protein–protein interaction and coexpressed mutually. The modules within these networks can be obtained by using the clustering algorithms. These module identification techniques belong to unsupervised learning. The modules can be defined as the dense region of a network that is not well connected with the other modules of the network. The module identification within the network can provide some new information regarding the new functional activity of the organism. The single node-type network is also used to find the connections between two entities. For the integration and network construction by using two or more data types, which are listed in [Table 25.1](#) and shown in [Fig. 25.1](#), the gene-set libraries can be used.

The gene-set library contains the family of the set and it is a type of data structure in systems biology and the label describes each set that is common about the genes within the set. Each set contains a different number of genes and varies in length. All sets are related to each other because labels cover a common resource belonging to a specific knowledge domain. The Kyoto Encyclopedia of Genes and Genomes have a gene-set library containing the sets of genes within each pathway with the pathway name ([Kanehisa & Goto, 2000](#)). These libraries are used to coordinate several individuals, such as illnesses, patients, medications, mouse phenotypes, and cell lines, and their side effects. The drug set libraries can be constructed by using FDA-approved drugs than genes. The drug-set enrichment analysis (DSEA) can be performed by using these types of drug-set libraries. The initial primary use of gene-set libraries was for GSEA ([Subramanian et al., 2005](#)). Provided a list of experimentally identified genes, GSEA uses gene-set libraries to classify and rank gene sets of prior information that are enriched for an experimentally identified gene input list. Therefore GSEA is a powerful tool for discovering the function of a newly discovered list of genes. However, for other kinds of purposes, these gene-set libraries may also be used. For exploring the unexpected relationship between two different types of data sources the two gene-set libraries can be combined. For example, a gene-set library has been created from the various tumor patient’s differentially expressed genes and another gene-set library is created by using the Encyclopaedia of DNA Elements (ENCODE) using ChIP-seq data, which profile the human TFs in the cancer cell lines ([ENCODE Project Consortium, 2011](#)). This ENCODE library can be used to assign putative TFs as possible drivers of gene-expression changes in the individual patient’s tumor ([Duan, Kou, Clark, Gordonov, & Ma’ayan, 2013](#)).

The bipartite or multipartite ([Fig. 25.3B](#)) is the third type of data structure that is used for network construction. This data structure fits well with the data integration goal. The bipartite graph can be used to make the connections



**FIGURE 25.4** A tripartite network that incorporates data on gene expression from cancer cell lines and drug response data for cancer cell lines from patient tumors. To recommend medications for patients, the network links groups of patients, cell lines, and drugs.

between drugs and their targets (Ma'ayan, Jenkins, Goldfarb, & Iyengar, 2007), cell lines, and drugs, which show the highest sensitivity toward the other drugs, side effects of the drugs, diseases and their associated genes, and drugs and their signature. The tri- and other multipartite graphs that allow the nontrivial identification of relationships can be identified with these bipartite graphs. For instance, bipartite drug-cell-line networks can be coupled with cell-line tumor networks, which can contribute to the possible allocation of drugs to specific patients (Duan et al., 2013) (Fig. 25.4).

The three data structures, such as gene-set library, single-node-type network, and bipartite graph, are related. For example, by the use of a gene-set library or a bipartite graph, the single-node-type network can be created. In addition, the bipartite graphs can be converted into the gene-set libraries and vice-versa. These gene-set libraries can be transported for another purpose.

## 25.6 Big data in drug repurposing and systems pharmacology

As already described in the above section, more data are better and effective to capture the status of human health and this big data needs the technology to handle, prioritize, and faster for analysis. This data and technology can be beneficial for researchers who are giving attention to drug repurposing. To make a drug is very complex, laborious, and tedious process. This complexity can be measured in the number, like for approval of new drugs it takes 10–15 years, and <12% of drugs can be approved by the clinical trial for public use. The development cost of the drugs includes the cost of failures also was 143 million dollars in 1980 and it increased to 2.6 billion dollars in the 2000s. For the same period, the industry investments in research passed from 2 to 50 billion dollars (source FDA). These data showed the complexity of the research that how traditional approaches are not working and more challenging in the drug designing pipeline. Hence, repurposing or repositioning of the drugs is came in to picture (Bansal, Srivastava, & Singh, 2018). It is a process to search the new targets for other diseases for the approved drug candidate; it does not need to conduct a clinical trial. Due to all these reasons, only \$40–\$80 million expense occurs for the repurposing of the drug, which is very low as compared to \$1–\$2 billion for developing a new drug. Drug repurposing can be done by using several computation techniques, such as computer-aided drug designing.

The failure of a drug in clinical trial stages cannot be described by simple biology while the human body is very complex. The drugs showed many off-target effects or can target multiple pathways and leads to toxicity. The concept of drug repurposing is originated from the side effect or off-target effect of the drugs. In this case, systems biology can

play a major role and find the novel targets for the approved drugs by using the network-based methods, and by this, the systems biology can play a major role in drug repurposing.

### 25.6.1 Network-based approaches for systems pharmacology

The network can better describe molecular and biochemical interactions, which are solved by experimental or hypothetical methods. We can also infer the drug and disease relationship with the network metrics (Yildirim, Goh, Cusick, Barabási, & Vidal, 2007). From the network, one can predict the novel targets against the drugs (Berger & Iyengar, 2009; Wu, Wang, & Chen, 2013). The network can be connected by many types, which we have described in an earlier section, and can make a complex visual structure that showed the interaction between the protein–protein or with other data types. Due to the invention of high graphics power, the visibility of the network is increased and can solve the purpose while the precise metrics that described the network nodes and edges that represent the gene, protein, and drugs in a complex network are also required for network analysis. Various types of metrics are developed, which can summarize the complexity of a network conveniently by using the graph theory. Some metrics are especially useful for proving the information about the individual nodes and others can well describe the edges of the network. Centrality metrics can identify the most important influential nodes of a network while every node influence can be also measured by the influence metrics (Dorogovisev & Mendes, 2014).

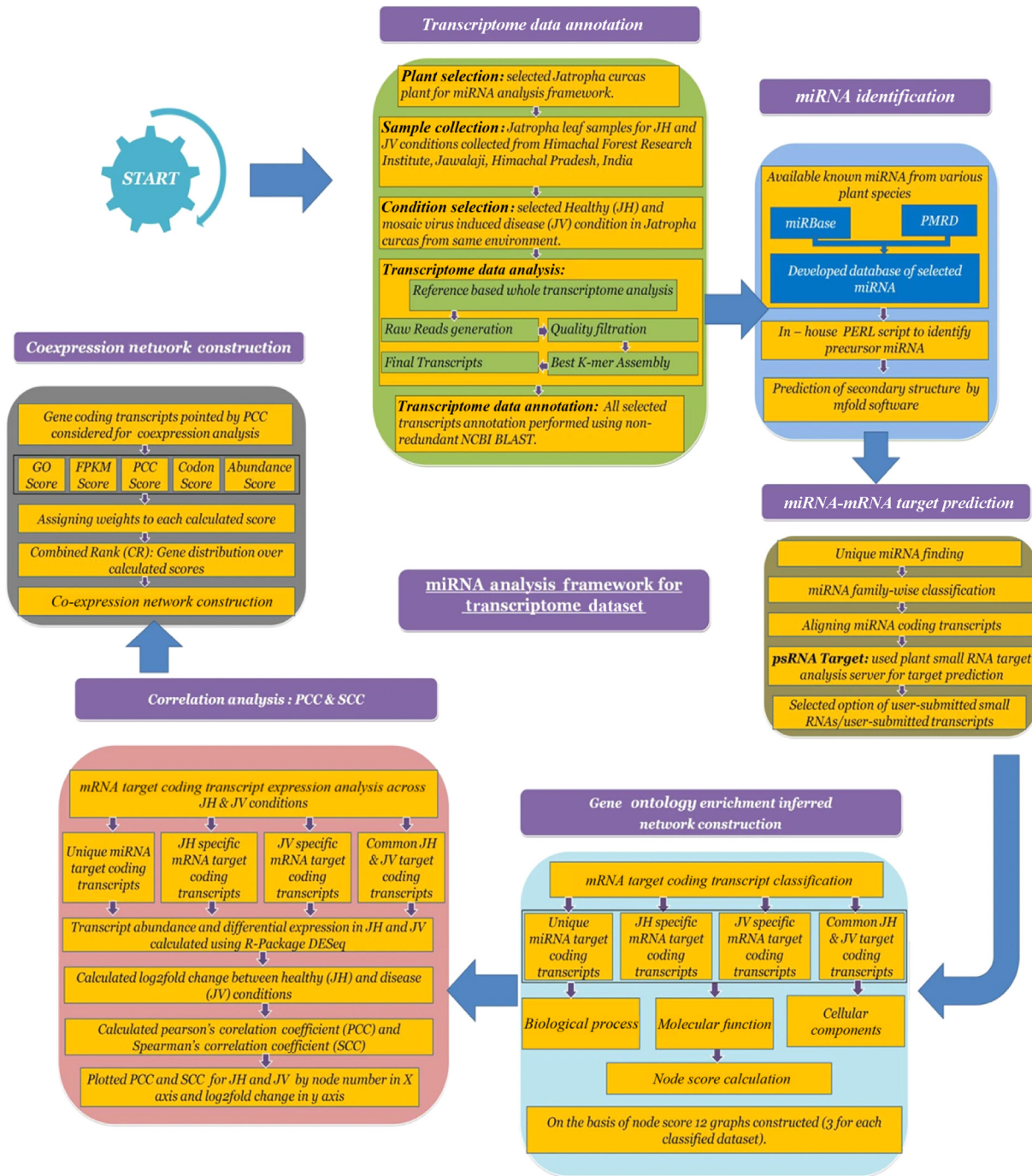
In the case of systems pharmacology, the protein interaction network or gene regulation interaction data is combined with the diverse drugs and a comprehensive network can be created to find the druggable targets. This is the very basic method to find probable targets from a network.

For a narrower emphasis, disease–gene networks are collected to classify potential targets for particular therapies by incorporating disease information in addition to biological datasets and drug information. The proteins that are involved in the disease process (Goh & Choi, 2012; Goh et al., 2007; Ozgur, Vu, Erkan, & Radev, 2008) and common biological process (Luo et al., 2007) frequently interact with each other. The disease-centric networks can be built and the information can be mined from those networks to find out the candidate target genes (Chen, Aronow, & Jegga, 2009; Köhler, Bauer, Horn, & Robinson, 2008). The other methods that do not focus on specific pathological domains can also be used for drug repositioning in which the refined or new network metrics should be developed, which use the unbiased way and can capture the essence of the potential drug targets. The traditional metrics use the actual topology of the network, such as common targets between drugs (Daminelli, Haupt, Reimann, & Schroeder, 2012) or the shortest path between the targets (Lee et al., 2012). These methods are biased due to the well-known nature of the interactomes toward the targets, which are already experimental validated and a lot of information about those genes or proteins is available in the literature.

## 25.7 Case study related to transcriptome data analysis

Bansal, Singh, and Chauhan (2017) used the systems biology approach for the analysis of miRNAs using RNAseq data. They did the module analysis, which regulates the overall pattern of the plant. They have collected the samples of *J. curcas* genotype IC561235 mature plants from the Himalayan Forest Research Institute at Jwalaji, Himachal Pradesh, India. They have taken the healthy and disease virus-infected younger apical leaves. Then, they have preprocessed the data using various steps. After that the reference genome of *J. Curcas* was used for the alignment. The Perl script was used for the miRNA identification, using miRbase and plant microRNA database. They have used a plant small RNA (psRNA) target analysis server for the miRNA target prediction. The generated result has crosschecked by using the TargetFinder Perl script. A miRNA–mRNA interaction network was constructed for further analysis. The Bipartite network analysis along with the gene ontology analysis was also done by the authors. Finally, the degree and correlation analysis was done, and a coexpression network was reconstructed by using Cytoscape software. The detailed methodology is shown in Fig. 25.5.

In the result analysis, they have identified 13 and 11 miRNAs in the *Jatropha* Viral (JV) and *Jatropha* healthy (JH), respectively, of which 8 are commons in both species. For a deeper understanding of the mechanism of resistance and disease, they identified the unique miRNAs in both the plant species where they have found miR-172, miR-414, and miR-529 in JH and miR-2910, miR-2914, miR-477, miR-f11953, and miR-f12158 in JV. For the identification of resistance mechanisms in the healthy tissue, the JH-specific miRNAs can be used. The miRNAs that are specific to JV can tell about the targets that are compromised during virus attack, such as miR-f12158, miR-f11908, and miR-f11953, which are novel miRNAs identified in this study in *Jatropha curcas*, and these miRNAs are also not experimentally validated in other plant species. The mRNA targets that correspond to the miRNAs are also identified by using the



**FIGURE 25.5** The brief methodology of the workflow for transcriptome data annotation (Bansal et al., 2017). This figure shows how big data can be used for systems biology analysis.

transcriptome analysis. A total of 39 and 61 targets are identified followed by KAAS annotation in JH and JV, respectively. For the quantification of the miRNAs target transcript in both plant species, such as JH and JV, the identified transcripts and their interactions are represented in nodes and edges respectively in a bipartite graph (Bansal et al., 2017). Only 74 and 50 nodes from JV and JH that showed the interaction were selected and a bipartite network according to miRNA–mRNA target distribution was constructed for further analysis. They have constructed the bipartite graph where they have found that the dominant effect is showed by some nodes as compared to other nodes (Fig. 25.6).

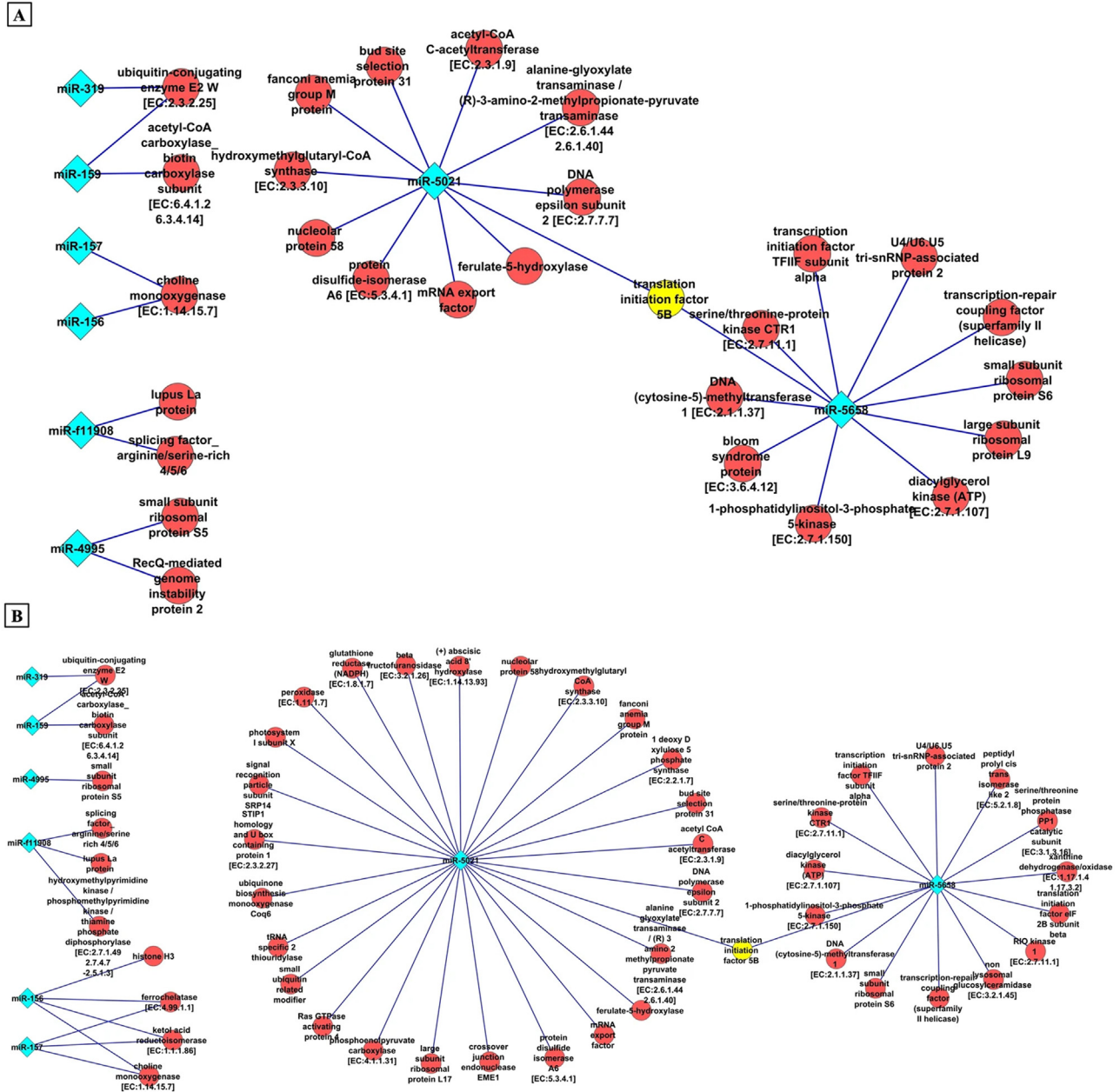


FIGURE 25.6 Bipartite network for common miRNA targets in (A) healthy and (B) diseased condition (Bansal et al., 2017).

Using these networks they did many analyses and found that miR-5021 and miR-5658 are the key miRNAs that regulate the expression of many transcripts in JV and JH.

### 25.8 Limitations in big data analysis

In addition to experimental, theoretical, and analytical sciences, data are considered the Fourth Paradigm. This is especially true in computational biology, where, for example, the “sequence first, think later” approach easily overcomes the hypothesis-driven approach. In this scenario, integration of big data is a critical process. Glue, which is called bioinformatics, can hold the various types of data together in the case of biomedical research. There are several constraints in the integration of various types of big data in the case of computational biology including systems biology. There are several types of problems in the data integration and network construction, such as a good storage capability, need a

good graphics card, improvement in the existing algorithm, construction of new algorithm, develop the kinetics laws, and set that kinetics laws in the appropriate equations, and develop the tools and plugins, which can directly integrate heterogeneous data and many more.

There is a need to develop a server that can automatically classify the information on the basis of their nature and store in a different location. For this purpose, the SQL databases are not a good choice and hence develop some non-SQL databases, which can easily store the data on the basis of their type. Hence, developing such type of automated programs that can directly classify and integrate the data is urgently required and beneficial for big data analysis in systems biology.

## 25.9 Conclusion

First, we have to understand the cell as a system and then only we can assume a whole living organism as a system. Hence, for understanding this thing the researcher should understand the function of gene/protein/metabolites in the whole system as well as at the individual level. Due to advancements in high-throughput sequencing technologies, a huge amount of biological big data are being produced, such as the whole genome of the organism, transcriptome, protein–protein interaction data, gene expression, and metabolite abundances. Due to the origination of a large amount of data, it is necessary to develop new methods and algorithms, which can handle all the data and make complex biological systems from a single cell. This requires high computational power as well as knowledge of other branches of science, such as physics, mathematics, statistics, and others. These big data analytics can be done by the use of graph theory and useful network metrics, which facilitates seeing the whole data as a biological system. However, there are a lot of methods and data resources are also available for systems biology analysis but we have to develop new and efficient methods for evaluation of new upcoming data and fit that new data in the puzzle of assays, cells, drugs, phenotypes, genes, proteins, metabolites, and so on. This chapter summarizes various systems biology resources and tools for the analysis of big data generated at an unprecedented rate at the global level. We assume that this chapter will help the students, academicians, and researchers to make a foundation in systems biology concerning various domains of data science and analysis.

## Acknowledgment

Rohit Shukla and Tiratha Raj Singh acknowledge the ICMR grant (ISRM/11(53)/2019) for providing the Senior Research Fellowship to Rohit Shukla.

## Conflict of interest

The authors declare that there are no competing interests.

## References

- Aderem, A., Adkins, J. N., Ansong, C., Galagan, J., Kaiser, S., Korth, M. J., . . . Katze, M. G. (2011). A systems biology approach to infectious disease research: Innovating the pathogen-host research paradigm. *mBio*, 2(1), e00325-10. Available from <https://doi.org/10.1128/mBio.00325-10>.
- Amberger, J., Bocchini, C., & Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human Mutation*, 32(5), 564–567. Available from <https://doi.org/10.1002/humu.21466>.
- Amberger, J., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Research*, 37(Database issue), D793–D796. Available from <https://doi.org/10.1093/nar/gkn665>.
- Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., . . . Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research*, 40(W1), W597–W603. Available from <https://doi.org/10.1093/nar/gks400>.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–29. Available from <https://doi.org/10.1038/75556>.
- Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., . . . National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. Available from <https://doi.org/10.1038/nature15393>.
- Ball, M. P., Thakuria, J. V., Zaranek, A. W., Clegg, T., Rosenbaum, A. M., Wu, X., . . . Church, G. M. (2012). A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(30), 11920–11927. Available from <https://doi.org/10.1073/pnas.1201904109>.
- Bansal, A., Singh, T. R., & Chauhan, R. S. (2017). A novel miRNA analysis framework to analyze differential biological networks. *Scientific Reports*, 7(1), 14604. Available from <https://doi.org/10.1038/s41598-017-14973-x>.

- Bansal, A., Srivastava, P. A., & Singh, T. R. (2018). An integrative approach to develop computational pipeline for drug-target interaction network analysis. *Scientific Reports*, 8(1), 10238. Available from <https://doi.org/10.1038/s41598-018-28577-6>.
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., ... Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391), 603–607. Available from <https://doi.org/10.1038/nature11003>.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomaszewski, M., ... Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Research*, 41(Database issue), D991–D995. Available from <https://doi.org/10.1093/nar/gks1193>.
- Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., ... Schreiber, S. L. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5), 1151–1161. Available from <https://doi.org/10.1016/j.cell.2013.08.003>.
- Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., ... Grosse, I. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics (Oxford, England)*, 21(11), 2657–2666. Available from <https://doi.org/10.1093/bioinformatics/bti410>.
- Berger, S. I., & Iyengar, R. (2009). Network analyses in systems pharmacology. *Bioinformatics (Oxford, England)*, 25(19), 2466–2472. Available from <https://doi.org/10.1093/bioinformatics/btp465>.
- 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. Available from <https://doi.org/10.1093/nar/28.1.235>.
- Blake, J. A., Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., & Mouse Genome Database Group. (2014). The Mouse Genome Database: Integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Research*, 42(Database issue), D810–D817. Available from <https://doi.org/10.1093/nar/gkt1225>.
- Boggess, B. (2001). Mass spectrometry desk reference (Sparkman, O. David). *Journal of Chemical Education*, 78(2), 168. Available from <https://doi.org/10.1021/ed078p168.2>.
- Bourne, P. E., Bonazzi, V., Dunn, M., Green, E. D., Guyer, M., Komatsoulis, G., ... Russell, B. (2015). The NIH Big Data to Knowledge (BD2K) initiative. *Journal of the American Medical Informatics Association*, 22(6), 1114. Available from <https://doi.org/10.1093/jamia/ocv136>.
- Chadwick, L. H. (2012). The NIH Roadmap Epigenomics Program data resource. *Epigenomics*, 4(3), 317–324. Available from <https://doi.org/10.2217/epi.12.18>.
- Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., ... Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1), D816–D823. Available from <https://doi.org/10.1093/nar/gks1158>.
- Chen, J., Aronow, B. J., & Jegga, A. G. (2009). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10, 73. Available from <https://doi.org/10.1186/1471-2105-10-73>.
- Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *The New England Journal of Medicine*, 372(9), 793–795. Available from <https://doi.org/10.1056/NEJMp1500523>.
- Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., ... D'Eustachio, P. (2014). The reactome pathway knowledgebase. *Nucleic Acids Research*, 42(Database issue), D472–D477. Available from <https://doi.org/10.1093/nar/gkt1102>.
- Daminelli, S., Haupt, V. J., Reimann, M., & Schroeder, M. (2012). Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network. *Integrative Biology: Quantitative Biosciences from Nano to Macro*, 4(7), 778–788. Available from <https://doi.org/10.1039/c2ib00154c>.
- Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., ... Crawford, D. C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)*, 26(9), 1205–1210. Available from <https://doi.org/10.1093/bioinformatics/btq126>.
- Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A protein – protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7), 1731–1737. Available from <https://doi.org/10.1021/ja026939x>.
- Dorogovisev, S. N., & Mendes, J. F. F. (2014). Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press.
- Duan, Q., Kou, Y., Clark, N. R., Gordonov, S., & Ma'ayan, A. (2013). Metasignatures identify two major subtypes of breast cancer. *CPT: Pharmacometrics and Systems Pharmacology*, 2(3), e35. Available from <https://doi.org/10.1038/psp.2013.11>.
- Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, 300(4), 1005–1016. Available from <https://doi.org/10.1006/jmbi.2000.3903>.
- ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, 9(4), e1001046. Available from <https://doi.org/10.1371/journal.pbio.1001046>.
- Erickson, B. K., Kinde, I., Dobbin, Z. C., Wang, Y., Martin, J. Y., Alvarez, R. D., ... Landen, C. N. (2014). Detection of somatic TP53 mutations in tampons of patients with high-grade serous ovarian cancer. *Obstetrics and Gynecology*, 124(5), 881–885. Available from <https://doi.org/10.1097/AOG.0000000000000484>.
- Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., ... Jensen, L. J. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database issue), D808–815. Available from <https://doi.org/10.1093/nar/gks1094>.
- Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., ... Wishart, D. S. (2010). SMPDB: The small molecule pathway database. *Nucleic Acids Research*, 38(Database issue), D480–D487. Available from <https://doi.org/10.1093/nar/gkp1002>.
- Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., ... Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391), 570–575. Available from <https://doi.org/10.1038/nature11005>.

- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., . . . Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue), D1100–D1107. Available from <https://doi.org/10.1093/nar/gkr777>.
- Genomics England. (2017a). The National Genomics Research and Healthcare Knowledgebase. <<https://doi.org/10.6084/m9.figshare.4530893.v5>>.
- Genomics England. (2017b, March 10). National Genomic Research Library. <<https://www.genomicsengland.co.uk/national-genomic-research-library/>>.
- Goh, K.-I., & Choi, I.-G. (2012). Exploring the human diseaseome: The human disease network. *Briefings in Functional Genomics*, 11(6), 533–542. Available from <https://doi.org/10.1093/bfpg/els032>.
- Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21), 8685–8690. Available from <https://doi.org/10.1073/pnas.0701361104>.
- GTE Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580–585. Available from <https://doi.org/10.1038/ng.2653>.
- Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., . . . Siuzdak, G. (2018). METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical Chemistry*, 90(5), 3156–3164. Available from <https://doi.org/10.1021/acs.analchem.7b04424>.
- Gupta, M. K., Singh, D. B., Shukla, R., & Misra, K. (2013). A comprehensive metabolic modeling of thyroid pathway in relation to thyroid pathophysiology and therapeutics. *OMICS*, 17(11), 584–593. Available from <https://doi.org/10.1089/omi.2013.0007>.
- Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology*, 210(Pt 9), 1518–1525. Available from <https://doi.org/10.1242/jeb.001370>.
- Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., . . . Apweiler, R. (2004). IntAct: An open source molecular interaction database. *Nucleic Acids Research*, 32(Database issue), D452–D455. Available from <https://doi.org/10.1093/nar/gkh052>.
- Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., . . . Sullivan, M. (2012). PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*, 40(Database issue), D261–D270. Available from <https://doi.org/10.1093/nar/gkr1122>.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuče, B. A., de Castro, E., . . . Sigrist, C. J. A. (2008). The 20 years of PROSITE. *Nucleic Acids Research*, 36(Database issue), D245–D249. Available from <https://doi.org/10.1093/nar/gkm977>.
- Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, 2, 343–372. Available from <https://doi.org/10.1146/annurev.genom.2.1.343>.
- Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS Biology*, 14(1), e1002333. Available from <https://doi.org/10.1371/journal.pbio.1002333>.
- Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T. (2001). Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of Molecular Evolution*, 53(4–5), 290–298. Available from <https://doi.org/10.1007/s002390010219>.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30. Available from <https://doi.org/10.1093/nar/28.1.27>.
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research*, 42(Database issue), D199–D205. Available from <https://doi.org/10.1093/nar/gkt1076>.
- Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2012). WikiPathways: Building research communities on biological pathways. *Nucleic Acids Research*, 40(Database issue), D1301–D1307. Available from <https://doi.org/10.1093/nar/gkr1074>.
- Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., . . . Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(Database issue), D841–D846. Available from <https://doi.org/10.1093/nar/gkr1088>.
- Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., . . . Pandey, A. (2009). Human Protein Reference Database—2009 update. *Nucleic Acids Research*, 37(Database issue), D767–D772. Available from <https://doi.org/10.1093/nar/gkn892>.
- Kirby, J. C., Speltz, P., Rasmussen, L. V., Basford, M., Gottesman, O., Peissig, P. L., . . . Denny, J. C. (2016). PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association: JAMIA*, 23(6), 1046–1052. Available from <https://doi.org/10.1093/jamia/ocv202>.
- Kitano, H. (2002). Systems biology: A brief overview. *Science (New York, N.Y.)*, 295(5560), 1662–1664. Available from <https://doi.org/10.1126/science.1069492>.
- Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4), 949–958. Available from <https://doi.org/10.1016/j.ajhg.2008.02.013>.
- Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., . . . Brazma, A. (2015). ArrayExpress update—Simplifying data submissions. *Nucleic Acids Research*, 43(Database issue), D1113–D1116. Available from <https://doi.org/10.1093/nar/gku1057>.
- Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., & Teufel, A. (2012). RNA-Seq Atlas—A reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics (Oxford, England)*, 28(8), 1184–1185. Available from <https://doi.org/10.1093/bioinformatics/bts084>.
- Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6, 343. Available from <https://doi.org/10.1038/msb.2009.98>.
- Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., . . . Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24), 13057–13062. Available from <https://doi.org/10.1073/pnas.94.24.13057>.

- Lee, H. S., Bae, T., Lee, J.-H., Kim, D. G., Oh, Y. S., Jang, Y., . . . Kim, S. (2012). Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Systems Biology*, 6, 80. Available from <https://doi.org/10.1186/1752-0509-6-80>.
- Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, 39(Database issue), D19–D21. Available from <https://doi.org/10.1093/nar/gkq1019>.
- Lemke, A. A., Wu, J. T., Waudby, C., Pulley, J., Somkin, C. P., & Trinidad, S. B. (2010). Community engagement in biobanking: Experiences from the eMERGE Network. *Genomics, Society, and Policy*, 6(3), 50. Available from <https://doi.org/10.1186/1746-5354-6-3-50>.
- Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schüssler-Fioreza Rose, S. M., . . . Snyder, M. P. (2017). Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biology*, 15(1), e2001402. Available from <https://doi.org/10.1371/journal.pbio.2001402>.
- Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., . . . Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(Database issue), D857–D861. Available from <https://doi.org/10.1093/nar/gkr930>.
- Linding, R., Jensen, L. J., Pasculescu, A., Olhovskiy, M., Colwill, K., Bork, P., . . . Pawson, T. (2008). NetworKIN: A resource for exploring cellular phosphorylation networks. *Nucleic Acids Research*, 36(Database issue), D695–D699. Available from <https://doi.org/10.1093/nar/gkm902>.
- Luo, F., Yang, Y., Chen, C.-F., Chang, R., Zhou, J., & Scheuermann, R. H. (2007). Modular organization of protein interaction networks. *Bioinformatics (Oxford, England)*, 23(2), 207–214. Available from <https://doi.org/10.1093/bioinformatics/btl562>.
- Ma'ayan, A., Jenkins, S. L., Goldfarb, J., & Iyengar, R. (2007). Network analysis of FDA approved drugs and their targets. *The Mount Sinai Journal of Medicine, New York*, 74(1), 27–32. Available from <https://doi.org/10.1002/msj.20002>.
- Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.-D., & Ritchie, D. W. (2010). HexServer: An FFT-based protein docking server powered by graphics processors. *Nucleic Acids Research*, 38(Web Server issue), W445–W449. Available from <https://doi.org/10.1093/nar/gkq311>.
- Maxmen, A. (2017). Google spin-off deploys wearable electronics for huge health study. *Nature*, 547(7661), 13–14. Available from <https://doi.org/10.1038/547013a>.
- Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., . . . Tissue source sites and disease working group. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), 330–337. Available from <https://doi.org/10.1038/nature11252>.
- Ozgur, A., Vu, T., Erkan, G., & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics (Oxford, England)*, 24(13), i277–i285. Available from <https://doi.org/10.1093/bioinformatics/btn182>.
- Palsson, B. O. (2006). *Systems biology: Properties of reconstructed networks*. Cambridge University Press.
- Pettersen, 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. Available from <https://doi.org/10.1002/jcc.20084>.
- Project Team, S. G. (2015). The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease. *IEEE Pulse*, 6(6), 22–26. Available from <https://doi.org/10.1109/MPUL.2015.2476541>.
- Reinharz, V., Major, F., & Waldspühl, J. (2012). Towards 3D structure prediction of large RNA molecules: An integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics (Oxford, England)*, 28(12), i207–i214. Available from <https://doi.org/10.1093/bioinformatics/bts226>.
- Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., . . . Kent, W. J. (2013). ENCODE data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Research*, 41(Database issue), D56–D63. Available from <https://doi.org/10.1093/nar/gks1172>.
- Rudnick, J., & Bruinsma, R. (1999). DNA-Protein Cooperative Binding through Variable-Range Elastic Coupling. *Biophysical Journal*, 76(4), 1725–1733. Available from [https://doi.org/10.1016/S0006-3495\(99\)77334-0](https://doi.org/10.1016/S0006-3495(99)77334-0).
- Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., . . . Mewes, H. W. (2008). CORUM: The comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(Database issue), D646–D650. Available from <https://doi.org/10.1093/nar/gkm936>.
- Sakaeda, T., Tamon, A., Kadoyama, K., & Okuno, Y. (2013). Data mining of the public version of the FDA Adverse Event Reporting System. *International Journal of Medical Sciences*, 10(7), 796–803. Available from <https://doi.org/10.7150/ijms.6048>.
- Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005). PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Research*, 33, W363–W367, Web Server issue. Available from <https://doi.org/10.1093/nar/gki481>.
- Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308–311. Available from <https://doi.org/10.1093/nar/29.1.308>.
- Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., . . . Robinson, G. E. (2015). Big Data: Astronomical or genetical? *PLoS Biology*, 13(7), e1002195. Available from <https://doi.org/10.1371/journal.pbio.1002195>.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. Available from <https://doi.org/10.1073/pnas.0506580102>.
- Tatonetti, N. P., Ye, P. P., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125), 125ra31. Available from <https://doi.org/10.1126/scitranslmed.3003377>.
- Thul, P. J., & Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein Science: A Publication of the Protein Society*, 27(1), 233–244. Available from <https://doi.org/10.1002/pro.3307>.

- Torres, T. T., Metta, M., Ottenwalder, B., & Schlotterer, C. (2008). Gene expression profiling by massively parallel sequencing. *Genome Research*, 18(1), 172–177. Available from <https://doi.org/10.1101/gr.6984908>.
- Tucker, T., Marra, M., & Friedman, J. M. (2009). Massively parallel sequencing: The next big thing in genetic medicine. *American Journal of Human Genetics*, 85(2), 142–154. Available from <https://doi.org/10.1016/j.ajhg.2009.06.022>.
- UniProt Consortium, T., Bateman, A., Martin, M. J., O’Donovan, C., Magrane, M., Alpi, E., . . . Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. Available from <https://doi.org/10.1093/nar/gkw1099>.
- Wang, J., & Song, Y. (2017). Single cell sequencing: A distinct new field. *Clinical and Translational Medicine*, 6(1), 10. Available from <https://doi.org/10.1186/s40169-017-0139-4>.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. Available from <https://doi.org/10.1038/sdata.2016.18>.
- Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., . . . Hassanali, M. (2008). DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Database issue), D901–D906. Available from <https://doi.org/10.1093/nar/gkm958>.
- Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., . . . Querengesser, L. (2007). HMDB: The Human Metabolome Database. *Nucleic Acids Research*, 35(Database issue), D521–D526. Available from <https://doi.org/10.1093/nar/gkl923>.
- Wu, H., Wang, C., & Wu, S. (2017). Single-cell sequencing for drug discovery and drug development. *Current Topics in Medicinal Chemistry*, 17(15), 1769–1777. Available from <https://doi.org/10.2174/1568026617666161116145358>.
- Wu, Z., Wang, Y., & Chen, L. (2013). Network-based drug repositioning. *Molecular Biosystems*, 9(6), 1268–1281. Available from <https://doi.org/10.1039/C3MB25382A>.
- Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y., & Johnson, C. H. (2013). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature*, 495(7439), 116–120. Available from <https://doi.org/10.1038/nature11942>.
- Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabasi, A.-L., & Vidal, M. (2007). Drug-target network. *Nature Biotechnology*, 25(10), 1119–1126. Available from <https://doi.org/10.1038/nbt1338>.
- Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3), 342–348. Available from <https://doi.org/10.1016/j.sbi.2008.02.004>.
- Zhu, J. (2012). A year of great leaps in genome research. *Genome Medicine*, 4(1), 4. Available from <https://doi.org/10.1186/gm303>.