Romana Ishrat *Editor*

Biological Networks in Human Health and Disease

Springer

Biological Networks in Human Health and Disease

Romana Ishrat
Editor

Biological Networks in Human Health and Disease

Springer

Editor
Romana Ishrat
Centre for Interdisciplinary Research in Basic Sciences
Jamia Millia Islamia University
New Delhi, India

ISBN 978-981-99-4241-1 ISBN 978-981-99-4242-8 (eBook)
https://doi.org/10.1007/978-981-99-4242-8

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Advances in high-throughput biotechnologies have led to the generation of huge amounts of biomedical data that open a new research horizon from a reductionist view to a more complex understanding of the biological systems. To describe complex interactions and regulatory mechanisms behind biological systems, biological networks are employed to represent all relevant interactions taking place in biological systems.

In networks, molecules (genes, proteins, metabolites) are reduced to a series of nodes connected by edges. Edges represent the pairwise relationships and interactions between two molecules within the same network. Molecular networks have become extremely popular and have been used in every area of biology to model, for example, transcriptional regulation mechanisms and physical protein-protein interactions. Network theory offers a versatile and general toolbox for a framework for investigating biological systems ranging from the molecular to the global scale level. A key factor for the success of network theory in biomedical applications is that many structural network characteristics can be related to the functional properties of the respective biological system. The biological network also leads to a wide range of applications, such as pathways related to a disease that can unveil how the disease acts and provide novel tentative drug targets. In addition, it can also help predict the responses to disease and can be useful for novel drug development and treatments.

In this book, the authors discuss various network theoretic and data analytics approaches used to analyze biological networks with respect to available tools, technologies, standards, algorithms, and databases for generating, representing, and analyzing graphical data.

New Delhi, India — Romana Ishrat

Contents

Editor and Contributors

About the Editor

Romana Ishrat is an Professor at the Centre for Interdisciplinary Research in Basic Sciences at Jamia Millia Islamia University in New Delhi, India. She has extensive experience in teaching and research in the fields of computer science, machine learning, bioinformatics, and related areas. Her research focuses on developing and utilizing novel statistical and computational methods to address genetics and genomics problems related to complex diseases. Dr. Ishrat has published over 60 research articles in prestigious journals such as Springer, Elsevier, Frontiers, Willey, and Oxford Press in the areas of network biology, transcriptional networks, systems biology, data and text mining.

Contributors

Mohd Murshad Ahmed Singapore Institute of Clinical Sciences (SICS), Singapore, Singapore

Aftab Alam Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Rafat Ali Department of Biosciences, Jamia Millia Islamia, New Delhi, India

Shahnawaz Ali CGTRM, Kings College London, London, UK

Sadik Bay Research Institute for Health Sciences and Technologies (SABITA), Istanbul Medipol University, Istanbul, Turkey

Anam Farooqui Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Nikhat Imam Department of Mathematics, Institute of Computer Science and Information Technology, Magadh University, Bodh Gaya, India

Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Riddhi Jangid School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

Nida Jamil Khan Department of Biosciences, Jamia Millia Islamia, New Delhi, India

Najma Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Faizan Siddiqui International Medical Faculty, Osh State University, Osh, Kyrgyzstan

Mohd Faizan Siddiqui International Medical Faculty, Osh State University, Osh, Kyrgyzstan

Gajendra Pratap Singh School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

Pallavi Somvanshi School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

Safia Tazyeen Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Okan Yildirim Department of Chemical Biology Otto-Hahn-Strasse, Max Planck Institute of Molecular Physiology, Dortmund, Germany

Graph Theory in the Biological Networks

1

Riddhi Jangid, Pallavi Somvanshi,
and Gajendra Pratap Singh

Abstract

Graph theory is a mathematical tool widely used to study many different areas today. In this chapter, we demonstrate how the basic graph theory concepts can be applied to study molecular biology and its insights. We first introduced graph preliminaries and then used the theory by representing molecular networks using graphs. The depiction has been made by using examples and figures for a better understanding of a beginner. The chapter gives a good connection and motivation to study biology and mathematics together.

Keywords

Biological data · Biological networks · Graph theory · Molecular interactions

1.1 Introduction

Graph theory is a branch of the study in discrete mathematics that expresses many different networks and structures in different fields like social network analysis and biological networks. We use interchangeably (very often) terms "graph" and "network" in the chapter. A graph is nothing but a set of points linked with a set of lines. These points in a graph depict the entities that we want to model and the lines in the graph represent the connection between the points in our network. For instance, suppose we want to understand the network topology of DNA then we use the concept of an n-dimensional De Bruijn graph in graph theory and can create a DNA

R. Jangid · P. Somvanshi · G. P. Singh (✉)
School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India
e-mail: psomvanshi@mail.jnu.ac.in; gajendra@mail.jnu.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_1

graph for any sequence (Jafarzadeh and Iranmanesh 2016). In the next section, we study the basic definitions of graph theory in order to understand how a graph can be studied in the context of Biology.

1.2 Basic Concepts of Graph Theory

1.2.1 Graph

A graph consists of a non-empty set of elements (called *vertices, points, junctions*) that are joined with directed/undirected lines (called *edges*). The collection of these lines forms an edge set that contains ordered/unordered pairs of vertices. Hence, we define a graph $G := (V(G), E(G))$ where $V(G)$ and $E(G)$ denotes the set of vertices and edges, respectively. $E(G)$ consisting of only ordered pair of vertices is called a *directed graph (digraph)* and with that of only unordered pair of vertices is called an *undirected Graph* (Barnes and Harary 1983; Gupta 2008).

Example:

In Fig. 1.1, the vertex set is $\{A, B, C, D, E, F\}$ while the edge set for Fig. 1.1a is $\{(A, D), (C, D), (A, B), (B, C), (A, E), (E, F), (C, F)\}$ where the order of elements can be altered while in Fig. 1.1b it is strictly in the order $\{(A, D), (D, C), (A, B), (B, C), (A, E), (E, F), (F, C)\}$. It is easy to note for the undirected graph that every edge can be written in two different orders unlike in directed graphs. For instance, (C, D) and (D, C) are equivalent in the undirected graph.

In biological context, directed graphs are used in the modeling of transcriptional regulatory networks and metabolic networks and in the study of neuronal networks. Directed graph because the nodes representing genes in the transcriptional regulatory networks will have a natural direction associated while modeling the interaction from genes X to Y. In the case of undirected edges, the study of protein–protein interaction networks describes the physical interactions among the organism's proteome.

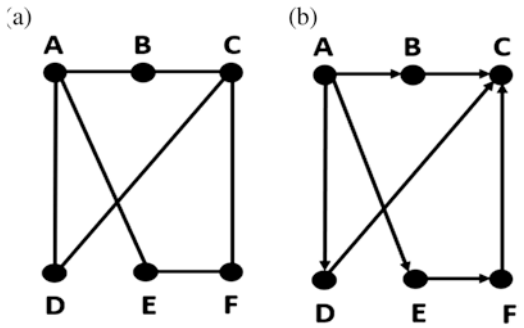

Fig. 1.1 (**a**) Undirected graph and (**b**) Directed graph with six vertices and seven edges

1.2.2 Degree of a Vertex

It is defined as the total number of edges associated with the vertex in a graph. In directed graphs, it is further classified as the Indegree or Outdegree of a vertex. We say Indegree and Outdegree are the number of edges incoming to the vertex and the number of edges outgoing from the vertex, respectively.

In the above example, the degree of every vertex in Fig. 1.1a can be written as: $deg(A) = deg(C) = 3$ and $deg(B) = deg(D) = deg(E) = deg(F) = 2$ while in Fig. 1.1b; $indeg(A) = 0$, $indeg(B) = 1$, $indeg(C) = 3$, $indeg(D) = 1$, $indeg(E) = 1$, and $indeg(F) = 1$. Also, $outdeg(A) = 3$, $outdeg(B) = 1$, $outdeg(C) = 0$, $outdeg(D) = 1$, $outdeg(E) = 1$ and $outdeg(F) = 1$.

1.2.3 Representation of a Graph

When it comes to representation, we represent any graph with its *Adjacency matrix* and *Incidence matrix*. If there is any graph with p vertices and q edges, then the adjacency matrix $M_A = [a]_{ij}$ is of order $p \times p$ while the incidence matrix $I_A = [b]_{ij}$ is of order $p \times q$. We define them as, wherever there exists any edge among the vertex, we place the value 1 otherwise, 0. In the case of an undirected graph, the adjacency matrix is obviously a symmetric matrix unlike in the digraph. For the incidence matrix in a digraph, +1 represents the direction of the edge outgoing from a vertex while −1 depicts the direction of the edge incoming to a vertex.

These adjacency matrices also give us information on the degree of vertices in the graph. For the example above, the adjacency matrices and incidence matrices for Fig. 1.1. are given in Figs. 1.2a, b and 1.3a, b. Here, we can also find the degree of respective vertices in an undirected graph by their row sums while for a directed graph, we can find the outdegree and indegree by row sums and column sums, respectively.

(a)

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Fig. 1.2 (**a**) Adjacency matrix and (**b**) Incidence matrix for Fig. 1.1a

(a)

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Fig. 1.3 (**a**) Adjacency matrix and (**b**) Incidence matrix for Fig. 1.1b

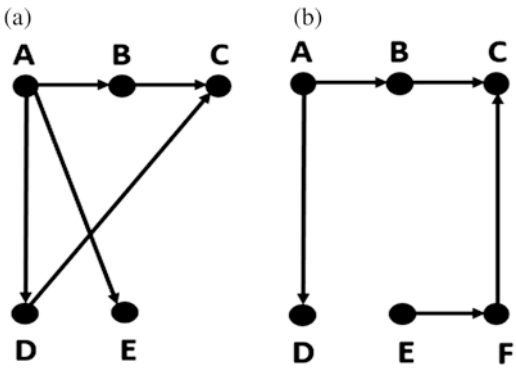

Fig. 1.4 (**a**) An example of a subgraph of *G*. (**b**) An example of a spanning subgraph of *G*

1.2.4 Subgraph

A graph $G' := (V(G'), E(G'))$ is said to be a subgraph of G if and only if the $V(G')$ and $E(G')$ of G' are the subsets of the $V(G)$ and $E(G)$ of G, respectively. Further, if the $V(G') = V(G)$ but the $E(G')$ of G' is a subset of the $E(G)$ of G, then it is called spanning subgraph (see Fig. 1.4a, b).

1.3 Graph Algorithms

Theorem 1 *The sum of degrees of all the vertices in any graph G, is even.*

Proof We know that every edge in a graph G contributes 2 degrees to the graph while 1 degree to every vertex to which it is adjacent. So, we can write:

$$2(q) = \deg(v_1) + \deg(v_2) + \ldots + \deg(v_p),$$

where p and q denotes the number of vertices and edges, respectively, in the graph G. Hence, $2(q)$ is always even.

Theorem 2 *In any graph, there are always even numbers of vertices of odd degrees.*

Proof Let us consider a graph G having both degrees of vertices, even and odd. Diving the vertices into two groups, one having even degree vertices and the other one with odd degree vertices such that $G_1 = deg\ (v_{1e}) + deg\ (v_{2e}) + \ldots + deg\ (v_{pe})$ and $G_2 = deg\ (v_{1o}) + deg\ (v_{2o}) + \ldots + deg\ (v_{po})$. By Theorem 1 we know that the sum of degrees of all the vertices is even, hence $G_1 + G_2$ is also even. This further implies that since G_1 is already even, G_2 must be even. Hence, proof.

1.4 Complex Graph Models

Graphs of different types are studied according to the complexity of literature requirements. These different types of graphs are defined as follows.

1.4.1 Bipartite Graph

Bipartite graph is that type of non-empty graph in which the set of vertices is partitioned into two disjoint subsets, say A and B such that every edge joins the vertex from set A to set B and vice versa. A directed bipartite graph is that type of bipartite graph which is directed also (see Fig. 1.5).

1.4.2 Complete Graph

Joining any vertex with every other vertex using an edge in a graph, such a graph is called a complete graph. We denote such graphs using K_p, where p denotes the number of vertices in the graph.

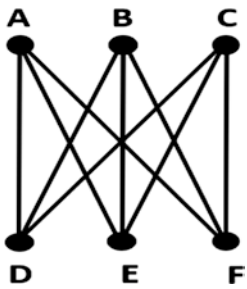

Fig. 1.5 Example of a complete Bipartite graph $K_{3,3}$

1.4.3 Weighted Graph

What if we assign some non-negative integer values to the edges of the graph? In that case, we will call such a graph a weighted graph in which non-negative integers (called weights) are attached to the edges.

1.4.4 Eulerian Graph

In a graph starting from a vertex, if we cover every edge exactly once even after repeating the vertices and end with the same initial vertex, then such type of graph is called Eulerian Graph. In Fig. 1.1, the graphs are not Eulerian. One important result for Eulerian graphs is—Every connected graph is Eulerian whose all the vertices are of even degree.

1.4.5 Hamiltonian Graph

If instead of edges we wish to cover every vertex exactly once in any graph, then such type of graph is called a Hamiltonian graph. In this case, all edges may not be included. For example, Fig. 1.1a is a Hamiltonian graph since moving from D→A→B→C→F→E covers all the vertices exactly once.

1.4.6 Regular Graph

A graph such that all the vertices in the graph have the same degree k is known as a k-regular graph.

1.4.7 Planar Graph

A graph that can be drawn on paper without crossing an edge and without lifting a pen even once is called a planar graph.

1.5 Fundamentals of Network Theory and Its Characteristics

We illustrate here the fundamentals of network theory with respect to the biological context.

1.5.1 Biological Networks

In the context of network biology, there are different questions that can be asked of graphs like:

1. Whether there is any way to produce a metabolite X from A?
2. How long is a particular chain to X from A?
3. Whether all the proteins are connected to others by any path?
4. Which is the most influencing protein in any network?

We will look for answers to such questions using the graph theoretical concepts that we discussed above. Showing the existence of a path in a graph/network solves our first query while the second problem is the shortest path problem in graph theory. For the third query, one can look for the identification of connected components in a graph/network. And in the last one, we can solve it using looking for the most connected nodes in the graph/network. In such problems, we study graph algorithms that are applicable and used in many different areas (Jafarzadeh and Iranmanesh 2016; Koutrouli et al. 2020; Mason and Verwoerd 2007; Pavlopoulos et al. 2011) as well. For example, the shortest path problem, traveling salesman problem, looking at connected components, minimum spanning tree problem, centrality measures, and Eulerian and Hamiltonian path problems.

1.5.2 Mathematical Concepts in Relation to Network Biology

Except from the basic terminology in graph theory, we study:

Density: It is defined as the ratio of edges that exists in a network. Mathematically, we write the formula for the directed graph $Den_D = \frac{q}{p(p-1)}$ while for the undirected graph $Den_U = \frac{2q}{p(p-1)}$.

Degree distributions: It is a histogram of degrees. It gives us the node fraction in any network having some degree x. Mathematically, $P(x) = \frac{p_x}{p}$ where p_x is the number of nodes with degree x while p is the nodes in the network.

Clustering Coefficients: Also called cliquishness, we determine the degree to which the vertices of a graph form clusters with each other. The clustering coefficient tells us how much of a clique we find in a neighborhood of a vertex. Mathematically, $C_u = \frac{2y}{x(x-1)}$ where u denotes the node with degree x and y are the number of edges among x neighbors of u in the graph.

1.6 Application of Graph Models in Biology

There are numerous applications of graph models in biology. We illustrate here by showing the examples of biological networks (see Fig. 1.6):

- Interaction networks for Protein–Protein
- Networks for Sequence similarity
- Networks for Gene regulatory system
- Network for Signal Transduction
- Network for Metabolic pathways
- Network for Gene co-expression

There are numerous types of data that can be presented in the theory of network biology and presenting the data as a network and their network analysis is an integral part of Systems Biology. When it comes to Protein–Protein Interactions, it is an undirected network as we can see in Fig. 1.7. When it comes to gene regulation networks, these are directed ones. Going from one gene to another is shown here by the green lines. In the case of the cell signaling pathway, the depicted path in the network is C→D→F. Metabolic biological pathways can be directional, and hence depicted using directed edges. Consider the network between proteins/genes A, B and metabolites m1, m2 and m3 where the directed edge set depicting metabolism pathways is {(A, m2), (m2, B), (m1, A), (m3, B)}, the concept of Bipartite graphs can be implemented here like metabolites interact with proteins/genes and two different kinds of nodes are used here to depict the interactions.

We discuss here one such property of the Protein-protein Interaction network (PPIN). We are basically looking at the study of how a protein interacts with the other proteins in a cell. While studying the interactions, we show here a small world effect.

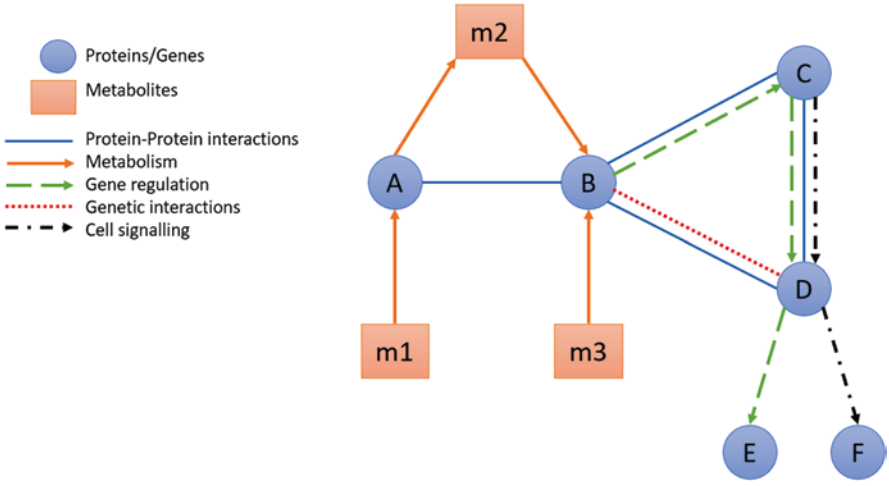

Fig. 1.6 Depiction of types of biological networks using a graph

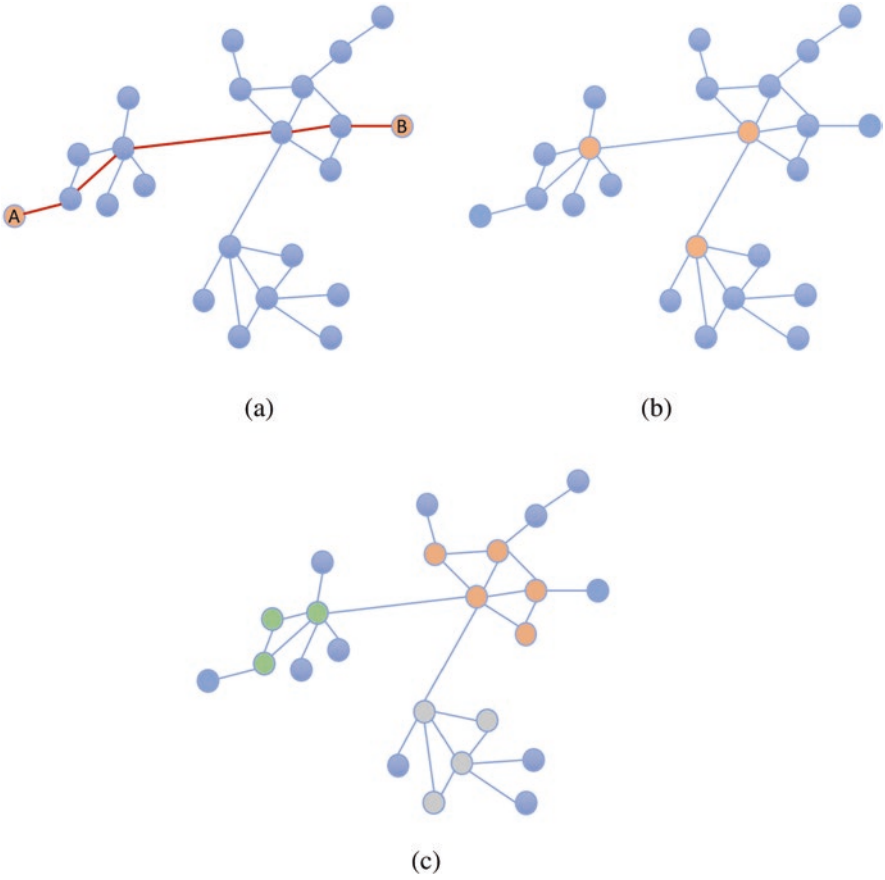

Fig. 1.7 (**a**) Protein–protein interaction network example of shortest path. (**b**) Protein–protein interaction network showing the proteins with maximum degrees. (**c**) Protein–protein interaction network showing the highly interconnected proteins in respective clusters

Looking at the figure in the PPINs example, there are majorly three sub-networks in whole. Another formal word we use here is that we have three clusters in this network. This is an undirected graph that has proteins as its nodes and their interconnections as edges. This shows good connectivity among the proteins. Consider looking for an optimal way to reach from protein A to protein B in the Fig. 1.7a. This is what we call the problem of finding the shortest path from one node to another. Here, the length of the path for moving from A to B is 5. The degree of node A and node B in the network is 1 in both cases which depicts that they both are connected to a single protein directly. We can see clearly from the figure that there is no Eulerian or Hamiltonian path that exists from A to B. Finding out the proteins that are important in the network can be studied using their centrality measures in the graph. This we can see in the next Fig. 1.7b, where we have the proteins in different colors showing high connectivity in the network. And in Fig. 1.7c is the nodes with high transitivity meaning the proteins that are densely connected in the respective clusters.

1.6.1 Petri Net Modeling Approach in Pathway Analysis

Petri nets, a formalized modeling language from around the 1960s, have been used to describe and model discrete event dynamic systems (DEDS) (Peterson 1981). These are structurally directed bipartite multigraphs, where we associate some defined set of rules with the graph and in turn help in analyzing DEDS. Bipartite in the sense as we have two disjoint sets of nodes depicted using circles (called places) and rectangles (called transitions) and directed arcs depicting the direction of process flow in the graph (Fig. 1.8). The arcs must be used to connect circles and rectangles but in a manner such that it joins no two circles or rectangles at a time. An edge that is directed from a circle to a rectangle is the input place of that rectangle and vice versa. For any graph, known as the Petri net graph, we assign the tokens denoted by dots inside the circles in the graph. Writing the number of tokens as a vector is known as marking the Petri net. These markings are used for further execution of the Petri net. Theoretically, we define marking as $n - vector = (M_1, M_2, M_3, \ldots, M_n)$, where n is number of places in the Petri net structure and each $M_i \in \{0, 1, 2, 3, \ldots, n\}$.

We use the Petri net modeling approach in order to find the possible pathways in biological networks given to us with certain conditions. Though there are some graph models that also study the pathway, there are some limitations to them. Hence we use the Petri net approach.

1.6.1.1 Petri Net Reachability-Based Analysis

Petri net firing rule: Removing $w(p, t)$ tokens from the input place of a transition and depositing $w(t, p)$ tokens into the output place of the transition causes the state change in a Petri net, which we call firing of that transition. Here, $w(p, t)$ denotes the weight of the arc from place to transition, and similarly $w(t, p)$ denotes the weight of the arc from transition to a place. In the example we are discussing, we have $w(p, t) = w(t, p) = 1$ for all the transitions (see Fig. 1.9) (Murata 1989).

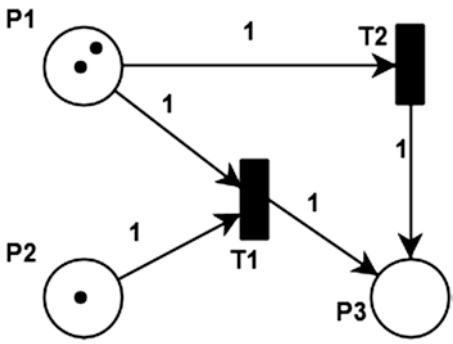

Fig. 1.8 Example of a Petri net with initial marking (2, 1, 0)

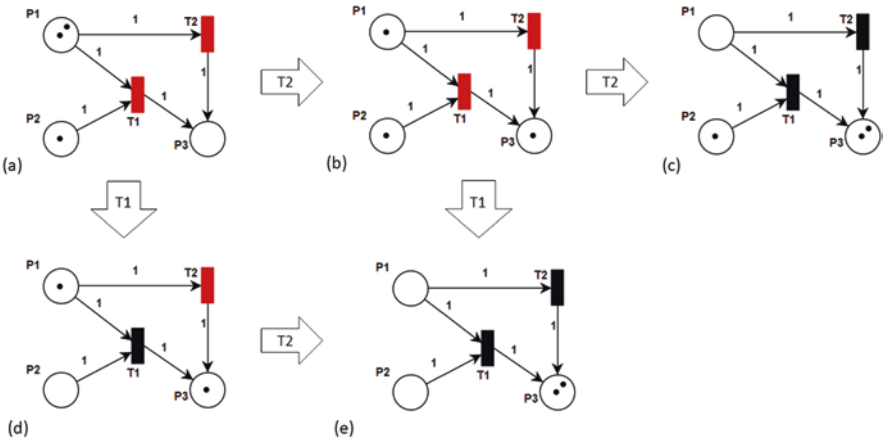

Fig. 1.9 Illustration of firing rule in Petri net depicted in Fig. 1.8

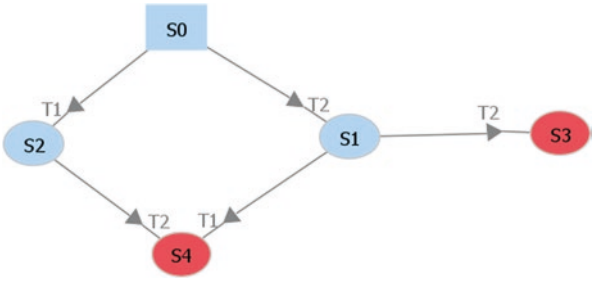

Fig. 1.10 Reachability/ Coverability graph of Petri net in Fig. 1.8

1.6.2 Petri Net Invariant Analysis

Another analyzing technique for Petri net models is by creating Incidence matrix and solving them. We will get some equations but these equations have limited solvability because of restrictions like the solution should be a non-negative integer. Hence, a Petri net model must be pure in order to use this analyzing technique. We write an incidence matrix where the rows and columns are the places and transitions, respectively. And based upon these matrices, we mathematically solve by creating state equations and studying the Invariants.

In Fig. 1.10, we have $S1$, $S2$, $S3$, $S4$ as the reachable markings from the initial marking $S0$. From Fig. 1.9, we can say that $S0 = (2, 1, 0)$ and $S1$, $S2$, $S3$, $S4$ equals (1, 1, 1), (1, 0, 1), (0, 1, 2), (0, 0, 2), respectively. The red color of $S3$ and $S4$ in Fig. 1.10 depicts the state of deadlock in the Petri net while the red color of transitions in Fig. 1.9 depicts that the transitions are enabled in the respective state of the Petri net.

In order to study the pathways in biological processes, we study with the reachability-based analysis (Jha et al. 2022; Mansoori et al. 2020) as well as the

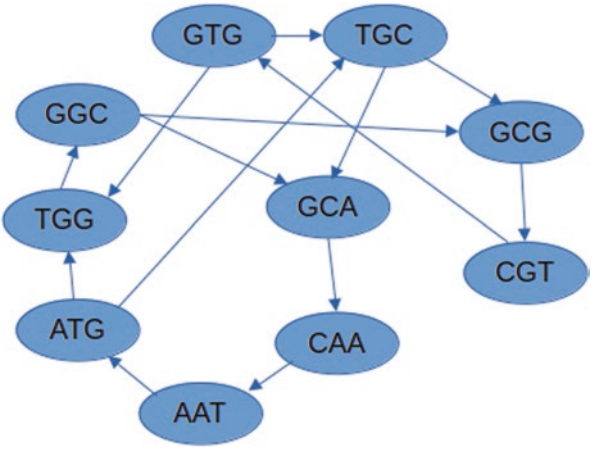

Fig. 1.11 Example of Hamiltonian path approach for DNA sequencing

Incidence matrix and state equations (Singh and Gupta 2019; Singh et al. 2022; Singh et al. 2020; Singh et al. 2021).

1.7 Case Study

As a case study, we look for a DNA sequencing example. In DNA sequencing, we shear the DNA into millions of small fragments. Now we read 500–700 nucleotides at a time from the small fragments (Sanger method). Our challenge is to assemble these short fragments (called reads) into a single genomic sequence (called "superstring").

Let us suppose that we are given data on some pieces of the Genome. If we wish to generate a DNA sequence from those genome pieces, then we have two graph theoretic approaches to solve this problem:

1. *Hamiltonian path approach*—Taking the given pieces as nodes (with length k) and these nodes are connected by an arc if the k-1 rightmost nucleotide of first vertex overlaps with the k-1 nucleotide of the second one.
2. *Eulerian path approach*—Here, we take suffix/prefix as a node. Each oligonucleotide becomes an arc in which its initial endpoint is k-1 rightmost nucleotide of the arc and its terminal endpoint is k-1 leftmost nucleotide.

For example (Fig. 1.11), we have the given pieces of genome—M = GTG, GCG, GCA, ATG, TGG, TGC, GGC, CGT, CAA, AAT. Hamiltonian path approach is used to generate a DNA sequence from these genome pieces and, we have the graph shown in the figure.

Using this approach, we have the following graph model where we follow the path in which every node is covered exactly once. The resulting Hamiltonian path will be—ATG(v_0)-TGG(v_1)-GGC(v_2)-GCG(v_3)-CGT(v_4)-GTG(v_5)-TGC(v_6)-GCA(v_7)-CAA(v_8)-AAT(v_{10}). Hence, we get the following sequence of Genome—ATGGCGTGCA.

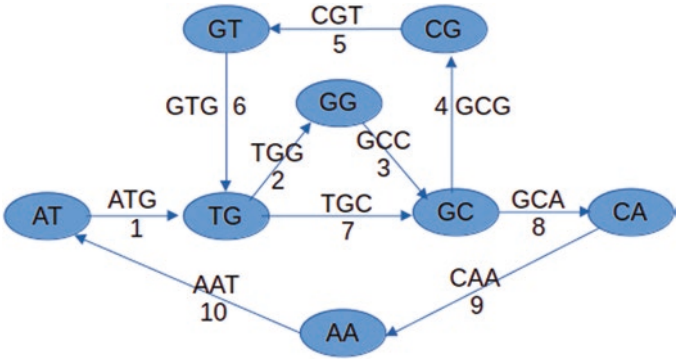

Fig. 1.12 Example of Eulerian path approach for DNA sequencing

But this approach of getting the sequence is time consuming when we have these pieces in millions or billions. So we use the second approach which is comparatively easier.

In the Eulerian path approach, we have the graph as shown in Fig. 1.12.

We follow the path as shown in the figure where every edge is covered exactly once and we got this as follows: ATG(v_0)-TGG(v_1)-GGC(v_2)-GCG(v_3)-CGT(v_4)-GTG(v_5)-TGC(v_6)-GCA(v_7)-CAA(v_8)-AAT(v_{10}). Hence, we get the following sequence of Genome—ATGGCGTGCA.

Clearly, both approaches give us the same solution but the second approach is less time consuming and easier to follow. So, in this way graph theory is used in DNA sequencing. Many other concepts can be studied in different other case studies as well.

1.8 Conclusion

We have discussed in the chapter only the basic mathematical concepts in graph theory that are used in biology. One can use the theory in many different ways according to their needs. In order to study complex systems in biology, there are many extensions available in the literature that we might have missed but this note definitely covers all the basics that are difficult to find at a place.

Acknowledgments The third author, Dr. Gajendra Pratap Singh is thankful to the Department of Science and Technology(DST)-Science and Engineering Research Board (SERB) Project (PID:M T R/2021/000378) and to the Indian Council of Medical Research (ICMR) project (PAC/SC&IS/GPS/ICMR/1511); and the authors are expressing their deep gratitude to anonymous reviewers and editors for their valuable suggestions and comments.

References

Barnes JA, Harary F (1983) Graph theory in network analysis. Soc Networks 5(2):235–244

Gupta SB (2008) Discrete mathematics and structures. Laxmi Publications, Ltd

Jafarzadeh, N., & Iranmanesh, A. (2016). Application of graph theory to biological problems. Studia Universitatis Babes-Bolyai, Chemia, 61(1)

Jha M, Singh M, Singh GP (2022) Modeling of second-line drug behavior in the treatment of tuberculosis using Petri net. Int J Syst Assur Eng Manag 13(2):810–819
Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA (2020) A guide to conquer the biological network era using graph theory. Front Bioeng Biotechnol 8:34
Mansoori F, Rahgozar M, Kavousi K (2020) A pathway analysis approach using Petri net. IEEE J Biomed Health Inform 25(3):874–880
Mason O, Verwoerd M (2007) Graph theory and networks in biology. IET Syst Biol 1(2):89–119
Murata T (1989) Petri nets: properties, analysis and applications. Proc IEEE 77(4):541–580
Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J et al (2011) Using graph theory to analyze biological networks. BioData mining 4(1):1–27
Peterson JL (1981) Petri net theory and the modeling of systems. Prentice Hall PTR
Singh GP, Gupta A (2019) A Petri net analysis to study the effects of diabetes on cardiovascular diseases. In: In *2019 6th international conference on computing for sustainable global development (INDIACom)*. IEEE, pp 481–488
Singh GP, Jha M, Singh M (2020) Modeling the mechanism pathways of first line drug in tuberculosis using Petri nets. Int J Syst Assur Eng Manag 11(2):313–324
Singh GP, Jha M, Singh M (2021) Petri net modeling of clinical diagnosis path in tuberculosis. In: Advances in interdisciplinary research in engineering and business management. Springer, Singapore, pp 401–412
Singh GP, Jangid R, Singh M (2022) Petri net modeling as an aid in bioprocess designing. In: Microbial products. CRC Press, pp 455–462

Biological Networks Analysis

2

Najma and Anam Farooqui

Abstract

Networks are widely recognized as a popular method for representing complex biological processes by capturing the interconnected relationships between different biological components using binary interactions or connections. This chapter discusses many forms of biological networks and network models that are important for understanding complicated networks. In addition, we describe different network measures that are quantifiable description of biological networks. We also discussed different methods for detection of community. We briefly mention about hubs and formation of rich-club, system-level organization in a hierarchical network, detection of network modules and motifs, and biomarkers. Lastly, we examine several databases that contain biological networks and explore how network modules are utilized in understanding the dynamics of diseases. We anticipate that a wide range of readers, from experts to newcomers, will benefit from this chapter and be influenced to advance the field.

Keywords

Biological networks · Graph theory · Protein–protein interaction networks (PPI) · Node · Edge · Genetic interaction networks · Degree distribution · Closeness centrality · Eigenvector centrality · Neighborhood connectivity · Betweenness centrality rich club · Transcriptional regulatory networks · Network motifs · Hierarchical network · Cell signaling networks · Hamiltonian energy · Clustering coefficient · Clustering methods · Biomarker · Metabolic networks · Network model · Scale-free network · Biological network databases · Disease dynamics

Najma · A. Farooqui (✉)
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: najma2300912@st.jmi.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_2

2.1 Overview of Biological Networks: Network Construction

For a layperson, genomes, in general, are thought of as components that are purely defined by their linear DNA sequences. However, the reality is far more complex. The DNA (Deoxyribonucleic Acid) serves as the primary unit of heredity in all types of organisms. The DNA helix folds itself into several layers of higher-order structures in a hierarchical fashion that forms a chromosome (Woodcock 2006). The DNA folded in this fashion eventually gives it a compact structure enabling it to get accommodated in the limiting space of the cell's nucleus. Besides this sophisticated arrangement of the genetic component itself, the cellular factors that are responsible for reading, copying, and maintaining the genome too are arranged in a complex and compartmentalized fashion within the cell's nucleus. The cellular organization of the genomes and the cellular factors give an architectural environment for them to function. Determining how these particular molecules contribute to various cellular processes is one of the main problems in current cell biology. Identification of these molecules and their interactions is critical for gaining complete knowledge of the complicated machinery within live cells. Exploring the answers to these questions may lead to insights into genome biology and how it works (Fig. 2.1).

We, as human beings, are surrounded by a network of networks. Our body functions through complex networks of networks working at different levels. Most biological functions arise from complex interactions from intricate interactions involving multiple cellular elements like nucleic acids, proteins, and small

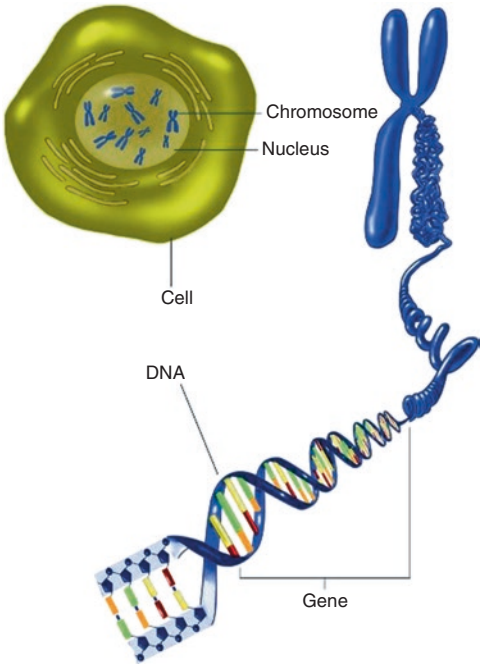

Fig. 2.1 DNA makes up genes and is coiled within chromosomes inside the nucleus of a cell. Credit: NIGMS.

molecules. It is exceedingly uncommon for a biological function to be attributed solely to a single molecule. These intercellular webs of interactions include PPI, metabolic networks, transcription-regulatory networks, and signaling networks. The technologies like PROTEIN CHIPS or YEAST TWO-HYBRID determine these molecular interactions.

Another key technique that is commonly used for analyzing complex systems in the physical, social, information, technology, and biological sciences is network theory. The application of network theory in biology is known as *Network Biology*. The development of network biology has revealed that there are universal laws that control cellular networks. This might provide a fresh conceptual framework for comprehending disease pathology in the twenty-first century. The biological systems are depicted in this as standard graphs with nodes and edges. The graphs' nodes or vertices stand in for various entities or agents, and the edges or links signify a connection between two nodes. In a more realistic and complicated framework, these edges may be heterogeneous, unweighted/weighted, and directional/unidirectional. Biological networks play a crucial role in examining and understanding the interconnected relationships among various components (interactome) inside a biological system. These interactomes are the collection of direct or indirect molecular linkages of the biological system. Thus, overall, a biological network signifies the cell's information processing system via the molecular wiring or molecular network (Kaviani and Sohn 2021).

2.1.1 Structure of Complex Networks and Notations

Because of the data explosion generated by the omics era of biological research, it was important to shift away from a single gene/protein perspective and develop more systematic data analysis methodologies. System biology seeks to explore biological processes on a systems level, not just as isolated components but also as interacting systems and their evolving characteristics. System biology, which uses graph theory approaches to describe and analyze biological systems, is connected to network biology. The study of graphs, which are mathematical constructions used to represent pairwise interactions between things, is known as graph theory. In this context, a graph is a collection of vertices or nodes interconnected by lines or edges. In actuality, it is a group of conceptual ideas and techniques for visualizing and analyzing networks. Nodes represent different entities (e.g., genes or proteins) and edges convey information about how the nodes are linked (Muzio et al. 2021).

Biological network analysis was driven by the methods and concepts of social network analysis, and the application of graph theory to the social sciences also contributed to its development. From the molecular to the ecosystem level, every biological entity interacts with other biological entities, giving us the opportunity to describe biology using a variety of networks such as ecological, neurological, metabolic, or molecular interaction networks. Complex interactions among biomolecules are usually described using network models when studying biological systems.

These networks are known as biological networks and they represent biological systems mathematically (Pavlopoulos et al. 2011).

For example, networks within cells typically exhibit the following characteristics:

1 They tend to be disassortative.
2 They possess structural and dynamic robustness.
3 Their degree distribution power law.
4 They exhibit a modular organization.
5 The average length of the shortest path between any two nodes is quite short, indicating a small-world feature.

2.2 Biological Networks and Types

2.2.1 Biological Networks

The idea of biological networks is based on the observation that multiple networks operate in living cells to carry out and regulate every element of cellular life, these elements are made up of a variety of components ranging from basic biomolecules to the entire organism, and these networks function within a highly organized framework. All biological activities are carefully and tightly controlled at the cellular level (Grigorov 2005).

The foundation of life processes for the entire structural range of living matter, including biomolecules like proteins and nucleic acids, cellular organelles like cytoskeleton and mitochondria, tissues, whole cells, and organs, is made up of biological networks, which are organized, deterministic systems. Finally, the entire organism makes up this structural spectrum. These networks collaborate to carry out certain physiological activities inside a particular cellular compartment. They describe how various biological processes, such as genetic regulation, cellular signaling, and metabolism are structured. These networks are frequently used to comprehend and analyze biological processes at the system level, find fascinating modules or sub-networks, and identify potentially important proteins based on their network characteristics. These networks are representations of biological systems and the relationships that exist between them. They provide insights into complicated biological systems, revealing information about them. They integrate biological omics data with biological interactome data to disclose information inside these systems (e.g., gene–gene associations and protein–protein interactions). The networks involved in transcription, protein–protein interaction, and metabolism are the most important for managing biological systems. These networks are studied using a combination of statistical methodologies, graph theory methods, mathematical models, and visualization tools. The study of biological networks is now an important part of systems and computational biology. Because such analysis provides a common language for describing relationships within complex systems, it has become more crucial in gaining a better understanding of physiological functions.

Many efforts have been made to study and analyze the topology and structure of cellular networks, as well as their relationship to cellular function and organization. These networks are often modular, have a small world property with minimal average path-lengths, and use scale-free topologies with power law degree distributions; this makes them resistant to adversity.

As a result, networks can enhance our knowledge of biological systems and their flow. Among other aspects, networks are used to analyze gene lists from high-throughput biology as well as data from post-genome-wide association studies. There are various approaches to using network information to enhance our knowledge of biological systems. One example is the development of new organizational assumptions based on network topology information. Using massive network data to test and confirm or refute current hypotheses is a complimentary strategy for a very long time, theorists have speculated about the connection between the evolution of genes and the networks that they produce. With the availability of large-scale quantitative data on the topology of molecular networks, it is now possible to pose explicit queries about the role of network structure in the evolutionary process and how evolution influences network structure. Several recent researches have demonstrated the effectiveness of this innovative approach to biology (Yu et al. 2013).

2.2.2 Types of Biological Networks

2.2.2.1 Different Types of Biological Networks Are Described below

1. Protein–protein interaction networks.
2. Metabolic networks.
3. Gene/transcriptional regulatory networks.
4. Genetic interaction networks.
5. Cell signaling networks.

Protein–Protein Interaction Networks

One of system biology's main objectives is to better understand protein–protein interactions. PPINs, or protein–protein interaction networks, are mathematical depictions of the physical interactions that take place between proteins in a cell. Understanding protein–protein interactions (PPIs) is critical for understanding both normal and pathological cell physiology because PPIs are required for understanding every process in a cell. It is also essential in drug development. PPIs are a type of molecular interaction data that is widely used. These interactions give both an experimental foundation for understanding cell modular architecture and important information for predicting the biological function of specific proteins. Many of the cell's most critical molecular activities, such as DNA replication, are carried out by complex molecular machines made up of many protein components that are organized by their protein–protein interactions. It is possible to represent a collection of pairwise interactions among a collection of proteins in a natural way by using a graph with proteins as its nodes and pairwise interactions as its edges. The collection of all interactions (entire set of PPINs that exist in a biological system) between the proteins of an organism is usually called the interactome (Jordán et al. 2012).

PPIs can be detected by: in vitro methods (like NMR spectroscopy, protein fragment complementation, phage display, tandem affinity purification, protein arrays, X-ray crystallography, and co-immunoprecipitation, affinity chromatography) and in vivo methods (like yeast two-hybrid (Y2H, Y3H) and synthetic lethality) based approaches). These experimental resources have been used to create extensive PPI networks. But on the other hand, the volume of PPI data is making laboratory validation difficult. Therefore, computational analysis of PPI networks using in silico methods (like in silico 2 hybrid structure-based approaches, gene expression, gene fusion, sequence-based approaches, chromosome proximity, phylogenetic tree, and mirror tree), is turning into a crucial tool for understanding the functions of yet undiscovered proteins. Protein–protein interaction (PPI) is currently one of the important areas of study for the advancement of contemporary system biology (Srinivasa Rao et al. 2014).

Metabolic Networks

Metabolic networks are built by collecting and linking biochemical data with genetic sequences. The mass flow in basic chemical pathways that generate vital components like amino acids, carbohydrates, and lipids, as well as the energy required by biological reactions, are usually the focus of the metabolic network. As a result, these networks frequently include information on both proteins and metabolites.

The metabolic network is made up of all chemical processes that utilize catalytic proteins to facilitate the metabolism of small molecules (metabolites). A metabolic pathway is a set of processes that transform one or more educts into one or more products. Metabolic pathways are involved in the storage and release of energy.

Metabolic networks are made up of very similar building components. The vertices in this network indicate the educts and products, while the connecting reaction is represented by an edge that will be tagged with the catalyzing enzyme and perhaps supplemented with cofactors. This representation solely depicts a reaction's connection (topology). The reaction kinetics, or the concentration gradients of the relevant molecular partners in terms of their starting concentrations, is required for modeling (Haggart et al. 2011).

Metabolic network databases include databases that focus on a single organism like HinCyc110 (for *H. influenzae*), EcoCyc109 (for *E. coli*), and PseudoCyc111 (for *Pseudomonas aeruginosa*), as well as databases that include a wide range of organisms like KEGG, MPW, and MetaCyc.

These databases contain information and graphics that depict such pathways and their connections. Computers are increasingly being used to process photos. These databases are large enough to compare metabolic pathway topologies in various animals (Succoio et al. 2022).

Gene Regulatory Networks

Gene regulatory networks, or GRNs, are essential for regulating how genes are expressed. These are collections of regulatory links between transcription factors (TFs) and TF-binding sites, or between genes and their regulators. Cis- and

trans-regulatory elements are the two primary regulatory types. Trans-regulatory elements can regulate away from the genes from which they were transcribed, whereas cis-regulatory elements are located close to the structural region of the gene they regulate. Moreover, GRNs are directional as TFs regulate their targets, dynamic, i.e., changing across different conditions, and can be visualized as bipartite graphs, which are crucial for understanding the underlying mechanisms of disease pathogenesis. GRNs are composed of "nodes," which represent the genes and their regulators, joined together by "edges," which represent physical and regulatory interactions (Ertaylan et al. 2014).

GRNs are bipartite because there are two types of nodes: Genes and TFs. The overall topology of GRNs has been analyzed in many systems. To create the most comprehensive and accurate GRN models possible, it is ideal to incorporate physical and regulatory links. The presence of TF and gene hubs indicates that GRNs are not random structures. GRN modules are highly interconnected. There are various databases present that hosts information about gene regulation, commonly used repositories are the KEGG, GTRD, and TRRUST. A comprehensive understanding of GRNs is a major challenge in the field of systems biology.

Reconstruction of regulatory networks is made possible by recent developments in high-throughput approaches, which offer a wealth of binding data from techniques like ChIP-Seq, miRNA-Seq, and ATAC-Seq along with expression data from RNA-Seq. Reconstruction of the gene regulatory interaction to form GRNs is one of the key tasks in systems biology. Gene expression data are frequently used in the creation of a GRN. For repairing GRNs in a true cellular context, a number of computational techniques and models have been developed to date. Nevertheless, each of them uses a unique set of presumptions and techniques to create unique blueprints. Several statistical and mathematical tools can be used to visualize and study it. For GRN visualization, Cytoscape is a widely used and simple-to-use application that has shown to be quite helpful (MacNeil and Walhout 2011).

Genetic Interaction Networks

A genetic interaction network is a collection of genes that are connected by edges and have functional interactions with one another. It is believed that these genes either physically interact with one another through the gene products they produce, such as proteins, or that one gene affects the activity of another gene. The phrase "genetic interaction" refers to a collection of functional connections between genes that cooperate to carry out a certain activity and are frequently physically linked to one another to create a more complicated structure.

Understanding biological activities requires information about interacting proteins, which may be easily attained by examining networks of these connections. These interactions are crucial to the majority of biological processes.

Genetic interactions require combination of two or more genes to generate an unexpected phenotype. These interactions are further categorized into two categories i) Negative genetic interaction) and ii) Positive genetic interactions. Negative genetic interactions are caused when two mutations, none of them lethal individually, combine to cause cell death. Positive genetic interactions occur when a lethal

variant of one gene is suppressed by the variant of another gene or their combined effect is less severe than expected. These types of interactions are essential in understanding pathways and regulation in model organisms, functional relationships between genes, also undiscovered genetics associated with complex human diseases (Boucher and Jenna 2013).

A resource for forecasting gene and pathway function is a global genetic interaction network, which shows the functional arrangement of a cell. This network highlights the ubiquity of genetic connections and how they can amplify the abnormalities brought on by a single mutation. They typically show pairs of genes active in parallel pathways or various molecular machinery (Costanzo et al. 2016).

Genetic interaction profile of a gene is made up of its unique set of positive and negative genetic interactions. Positive interactions may provide links associated with problems in cellular proteostasis and cell cycle progression or insights into general mechanisms of genetic suppression. Negative genetic interactions between functionally related genes, mapped key bioprocesses, and recognized pleiotropic genes can be inferred using alternative functional information. Genetic interaction profiles provide a quantifiable measure of functional similarity, and similar networks created by correlating large-scale genetic interaction profiles organize genes into clusters that highlight biological processes. When genetic interaction profiling networks are coupled with other types of interactome networks, predictive models of biological processes can be created, resulting in potentially powerful models which can form a potentially powerful model by using different datasets (Wiredja and Bebek 2017).

Cell Signaling Networks

A crucial regulatory system that is vital for all life activities in living creatures is cell signaling. Cell signaling networks control and direct cellular functions, intercellular interactions, and reactions to the environment. The various signaling pathways that signaling networks use to carry out their functions each represent an ordered series of reactions that are elicited and started by signal molecules, primarily proteins that turn on receptor proteins, and that lead to biochemical or biophysical modifications in the pathway. A primary mechanism to control the number of intracellular components occurs at different levels, i.e., the post-transcriptional mRNA processing level (alternative splicing) than on next level which includes post-translational modifications (such as phosphorylation, acetylation, methylation, and so on).

Signaling networks can be visualized as directed graphs with edges pointing in the direction of signal propagation. There are one or more starting nodes in a signaling network that represent the binding of the initial signal(s) to receptor(s) and one or more output nodes that indicate the cellular responses to the signal(s). Along with these nodes, there are several intermediary nodes that involve ions, enzymes, genes, secondary messengers, kinases, proteins, metabolites, and other chemicals in signal transmission. The edges of a signaling network indicate many interactions between signaling components such as transcription, protein phosphorylation, protein binding, enzymatic catalysis, complex formation, and regulation (Vieira and Vera-Licona 2019).

In living organisms, there are numerous cellular signaling networks, both extracellular and intracellular. Some important cellular signaling networks are the PI 3-Kinase/Akt Signaling pathway, MAPK/ERK pathway, Wnt signaling pathways, Apoptosis signaling pathways, cAMP-dependent signaling pathway, Retinoic acid signaling pathway, Calcium signaling pathways, Delta-Notch signaling pathway, Hedgehog signaling pathway, Insulin signaling pathways, VEGF mediated signaling pathway, Oscillatory rhythm network, and many more. At each step of the cellular signaling process, feedbacks are possible which forms a signal transduction network.

Cellular signaling network reconstruction includes genome annotation, expression arrays, cell physiology characterizations, biochemical experimentation, and other such data sources. Data for reconstructing networks can be found in renowned pathway databases like (Reactome, KEGG), specific repositories such as the NetPath, MiST, or Human-gpDB are also present (Papin et al. 2005).

Modeling cellular signaling networks is quite challenging as they involve the interactions of components from different levels such as transcriptome, metabolome, and proteome. Mathematical modeling of cellular networks enables us to figure out the basic design of cell signaling networks from a system-level perspective and how transmission of information affects the network (Albert and Wang 2009). Large-scale networks can be structurally analyzed in their totality since it does not require a comprehensive understanding of the parameters that have been obtained via thorough experimentation.

Complex networks being ubiquitous have gained extensive attention from the scientific community. Complex systems' internal states can help us gain a general understanding of their biological, technical, and social surroundings. The key component of network biology is the behavior of a complex network. The characteristics of complex networks can be understood through network models. Three main models have been studied that have direct control of biological networks.

2.3 Network Topological Properties

From living cells to the Internet, complex systems work in synchronization with its components through pairwise interactions. As mentioned earlier, these components can be described as a set of interconnected nodes. Here, each edge represents the interactions between two nodes. Altogether, these nodes and edges form a network or a graph. Such networks/graphs are studied and conceptualized through Graph Theory.

Depending on the type of interactions, networks can be classified as directed or undirected. A directed network's connections have a specific direction between its nodes, such as when information flow is regulated from a transcription factor to a gene. However, the interactions in undirected networks have no assigned direction. In protein interaction networks, for example, a link represents a mutually binding relationship (Fig. 2.2).

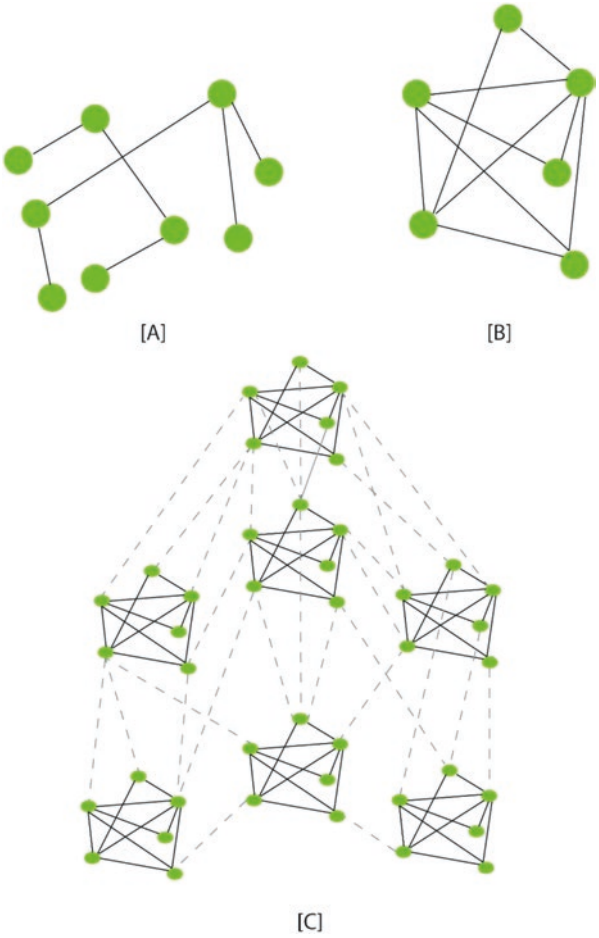

Fig. 2.2 Types of graphs (**a**) Graph representing edges and nodes (**b**) Directed graph (**c**) Undirected graph (**d**) Weighted graph

The structural properties of complex networks are understood by examining their topological parameters. The fundamental network measures that provide insights into the significant behaviors of the network are Degree distribution, Clustering coefficient, Neighborhood connectivity, Eigenvector centrality, Closeness centrality, and Betweenness centrality. These measures help us gain a deeper understanding of the network's important properties and how its nodes are connected.

Degree distribution: In a network, the degree k refers to a measure of centrality that indicates the number of connections a node has with other nodes. In a network represented by a graph G = (N, E), where N and E represent the number of nodes and edges, respectively, the probability of the degree distribution (P(k)) of the network can be defined as the proportion of nodes with a particular degree relative to the total size of the network:

$$P(k) = \frac{n_k}{N} \tag{2.1}$$

Here, nk depicts the count of nodes that have a degree of k, while N denotes the total number of nodes in the network. The probability distribution P(k) provides insight into the significance of hubs or modules within the network. It follows a power law pattern, where P(k) is approximately proportional to the inverse of k raised to the power γ. The value of γ determines the level of importance attributed to hubs or modules in the network, which can vary depending on whether the network exhibits characteristics of scale-free or hierarchical structures.

Neighborhood connectivity: A node's connectivity is determined by the number of its neighbors. The neighborhood connectivity of a node "n" is calculated by finding the average connectivity of all of its neighboring nodes. In the network, the neighborhood connectivity (CN(k)) can be expressed as follows:

$$C_N(k) = \sum_q qP\left(\frac{q}{k}\right) \tag{2.2}$$

In the network's context, the conditional probability (P(q/k)) represents the probability that a link originating from a node with connectivity k will connect to a node with connectivity q. The presence of a positive power relationship in the neighborhood connectivity (CN(k)) can serve as an indicator of assortativity within the network's topology.

Clustering coefficient: The clustering coefficient is a topological parameter that represents the extent of interconnectivity and strength of connections between a node and its neighboring nodes in a network. It is determined by calculating the ratio of the number of edges that exist between the node's nearest neighborhood edges (ei) to the total possible number of edges for a node of degree ki. In an undirected network, the clustering coefficient (C(ki)) of the "*i*th" node can be calculated using the below formula:

$$C(k_i) = \frac{2e_i}{k_i(k_i - 1)} \tag{2.3}$$

Betweenness centrality: Betweenness centrality (CB) is a measure of a node's importance in controlling the flow of information within the network. It reflects the extent to which a node can control other nodes by acting as a bridge between them. To compute the betweenness centrality of a node "v", we consider the number of geodesic paths between every pair of nodes "i" and "j" that pass through "v", represented by dij(v), and the total number of geodesic paths between "i" and "j", represented by dij. The formula for calculating the betweenness centrality (CB(v)) of a node "v" is:

$$C_B(v) = \sum_{i,j,i\neq j\neq k} \frac{d_{ij}(v)}{d_{ij}} \tag{2.4}$$

Closeness centrality: Closeness centrality (CC) quantifies the efficiency of information dissemination from a node to other nodes that it can reach within the

network. It is defined as the reciprocal of the average geodesic distance between the node and all other connected nodes in the network. In other words, the closeness centrality of a node "i" is calculated as the inverse of the average shortest path distance between node "i" and its neighboring nodes. The formula to determine the closeness centrality (CC) of a node "i" is as follows:

$$C_C(i) = \frac{n}{\sum_j d_{ij}} \tag{2.5}$$

where d_{ij} represents the geodesic path length from nodes i to j, and n is the total number of vertices in the graph reachable from node i.

Here, dij denotes the geodesic path length from node i to j, while n represents the total number of vertices in the graph that are accessible from the node i.

Eigenvector centrality: The eigenvector centrality (CE) of a node "i" in a network is directly related to the collective centrality of its neighboring nodes. The formula for calculating the eigenvector centrality (CE) of node "i" is as follows:

$$C_E(i) = \frac{1}{\lambda} \sum_{j=nn(i)} v_j \tag{2.6}$$

Here, the term "nn(i)" refers to the nearest neighbors of node "i" within the network. The eigenvalue (λ) and eigenvector (vi) of the node are determined by the equation Avi = λvi, where "A" represents the adjacency matrix of the network or graph. The principal eigenvector, associated with the highest eigenvalue (λmax), is considered to have a positive eigenvector centrality score. It can be used as an indicator of spreading power a node in the network (Farooqui et al. 2018).

2.4 Detection of Network Module and Motifs

Network motifs refer to small, interconnected sub-graphs that are frequently observed in networks. These motifs represent recurring patterns within the network structure. However, identifying network motifs requires significant computational effort as they need to be tested for similarity multiple times (Patra and Mohapatra 2020). The various tools and algorithms used in the process of finding network motifs are briefly described below:

NeMoFinder: Mesoscale network motifs are discovered using this algorithm.
Network motif detection (NetMODE): It is a network motif detection software package to improve runtime efficiency.
Grochow and Kellis: It is a motif-centric algorithm, where frequency is counted on the basis of a particular isomorphic class.
Kavosh: It is a network-centric algorithm to improve runtime efficiency.
Elhesha–Kahveci: It is a motif-centric algorithm for finding disjoint network motifs in a target network.
MODA: It is a motif-centric algorithm based on a pattern growth methodology.

CoMoFinder: An algorithm to accurately and efficiently identify composite network motifs in genome-scale coregulatory networks.

MODET:—It is an motif-centric algorithm based on a static ET.

Accelerated motif (Acc-Motif) and QuateXelero: These are network motif discovery algorithms.

MDET: It is a Fast and scalable network motif discovery algorithm for exploring higher-order network organizations.

2.5 Rich Club and Community Finding Algorithm

2.5.1 Rich-Club Finding Algorithm

In a network, the nodes that have a large endowment to the overall topological organization are called hubs. The most generally used centrality in identifying hubs is degree centrality, and the other centralities (like closeness, eigenvector, and betweenness) also give useful detail on each node. Other measures like vulnerability, are used to identify the vitality of a node in a network. Since there is no single direct way to identify hubs, each measure is very much correlated to one another. Therefore, it is always good to use them together to rank the important nodes and identify the candidate hubs (Rubinov and Sporns 2010).

Due to their central role, hubs perform many important roles in the network. These hubs are accounted for effective communication, as well as the inclusion of information, among different nodes or modules in a network. Many real-world complex networks, such as social and World Wide Web, are observed to have dispersed topologies consisting of an organization of functional modules. In such networks, the importance and roles of hubs are very illuminative, and these hubs have a tendency of interconnecting among themselves further forming a highly powerful group of specialized hubs, known as a rich club. The formation of rich clubs increases the robustness, efficient communication, propagation of the signal, and integrability of a complex network (van den Heuvel and Sporns 2011).

In a modular network, the hubs can be further classified using the participation coefficient measure (P_i) defined as (Guimerà et al. 2007):

$$P_i = 1 - \sum_{N}^{j=1} \left(\frac{k_{ij}}{k_i} \right)^2 \tag{2.7}$$

Here N stands for the total number of modules, ki denotes the hub node's degree, and kij denotes the number of links the hub node possesses with other nodes within a specific module (j).

After identifying the degree range k for hub nodes, the existence of rich clubs is studied by measuring the rich-club coefficient (k) over the degree range (Zhou and Mondragon 2004).

2.5.1.1 System-Level Organization in a Hierarchical Network

System-level organization of fundamental functional components of a hierarchical network at different levels is maintained to perform important specific tasks at those various levels in the self-organized fashion of the components (System-Level Organization in a Hierarchical Network n.d.) and the emergence of important regulators at a local and global level referred to as hubs (Albert and Barabási 2002) are some of the fundamental features of most of the natural and artificial networks. This hierarchical organization of the network extends from how cells function, brain organization, and how proteins cross-talk at the molecular level (Stelzl et al. 2005), the evolution of species during the prebiotic era (Jain and Krishna 2001) to inter and intracellular talks in tissue network. Traditionally, biology has been based on the central idea that life processes are hierarchically organized and indicates that it is this structure that controls the system's dynamics. Surprisingly, we lack an objective manner to assess how real this hierarchical organization is, even if we are given different levels and their interactions in the hierarchy (Pennisi 2005). Among the topological characteristics of a network one is functional organization of it via various fundamental functional units known as network motifs (Milo et al. 2002) and their roles in building up the network organization/reorganization that led to the network complex in nature. Motifs in a network may be of many types, and every network motif performs a well-defined function within the network (Alon 2019) and most motifs in the network overlap to process information among them. Clustering of motifs (similar types or different) by overlapping structural and functional modules of various topologies, clustering modules form super-modules that cross-talk among them to organize the whole network. Cancer networks are observed to have self-similar organizational characteristics like rich-club formation, modular structure, hubs, etc. This small-world structure generates nonlinear dynamical behavior and the rich-club formation supports the flexible integration of the individual modules. Complex network theory is a useful tool to study the organizational structure of complex systems like cancer. The exploration of self-organizing properties can be done by observing fractal growth mechanisms (Song et al. 2006).

2.5.2 Community Finding Algorithm

Clustering: detection of community: Networks that are created from the real world have distinctive properties (topological) which are dissimilar from random networks given by Pál Erdős and Alfréd Rényi in 2002. They generally show dependence on degree for clustering coefficients with biased diversity in degree, which is absent in random networks. These networks (real-world) show edges heterogeneously on a large scale in distribution, with nodes forming dense sub-regions of closely connected groups resulting in a modular structure in the network calling it a community. As seen clearly from fractal studies and scaling behaviors, many of the real-world complex networks have shown to be organized as hierarchical with self-similar sub-units (sub-communities) present in larger communities (Ravasz and Barabási 2003).

Finding communities is, however, not a simple problem. Primary goal of this is to extract useful information to efficiently perceive the functional organization in the system that it represents, however, it may indirectly help in identifying the abnormalities in the system and also a possible target. A community can be perceived as the existence of densely connected nodes into a consistent group with relatively fewer connections with the outside group. Therefore, it is required to apply efficient algorithmic methods, which is a great challenge considering the huge amount of computational complexity (Fortunato 2010). This computational complexity estimates the maximal number of computational steps required to perform an assigned task that accounts for the worst-case scenario(s) and is affected by the way a network is stored in the working memory of the computer. In network-based algorithms, networks are usually stored in either adjacency list format or adjacency matrix (adjacency list occupies less space). The long history of dividing a network into modules starts from the work of Stuart A. Rice in the 1930s who grouped people on voting patterns similarities, by studying their working relationships using personal interviews with Weiss and Jacobson. In 1962, H. Simon discussed the importance of maintaining hierarchy, self-organization, evolution in complex systems, and many more. The first algorithm called Kernighan–Lin algorithm was developed in the 1970s for dividing resources to have effective parallel computing. However, being specific, it was a graph partitioning algorithm. While the idea behind finding communities is to identify the natural fault lines between the communities without having previous knowledge. Therefore, it is natural to come up with so many ways for finding a community based on their definition of a community. The techniques for finding community can be classified into four major types: (1) Agglomerative that relies on some similarity measures: such as Jaccard similarity, cosine similarity, correlation coefficient, Katz similarity, or Euclidean distance; (2) Identification of intermodular edges: methods based on fluid-flow, betweenness centrality measures, current-flow analogies, etc.; (3) Quality function optimization of partitions, like likelihood-based measures, modularity, and Hamiltonian in Potts model; and (4) statistical inference methods such as block modeling (Bickel and Chen 2009).

The most commonly used quality function is the modularity that compares an observed edge density within a partition. Modularity Q can be measured by using:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(C_i, C_j\right) \tag{2.8}$$

Here m stands for a total number of edges in the community, k represents degrees, A_{ij} denotes adjacency matrix of size i × j, and the δ function yields 1 if nodes i and j are in the same community. Some well-known community finding algorithms are discussed below.

Hierarchical clustering method: With the rise of social network analysis, this clustering method focuses on the hierarchical decomposition of a network aggregating nodes based on predefined topological similarity measures. There are varieties of clustering (hierarchical) algorithms that emerge from the pliability in the choice

of the "similarity measure" as well as having different rules for assigning "similarity score" to a collected mass of nodes. Therefore, a hierarchical clustering method begins with a commonly used measure, i.e., cosine similarity σ_{ji}:

$$\sigma_{ij} = n_{ij} / \sqrt{k_i k_j} \tag{2.9}$$

where k_j and k_i are the respective degree of nodes j and i. The algorithm then generates an $n \times n$ similarity matrix, from this similarity matrix the nodes bearing the highest similarity values are agglomerated to form group(s) of size = 2. The next step of agglomeration involves a choice from three methods: (1) single-linkage, (2) complete-linkage, and (3) average-linkage clustering. Let us suppose in the general form we have two groups of size v_i and v_j so there will be $v_i v_j$ possible pairs such that one node falls into group vi and another node into group v_j. From these pairs, highest similarity pair is considered in the case of single-linkage clustering, while the least similar pair is considered for complete-linkage clustering. Finally, the mean similarity from all the pairs is considered as an index in the case of average-linkage clustering. Finally, we obtain a complete hierarchical picture of the network, starting from single nodes and successively grouping them into clusters in the hierarchical order (Newman 2012). This method is still widely used, as it tends to have clusters of high-degree nodes leaving behind low-degree nodes.

Betweenness-based method: In 2002, M. Girvan and M. E. J. Newman suggested the most popular method for recognizing communities by performing topological measures of the network (Girvan and Newman 2002). They used the concept of betweenness centrality and formulated the concept of edge betweenness which is described as the number of geodesic paths that run through an edge. It further considers the emergence of compact modules expected to have a higher edge betweenness. The algorithm first calculates the edge betweenness and search for the highest score. Then remove the corresponding highest-scoring edge and on each edge removal, the edge betweenness scores are again calculated, this process iterates, then, in the end, the network starts to split into parts until the nodes are separated. Finally, the network is decomposed hierarchically and can be represented by a dendrogram which is left at the bottom representing the individual nodes.

One disadvantage of this algorithm is its high computational complexity which means it is a slow algorithm. The method was further extended by using modularity maximization to identify the best partition. Another vital variation of this model was proposed by Radicchi et al. in which the intermodular edges are identified in dense modules that can have the formation of short loops as compared to the edges between modules. Other related models that use the "fluid-flow" can successfully identify intermodular edges and also improve the speed of the algorithm (Wu and Huberman 2004).

Optimization method: Optimization method is based on heuristics algorithms involving the calculation of an approximate solution by assigning a quality function. Modularity, Hamiltonian in Potts model, E/I ratio, likelihood base measures, etc., are some classes of the quality function used in optimizing problems. From these, modularity is mostly used because of its innate ability to interpret communities based on a null model. There is a variety of algorithms based on approximation

techniques from various disciplines such as physics, mathematics, biology, and computational science. Some common methods are simulated annealing (Guimerà et al. 2004), greedy algorithm, spectral optimization techniques, etc.

For instance, simulated annealing is a global optimizing strategy based on the concept of slow cooling of solids, in physics. It uses a probabilistic strategy to provide a good approximation for a global optimum solution (Medus et al. 2005). It starts by arbitrarily assigning some communities to the network. Then, randomly moves node i into another community with a condition for the new community must have at least one edge already connected to node i. If the modularity Q increases the move is accepted, otherwise a probability eβ_Q is assigned to the move. The algorithm stops when no further improvement on modularity is possible, and finally, the state reached is the best-approximated state. The algorithm has been reported to give good results (Danon et al. 2005); however, it is slow. While greedy algorithm on the other hand starts considering each node as a partition. Then in every followed step, a single edge is added until all edges are added. Then, the optimized configuration is chosen by selecting the maximum modularity state. In addition to it, Schuetz and Caflisch (Schuetz and Caflisch 2008) suggested instead of allowing one community pair, more pairs should be allowed to avoid large accumulation into large communities. Other important community detection techniques include Modularity optimization and methods like Infomap algorithm a non-modularity-based, statistical inference and distance-based clustering algorithm, and other statistical inference methods.

2.6 Biomarker Discovery

Biomarker was defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" by a working group of National Institutes of Health Biomarkers in 1998. According to WHO, definition of biomarkers includes "almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological? The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction." Biomarkers are a measurement tool for assessing the health status of an individual measured from outside of the patient, i.e., they provide reliable and accurate indications of the state of their health (Strimbu and Tavel 2010) Currently, it is used as a chemical that is administered into an organism to check organ function or other health-related characteristics whose detection signals a certain disease's development, progression, and treatment effectiveness. For instance, an infection may be indicated by the presence of an antibody. More specifically, biomarkers can include certain cells, chemicals, genes, gene products, enzymes, or hormones. They can also include anything from blood pressure pulse and to more comprehensive laboratory testing of blood, urine, tissues, and other body fluids. Biomarkers can identify both complex organ processes and distinctive alterations in biological systems.

Biomarkers are used in the pharmaceutical industry to determine a drug's efficacy and safety, which can reduce development costs and lead to additional therapeutic targets. By using biomarkers, scientists can increase their precision in addition to providing more accurate diagnoses and recommending more efficient therapies.

The need for novel therapeutic approaches, such as regenerative medicine, the availability of potent new "omics" technologies, the rise of novel and unproven targets in the pharmaceutical industry, and the chance to use improved and well-validated biomarker assays have all played a role in the dramatic rise in the discovery of biomarkers over the last 10 years. In order to encourage the early removal of unpromising compounds and accelerate the release of breakthrough medicines and technology, the tool Support Industry has prioritized the safety and mode of action, i.e., potency-related indicators over disease-related biomarkers. Surrogate markers are required to replace inaccurate clinical end objectives, particularly in fields of newly evolving technologies such as regenerative medicine. Precisely, biomarkers are required to determine which treatments are most suited to each individual's needs (Krzyszczyk et al. 2018). The use of new and complicated technologies is improving biomarker discovery and development, raising the possibility that additional clinical uses of biomarkers will be applied to improve illness diagnosis, prognosis, and monitoring. Biomarker development is a multi-step process that includes basic research, validation, and clinical application.

2.7 Identification of Key Regulators

All hubs within a network play crucial roles in regulating its functions, but the most powerful and impactful genes are those that govern the network's operations at both the overall and motif-specific levels. These genes, referred to as the "Key Regulators," were identified through gene tracing techniques. By employing Newman and Girvan's community detection or clustering method, gene tracing was performed within numerous communities or sub-communities, extending up to the motif level (Tazyeen et al. 2022). By employing tracing techniques, it is possible to identify most influential and significant genes in the network that control the network.

Other Methods like *Rich-club finding algorithm* (Sect. 2.5.1), *Community finding algorithm* (Sect. 2.5.2), and *Biomarker Discovery* (Sect. 2.6) which we have discussed earlier in this chapter are also used for identification of Key regulators.

2.8 Statistical Properties and Models of Biological Network

2.8.1 Random Network Model

The Random Network Model was first developed by Pál Erdős and Alfréd Rényi. Since the large random networks follow the Poisson degree distribution, it is often called the Poisson random network model (Wu et al. 2017). Generally, a network

consists of only nodes and links. However, to reproduce the complexity of a real system, the links of the network must be placed appropriately. In the case of Random Networks, the links are randomly placed between the nodes (Fig. 2.3a). Two different models define the Random Network Model. The first is G(N, M) model in which a random network G is connected randomly with N vertices and M edges. Another model is G(N, p) in which instead of specifying M edges, the nodes are specified with probability p in the random network. Thus, the size of the preferred random network is produced both by specifying the number of edges, or the probability of observing the links between the edges. The random networks can be analyzed by investigative multiple graphs, where the number of nodes/vertices remains identical, but their links are randomly changed (Fig. 2.3a). The Random Network Model has led to several significant findings for network structure. The Random Network follows Poisson degree distribution and is completely connected for fairly small values of average degree. This signifies that for a random network each node does not need to be connected to too many other nodes. Another significant feature of a random network is that the connected random network is slightly compact even for large networks.

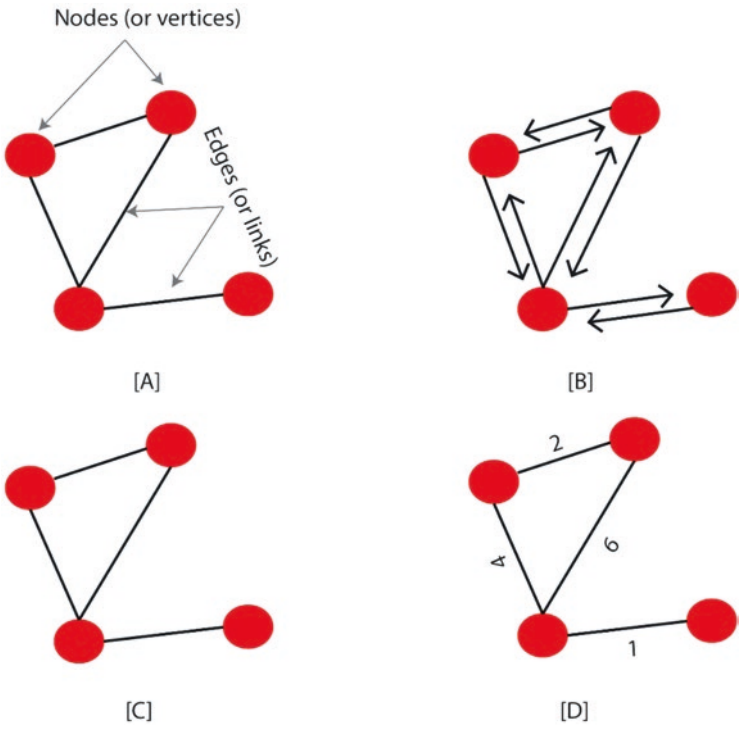

Fig. 2.3 (**a**) Random network. (**b**) Scale-free network. (**c**) Hierarchical network

2.8.2 Scale-Free Network Model

Despite the success of the random network model, there remains uncertainty whether real networks such as biological networks or social networks are truly random. This led to the discovery of the scale-free property which describes the structure and dynamics of complex systems (Barabási 2009). Since then, researchers have revealed scale-free structures in a wide range of systems. In the case of the Barabási–Albert model of a scale-free network, the network is governed by high-degree nodes which means nodes that are highly connected (Fig. 2.3b). These highly connected nodes are called hubs (Fornito et al. 2016). For the reason that the real-world networks are scale-free, some degree of nodes with degree k follows a power law $k^{-\alpha}$, where $\alpha > 1$. Scale-free networks are robust against accidental failures but susceptible to coordinated attacks. The links in scale-free networks are established based on two different mechanisms, i.e., growth and preferential attachment. A new node prefers to get attached to a node that already has many connections. This will ultimately result in a network that is dominated by a few highly connected nodes called hubs.

2.8.3 Hierarchical Network Model

Several real-life networks like protein–protein interaction networks, metabolic networks, or some social networks exhibit scale-free properties along with high clustering. This led to the evolution of hierarchical network models which incorporate the scale-free topology and high clustering into one single model (Ravasz and Barabási 2003). For generating the hierarchical network models, it is expected that clusters associate in an iterative style for the co-occurrence of local clustering, modularity, and scale-free topology in various real systems (Fig. 2.3c). We can also say that the hierarchical network model shares its foremost property of having a greater number of hubs in the network with the scale-free model family. However, unlike other alike models (Barabási–Albert, Watts–Strogatz), in hierarchical models, the nodes with a greater number of connections are likely to have a lesser clustering coefficient. In the case of the Barabási-Albert model, with the increase in the number of nodes the average clustering coefficient decreases. On the contrary, in a hierarchical network, there is no exact pattern and relationship between the network size and its average clustering coefficient.

2.9 Biological Network Databases

The analysis and modelling of biological networks, as well as their investigation, are crucial tasks in modern life sciences. The majority of biological networks are still far from full, and because of the complexity of the relationships and the unique characteristics of the data, they are frequently challenging to understand (Zhang and Itan 2019). Therefore, a major problem for bioinformatics is the creation of

sufficient storage and querying technologies. Initially, flat files and relational databases were used to construct storage systems. Despite their simplicity, both types have limitations in terms of access times and querying capability. As a result, a lot of databases have come onto the scene recently (Guzzi and Roy 2020).

Here, we will talk about categories of biological network databases as well as some helpful online resources that have biological network information.

2.9.1 Biological General Repository for Interaction Datasets

It is a free database that consists manually curated proteins from several species, like yeast, mouse, fly, worm, and human. (Oughtred et al. 2021a). In Biological General Repository for Interaction Dataset (BioGRID), curated interactions can be used to create complicated networks capable of speeding up the discovery of new biomedical treatments, especially for conditions affecting human health and disease. Its data is derived solely from primary experimental data found in the biomedical literature, and it includes both narrowly focused low-throughput trials and large high-throughput datasets. It also tracks how proteins change post-translationally and how proteins or genes interact with bioactive small molecules, including many well-known medications. All annotations are incorporated using an integrated network visualization tool that allows users to generate network graphs of protein, chemical, and genetic linkages (Oughtred et al. 2021b).

2.9.2 The Database of Interacting Proteins

This database contains a list of protein interactions that have been verified through experimentation. It integrates data from several sources to create a single, reliable list of protein–protein interactions. The information on protein–protein interaction networks was retrieved from the most trustworthy, core subset of the Database of Interacting Proteins (DIP) data and used in computational ways to both manually and automatically curate the data housed in the DIP database (DIP:Home n.d.). DIP contains PPI from various organisms that have been experimentally confirmed. It is implemented as a relational database. Each DIP entry comprises generic protein information (e.g., gene name and cellular location), as well as cross-references to other databases and information about experimental methods and specific experiments. Each interaction is issued a unique code. DIP interactions must be detailed in peer-reviewed papers, and the entry process is manual. A web interface can be used to create queries in both interactive and batch modes. A user can download a subset of DIP in many formats in batch mode.

2.9.3 Biomolecular Interaction Network Database

Biomolecular Interaction Network Database (BIND) comprises protein interactions that have been annotated with molecular function information gathered from the literature. It is based on three types of data records: pathways, molecular complexes, and interactions. The database allows various types of searches, including those that use identifiers from other biological databases or those that focus on certain fields like literature data, molecule structure, and gene data, including functions. With the use of a BIND interaction viewer, the extracted data may be seen. Graphs are used to describe networks, and the nodes in the graphs, which represent molecules, are labeled with various pieces of ontological data (Bader et al. 2001). A web-based system is readily available for searching, examining, and submitting records. Individual contributions, interaction data from the PDB, and various large-scale interaction and complex mapping studies using mass spectrometry, yeast two hybrid, genetic interactions, and phage display have all been added to BIND. The graphical analysis tool, which helps connect functional domains to protein interactions, allows users to visualize the domain composition of proteins in interaction and complex data. In addition, a tool for grouping interaction networks has been developed to aid in focusing on crucial areas. (Bader et al. 2003).

2.9.4 IntAct

IntAct is a free database and toolbox for storing, presenting, and analyzing protein interactions. It contains not only protein interactions data but also DNA and molecular interactions data (Orchard et al. 2014). All interactions are either submitted directly by the user or derived from literature curation. The web interface of IntAct provides users with textual and visual depictions of protein interactions, allowing exploration of interaction networks in relation to the Gene Ontology (GO) annotations of the involved proteins. Additionally, a web service is available to facilitate computational retrieval of interaction networks in XML format. Currently, IntAct contains approximately 2200 binary and complex interactions, which have been meticulously curated in collaboration with the Swiss-Prot team and extensively annotated using controlled vocabularies to maintain data consistency (Hermjakob et al. 2004).

It establishes a system for semantically consistent annotation by utilizing regulated vocabularies and ontologies. Researchers can submit PSI-MI interactions to the database curators via email.

2.9.5 Online Predicted Human Interaction Database

It contains predicted interactions between human proteins. It mixes PPI obtained from databases and books with hypotheses generated by other organisms. Online Predicted Human Interaction Database (OPHID) graph visualization tool allows

users to view the results of queries made using one or more protein IDs. For academic use, the database is available without charge (Brown and Jurisica 2005).

2.9.6 Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)

It is a repository that intends to collect biochemical relationships between proteins, proteins and DNA, and DNA and DNA. On the website, you can access the database by entering a protein identification or the main sequence of a protein. The STRING database's goal is to gather, score, and integrate all publicly available sources of protein–protein interaction data, as well as to supplement them with computational predictions. Its ultimate goal is to create a comprehensive and objective global network that includes both direct (physical) and indirect (functional) interactions. The most crucial feature is the ability to upload complete genome-wide datasets as input, which allows users to visualize subsets as interaction networks and do gene set enrichment analysis on the entire input (Szklarczyk et al. 2019).

2.9.7 Molecular Interaction

It is a curated database of PPIs for multiple model organisms. Molecular Interaction's (MINT's) objective is to systematically collect and organize information regarding molecular interactions by retrieving experimental details from peer-reviewed journal publications. Over time, the number of curated physical interactions in the database has increased to approximately 95,000. This comprehensive dataset is publicly accessible online through web-based interfaces and an FTP server, enabling users to explore the data interactively or in batch mode. MINT also contains HomoMINT, a database that focuses on interactions between human proteins discovered through analysis of orthologous proteins in model species (Chatr-aryamontri et al. 2007).

2.9.8 Regulatory Network Repository

This resource focuses on five main types of post-transcriptional regulatory connections in humans and mice. It contains in-depth details on numerous combinations of synergistic organizational interactions between TFs, miRNAs, and genes. The flexible architecture of Regulatory Network Repository's (RegNetwork's) database allows for future expansions to encompass gene regulatory networks of other organisms. It encompasses a comprehensive compilation of experimentally observed or predicted regulatory interactions at the transcriptional and post-transcriptional levels. Utilizing RegNetwork, researchers can delve into context-specific investigations of transcriptional and post-transcriptional regulatory interactions by leveraging domain-specific experimental data (Liu et al. 2015).

2.9.9 Transcriptional Regulatory Relationships Unraveled by Sentence-Based Text Mining

It is a repository of human and mouse transcriptional regulatory networks. It contains 8444 TF-target regulatory interactions from 800 human TFs and 6552 interactions from 828 mouse TFs. It is now the most comprehensive publicly accessible database of human transcription factor (TF)–target interactions. It enriches TF–target pairs significantly, especially in highly ranked interactions inferred from high-throughput data. This shows that Transcriptional Regulatory Relationships Unraveled by Sentence-Based Text Mining (TRRUST) can be used to reconstruct human transcriptional regulatory networks (TRNs) computationally (Han et al. 2018).

2.9.10 miRTarBase

This database collects verified interactions between microRNA and its targets. Six hundred fifty seven miRNAs and 2297 target genes from 17 species, including human, mouse, chicken, sheep, and others, are stored in this database. It compiled a database of around 3500 MTIs by manually examining relevant literature and painstakingly data mining the text to filter research articles linked to functional investigations of miRNAs. The MTIs collected in the miRTarBase can also provide a considerable number of positive samples for the development of computational algorithms capable of finding miRNA–target interactions. This database also makes use of gene ontology and KEGG pathway enrichment annotation to investigate the functionality of target genes involved in human MTIs (Hsu et al. 2011).

2.9.11 BioCyc Pathway/Genome Databases (PGDBs)

Thousands of sequenced organisms' metabolic pathways and genome information are available in this database. These are created by the same software that anticipates the metabolic pathways of completely sequenced organisms and identifies the operons and genes that provide the necessary enzymes. Additionally, information from other bioinformatics databases is incorporated, including Gene Ontology details and the protein features from UniProt. On its website, a variety of software applications are available for database searching and visualization, omics data processing, comparative genomics, and comparative pathway research (Karp et al. 2017).

2.9.12 MetaCyc

It is a reference database that includes metabolic processes and enzymes from different domains of life. It lists the primary and secondary metabolic pathways as well

as the metabolites, operations, enzymes, and genes associated with them. It is the most extensive curated compilation of metabolic pathways, with 2749 pathways drawn from over 60,000 publications. Its content is carefully selected and evidence-based, resulting in an encyclopedic reference tool for metabolism. It is also used to produce several hundred organism-specific Pathway/Genome Databases (PGDBs) that can be assessed from BioCyc.org. MetaCyc aims to create a database of the entire world of metabolism by conserving a representative sample of each experimentally elucidated clarified pathway (Caspi et al. 2020).

2.9.13 ENZYME

The ENZYME database is a library of information about the nomenclature of enzymes. It has been recently developed into a crucial tool for the creation of metabolic datasets. In the most recent version, there are details on 3705 enzymes. It is mostly based on recommendations made by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB), and it describes every type of recorded enzyme for which an EC (Enzyme Commission) number has been assigned (Bairoch 2000).

2.9.14 Reactome

It is a freely available and reviewed database having knowledge about pathways in living organisms. It shows how different molecules, like DNA, proteins, and small molecules, work together in biological processes such as traditional intermediate metabolism, signaling, innate and adaptive immunity, and apoptosis. Its objective is to offer bioinformatics tools that will facilitate fundamental research, genomic analysis, modelling, systems biology, and education by visualizing, interpreting, and analyzing pathway knowledge. With a focus on producing precise and reliable answers for genome-wide datasets with interactive reaction times, it offers a variety of pathway analysis tools. Pathway analysis techniques are commonly used to analyze Omics data generated by high-throughput technology (Fabregat et al. 2017).

2.9.15 KaPPA-View4

It is a database that stores metabolic pathways and enables the visual representation of gene-to-gene and metabolite-to-metabolite relationships as curves on a metabolic pathway map or a combination of up to four maps. This illustration is helpful in uncovering new roles for transcription factors that control the genes in a metabolic pathway. Its website http://kpv.kazusa.or.jp/kpv4-kegg/ provides access to KEGG pathway maps and maps produced from their gene classifications (Sakurai et al. 2011).

2.9.16 Netpath Pathway

NetPath is a database of carefully selected human signaling pathways. NetPath provides extensive maps of several immunological signaling pathways, including over 2800 examples of transcriptionally regulated genes and roughly 1600 reactions retrieved from the literature all of which are connected to more than 5500 research publications. It is the result of collaboration between Johns Hopkins University's Pandey Lab and the Institute of Bioinformatics. NetPath's all pathways can be downloaded free of cost in SBML version 2.1 and BioPAX level 3.0, PSI-MI version 2.5 (Kandasamy et al. 2010).

2.9.17 TRANSFAC

The database TRANSFAC offers details on the genomic binding locations and DNA-binding preferences of eukaryotic transcription factors. It has a library of positional weight matrices that includes a unique collection of DNA-binding models that can be used to conduct a complete examination of genomic sequences for probable transcription factor binding sites (TFBSs). It can be used as a transcriptional regulatory encyclopedia or as a tool to identify possible TFBSs (Matys et al. 2003).

2.9.18 Human Protein Reference Database

Human Protein Reference Database (HPRD) is a unified platform for visually representing and integrating information about domain architecture, post-translational modifications, interaction networks, and disease associations for each protein in the human proteome.

HPRD serves as a consolidated platform that visually presents and integrates various data about post-translational modifications, domain architecture, interaction networks, and disease associations for every protein present in the human proteome. The data within this database is meticulously curated from scientific literature by proficient biologists who carefully read, comprehend, and analyze the published information. HPRD was constructed using an object-oriented database in Zope, an open-source web application server that provides flexible querying capabilities and enables dynamic display of data (Peri et al. 2004).

2.9.19 DisGeNET

DisGeNET is among the most complete databases of its sort currently available, containing over 38, 0000 linkages between more than 16,000 genes and 13,000 variants and disorders of genes implicated in human disorders. It incorporates data from GWAS catalogs, animal models, scholarly literature, and expert-curated sources. Its

data is routinely tagged with community-driven ontologies and controlled vocabularies. It is a platform that can be utilized in a number of research tasks, including investigating the molecular causes of particular human diseases and their comorbidities, analyzing disease gene characteristics, developing hypotheses about drug therapeutic effects and side effects, confirming computationally predicted disease genes, and evaluating the effectiveness of text mining techniques.

The data can be accessed using Cytoscape, a web interface application.

2.9.20 Drug Bank

Drug Bank is an extensively annotated database that integrates comprehensive drug data with detailed information about drug targets and drug actions. It has been extensively employed in various applications such as identifying in silico drug targets, conducting drug docking or screening experiments, predicting drug metabolism and interactions, facilitating drug design, as well as supporting pharmaceutical education in a comprehensive manner (Wishart et al. 2008).

2.9.21 The Molecular Signatures Database (MSigDB)

It is one of the gene set databases for gene set enrichment analysis. It is among the most popular and extensive repositories of gene sets primarily used for gene set enrichment analysis. Initially focused on metabolic disorders and cancer, it has significantly grown over time and now encompasses over 10,000 sets of genes.

This database provides a range of gene expression signatures, including empirically obtained signatures and signatures describing pathways and ontologies from other curated resources.

2.9.22 Kyoto Encyclopedia of Genes and Genomes

Kyoto Encyclopedia of Genes and Genomes (KEGG) serves as a valuable resource for conducting systematic analyses of gene activities through the exploration of gene and molecular networks. The primary component of KEGG is the PATHWAY database, which provides graphical representations of biochemical pathways, including a wide range of metabolic pathways and selected regulatory processes. Another essential aspect of KEGG is the ortholog group tables, which present information on orthologous and paralogous gene groups across multiple organisms. These tables contribute to the expression of pathway information within the KEGG database (Kanehisa and Goto 2000).

KEGG handles the GENES database, which includes gene catalogs for organisms with complete genomes and some with incomplete genomes. Apart from collecting data, KEGG also provides computational tools. These tools help in

reconstructing biochemical pathways from entire genome sequences and predicting gene regulatory networks.

2.9.23 NCBI Gene Expression Omnibus

Established in 2000, the Gene Expression Omnibus (GEO) database is a freely accessible repository that serves as a global resource for gene expression studies. Its primary purpose is to preserve and distribute high-throughput gene expression data sets and other functional genomics data. Over time, GEO has expanded to accept high-throughput data related to various applications, such as genome–protein interactions, genome methylation, chromatin structure, and keeping pace with advancing technologies. The database contains a vast collection of tens of thousands of research data sets and offers web-based tools and techniques for users to discover data relevant to their specific interests. Furthermore, users can leverage these tools to visualize and analyze the data within the database (Clough and Barrett 2016).

2.9.24 EBI Array Express

Array Express serves as a publicly accessible repository dedicated to microarray-based gene expression data. It stores meticulously annotated raw and normalized data, which can be submitted online in a standardized format or directly from local databases or LIMS (Laboratory Information Management Systems). Reviewers and writers are granted password-protected access to prepublication data. Access to the stored data can be obtained through accession numbers or by employing various search criteria like species, author, and array platform. Additionally, a subset of curated data deposited in the Array Express data warehouse offers the capability to query experiments based on gene and sample attributes. For further analysis and visualization of the data, Array Express provides an integrated data analysis tool called Expression Profiler (Rocca-Serra et al. 2003).

2.10 Role of Network Modules in Disease Dynamics

In molecular networks, the concept of modularity is widely accepted. Module-based approaches have a number of advantages, including improved disease classification and robustness in the discovery of dysregulated pathways. Module-centric techniques are particularly promising in their investigation because it is thought that complex disorders are caused by a variety of genetic changes altering a common element of the biological system. How can disease-related modules and sub-networks be found? Sub-networks impacted by a specific disease can be differentiated from the broader network by combining the interaction data with extra information available on the disease state.

To detect modified network modules and elucidate the correlation between phenotypic and genotypic data (Cho et al. 2012), scientists have employed a combination of molecular phenotypic data, such as gene expression patterns observed in diseased samples, and genotypic information such as single nucleotide polymorphisms (SNPs) and copy number alterations.

By mapping the genes exhibiting alterations in diseases onto a protein–protein interaction (PPI) network and subsequently identifying network modules enriched with these altered genes, it becomes possible to uncover dysregulated pathways. This approach operates under the hypothesis that complex diseases arise from a collection of mutations that, while varying considerably among patients, tend to dysregulate shared pathways.

On the other hand, molecular-level alterations like changes in gene expression are directly associated to organismal-level phenotypes like diseases. Therefore, the modules that are enriched with differentially expressed genes are taken into account by a different group of methodologies and it is possible to think of molecular pathways as informational channels.

For example, the activation of the EGFR (epidermal growth factor receptor) by its receptor triggers the activation of various downstream signaling proteins, initiating multiple signal transduction cascades such as the MAPK, Akt, and JNK pathways, ultimately leading to cell proliferation. Consequently, the third category of methods focuses on predicting the molecules and modules that govern the transfer of such specific information (Wee and Wang 2017).

What advantages do phenotypic and genotypic variations in disease have in relation to their molecular interactions? Firstly, through a groundbreaking approach, *Ideker* et al. (Ideker et al. 2002) integrated yeast protein–protein and protein–DNA interactions with gene expression changes resulting from perturbations in the yeast galactose utilization pathway. They successfully identified active sub-networks, which encompassed interconnected genes with significant differential expression. Interestingly, these sub-networks included common transcription factors that exhibited modest changes in gene expression but played a crucial role in connecting other dysregulated genes (Hughes and de Boer 2013). Secondly, employing a module-based method enhances statistical power, enabling the identification of perturbed modules even when individual gene perturbations are statistically insignificant. This is particularly relevant in genetic disorders such as autism and schizophrenia, where rare germline alterations pose challenges in distinguishing them from background noise. Recent studies have shown that many of the altered genes are part of highly interconnected protein networks. Consequently, a network-based approach becomes more effective in identifying the causal genes in such scenarios.

Thirdly, recognized network modules provide a deeper understanding of the underlying biological mechanisms of diseases, leading to more precise indicators for disease diagnosis and treatment. This enables the identification of specific targets and strategies for intervention. This holistic perspective offered by network modules contributes to improved accuracy and effectiveness in disease management.

2.11 Case Study Based on Inference and Analysis of Network Related to Disease Dynamics

When used to analyze breast cancer metastasis, *Chuang* et al. found that dysregulated network modules provide predictions that are more reliable and precise than single gene-based classifications (Chuang et al. 2007). This research established the efficacy of using network modules to classify diseases. Later on, various modifications and changes were proposed to their work. Like (Lee et al. 2008), curated pathways and a subset of genes having characteristics that may be used to distinguish between different disease phenotypes were used by *Lee* et al. and *Dao* et al. (Wu et al. 2013) created new network-based methods for classifying cancer subtypes by locating highly connected sub-networks using randomized algorithms and for ideal marker identification methods like bottom-up enumeration and set cover were also presented.

Kim et al. found gene modules using a module cover technique to identify disease heterogeneity in Rembrandt brain cancer data and TCGA ovarian cancer samples (Kim et al. 2013). These selected modules were then integrated with the results of an independently proposed classification scheme and this led to the discovery of disease type characterization on the basis of module combinations.

Disease homology can also be discussed using network modules. Overlaps of dysregulated network modules help to explain why some complicated diseases have similar phenotypic characteristics. A PathBlast model was used by Suthram et al. (Sharan et al. 2005) to locate dense sub-networks. By evaluating the expression patterns of various diseases in the modules, analysis of disease homology was performed. It was shown that some dysregulated modules are shared by various diseases, this reveals why some medications are effective for a variety of ailments.

The PPI network approaches can also be used to identify key regulatory genes of the network that regulates the activities and signal flow of the complete network. In one of our studies, we analyzed the Turner Syndrome (TS) regulatory network constructed from manually curated genes that were involved in TS from published literatures. Protein–protein interactions, functions, TS networks, and orthologs were all combined to accomplish this. It was found that the TS network shows hierarchical features, signifying system-level organization involving the occurrence of modules/communities interrelated in a certain fashion. Two key regulators namely, KDM6A and BDNF were identified. These key regulators serve as the backbone of the network and are deeply rooted. Any network activities are regulated by these genes and could be possible target genes for resisting the syndrome. Because the network is hierarchical, removing KDM6A and BDNF does not cause the network to break down; rather, the network reorganizes itself to accommodate the alteration. We combined the network-based study with phylogenetic study to find a few essential and conserved interactions (interologs). KDM6A-WDR5, KDM6A-ASH2L, and WDR5-ASH2L are three significant interologs (evolutionarily conserved protein–protein interactions) that we identified, constituting a motif (Farooqui et al. 2018).

In another study, we examined the Turner Syndrome network created from the microarray expression data of TS. We again identified the key regulators of TS through microarray data of TS by combining functions, protein interactions, orthologs, disease networks, and its correlation with two common comorbidities, Recurrent Miscarriage (RM) and Diabetes Mellitus Type 2 (T2DM). Some important signature genes of the above comorbidities were identified. It was found that the TS network shows hierarchical features, signifying system-level organization involving occurrence of modules/communities interrelated in a certain fashion. POU2F1, LCP2, CCL22, ENAM, PTPN22, CXCL5, S1PR4, FAM20A, and EFNB3 were identified as important regulators (Farooqui et al. 2021). These key regulators act as the backbone of the network and are deeply rooted (Ideker et al. 2002). Any network activities are regulated by these genes and could be possible target genes for resisting the syndrome.

The study of complicated diseases can be benefited by the effective tools provided by network biology. The concept behind network-based techniques is that rather than looking at individual genes, dysregulated modules can provide a more comprehensive understanding of complicated disorders.

2.12 Conclusion

Great efforts have been made over the last two decades to extract the dependence and interplay between structure and function in biological networks because they are highly relevant to biological processes. To summarize, biological networks provide a conceptual and intuitive framework for studying, modelling, characterizing, and comprehending complicated interactions between various components of a biological system. Network biology studies the "interactome," which is a collection of direct and indirect molecular connections in biological systems. Many areas of biomedical science have benefited from network biology. This simple but powerful concept enables us to extract the essence of gene–protein interactions, predict drug interactions, study disease comorbidity, and discover important associations. Of course, identifying an association is only the first step toward identifying a mechanism, but it is frequently a critical step. Because of significant advancements in data capture, computational tools, and network models, network biology has emerged as the first and most crucial step in bioinformatics, providing an approach for understanding the structure–function relationship in biological systems. Recent developments in this subject demonstrate that it can be used to infer biological organization, function, and evolution.

References

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97. https://doi.org/10.1103/RevModPhys.74.47

Albert R, Wang RS (2009) Discrete dynamic modeling of cellular signaling networks. Methods Enzymol 467:281–306

Alon U (2019) An introduction to systems biology: design principles of biological circuits, 2nd edn. CRC Press, Boca Raton
Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV (2001) BIND—the biomolecular interaction network database. Nucleic Acids Res 29(1):242–245
Bader GD, Betel D, Hogue CWV (2003) BIND: the biomolecular interaction network Db. Nucleic Acids Res 31(1):248–250. https://bio.tools/bind. Accessed 06 Jul 2022
Bairoch A (2000) The enzyme database in 2000. Nucleic Acids Res 28(1):304–305. https://doi.org/10.1093/nar/28.1.304
Barabási A-L (2009) Scale-free networks: a decade and beyond. Science 325(5939):412–413. https://doi.org/10.1126/science.1173299
Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman–Girvan and other modularities. Proc Natl Acad Sci 106(50):21068–21073. https://doi.org/10.1073/pnas.0907096106
Boucher B, Jenna S (2013) Genetic interaction networks: better understand to better predict. Front Genet 4:290. https://doi.org/10.3389/fgene.2013.00290
Brown KR, Jurisica I (2005) Online predicted human interaction database. Bioinformatics 21(9):2076–2082. https://doi.org/10.1093/bioinformatics/bti273
Caspi R et al (2020) The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res 48(D1):D445–D453. https://doi.org/10.1093/nar/gkz862
Chatr-aryamontri A et al (2007) MINT: the molecular INTeraction database. Nucleic Acids Res 35(Database):D572–D574. https://doi.org/10.1093/nar/gkl950
Cho D-Y, Kim Y-A, Przytycka TM (2012) Chapter 5: network biology approach to complex diseases. PLoS Comput Biol 8(12):e1002820. https://doi.org/10.1371/journal.pcbi.1002820
Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140. https://doi.org/10.1038/msb4100180
Clough E, Barrett T (2016) The gene expression omnibus database. Methods Mol Biol Clifton NJ 1418:93–110. https://doi.org/10.1007/978-1-4939-3578-9_5
Costanzo M et al (2016) A global interaction network maps a wiring diagram of cellular function. Science 353(6306):aaf1420. https://doi.org/10.1126/science.aaf1420
Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech Theory Exp 2005(09):P09008. https://doi.org/10.1088/1742-5468/2005/09/P09008
DIP:Home (n.d.). https://dip.doe-mbi.ucla.edu/dip/Main.cgi. Accessed 6 Jul 2022
Ertaylan G, Okawa S, Schwamborn JC, del Sol A (2014) Gene regulatory network - an overview. Front Cell Neurosci 8:437. https://www.sciencedirect.com/topics/neuroscience/gene-regulatory-network. Accessed 27 Jul 2022
Fabregat A et al (2017) Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinformatics 18(1):142. https://doi.org/10.1186/s12859-017-1559-2
Farooqui A et al (2018) Assessment of the key regulatory genes and their interologs for turner syndrome employing network approach. Sci Rep 8(1):10091. https://doi.org/10.1038/s41598-018-28375-0
Farooqui A, Alhazmi A, Haque S et al (2021) Network-based analysis of key regulatory genes implicated in Type 2 Diabetes Mellitus and Recurrent Miscarriages in Turner Syndrome. Sci Rep 11:10662. https://doi.org/10.1038/s41598-021-90171-0
Fornito A, Zalesky A, Bullmore E (2016) Centrality and hubs. In: Fundamentals of brain network analysis. Elsevier, pp 137–161. https://doi.org/10.1016/B978-0-12-407908-3.00005-4
Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799
Grigorov MG (2005) Global properties of biological networks. Drug Discov Today 10(5):365–372
Guimerà R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70(2):025101. https://doi.org/10.1103/PhysRevE.70.025101
Guimerà R, Sales-Pardo M, Amaral LAN (2007) Classes of complex networks defined by role-to-role connectivity profiles. Nat Phys 3(1):63–69. https://doi.org/10.1038/nphys489

Guzzi PH, Roy S (2020) Biological network databases. In: Biological network analysis. Academic, pp 77–93. https://www.researchgate.net/publication/341704175_Biological_network_databases. Accessed 06 Jul 2022

Haggart CR, Bartell JA, Saucerman JJ, Papin JA (2011) Whole-genome metabolic network reconstruction and constraint-based modeling. Methods Enzymol 500:411–433. https://doi.org/10.1016/B978-0-12-385118-5.00021-9

Han H et al (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 46(D1):D380–D386. https://doi.org/10.1093/nar/gkx1013

Hermjakob H et al (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32(Database issue):D452–D455. https://doi.org/10.1093/nar/gkh052

Hsu S-D et al (2011) miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic Acids Res 39(Database issue):D163–D169. https://doi.org/10.1093/nar/gkq1107

Hughes TR, de Boer CG (2013) Mapping yeast transcriptional networks. Genetics 195(1):9–36. https://doi.org/10.1534/genetics.113.153262

Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(Suppl 1):S233–S240. https://doi.org/10.1093/bioinformatics/18.suppl_1.S233

Jain S, Krishna S (2001) A model for the emergence of cooperation, interdependence, and structure in evolving networks. Proc Natl Acad Sci 98(2):543–547. https://doi.org/10.1073/pnas.98.2.543

Jordán F, Nguyen PT, Liu W-C (2012) Studying protein-protein interaction networks: a systems view on diseases. Brief Funct Genomics 11(6):35. https://doi.org/10.1093/bfgp/els035. https://www.researchgate.net/publication/230717082_Studying_protein-protein_interaction_networks_A_systems_view_on_diseases. Accessed 27 Jul 2022

Kandasamy K, Mohan SS, Raju R et al (2010) NetPath: a public resource of curated signal transduction pathways. Genome Biol 11:R3. https://doi.org/10.1186/gb-2010-11-1-r3. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-1-r3. Accessed 07 Jul 2022

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30. https://doi.org/10.1093/nar/28.1.27

Karp PD et al (2017) The Biocyc collection of microbial genomes and metabolic pathways. Brief Bioinform 20(4):1085–1093. https://doi.org/10.1093/bib/bbx085

Kaviani S, Sohn I (2021) Complex network theory - an overview. Expert Syst Appl 180:115073. https://www.sciencedirect.com/topics/computer-science/complex-network-theory. Accessed 27 Jul 2022

Kim Y-A, Salari R, Wuchty S, Przytycka TM (2013) Module cover - a new approach to genotype-phenotype studies. Pac Symp Biocomput 2013:135–146

krzyszczyk P et al (2018) The growing role of precision and personalized medicine for cancer treatment. Technology 6(3–4):79–100. https://doi.org/10.1142/S2339547818300020

Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási A-L (2008) The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci 105(29):9880–9885. https://doi.org/10.1073/pnas.0802208105

Liu Z-P, Wu C, Miao H, Wu H (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015:bav095. https://doi.org/10.1093/database/bav095

MacNeil LT, Walhout AJM (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res 21(5):645–657. https://doi.org/10.1101/gr.097378.109

Matys V et al (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31(1):374–378. https://doi.org/10.1093/nar/gkg108

Medus A, Acuña G, Dorso CO (2005) Detection of community structures in networks via global optimization. Phys Stat Mech Its Appl 358(2–4):593–604. https://doi.org/10.1016/j.physa.2005.04.022

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827. https://doi.org/10.1126/science.298.5594.824

Muzio G, O'Bray L, Borgwardt K (2021) Biological network analysis with deep learning. Brief Bioinform 22(2):1515–1530. Oxford Academic. https://academic.oup.com/bib/article/22/2/1515/5964185. Accessed 27 Jul 2022

Newman MEJ (2012) Communities, modules and large-scale structure in networks. Nat Phys 8(1):25–31. https://doi.org/10.1038/nphys2162

Orchard S et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(D1):D358–D363. https://doi.org/10.1093/nar/gkt1115

Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-aryamontri A, Dolinski K, Tyers M (2021a) BioGRID database of protein, chemical, and genetic interactions. Protein Sci 30(1):187–200. https://thebiogrid.org/. Accessed 06 Jul 2022

Oughtred R et al (2021b) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci Publ Protein Soc 30(1):187–200. https://doi.org/10.1002/pro.3978

Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6(2):99–111

Patra S, Mohapatra A (2020) Review of tools and algorithms for network motif discovery in biological networks. IET Syst Biol 14(4):171–189. https://doi.org/10.1049/iet-syb.2020.0004. https://pubmed.ncbi.nlm.nih.gov/32737276/. Accessed 27 Jul 2022

Pavlopoulos GA et al (2011) Using graph theory to analyze biological networks. BioData Min 4(1):10. https://doi.org/10.1186/1756-0381-4-10

Pennisi E (2005) How will big pictures emerge from a sea of biological data? Science 309(5731):94–94. https://doi.org/10.1126/science.309.5731.94

Peri S et al (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32(Database):D497–D501. https://doi.org/10.1093/nar/gkh070

Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. Phys Rev E Stat Nonlinear Soft Matter Phys 67(2 Pt 2):026112. https://doi.org/10.1103/PhysRevE.67.026112

Rocca-Serra P et al (2003) ArrayExpress: a public database of gene expression data at EBI. C R Biol 326(10):1075–1078. https://doi.org/10.1016/j.crvi.2003.09.026

Rubinov M, Sporns O (2010) Complex network measures of brain connectivity: uses and interpretations. NeuroImage 52(3):1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

Sakurai N, Ara T, Ogata Y, Sano R et al (2011) kaPPA-view4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. Nucleic Acids Res 39(Database issue):D677–D684. Oxford Academic. https://academic.oup.com/nar/article/39/suppl_1/D677/2509108?login=true. Accessed 07 Jul 2022

Schuetz P, Caflisch A (2008) Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. Phys Rev E 77(4):046112. https://doi.org/10.1103/PhysRevE.77.046112

Sharan R et al (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci 102(6):1974–1979. https://doi.org/10.1073/pnas.0409522102

Song C, Havlin S, Makse HA (2006) Origins of fractality in the growth of complex networks. Nat Phys 2(4):275–281. https://doi.org/10.1038/nphys266

Srinivasa Rao V, Srinivas K, Sujini GN, G. N. (2014) Sunand Kumar Protein-protein interaction detection: methods and analysis. Int J Proteomics 2014:147648. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3947875/. Accessed 27 Jul 2022

Stelzl U et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957–968. https://doi.org/10.1016/j.cell.2005.08.029

Strimbu K, Tavel JA (2010) What are biomarkers? Curr Opin HIV AIDS 5(6):463–466. https://doi.org/10.1097/COH.0b013e32833ed177

Succoio M, Sacchettini R, Rossi A, Parenti G, Ruoppolo M (2022) Metabolic network - an overview. Biomol Ther 12(7):968. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/metabolic-network. Accessed 27 Jul 2022

System-Level Organization in a Hierarchical Network

Szklarczyk D et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(Database issue):D607–D613. https://doi.org/10.1093/nar/gky1131

Tazyeen S et al (2022) Identification of key regulators in Sarcoidosis through multidimensional systems biological approach. Sci Rep 12(1):1. https://doi.org/10.1038/s41598-022-05129-7

van den Heuvel MP, Sporns O (2011) Rich-club organization of the human connectome. J Neurosci 31(44):15775–15786. https://doi.org/10.1523/JNEUROSCI.3539-11.2011

Vieira LS, Vera-Licona P (2019) Computing signal transduction in signaling networks modeled as Boolean networks, petri nets, and hypergraphs. Biorxiv 2:272344. https://doi.org/10.1101/272344

Wee P, Wang Z (2017) Epidermal growth factor receptor cell proliferation signaling pathways. Cancer 9(5):52. https://doi.org/10.3390/cancers9050052

Wiredja D, Bebek G (2017) Identifying gene interaction networks, in *statistical human genetics: methods and protocols*. In: Elston RC (ed) Methods in molecular biology. Springer, New York, pp 539–556. https://doi.org/10.1007/978-1-4939-7274-6_27

Wishart DS et al (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(Database issue):D901–D906. https://doi.org/10.1093/nar/gkm958

Woodcock CL (2006) Chromatin architecture. Curr Opin Struct Biol 16(2):213–220. https://doi.org/10.1016/j.sbi.2006.02.005

Wu F, Huberman BA (2004) Finding communities in linear time: a physics approach. Eur Phys J B - Condens Matter 38(2):331–338. https://doi.org/10.1140/epjb/e2004-00125-x

Wu M-Y, Dai D-Q, Xiao-Fei Z, Zhu Y (2013) Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm. PLoS One 8:e66256. https://doi.org/10.1371/journal.pone.0066256

Wu J, Tan S-Y, Liu Z, Tan Y-J, Lu X (2017) Enhancing structural robustness of scale-free networks by information disturbance. Sci Rep 7(1):7559. https://doi.org/10.1038/s41598-017-07878-2

Yu D, Kim M, Xiao G, Hwang TH (2013) Review of biological network data and its applications. Genomics Inform 11(4):200–210. https://doi.org/10.5808/GI.2013.11.4.200

Zhang P, Itan Y (2019) Biological network approaches and applications in rare disease studies. Gene 10(10):797. https://doi.org/10.3390/genes10100797

Zhou S, Mondragon RJ (2004) The rich-club phenomenon in the internet topology. IEEE Commun Lett 8(3):180–182. https://doi.org/10.1109/LCOMM.2004.823426

3 Network Analysis Based Software Packages, Tools, and Web Servers to Accelerate Bioinformatics Research

Nikhat Imam, Sadik Bay, Mohd Faizan Siddiqui, and Okan Yildirim

Abstract

Biological networks are the best way to represent complex biological systems in terms of sets of interactions between biological entities (e.g., genes, proteins, taxa, and metabolites). Now, with the availability of large-scale multi-omics data, the biological system has expanded from basic to advanced levels (like PPI connectivity and functional changes in disease stages, disease-specific interactome, genetic perturbations, and network dysfunction). In this chapter, we discuss the fundamental concept of network theory and various network types such

Sadik Bay and Okan Yildrim are both currently affiliated with the Memorial Sloan Kettering Cancer Center in New York City, USA.

N. Imam (✉)
Department of Mathematics, Institute of Computer Science and Information Technology, Magadh University, Bodh Gaya, India

Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

S. Bay
Research Institute for Health Sciences and Technologies (SABITA), Istanbul Medipol University, İstanbul, Türkiye

M. F. Siddiqui
International Medical Faculty, Osh State University, Osh, Kyrgyzstan

O. Yildirim
Department of Chemical Biology Otto-Hahn-Strasse, Max Planck Institute of Molecular Physiology, Dortmund, Germany

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_3

as Protein–protein interaction networks, metabolic networks, genetic interaction networks, gene/transcriptional regulatory networks, and cell signaling networks. In addition, we describe network topological properties that help to understand the structure of a network which facilitates understanding the hidden mechanisms. Finally, we discuss a variety a number of online and stand-alone tools that exist for network construction, analysis, Network functional module Identification, and visualization. Overall, this chapter reaches a very broad spectrum of researchers varying from experts to beginners and they can look up a genetic universe virtually via biological networks.

Keywords

Network biology · Network visualization · Tools/software

3.1 Introduction

A biological network serves as a representation of complex systems, encompassing binary interactions between different biological entities, such as genes, proteins, taxa, and metabolites. The advancement of multi-omics data has expanded our understanding of biological systems, including PPI connectivity and functional changes during disease progression, disease-specific interactomes, genetic perturbations, and network dysfunction.

Protein–protein interaction networks (PINs) depict the physical connections between proteins within a cellular context. In these networks, proteins are represented as nodes, and their interactions are depicted as undirected edges. PPIs play a critical role in various cellular processes and have been extensively investigated in biological research. Experimental techniques, including the yeast two-hybrid system and mass spectrometry, have been utilized to discover and identify large sets of protein interactions.

In recent decades, numerous international initiatives have led to the development of databases that compile experimentally determined protein–protein interactions. Examples of these databases include the Munich Information Center for Protein Sequence (MIPS) protein interaction database (Schoof et al. 2005), the Biomolecular Interaction Network Database (BIND) (Bader 2003), the Database of Interacting Proteins (DIP) (Xenarios 2000), the Molecular Interaction database (MINT) (Chatraryamontri et al. 2007), and the protein Interaction database (IntAct) (Hermjakob 2004). These databases are classified into primary and secondary databases based on their interaction prediction method but now one new term introduced is “Meta-database” which is a combination of different primary and secondary databases to get new and maximum protein–protein interactions in the network (Fig. 3.2). The list of protein interaction databases is given in Table 3.1.

Table 3.1 Protein–protein interaction databases

Primary database			
Databases	URL	Experimental/ Predicted	References
BioGrid	https://thebiogrid.org/	Exp.	Oughtred et al. (2019)
HPRD	http://www.hprd.org/	Exp.	Goel et al. (2012)
IntAct	https://www.ebi.ac.uk/intact/home	Exp.	Hermjakob (2004)
MINT	https://mint.bio.uniroma2.it/	Exp.	Chatr-aryamontri et al. (2007)
HuRI	http://www.interactome-atlas.org/	Exp.	Luck et al. (2020)
Secondary database			
STRING	https://string-db.org/	Exp. & pred.	Szklarczyk et al. (2019)
UniHI	http://www.unihi.org/	Exp. & pred.	Kalathur et al. (2014)
Mentha	http://mentha.uniroma2.it/	Exp.	Calderone et al. (2013)
APID	http://cicblade.dep.usal.es:8080/APID/init.action	Exp.	Alonso-López et al. (2019)
HIPPIE	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/index.php	Exp.	Alanis-Lobato et al. (2017)
HitPredict	http://www.hitpredict.org/	Exp.	Patil et al. (2011)
IID	http://iid.ophid.utoronto.ca/	Exp. and pred.	Kotlyar et al. (2016)
HINT	http://hint.yulab.org/	Exp.	Das and Yu (2012)
GPS-Prot	http://gpsprot.org/	Exp.	Fahey et al. (2011)

- primary protein interaction databases containing literature-curated PPIs for human proteins
- secondary protein interaction databases containing predicted and curated from primary databases

3.2 Types of Biological Networks

There are various types of networks (Fig. 3.1). The details are given below:

- *Protein–protein interaction (PPI) networks.*
- Protein–protein interaction networks (PINs) serve as mathematical models that depict the direct interactions between proteins. These interactions, known as PPIs, play a fundamental role in nearly every cellular process, making it imperative to comprehend them in order to understand cell physiology under both normal and disease conditions.
- *Metabolic Networks.*
- Metabolic networks illustrate the relationships between enzymes and small biomolecules, referred to as metabolites, which are proteins that facilitate biochemical reactions. These networks illustrate the interactions that occur during the catalysis of these reactions.

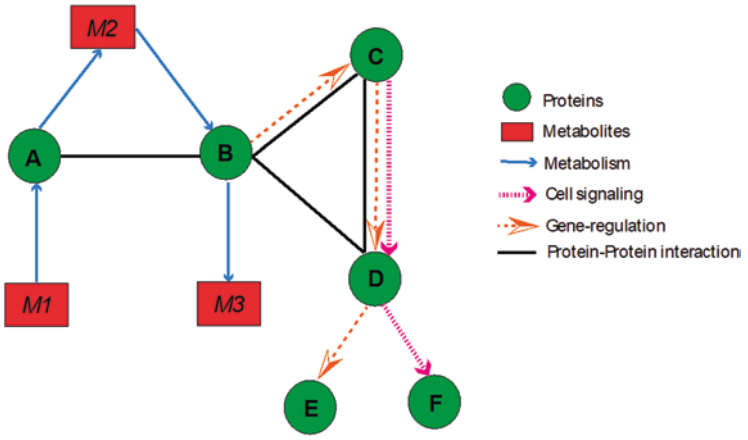

Fig. 3.1 Types of biological interactions that networks can represent

- *Genetic Interaction Networks.*
- Genetic interaction networks provide insights into the functional relationships between pairs of genes within an organism, facilitating our understanding of the connection between genotype and phenotype.
- *Gene/transcriptional Regulatory Networks.*
- A genetic regulatory network (GRN), also known as a gene regulatory network, comprises a set of molecular regulators that interact with each other and other substances within a cell. Its purpose is to control the levels of gene expression for mRNA and proteins.
- *Cell Signaling Networks.*
- Cellular signaling networks emerge through the interaction of various cell signaling pathways and are typically identified through a combination of experimental and computational techniques.

3.3 Network Topology

The properties of networks are valuable in extracting meaningful information. Network analysis aims to utilize the complexity of networks to uncover insights that would be challenging to obtain by examining individual components in isolation. Topological properties, such as the arrangement of nodes and edges within a network, are instrumental in identifying significant sub-structures. These properties can be applied to the network as a whole or to individual nodes and edges. Various topological properties and concepts are commonly used in network analysis (Fig. 3.2).

- *The Degree of a Network*
- The degree of a node indicates the number of edges connected to that specific node. This parameter is highly significant as it influences various characteristics, such as node centrality. By examining the degree distribution across all nodes in

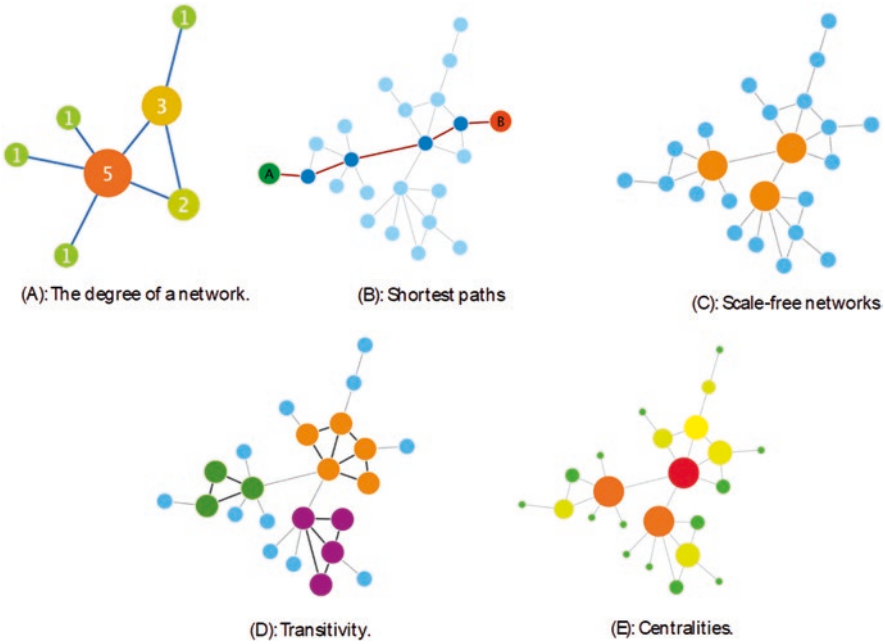

Fig. 3.2 Topological properties of a network (figure adopted from EMBL-EBI Training)

a network, we can determine whether the network exhibits a scale-free pattern, which we will explore in more detail. In visual representations, such as Fig. 3.2a, the size and color of each node correspond to its degree. In directed networks, nodes possess two types of degrees: out-degree, which represents edges leaving the node, and in-degree, which represents edges entering the node.

- *Shortest Paths*
- Shortest paths, which represent the minimum distance between any two nodes, provide a model for information flow within networks. This concept holds particular relevance in numerous biological networks. In Fig. 3.2b, the shortest path between nodes A and B is highlighted and spans five steps.
- *Scale-Free Networks*
- In scale-free networks, most nodes have connections to only a limited number of neighboring nodes. However, a small subset of high-degree nodes, known as hubs and highlighted in orange, play a crucial role in ensuring the network's overall connectivity. This configuration is depicted in Fig. 3.2c.
- *Transitivity*
- Transitivity refers to the existence of closely interconnected nodes within a network, often referred to as clusters or communities. These groups of nodes exhibit stronger internal connections compared to the connections with the rest of the network, as depicted in Fig. 3.2d. Alternatively, these groups are also known as topological clusters.
- *Centralities*

- Centrality measures encompass different concepts and can be assessed for both nodes and edges to determine their significance in terms of network connectivity and information flow. The degree of a node directly affects various centrality measures, such as "degree centrality," but its influence decreases in more advanced measures like "betweenness centrality." Figure 3.2e highlights nodes with higher betweenness centrality, indicating greater centrality, using warm colors. The size of each node in the figure corresponds to its degree.

3.4 Network Representation and Analysis Tools

There are multiple tools that can be used to construct, integrate, and analyze PPI data to understand its biological meaning of the network.

3.4.1 Cytoscape

It is widely regarded as a highly popular open-source software tool that facilitates the visual exploration of diverse biomedical networks, encompassing interactions involving proteins, genes, and various other types. Initially, it was designed for biological networks, but it also can be used for other purposes where show the relationship between two entities (Fig. 3.3). Cytoscape is popular among the network biologist to its working with a large variety of apps and plugins for various network analysis tailored for specific purposes, such as community search (e.g., MCODE, clusterMaker2, and JActiveModules), or for conducting Gene Set Enrichment Analysis (BiNGO, ClueGO, EnrichmentMap), are available within the Cytoscape software ecosystem.

3.4.2 GeNeCK

GeNeCK is a user-friendly web server with a graphical interface, accessible at http://lce.biohpc.swmed.edu/geneck. Users can easily upload their data and submit it through the provided interface. One recommended approach for most users is to utilize ENA (Edge Neighborhood Aggregation), as it typically performs well in various scenarios without the need for manually selecting tuning parameters. ENA also provides p-values for each connection, indicating the statistical significance of the associations. Once the job is complete, the constructed network can be viewed on the GeNeCK website.

3.4.3 GeneMANIA

GeneMANIA is an effective tool that utilizes a vast collection of functional association data to identify genes that are related to a given set of input genes. This comprehensive data encompasses protein and genetic interactions, pathways,

NODE-1		NODE-2
A	interact with	O
B	interact with	P
C	interact with	Q
D	interact with	R
E	interact with	S
F	interact with	T
G	interact with	U
H	interact with	V
I	interact with	W
J	interact with	X
K	interact with	Y
L	interact with	Z
M	interact with	Z
N	interact with	U
O	interact with	A
P	interact with	R
Q	interact with	T
R	interact with	W
S	interact with	R
T	interact with	Y
U	interact with	G
V	interact with	B
W	interact with	C
X	interact with	D
Y	interact with	X
Z	interact with	A

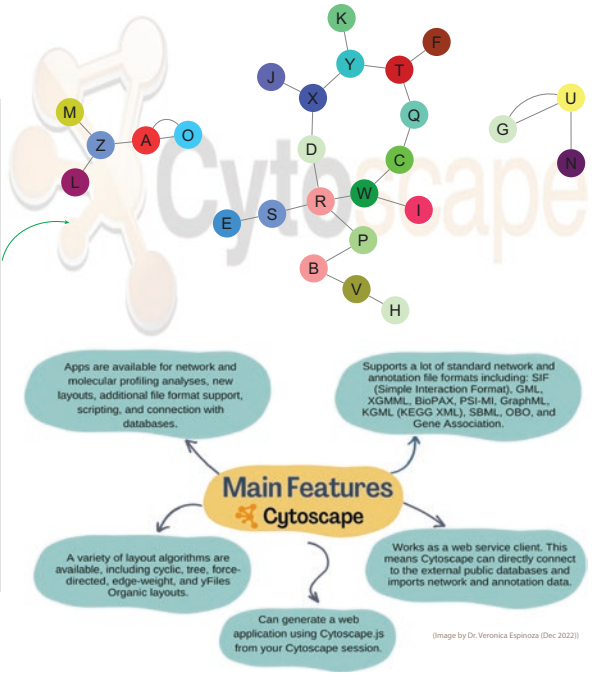

Fig. 3.3 Cytoscape is a popular tool for network analysis that represents the relationship between two entities

co-expression, co-localization, and protein domain similarity. With GeneMANIA, you can explore additional genes associated with specific pathways, complexes, or functions, such as protein kinases. The results generated by GeneMANIA depend on the genes you provide as input. For example, if your gene list represents a protein complex, GeneMANIA will suggest potential additional members of that complex. Similarly, when you input a gene list, GeneMANIA will identify connections between the genes within the selected datasets.

3.4.4 STRING

STRING serves as an extensive database that encompasses a wide range of known and predicted protein–protein interactions. It covers both direct (physical) and indirect (functional) associations between proteins. The data in STRING is obtained from diverse sources, including computational predictions, knowledge transfer between organisms, and the aggregation of interactions from other primary databases. Presently, the STRING database offers valuable information on 24,584,628 proteins across 5090 organisms.

3.4.5 FunCoup

FunCoup utilizes a range of evidence and gold standards, including protein complexes, physical protein interactions, metabolic pathways, and signaling pathways, to predict four distinct classes of functional couplings. It integrates multiple types of evidence, such as co-expression, protein–protein interactions (PPIs), genetic interactions, PHP similarity, and co-regulation, across 11 model organisms, to construct comprehensive genome-wide networks. To facilitate the transfer of evidence between species, FunCoup relies on ortholog assignments from in paranoid.

3.5 Network Visualizer Tools

3.5.1 Arena3Dweb

Arena3Dweb is the first, fully interactive and dependency-free, web application which allows the visualization of multilayered graphs in 3D space. This tool helps users to integrate multiple networks in a single view along with their intra-layer and inter-layer links. Users also can align networks and highlight the network's topological properties, highlighting the edges and important paths. The current version supports the weighted and unweighted undirected networks and it is written in R, Shiny, and Javascript (Karatzas et al. 2021). It can be accessed by using link (https://bib.fleming.gr:8084/app/arena3d).

3.5.2 GEPHI

Gephi exhibits the ability to handle extensive networks, encompassing thousands of nodes and millions of edges, necessitating substantial computational resources. Being open source and multi-platform, it offers a diverse array of advanced network-related algorithms available as plugins, such as NET-EXPO and DyCoNet (Bastian 2009). It can be downloaded from Gephi - The Open Graph Viz Platform than can used in Windows, Mac OS X and Linux system in local machine.

3.5.3 Igraph

Igraph is a versatile collection of libraries designed for graph creation, manipulation, and network analysis. Originally written in C, Igraph is also available as package for Python and R programming languages. Igraph can be installed by using R studio. Just type "install.packages("igraph")".

3.5.4 Pathview

The Pathview R package is a versatile toolset designed for integrating and visualizing pathway-based data. It facilitates the mapping and visualization of user data

onto pathway graphs. Users simply provide a list of gene or compound data along with the desired pathway, and Pathview takes care of automatically downloading pathway network data, parsing the data file, mapping the user data to the pathway, and generating pathway graphs with the mapped data. While Pathview can function as a standalone program, it also offers seamless integration with other pathway analysis tools, enabling large-scale and fully automated analysis pipelines (Luo and Brouwer 2013).

3.5.5 VisANT

It is a specialized tool designed for visual data mining of biological networks and pathways. It offers functionalities such as integrating disease and therapy hierarchies, associations between diseases and genes, associations between therapies and drugs, and interactions between drugs and targets. The latest version of VisANT incorporates disease and drug hierarchies, disease–gene associations, therapy–drug associations, and drug–target interactions. It allows for gene and drug annotation based on disease and therapy information. Additionally, it enables the prediction of associated diseases and therapies through enrichment analysis using user-provided gene or drug sets. VisANT supports network transformation and provides a user-friendly web interface for customizing node and edge properties. It is freely available at http://visant.bu.edu (Hu et al. 2013).

3.5.6 BioNetStat

This tool is accessible through the Bioconductor package in R, which provides a user-friendly graphical interface. It allows users to compare two or more correlation networks based on the probability distribution of network centrality measures. BioNetStat specializes in conducting differential network analysis, examining network features, and highlighting significant differences between disease and normal conditions using statistical significance scores. It is worth noting that this tool is not restricted to gene expression networks alone; it can also be applied to various data types such as proteomics, phenomics, metabolomics, as well as economic and social network data (Jardim 2019).

3.5.7 NetworkAnalyst

The increasing use of gene expression profiling within the context of protein–protein interaction (PPI) networks requires robust and user-friendly bioinformatics tools for understanding systems-level data. This tool assists in visualizing and comparing multiple gene lists through interactive heatmaps, enrichment networks, Venn diagrams, or chord diagrams (Zhou et al. 2019). It can be accessed at https://www.networkanalyst.ca/.

3.6 Network Clustering Tools

3.6.1 NeAT

The Network Analysis Tools (NeAT) is a series of modular computer programs (Fig. 3.4) specifically designed for the analysis of biological networks via integration of many algorithms for the analysis of biological networks: networks comparison, clustering and pathfinding, network randomization, and network topological properties analysis (Brohée et al. 2008). NeAT user-friendly web interface to allow easy access to the tools and data processing (http://rsat.sbroscoff.fr/index_neat.html).

3.6.2 clusterMaker

clusterMaker is a plugin available for Cytoscape that provides a unified interface for various clustering techniques and visualization options. It supports multiple clustering algorithms, such as hierarchical, k-medoid, AutoSOME, and k-means, which are useful for clustering expression or genetic data. Additionally, it offers partitioning algorithms like MCL, transitivity clustering, affinity propagation, MCODE, community clustering (GLAY), SCPS, and AutoSOME, enabling network partitioning based on similarity or distance values. The clustering results can be visualized as hierarchical groups of nodes or heat maps for hierarchical, k-medoid, AutoSOME, and k-means algorithms. The plugin introduces collapsible "meta nodes" that aid in interactive exploration of putative family associations within the Cytoscape network. Moreover, the results can be displayed as a separate network containing only intra-cluster edges or with inter-cluster edges included. clusterMaker is compatible with Cytoscape version 2.8.2 or newer and can be accessed through the Cytoscape plugin manager under the Analysis category.

3.6.3 GephiCrunch

It utilizes a range of techniques to compare the structures of two networks, such as average clustering coefficients, average path lengths, diameters, degree distributions, clustering and eccentricity spectra, and graphlet-based heuristics. These methods enable more precise and rigorous comparisons between the networks.

3.7 Other Miscellaneous Tools

3.7.1 Network BLAST

This tool enables the identification of protein complexes within protein–protein interaction networks. It has the capability to analyze either a single network or two networks from different species. In the case of analyzing two networks, Network

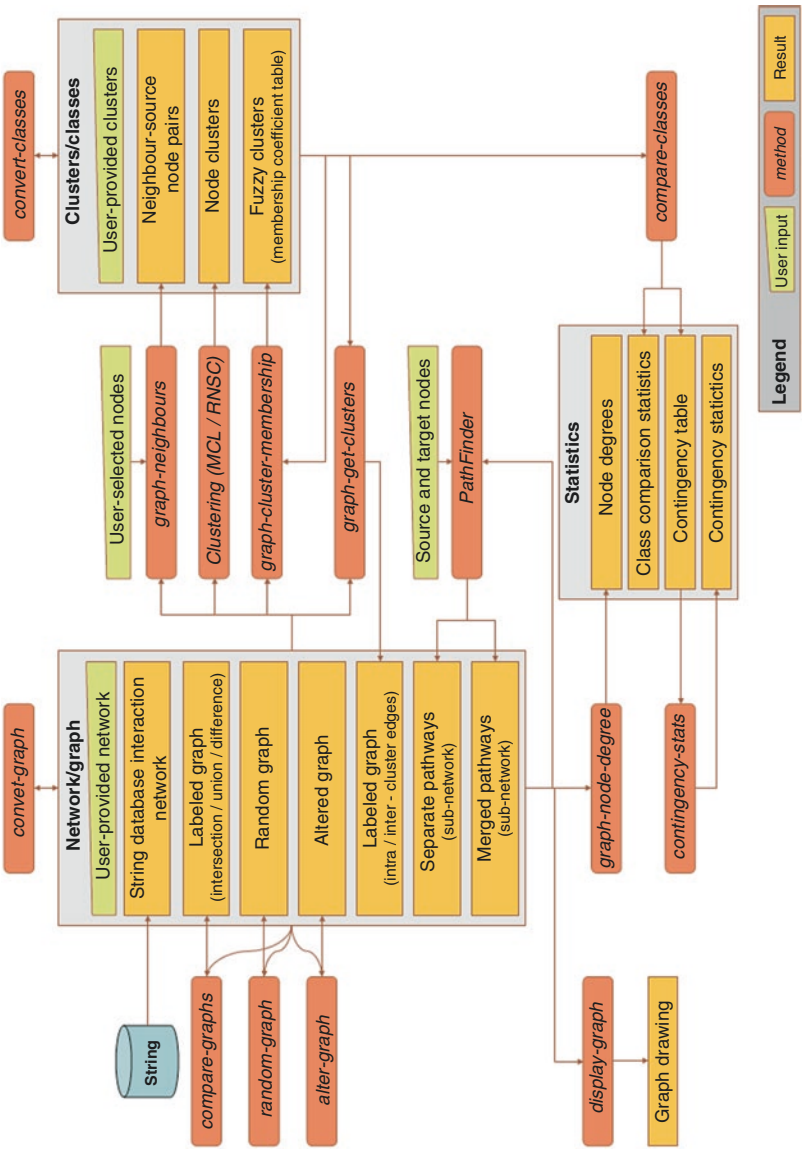

Fig. 3.4 Network Analysis Tools (NeAT) tool map (Figure adopted from NeAT web page)

BLAST generates a collection of complexes that exhibit evolutionary conservation across the two networks. The tool is available for download on a local Linux machine and can be installed with minimal hardware requirements (AMD XP2500+ 1 GB, Intel XEON 3 Ghz 3 GB).

3.7.2 SpectralNET

A tool to analyze the networks and node interactions, such as chemical–genetic networks. It is available as a standalone. Spectral NET offers a user-friendly approach for analyzing graph-theoretic metrics in data modeling and dimensionality reduction (Forman et al. 2005).Users can conveniently access it through either a.NET application or an ASP.NET web application, available at http://chembank.broad.harvard.edu/resources/.

3.8 Conclusion

A biological network is a way to represent the complex system of binary interactions or links between various biological entities. In this chapter, we have focused on Network analysis and visualization tools (Offline/Online) which make it easy for us to understand the broad maps of cellular organization. There are many network analysis visualization tools are exist, but we mainly discussed here only 11 tools that are widely used and popular among researchers. Most of the tools support a GUI (graphic user interface) where users can process tire data by a few mouse clicks, simple dialog boxes and data imports allow most functionality to be accessed. However, Several tools need some basic programing skills in R and Python to read, process, and write the data. Plugins are an important way for advanced users to customize and extend an application (Cytoscape and VisANT) based on third parties generic software to develop new functionality and integrate it directly within the tool. As we know that each tool is designed differently (different algorithm used) so the result would be different from other tools, so before using any tool, the user must be sure about what is aim of study, what type of result they want, and cross-check and validate the result with other tools. All the tools/software introduced in this chapter are freely accessible to foster collaboration among researchers and accelerate advancements in the field of systems biology.

References

Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH (2017) HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res 45(D1):D408–D414. https://doi.org/10.1093/nar/gkw985

Alonso-López D et al (2019) APID database: redefining protein–protein interaction experimental evidences and binary interactomes. Database 2019:baz005. https://doi.org/10.1093/database/baz005

Bader GD (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31(1):248–250. https://doi.org/10.1093/nar/gkg056

Mathieu Bastian, "Gephi: an open source software for exploring and manipulating networks," Third International AAAI conference on weblogs and social media, 2009

Brohée S, Faust K, Lima-Mendez G, Vanderstocken G, van Helden J (2008) Network analysis tools: from biological networks to clusters and pathways. Nat Protoc 3(10):1616–1629. https://doi.org/10.1038/nprot.2008.100

Calderone A, Castagnoli L, Cesareni G (2013) Mentha: a resource for browsing integrated protein-interaction networks. Nat Methods 10(8):690–691. https://doi.org/10.1038/nmeth.2561

Chatr-aryamontri A et al (2007) MINT: the molecular INTeraction database. Nucleic Acids Res 35, no. Database:D572–D574. https://doi.org/10.1093/nar/gkl950

Das J, Yu H (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 6(1):92. https://doi.org/10.1186/1752-0509-6-92

Fahey ME et al (2011) GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics 12(1):298. https://doi.org/10.1186/1471-2105-12-298

Forman JJ, Clemons PA, Schreiber SL, Haggarty SJ (2005) SpectralNET–an application for spectral graph analysis and visualization. BMC Bioinformatics 6(1):260. https://doi.org/10.1186/1471-2105-6-260

Goel R, Harsha HC, Pandey A, Prasad TSK (2012) Human protein reference database and human Proteinpedia as resources for phosphoproteome analysis. Mol BioSyst 8(2):453–463. https://doi.org/10.1039/C1MB05340J

Hermjakob H (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32(90001):452D–4455D. https://doi.org/10.1093/nar/gkh052

Hu Z et al (2013) VisANT 4.0: integrative network platform to connect genes, drugs, diseases and therapies. Nucleic Acids Res 41, no. Web Server issue:W225–W231. https://doi.org/10.1093/nar/gkt401

Jardim VC, Santos S d S, Fujita A, Buckeridge MS (2019) BioNetStat: a tool for biological networks differential analysis. Front Genet 10:594. https://doi.org/10.3389/fgene.2019.00594

Kalathur RKR et al (2014) UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. Nucleic Acids Res 42(D1):D408–D414. https://doi.org/10.1093/nar/gkt1100

Karatzas E, Baltoumas FA, Panayiotou NA, Schneider R, Pavlopoulos GA (2021) Arena3Dweb: interactive 3D visualization of multilayered networks. Nucleic Acids Res 49(W1):W36–W45. https://doi.org/10.1093/nar/gkab278

Kotlyar M, Pastrello C, Sheahan N, Jurisica I (2016) Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res 44(D1):D536–D541. https://doi.org/10.1093/nar/gkv1115

Luck K et al (2020) A reference map of the human binary protein interactome. Nature 580(7803):402–408. https://doi.org/10.1038/s41586-020-2188-x

Luo W, Brouwer C (2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29(14):1830–1831. https://doi.org/10.1093/bioinformatics/btt285

Oughtred R et al (2019) The BioGRID interaction database: 2019 update. Nucleic Acids Res 47(D1):D529–D541. https://doi.org/10.1093/nar/gky1079

Patil A, Nakai K, Nakamura H (2011) HitPredict: a database of quality assessed protein–protein interactions in nine species. Nucleic Acids Res 39, no. suppl_1:D744–D749. https://doi.org/10.1093/nar/gkq897

Schoof H et al (2005) Munich information Center for Protein Sequences Plant Genome Resources. A framework for integrative and comparative analyses. Plant Physiol 138(3):1301–1309. https://doi.org/10.1104/pp.104.059188

Szklarczyk D et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1):D607–D613. https://doi.org/10.1093/nar/gky1131

Xenarios I (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28(1):289–291. https://doi.org/10.1093/nar/28.1.289

Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J (2019) NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res 47(W1):W234–W241. https://doi.org/10.1093/nar/gkz240

Networks Analytics of Heterogeneous Big Data

4

Rafat Ali and Nida Jamil Khan

Abstract

Day by day exponential growth in experimentation is generating staggering amounts of data in the different fields of biological research. Generating this high volume of data brought the term "Big Data." The term "Big Data" has several dimensions which contain a huge amount of unstructured and structured data sets. Analysis, interpretation, and complete extraction of meaningful information from the raw, structured, and unstructured datasets lead to big scientific inventions, open the routes of progress in industry and economic development. Due to bulky data set makes it is very complex and complicated in the sense of co-relationship and connections among them, managing the hierarchy level, and many data linkages. It is extremely difficult to easily manage data quality with proper security and privacy with such a large number of multivariate raw data. Since "Big Data" encompasses both organized and unstructured data sets, therefore storing, transferring, processing, and searching the raw data is extremely challenging, and it cannot be managed using traditional database systems and software tools. The heterogeneous "Big data" set contains petabytes or exabytes of raw data, with billions to trillions of archived records. To gain hidden information from the big, archived records complex network biology approach has a significant role. Biological complex networks like gene–gene and protein–protein interaction networks have been appreciated for finding genes and pathways associated with diseases. These complex networks could provide significant insights into the mechanisms of complex diseases like cancer.

R. Ali · N. J. Khan (✉)
Department of Biosciences, Jamia Millia Islamia, New Delhi, India
e-mail: njkhan@jmi.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_4

Keywords

Big data · Network analytics · Genetic profile · Next generation sequencing · Data quality · Genomic databases

4.1 Big Data Analytics for Network Biology

In the various domains of biological study, exponential increase in experimentation has produced huge amounts of data. The phrase "Big Data" was coined to describe the emerging field of modern biology due to the huge amount of data generated. The phrase "Big Data" refers to a number of dimensions that include a sizable number of unstructured and organized data sets. Big scientific discoveries are made possible by the analysis, interpretation, and thorough extraction of useful data from large raw, structured, and unstructured databases. This also paves the way for advancements in business and economic growth. Because of the data set's weight, managing the hierarchy level and numerous data links, as well as its co-relationships and linkages, makes it extremely difficult. Easily managing such a massive volume of multivariate raw data is challenging. Because "Big Data" includes both structured and unstructured data sets, managing the raw data is exceedingly difficult. It cannot be done with the help of conventional database management systems and software tools. The heterogeneous "Big data" collection might include trillions of old records together petabytes or exabytes in size (Altaf-Ul-Amin et al. 2014). The era of Big Data is closely related to the era of Omics, and omics technologies and assays, in particular, omics technologies and assays are key to producing enormous amounts of data. So, omics is a significant stakeholder in big data.

Future generations will use emerging technologies like imaging systems, spectroscopy-based flow cytometry, different sequencing techniques, and others to produce enormous amounts of data (Nielsen et al. 2010). Big challenges of big data in biology to store, analyze, and interpret have been brought in by the greater volume of nucleic acid sequencing data and output of advanced high-throughput techniques in system biology, whether it is cell biology or molecular biology, producing piles of omics data in different omics-based fields like-genomes, transcriptomes, proteomes, metabolomes, interactomes, and so on (Pal et al. 2020). Additional sources that are expanding exponentially are also producing a significant amount of raw data. Analysis of a big amount of data is really a concern of the new scientific era. Big data analytics is a fast-growing technology that has been used in various areas, including network biology. The complex network biology approach plays a vital role in extracting hidden information from large stored records. To uncover disease-associated genes and pathways complex networks such as gene–gene or protein–protein interaction networks have gained importance. It is possible that these intricate networks may reveal important information about how complicated diseases like cancer function. Since every gene, protein, piece of DNA, and pathway in the human body is somehow related to one another, a complex network approach offers a simulation method for learning actual information about the gene, protein, etc. plays a vital role in extracting hidden information from the big stored biological data (Fig. 4.1).

4.2 Genetic Profiling Data

The word "genetic profile" is typically used to refer to a composition of genetic traits that are associated with a human individual, such as genetic signals or information. It is crucial to emphasize that profiling is the method by which such a collection of traits is linked to the target qualities considered when making decisions, such as any disease occurrence (European Commission 2018).

In healthcare and scientific research, genetic profiling is used to link particular genetic traits to increased or decreased risk of detecting and curing specific diseases. Rarely diseases are monogenic or caused by genetic variations at a single locus that have a significant impact on the disease's course. An illustration of such a monogenic disease is cystic fibrosis. Genetic diseases are often polygenic, which means they are probably caused by a combination of various genes, lifestyle choices, and environmental factors. Even though several DNA-related markers are important for identifying individuals and particular characteristics, in general, only a very small portion of DNA is important for medical treatment and study, such as genes, SNPs, Short Tandem Repeats (STR), and whole genome sequences. In contrast to, RNA variations, these markers are directly associated with the DNA without any translation step in between, allowing one to concentrate on stable molecular properties. Because STRs are typically employed to create an individual's genetic fingerprint, they are extremely important (Sariyar et al. 2017). Profiling may be done in various situations, including forensic science, marketing, healthcare settings, and information science. Big data expands how profiling may be used to find new patterns and make automated decisions (Hildebrandt 2008). Healthcare profiling allows for more tailored decisions than those based on average traits, which is closely connected to personalized therapy. Its potential applications include prognostication, resistance detection, disease monitoring, risk assessment, recurrence

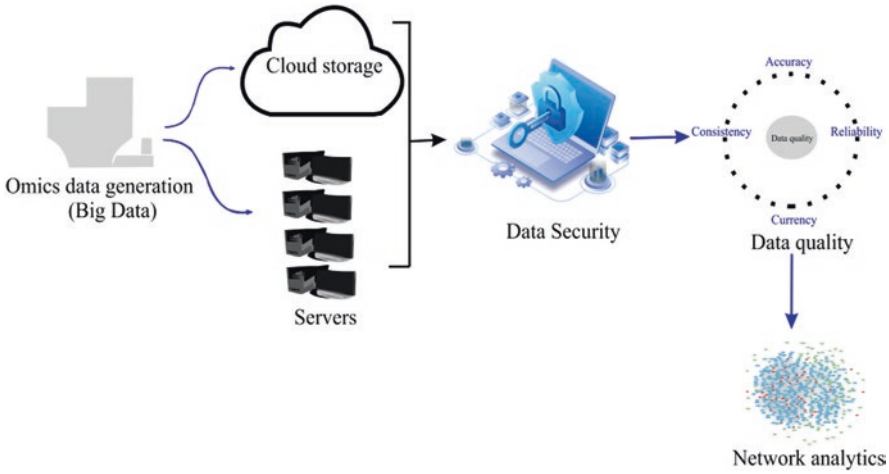

Fig. 4.1 Figure illustrating big data handling and analytical approach

detection, and early detection. Biomarker-guided diagnostic decisions and therapy selection are significant (Wjst 2010).

4.3 Data Quality

By 2025, it is estimated that between 100 million and 2 billion human genomes will have been sequenced, making genomics the world's biggest "big data" challenge (Stephens et al. 2015). Increasing amounts of genomic data of various types have been generated by high-throughput technologies and, more recently, Next Generation Sequencing (Schuster 2008), making significant advancements in the understanding of human genome mechanisms and their application to previously unheard-of personalized medicine outcomes. Data and metadata integration is regarded as an activity of unquestionable importance prior to data analysis and biological knowledge discovery, with pressing demands for improved data extraction, matching, normalization, and enrichment methodologies to enable building multiple perspectives over the genome; these can lead to the identification of meaningful relationships, which are otherwise not perceptible when using incompatible data representations (Samarajiwa et al. 2018). The way that bioinformatics and genomics, in past years operated by utilizing the extensive fieldwork done by its practitioners in terms of data collection, management, and analysis. Best practices are gathered from many laboratories and projects, discussed on discussion boards (like https://www.biostars.org, http://seqanswers.com, and https://www.researchgate.net), and compiled in documentation or wiki guides in tool and software source repositories. The quality of the experimental data generated by these procedures, for which standard protocols are frequently made up of several scripts, are accessible, and become a primary interest to bioinformaticians. In contrast, less focus has been placed on the high-quality actions that may be taken when combining several experimental data in systematic methods. Handling data standards in both schemata and values, as well as applying integration methods that enhance data quality, become more crucial with the development of a culture of data FAIRness (Wilkinson et al. 2016) and of open and sharable research, which is supported by efforts like FAIRsharing (Sansone et al. 2019) focusing on accuracy, consistency, currency, and reliability (Hedeler and Missier 2008). One out of three business leaders do not trust the data they use to make decisions, according to IBM's 2012 report on data quality (https://opensistemas.com/en/the-four-vs-of-big-data); this ratio is unacceptable in industries like healthcare and precision medicine, which are heavily reliant on genomic databases and decision-making techniques.

"Quality" in genomics has typically been used to refer to "quality control" steps on sequences, typically a pre-processing activity aimed at removing adapter sequences, low-quality reads, uncalled bases, and contaminants. This usage dates back to DNA microarrays (Ji and Davis 2006) and Next Generation Sequencing (Schuster 2008). Instead, we use the term "Data Quality" (DQ) in this study to refer to the larger definition given by Wang and Strong (Wang and Strong 1996), which is typically encapsulated by the phrase "fitness for use," i.e., the capacity of datasets

to satisfy the needs of their users. DQ is assessed using many quality factors (i.e., single aspects or components of a data quality concept (Stvilia et al. 2007). A summary of cutting-edge methods for resolving problems with data quality in general databases is given in (Fan 2015), under the heading "data cleaning." Data of low quality in genetic databases have a huge economic and medical impact on their users/customers. For instance, incorrect target selection for medical investigations or pharmaceutical research may result from inaccuracies in genomic data. Pharmaceutical corporations invest billions of dollars in research just to release a few novel medications (Hensley 2002). Only a small number of the hundreds of potential leads generated from experimental genetic data make it to clinical trials, and only one medicine is commercially viable. It is of considerable importance to base these far-reaching judgments on good-quality data (Shankaranarayanan et al. 2000).

4.4 Major Public Databases

Genome mapping and sequencing technologies are producing a huge amount of data; as a result, dependable and effective storage methods are needed. The first organization of genome sequencing and mapping data involved manual data collection and deposit in a central location, such as a table with columns and rows. This uses resources in a laborious, time consuming, and ineffective manner. A drawback was rapidly identified as the absence of synergy between big independent collections of different sorts of data. Data are combined through the science of informatics developed by information technology experiments in the lab that enables the gathering, analysis, and distribution of beneficial combinations of pertinent data. In order to minimize mistakes, eliminate redundancy across comparable data sources, and make well-informed judgments regarding outcomes, informatics offers a semi-automated method for retrieving, filtering, and making comparisons and contrasts of data in an electronic format. Many databases are created using tables or relational databases as their foundation. Alternatives include object-oriented techniques that allow for flexible data categorization and analysis based on this classification, as well as data storage and retrieval (Carroll et al. 2001).

Biology is undergoing a revolution as a result of genome-wide analyses of gene function and expression as well as a genomic structure made possible by genomic sequence data. The application of human genetic data is anticipated to have profound effects on pathology and the creation of individualized treatments. Online genomic databases provide free access to genome reference sequences for thousands of species. Genome browser's user-friendly software that produces interactive, graphical representations of key chromosomal areas with comprehensive annotations, including genes, epigenetic information, and sequence variations, may be used to immediately download or search sequence data. In addition to detailing several methods for searching them, such as employing IDs for genes and chemicals, karyotype bands, chromosomal locations, sequences, and motifs, this chapter enlisted Table 4.1 major genomic databases and genome browsers. The human

genome is highlighted, along with methods for visualizing and retrieving data related to genome plasticities, such as sequence and structural variations (Hutchins 2020).

As the principal public repository for genomic sequence data, the National Center for Biotechnology Information (NCBI) gathers and preserves vast quantities of diverse data. The NCBI website's text-based search and retrieval system, which enables quick and simple access to a variety of biological databases, integrates data on genomes, genes, gene expressions, gene variation, gene families, proteins, and protein domains with analytical, search, and retrieval capabilities. The speed of discovery is accelerated by the use of comparative genomic analysis methods to gain a deeper knowledge of evolutionary processes. Our knowledge of the biology of living things has undergone a fundamental transformation as a result of the increase in genome sequencing that has been sparked by recent technological advancements. The information management system and the visualization tools will now face additional difficulties as a result of the enormous growth in DNA sequence data. To organize this genomic sequence flood and enhance the usefulness of the related data, new approaches need to be developed (Tatusova 2016).

4.5 Challenges of Handling Genomics' Data

We reported here a number of difficulties that data scientists use open datasets to address biological and clinical issues confront. Despite the significant effort put into creating such public datasets, they are spread across several sources, varied in terms of their formats, and frequently fulfill extremely varying quality criteria (Ceri and Pinoli 2020). The development of high-throughput technologies at much lower prices has led to a massive data explosion in genomics in recent years. The age of millions of accessible genomes is about to begin. Notably, each genome can include billions of nucleotides that are encoded in terabytes of plain text (GBs). It is obvious that these genetic data provide unaware of data difficulties (Wong 2019). There has been an explosion of genetic data created as a result of recent advancements in biotechnology, particularly next generation sequencing in genomics. Both in terms of volume and variety, the findings are substantial. Much more information is contained in big data, which also presents data analysis with novelty. The high-throughput data produced by NGS and similar technologies is called big data. Volume, Velocity, Variety, Veracity, and Value, sometimes known as the 5Vs, are characteristics of Big Data that make it distinctive in terms of difficulties and potential. Big Data in genomics is being produced at a pace and scale never before seen because of new technologies like NGS. Data analysis is challenging because of Big Data. Regarding the study of Big Data in genomics, we talked about the difficulties with data integration, data administration, computer infrastructure, dimension reduction, data smoothing, and data security. The analysis of single-cell sequencing data, de-novelization of sequencing reads, and the study of rare genetic variations are a few more issues that are more data specific. These fields are the subject of active research and will deliver crucial data and tools for comprehending genomics

Table 4.1 Genomic databases and genome browsers

Resource	Website	References
NCBI resources		Sayers et al. (2019)
RefSeq homepage	https://www.ncbi.nlm.nih.gov/refseq/	O'Leary et al. (2016)
Genomes homepage	https://www.ncbi.nlm.nih.gov/genome	
Genome data download (FTP)	http://ftp//ftp.ncbi.nlm.nih.gov/genomes/	
Human genome page	https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml	
Genome Data Viewer (GDV)	https://www.ncbi.nlm.nih.gov/genome/gdv/	
GDV tutorial	https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/	
Viral genomes	https://www.ncbi.nlm.nih.gov/genome/viruses/	Brister et al. (2015)
Ensembl		Cunningham et al. (2019)
Homepage	https://www.ensembl.org/	
Genome data download (FTP)	http://ftp//ftp.ensembl.org/pub/	
Human genome page	https://www.ensembl.org/Homo_sapiens/Info/Index	
Ensembl genomes	http://ensemblgenomes.org/	Kersey et al. (2018)
Ensembl bacteria	https://bacteria.ensembl.org/	
Ensembl fungi	https://fungi.ensembl.org/	
Ensembl metazoa	https://metazoa.ensembl.org/	
Ensembl plants	https://plants.ensembl.org/	
Ensembl protists	https://protists.ensembl.org/	
Ensembl tutorials	https://www.ensembl.org/info/website/tutorials/index.html	
UCSC Genome Browser		Haeussler et al. (2019)
Homepage	https://genome.ucsc.edu/	
Genome data download (FTP)	http://ftp//hgdownload.soe.ucsc.edu/goldenPath/currentGenomes/	
Genome Browser User Guide	https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html	
Stand-alone genome browsers		
Integrated Genome Browser (IGB)	https://bioviz.org/	Freese et al. (2016)
Integrative Genomics Viewer (IGV)	http://software.broadinstitute.org/software/igv/	Thorvaldsdottir et al. (2013)
NCBI Genome Workbench	https://www.ncbi.nlm.nih.gov/tools/gbench/	

and life in general (Xu 2020). High dimensionality, which relates to both the sample size and the number of variables and their structures, distinguishes big data in genomics. The sheer amount of data presents problems for computing and data storage. For only the raw data of each sample, the volume of data might be terabytes. It is a good idea to maintain the raw data for the various forms of genomic data, frequently in the form of picture files, so that when more advanced base calling algorithms become available, they may be used to increase accuracy (Barrett et al. 2013).

The computation in genomics is impractical with current computer infrastructure due to the massive amounts of Big Data. For research with huge sample sizes, it can take months to complete the alignment and annotation of NGS reads utilizing desktop PCs. Utilizing high-performance computing resources like computer clusters is one way to solve this issue. The goal is to divide the large computing task into smaller tasks and distribute them across the cluster's compute nodes. The result is highly parallel computing, which enables quick completion of large tasks (Almasi and Gottlieb 1989). In the field of processing large data, there are several problems that need to be solved. These problems need to be solved holistically, utilizing the knowledge of several computer science disciplines. Data Cleansing/Acquisition/Capture, Data Storage/Sharing/Transfer, Data Analysis, as well as certain ethical considerations that come from the exposure and processing of large data, are some big data management challenges that scientists should address (Jagadish et al. 2014).

4.6 Security of Genomics' Data

A wealth of genetic data has recently been generated due to the lower cost of DNA sequencing, which is used to promote scientific research, enhance clinical practices, and enhance healthcare delivery. Genome-wide association studies (GWASs), diagnostic testing, personalized medicine, and drug discovery are all being revolutionized by these developments. The human genome is complex in nature and uniquely identifies an individual. Therefore this presents security and privacy problems. In this chapter, we discuss the issue of genomic privacy and evaluate pertinent privacy assaults that have been utilized to invade someone's privacy. These attacks may be categorized as identity tracking, attribute disclosure, and completeness attacks (Abukari and Chen 2020). The development of genome sequencing methods promotes the accessibility of genetic data and its gathering for processing, sharing, and storage. Despite these advantages, there are still significant worries regarding how genetic data is stored, shared, transported, and processed. DNA donors occasionally inquire about the storage of their genetic information. Who can access it? What are safety precautions in place to safeguard my privacy? These worries result from the fact that (Australian Genomics 2016) Applications of genetic data now and in the future might lead to ethical and privacy issues. The danger of misuse by prospective criminals is increased by the processing and storing of this data since the human genome contains sensitive personal data.

4.7 Conclusion

We are currently in the era of "big data," in which big data technology is being rapidly applied to biomedical and healthcare fields. We might face big problems if we do not have big solutions for big amounts of data. With sophisticated platforms and tools like gene sequencing mapping tools now in use to assist in analyzing biological data, big data application in bioinformatics is still in a relatively early stage of development. To analyze big data network biology approach is a big solution. However, before analyzing the big amount of data, researchers need to improve several things. The absence of consistency in laboratory practices and values makes it difficult to integrate data. For instance, imaging data that originates from many laboratories using various techniques can experience technological batch effects. When there is a batch effect, efforts are made to normalize the data; while this may be simpler for image data, normalizing laboratory test data is inherently more challenging. Big data integration and usage in all domains continue to be hampered by security and privacy issues; as a result, secure platforms with improved communication standards and protocols are urgently required. Big Data involves complex systems, revenue, and difficulties. To increase the effective evaluation online as well as the display, analysis, and storage of Big Data, more research is required. Big Data's main security issues are privacy, integrity, availability, and confidentiality of outsourced data.

References

Abukari MY, Chen Y-PP (2020) Ensuring privacy and security of genomic data and functionalities. Brief Bioinform 21(2):511–526

Almasi GS, Gottlieb A (1989) Highly parallel computing. Benjamin-Cummings Publishing Co., Inc., Redwood City

Altaf-Ul-Amin M et al (2014) Systems biology in big data and networks. Biomed Res Int 2014:428570

Australian Genomics. A National approach to data federation and analysis. 2016

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995

Brister JR, Ako-Adjei D, Bao Y, Blinkova O (2015) NCBI viral genomes resource. Nucleic Acids Res 43(Database issue):D571–D577

Carroll ML, Nguyen SV, Batzer MA (2001) Genome databases. In: e LS

Ceri S, Pinoli P (2020) Data science for genomic data management: challenges, resources, experiences. SN Computer Science 1(1):1–7

Cunningham F, Achuthan P, Akanni W et al (2019) Ensembl 2019. Nucleic Acids Res 47(D1):D745–D751

European Commission (2018). Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (wp251rev.01)

Fan W (2015) Data quality: from theory to practice. ACM SIGMOD Rec 44(3):7–18

Freese NH, Norris DC, Loraine AE (2016) Integrated genome browser: visual analytics platform for genomics. Bioinformatics 32(14):2089–2095

Haeussler M, Zweig AS, Tyner C et al (2019) The UCSC genome browser database: 2019 update. Nucleic Acids Res 47(D1):D853–D858

Hedeler C, Missier P (2008) Information quality management challenges for high-throughput data. In: Biological database model, p 81
Hensley S (2002) Death of Pfizer's 'youth pill' illustrates drug makers woes. The Wall Street Journal online
Hildebrandt M (2008) Defining profiling: a new type of knowledge? In: Hildebrandt M, Gutwirth S (eds) Profiling the European citizen: cross-disciplinary perspectives. Springer Netherlands, Dordrecht, pp 17–45. https://doi.org/10.1007/978-1-4020-6914-7_2
Hutchins JR (2020) Genomic databases. In: Genome plasticity in health and disease. Academic Press, pp 47–62
Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. Commun ACM 57(7):86–94
Ji H, Davis RW (2006) Data quality in genomics and microarrays. Nat Biotechnol 24(9):1112–1113
Kersey PJ, Allen JE, Allot A et al (2018) Ensembl genomes 2018: an integrated omics infrastructure for nonvertebrate species. Nucleic Acids Res 46(D1):D802–D808
Nielsen CB et al (2010) Visualizing genomes: techniques and challenges. Nat Methods 7(3):S5–S15
O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745
Pal S et al (2020) Big data in biology: the hope and present-day challenges in it. Gene Reports 21(4):100869
Samarajiwa SA, Olan I, Bihary D (2018) Challenges and cases of genomic data integration across technologies and biological scales. In: Advanced data analytics in health. Springer, pp 201–216
Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M (2019) Fairsharing as a community approach to standards, repositories and policies. Nat Biotechnol 37(4):358–367
Sariyar M, Suhr S, Schlünder I (2017) How sensitive is genetic data? Biopreserv Biobank 15:494–501. https://doi.org/10.1089/bio.2017.0033
Sayers EW, Agarwala R, Bolton EE et al (2019) Database resources of the national center for biotechnology information. Nucleic Acids Res 47(D1):D23–D28
Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5(1):16–18
G. Shankaranarayanan, R.Y. Wang, M. Ziad, IP-MAP: representing the manufacture of an information product. In proceedings of the International conference on information quality (IQ), Cambridge, 2000, 1–16
Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE (2015) Big data: astronomical or genomical? PLoS Biol 13(7):e1002195
Stvilia B, Gasser L, Twidale MB, Smith LC (2007) A framework for information quality assessment. J Am Soc Inf Sci Technol 58(12):1720–1733
Tatusova T (2016) Update on genomic databases and resources at the national center for biotechnology information. In: Data mining techniques for the life sciences. Humana Press, New York, pp 3–30
Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14(2):178–192
Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data con-sumers. J Manag Inf Syst 12(4):5–33
Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE et al (2016) The fair guiding principles for scientific data management and stewardship. Sci Data 3:160018
Wjst M (2010) Caught you: threats to confidentiality due to the public release of large-scale genetic data sets. BMC Med Ethics 11:21. https://doi.org/10.1186/1472-6939-11-21
Wong KC (2019) Big data challenges in genome informatics. Biophys Rev 11(1):51–54
Xu H (2020) Big data challenges in genomics. In: Handbook of statistics, vol 43. Elsevier, pp 337–348

Network Medicine: Methods and Applications

5

Aftab Alam, Okan Yildirim, Faizan Siddiqui, Nikhat Imam, and Sadik Bay

Abstract

Network medicine (NM) is a developing field within network science that focuses on molecular and genetic interrelationships, disease network biomarkers, and the discovery of therapeutic targets, and it is a rapidly growing arena for medical science and research, which comes with the possibilities to reform the system of disease diagnosis and its treatment. The NM uses topological and dynamic properties of the biological networks (protein–protein interactions and metabolic pathways) to distinguish the disease patterns (characterizing the behavior of disease genes) and associated drugs. Biomedical data provide a base to develop a significant model and get potential results at the network level. In

Aftab Alam has become affiliated with the United Arab Emirates University, UAE. Okan Yildirim and Sadik Bay are both currently affiliated with the Memorial Sloan Kettering Cancer Center in New York City, USA.

A. Alam
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

O. Yildirim
Department of Chemical Biology Otto-Hahn-Strasse, Max Planck Institute of Molecular Physiology, Dortmund, Germany

F. Siddiqui
International Medical Faculty, Osh State University, Osh, Kyrgyzstan

N. Imam
Department of Mathematics, Institute of Computer Science and Information Technology, Magadh University, Bodh Gaya, India

S. Bay (✉)
Research Institute for Health Sciences and Technologies (SABITA), Istanbul Medipol University, Istanbul, Turkey

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_5

this chapter, we have discussed the integrative use of emerging tools and the databases of NM, which provide a basic platform to systematically investigate the molecular complexity of diseases, dominant disease genes (modules), drug targets, disease-enriched pathways (altered pathways), and molecular interactions between apparently distinct phenotypes. The field of network medicine and its implications for diagnostics, prognosis, and therapeutics with unprecedented breadth and precision have great potential in the future.

Keywords

Network medicine · System pharmacology · Precision medicine · Disease–gene relationship · Drug repurposing

5.1 Introduction

"Interaction" is a single word that makes communication among the physical body (human beings), and even cellular components activate its functional switches by "interactions" with another component of the cell; overall, these interactions represent the human interactome. Interactome assessment is exclusively a tool-based approach to the theoretical paradigm and methodological tools utilized, describe, investigate, and comprehend structural and relational features of human health and disorders. Network-based research is becoming a crucial technique for identifying disease susceptibility genes and their associations with various diseases (Alam et al. 2022). Additionally, this research has enhanced our comprehension of drug targets and their results and proposed new drug targets, treatments, and therapeutic management strategies for serious disorders (Fig. 5.1).

Network-based approaches to human disease offer a variety of biological and therapeutic uses. Indeed, a better comprehension of the consequences of (1) cellular interconnections failure or (2) rewiring in cellular interconnections on disease progression could lead to the identification and classification of disease-associated genes and pathways which could provide better targets for drug development. Advancement in these fields could also restructure the clinical application and practice, from the development of improved and more precise biomarkers that monitor the functional integrity of the network that is perturbed by the diseases as well as improvements in disease classification and pave the way to personalized therapies and treatment.

In a recent study, Gysi et al. employed network medicine and drug repurposing methods to distinguish the repurposable drugs for COVID-19 (Morselli Gysi et al. 2021). Similarly, a multi-target herb called *Caesalpinia pulcherima* (CP) is used therapeutically to treat breast cancer (Sakle et al. 2020). Furthermore, Azuaje et al. contributed to our understanding of the cardiovascular effects of non-cardiovascular medications by integrating multiple sources of drug and target interaction data. They constructed the myocardial infarction (MI) drug-target interactome network, offering systemic insights into the topic (Azuaje et al. 2011). In a related study, Kim et al. have proposed that examining the network-based drug–disease intimacy can offer a novel perspective on the therapeutic effects of drugs in the context of

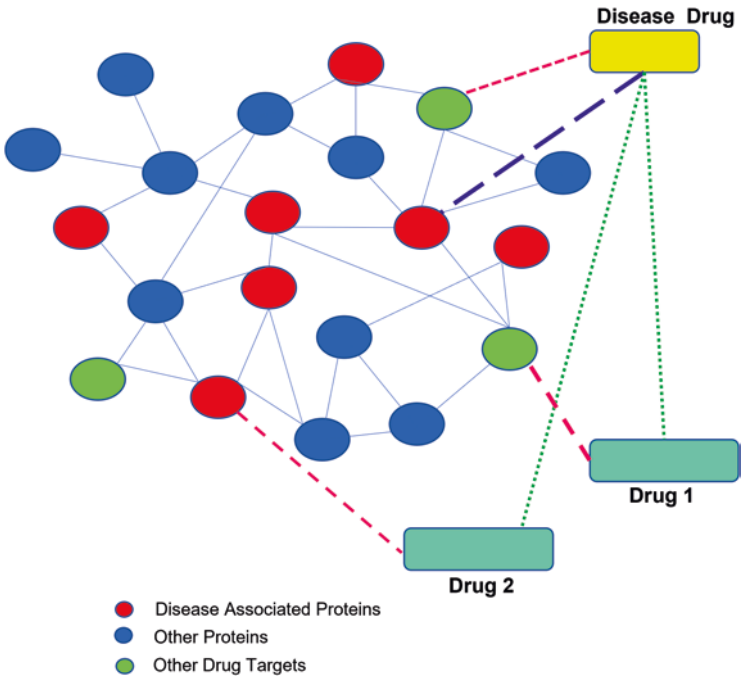

Fig. 5.1 Overview of network medicine approach

Systemic Sclerosis (SSc) disease. This approach may provide insights into drug combinations or drug repositioning strategies (Kim et al. 2020).

In this chapter, we focused on networks and their role in disease, System Pharmacology, Pharmacogenomics in Precision Medicine, Drug–Target Interaction, Drug–Drug Interaction, Drug Repurposing Opportunities, Drug Side Effects, and Integrating Omics data with Networks: Challenges and Ways. We hope this new chapter provides the same platform for scholars from biological science backgrounds to work on interdisciplinary research areas that can be useful for describing the causes of disease and pinpointing potential treatment targets, which will improve preventative healthcare and have a knock-on effect on personalized therapy.

5.2 Basic Principles and Key Components of Network Medicine

5.2.1 Systems Pharmacology

The early understanding of the molecular mechanisms behind pharmacological action was provided by classical investigations, such as the development of the receptor hypothesis that distinguished between competitive and noncompetitive inhibition. The prevalence of medications that target membrane receptors (mostly GPCRs) explains the influence and applicability of receptor theory in contemporary

pharmacology. The majority of the remaining drugs are enzyme inhibitors, which are often analogs based on substrates, analogs based on transition states, or allosteric inhibitors that are designed to bind reversibly or irreversibly based on the substrate and substrate-binding pocket configuration. The genomic, proteomic, network, and other high-throughput investigations have produced a wealth of "systems-level" knowledge over the last 10 years. The number of well-characterized, druggable targets is continuously rising through the use of high-throughput technology, structural and biochemical studies, and human genome research. Targeted therapies and biological medicines have thus emerged as a result of the expansion of the pharmacological pipeline and concentration on complicated, multigenic disorders. In recent studies, regulatory network analysis and structural analysis were used to anticipate the therapeutic benefits of medications for complicated disorders as well as potential off-target consequences (Boran and Iyengar 2010; Black and Leff 1983; Maehle et al. 2002; Colquhoun 2006).

5.2.2 Pharmacogenomics in Precision Medicine

One of the fundamental components of personalized treatment is pharmacogenomics (PGx). In personalized medicine, also known as precision medicine, patients are given prescriptions for drugs that are right for them based on their genetic, environmental, and lifestyle characteristics. Two key functions of pharmacogenomics in precision medicine. It first directs pharmaceutical firms in drug development and discovery. Second, it helps doctors choose the best medication for patients based on their genetic makeup, avoid adverse drug reactions, and maximize drug efficacy by providing the appropriate amount. Based on the understanding of pharmacogenomics, personalized/precision medicine has significant potential benefits. Precision medicine is the future of healthcare, and it will eventually become the standard of care.

5.2.3 Biological Networks and Important Databases

A method of expressing systems as complicated sets of binary interactions or relations between distinct biological components is called a biological network (e.g., genes, proteins, taxa, and metabolites). Now, with the availability of large-scale multi-omics data, the biological system has expanded from basic to advanced levels (like PPI connectivity and functional changes in disease stages, Disease-specific interactome, genetic perturbations, and network dysfunction).

Protein–protein interaction networks (PINs), in which proteins act as nodes and interactions as undirected edges, depict the physical connections between the proteins that are present in a cell. Protein–protein interactions (PPIs) are the most thoroughly studied networks in biology and are crucial to cellular functions. PPIs can be found using a variety of experimental methods, with the yeast two-hybrid system being one of the more popular methods for studying binary interactions. Mass

spectrometry-based high-throughput research have recently uncovered numerous sets of protein interactions.

The databases that catalog experimentally determined protein–protein interactions have been the result of numerous international efforts over the past few decades, such as the Munich Information Center for Protein Sequence (MIPS) protein interaction database (Schoof et al. 2005), Biomolecular Interaction Network Database (BIND) (Bader 2003), Database of Interacting Proteins (DIP) (Xenarios 2000), Molecular Interaction database (MINT) (Chatr-aryamontri et al. 2007), and Protein Interaction database (IntAct) (Hermjakob 2004). These databases are classified into primary and secondary databases based on their interaction prediction method but now one new term introduced "Meta-database" which is a combination of different primary and secondary databases to get new and maximum protein–protein interactions in network (Fig. 5.2). The list of protein interaction databases is given in Table 5.1.

Moreover PPI, the Metabolic networks explain the associations between small biomolecules (metabolites) and the enzymes (proteins) that interact with them to catalyze a biochemical reaction. Genetic interaction networks are valuable for understanding the relationship between genotype and phenotype because they relate to the functional interactions between pairs of genes in an organism.

Similarly, a gene regulatory network (GRN) is a collection of molecular regulators that interact with one another and with other components of the cell to regulate the expression level of mRNA and protein. Further, when different cell signaling pathways interact, cellular signaling networks are built, and they are identified by a combination of experimental and computational techniques.

A complex biological network consists of various nodes (including genes, proteins, and metabolites.) and connections among nodes are represented by joining lines, called "edges." The hubs are nodes that connect with other nodes frequently and play important roles in biological processes. The total number of edges that a node is connected to is referred to as its degree. Finding the node with the top degree can support to distinguish a biological-entity that plays the most important role within the network. For more details about the basic biological network and its

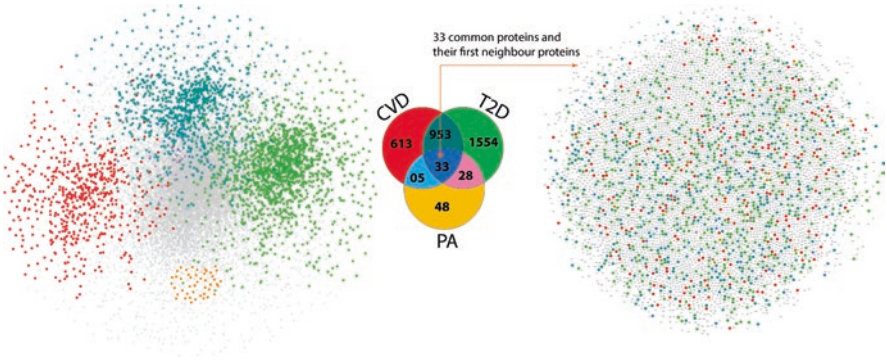

Fig. 5.2 An example of human disease networks

Table 5.1 Protein–protein interaction databases

Primary database			
Databases	URL	Experimental/ Predicted	Reference
BioGrid	https://thebiogrid.org/	Exp.	Oughtred et al. (2019)
HPRD	http://www.hprd.org/	Exp.	Goel et al. (2012)
IntAct	https://www.ebi.ac.uk/intact/home	Exp.	Hermjakob (2004)
MINT	https://mint.bio.uniroma2.it/	Exp.	Chatr-aryamontri et al. (2007)
HuRI	http://www.interactome-atlas.org/	Exp.	Luck et al. (2020)
Secondary database			
STRING	https://string-db.org/	Exp. & Pred.	Szklarczyk et al. (2019)
UniHI	http://www.unihi.org/	Exp. & Pred.	Kalathur et al. (2014)
Mentha	http://mentha.uniroma2.it/	Exp.	Calderone et al. (2013)
APID	http://cicblade.dep.usal.es:8080/APID/init.action	Exp.	Alonso-López et al. (2019)
HIPPIE	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/index.php	Exp.	Alanis-Lobato et al. (2017)
HitPredict	http://www.hitpredict.org/	Exp.	Patil et al. (2011)
IID	http://iid.ophid.utoronto.ca/	Exp. & Pred.	Kotlyar et al. (2016)
HINT	http://hint.yulab.org/	Exp.	Das and Yu (2012)
GPS-Prot	http://gpsprot.org/	Exp.	Fahey et al. (2011)

- Primary protein interaction databases containing literature-curated PPIs for human proteins.
- Secondary protein interaction databases containing predicted and curated from primary databases.

topological properties, importance can be read in our previous articles (Alam et al. 2019; Alam et al. 2021).

5.2.4 Human Disease Networks

A disease gene network involves the incorporation of genes linked to specific disease phenotypes into an established human interactome. The hypothesis posits that genes and their corresponding products associated with a given disease have the propensity to interact and form clusters within a localized sub-network, as opposed to being randomly distributed across the entire human interactome (Lee and Loscalzo 2019). The unbiased analysis of pathobiological linkages between various disease processes has also been made possible by disease networks. To construct the human disease network, there are many ways including literature survey to find the disease-associated genes and construct network. Further, there are many good, comprehensive databases available DisGeNet, an extensive and carefully curated database, integrates data from multiple publicly accessible databases, such as "UniProt/

SwissProt," "Cancer Genome Interpreter (CGI)," "Comparative Toxicogenomic DatabaseTM (CTDTM)," "Orphanet," "Mouse Genome Database (MGD)," "PsyGeNET," "Genomics England," "ClinGen," and "Rat Genome Database (RGD)." These databases were utilized to collect information on genes associated with various diseases, which were subsequently compiled and organized within the DisGeNet database (Pinero et al. 2015).

Recently, we analyzed three diseases (including Cardiovascular disease, Diabetes type-2, and Parathyroid adenoma) and found that 33 genes are common in these diseases (Fig. 5.2). Similarly, in our one more previous articles (Alam et al. 2022) where we identified 33 high-scoring significant modules that hub genes that are shared between tuberculosis (TB) and overlapping non-communicable diseases (NCDs) (lung cancer, rheumatoid arthritis, diabetes mellitus, Parkinson's disease, and cardiovascular disease).

5.2.5 Drug–Target Interactions

The conventional "one disease–one target reductionist approach" is not widely applicable since the majority of drugs exert their effects by targeting multiple proteins, resulting in a net pharmacological impact that encompasses both therapeutic and adverse effects. Network techniques provide valuable tools for predicting drug actions in complex biological contexts. In this context, drugs can be mapped onto the human interactome through their identified targets, where the effects of drugs on specific nodes or targets are represented by edges in the biological network. In our previous study, when the common disease genes related to TB and non-communicable disease (NCDs) that include lung cancer, rheumatoid arthritis, diabetes mellitus, Parkinson's disease, and cardiovascular disease; and built a bipartite network (drugs and targets) by mapping the hub genes of the modules to their corresponding drugs using the DGIdb database, the results revealed that a significant portion of the target genes had multiple hits (as shown in Fig. 5.3). This indicates that genes interacting with a greater variety of drugs may be more intricately connected to the underlying mechanisms driving the pathological phenotype associated with these drugs. The presence of multiple drug interactions with a gene suggests its involvement in multiple pathways or biological processes relevant to the disease phenotype, emphasizing its potential significance in the context of therapeutic interventions and understanding disease mechanisms (Alam et al. 2022).

5.2.6 Drug–Drug Interaction

A change in a drug's impact on the body when it is combined with another drug. The absorption of either drug can be sped up, slowed down, or improved by a drug–drug interaction. This could alter the way one or both drugs work, increase or decrease their effects, or have negative consequences. Studies have demonstrated that a substantial proportion, ranging from approximately 37–60%, of hospitalized patients may have one or more potentially interacting drug combinations upon admission

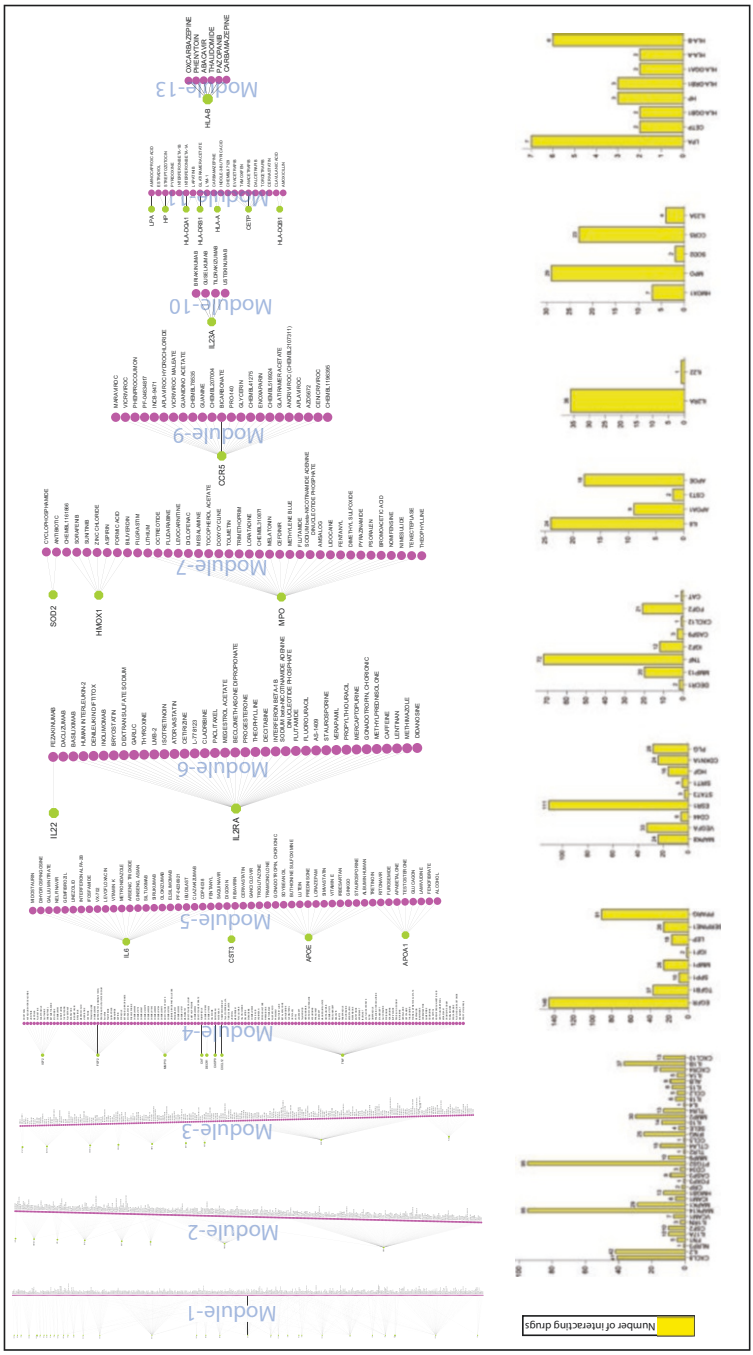

Fig. 5.3 Drug–target interaction network: all the 13 modules' targets (Green) are mapped with their respective drugs (Magenta). The bar graph illustrating the number of interacting drugs with key targets (Ref: doi:10.3389/fphar.2021.770762)

(Costa 1991). Digoxin, beta-blockers, estrogen, oral hypoglycemic medicines, and diuretics were the five drug classes most likely to be involved in possible drug interactions. In this way, Cao et al. have developed a database, e.g., *DDInter* (http://ddinter.scbdd.com) (Xiong et al. 2022), The curated Drug–Drug Interaction (DDI) database offers extensive data, practical medication guidance, an intuitive functional interface, and powerful visualization tools to cater to the needs of the scientific community. The database currently includes approximately 0.24 million DDI associations. Figure 5.3 showcases the Drug–Target Interaction Network, where all 13 modules' targets (in green) are mapped with their respective drugs (in magenta). The illustration on the right demonstrates the number of interacting drugs with key targets, the database connects 1833 approved drugs, encompassing 1972 entities. Each drug in the database is accompanied by essential chemical and pharmacological information, along with its interaction network. This comprehensive annotation allows for a deeper understanding of the drug's characteristics and its interactions with other entities within the network (Fig. 5.4).

5.2.7 Functional Modules in Molecular Networks

Protein clusters or sub-networks that exhibit dense connections are often regarded as potential functional modules within a given network. Another hypothesis proposes that proteins interacting with similar groups of other proteins in the network

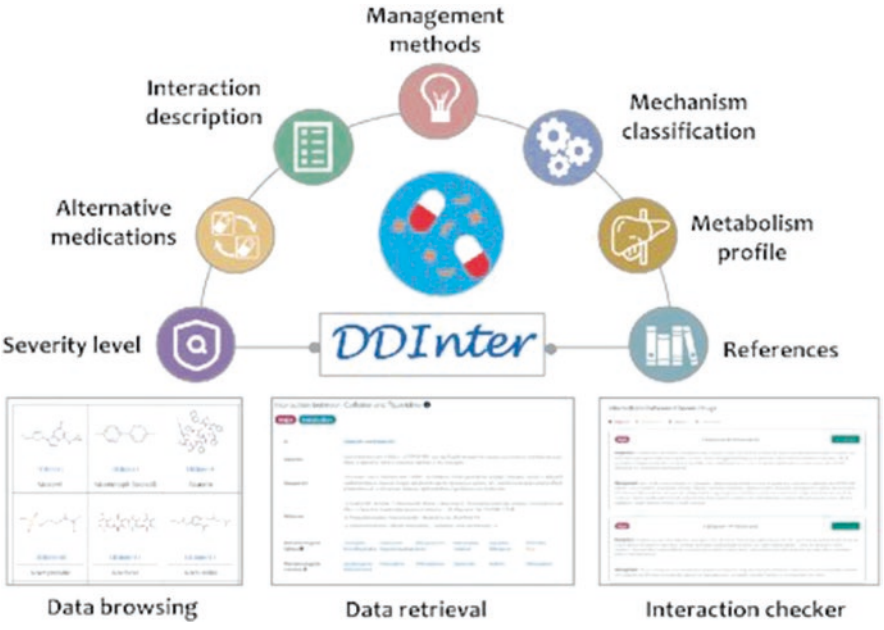

Fig. 5.4 DDInter provides detailed annotations of each DDI association and enables users to conduct data query (image courtesy, please refer: https://doi.org/10.1093/nar/gkab880)

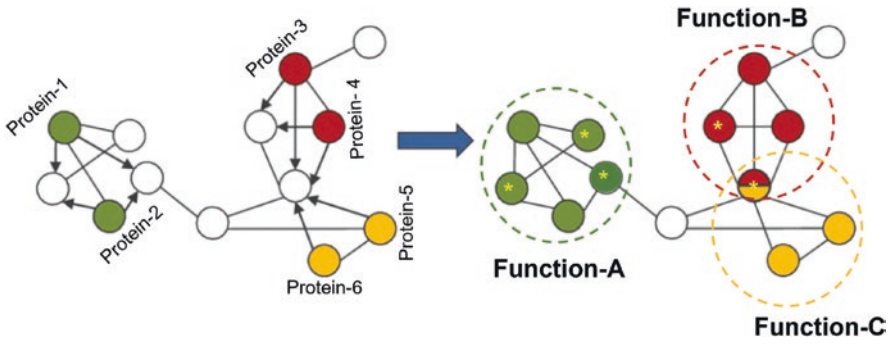

Fig. 5.5 Guilt-by-association approach: different colors are used to denote proteins with known functions, while proteins with unknown functions are left uncolored. Transferring functional annotation from directly interacting proteins allows for the inference of protein function (indicated by arrows)

tend to have comparable functions. Even in cases where direct interactions are absent, these proteins contribute to similar biological functions and should be included in the same modules. However, both definitions of modules rely on a guilt-by-association approach, as illustrated in Fig. 5.5. For instance, if two proteins interact with each other, it is more likely that they share the same cellular functionalities compared to proteins that do not interact (Wang and Qian 2014).

There are many tools for the identification of functional modules within the network, but few are widely used among the researcher including the *LEV* (leading eigenvector) method that detects the communities in network from package "igraph" in R (Newman 2006). Another approach called MCODE (Molecular Complex Detection) focuses on identifying densely connected regions within a network. MCODE aims to identify molecular complexes or clusters characterized by high connectivity and close proximity of nodes (Bader and Hogue 2003).

Another is the *DIAMOnD* (DIseAse MOdule Detection), which we can investigate the local network neighborhood (LNN) near a particular set of known disease proteins, and this helps us identify potential new disease target protein (Ghiassian et al. 2015) and last one which I found very useful and accurate that is MTGO (Module detection via Topological information and Gene Ontology knowledge) is a method that identifies modules within a network using both network topology and the biological role of proteins based on Gene Ontology (GO) terms (Vella et al. 2018).

5.3 Drug–Repurposing Opportunities

Network-based drug repurposing methodologies are based on the premise that drugs capable of interacting with multiple targets can demonstrate efficacy against specific diseases. Furthermore, these methodologies take into account the potential therapeutic implications when two drugs target the same protein. By leveraging the interconnected nature of biological networks, these approaches offer promising

avenues for identifying new therapeutic uses for existing drugs. This approach proves to be effective in the discovery and development of drug molecules with new pharmacological or therapeutic indications. By targeting specific proteins through therapeutic combinations or drug repurposing, it becomes possible to improve clinical conditions in cases of comorbidity, enhance the potency of certain drugs, and achieve synergistic effects for better treatment outcomes (Imam et al. 2023).

5.4 Drug Side Effects

Over a million significant injuries and fatalities are caused annually by adverse drug reactions (ADRs), which are very common. Currently, many machine learning and network-based approaches are used to predict the adverse drug. By leveraging features such as composition, structure, and binding affinity, researchers have employed methods that involve machine learning (ML) and deep learning techniques (Dara et al. 2022). As part of these ongoing efforts, a recent development is the T-ARDIS database (Galletti et al. 2021). T-ARDIS is a carefully curated compilation of relationships between proteins and Adverse Drug Reactions (ADRs). These associations are statistically evaluated and sourced from existing databases of drug–target and drug–ADR associations.

5.5 Integrating Omics Data with Networks: Challenges and Ways

Multi-omics data will soon be regularly used in preclinical and clinical contexts, hence further constraints should apply standardized, rigorous bioinformatics techniques to process, normalize, and analyze the massive datasets from various study modalities. From this perspective, it should soon be possible to build networks of networks to determine how various biological aspects are interconnected. Beyond intracellular molecular networks, we are now combining the connections between various cell-types, organ-systems, hosts and microorganisms, hosts and environmental exposures, as well as other interconnections. It would also be essential to incorporate psychosocial components when defining disease networks in order to link all contributing factors to clinical results. As demonstrated, network controllability analysis continues to be useful for identifying important disease genes and prioritizing potential targets.

The existing human interactome is incomplete, which restricts network analysis of complicated human disorders. The interactome is expected to keep growing as high-throughput technology and bioinformatics continue to progress. The existing interactome is based solely on the binary biophysical interactions between the curated PPIs, with little knowledge of or annotations for protein-binding patterns or domains. There is an ongoing attempt to develop a “domain-specific interactome (DSI)” where commonly shared domains or motifs, such “SH3” and “PDZ domain” are two examples of frequently shared domains or motifs that are being screened

and cloned for their interacting partners. The possibility of modifications to the physical and metabolic characteristics in a physiological context that changes protein binding at these domains is a limitation of this strategy.

In addition, the majority protein–protein interactions (PPIs) within the current interactome are predicted based on the induced protein expression levels observed in experimental yeast cells, which could be very different from the endogenous ecosystem where key targets are normally expressed. An area of active research is the integration of PPIs with gene expression data and further techniques to provide tissue and disease-specific perspective to the existing interactome. Notably, the present interactome only contains one isoform of each gene product and provides scant annotations regarding the splice isoforms that are being considered. It is generally accepted that distinct spliced forms can results in significant changes in phenotypic variations.

Reticulocyte analysis is a recently developed idea in which a patient's unique integrative biological network (reticulome) and a set of molecular-mutants/variants are investigated. Each person has a unique set of biological networks. Biological network environment within an individual unquestionably influences the final result (phenotype) of a certain set of genetic variations (genotype) and therefore, should be an essential element of any patient-specific data assessment. Therefore, customized reticulotype-based network studies have potential to strengthen the current genotype/phenotype correlation attempts and may make the search for customized targeted therapies.

5.6 Case Studies

5.6.1 Case Study: 1

Imam et al. conducted a recent study titled "Network-Medicine Approach for the Identification of Genetic Association of Parathyroid Adenoma with Cardiovascular Disease and Type-2 Diabetes." This study delves into unexplored dimensions of diseases by specifically investigating distantly related protein sets associated with other diseases that have not been previously studied. The aim is to understand their collective physiological impact on the pathological phenotype through network analysis.

The researchers performed a comparative analysis of disease-associated proteins in Parathyroid Adenoma, Cardiovascular Disease, and Type-2 Diabetes with the aim of identifying shared genetic factors. Utilizing network analysis methods, they investigated functional modules within the protein-disease network that exhibited dense internal connections but sparse connections with the rest of the network. As a result, they discovered 13 target proteins that were found to be common to parathyroid adenoma, cardiovascular disease, and type 2 diabetes. These proteins were subsequently organized into hierarchical modules and sub-modules within the Protein–Protein Interaction (PPI) network. In their study, the researchers also employed the concept of drug repurposing and drug combinations. They utilized

target proteins and associated drugs in a drug–target bipartite network, which included experimentally verified drug–target binary connections. They found that 36 drugs were common to both target-associated drugs (TAD) and disease-associated drugs (DAD), supporting the effectiveness of a multi-target drug approach.

This network-based analysis presents promising avenues for personalized treatment and the repurposing of drugs. It allows for the exploration of new targets and combinations of multiple drugs, facilitating a comprehensive understanding of protein–disease associations and disease–disease relationships. By utilizing advanced computational techniques, the study prioritizes drug–target interactions and examines disease–disease connections, thereby enhancing the selection of potential therapeutic targets based on efficacy and safety in complex disease scenarios (For more details, read the full article: https://doi.org/10.1093/bfgp/elac054).

5.6.2 Case Study: 2

A study conducted by Aftab et al. aimed to construct a disease network by investigating the overlap between Tuberculosis (TB) and other Non-Communicable Diseases (NCDs) such as Parkinson's Disease (PD), Cardiovascular Disease (CVD), Diabetes Mellitus (DM), Rheumatoid Arthritis (RA), and Lung Cancer (LC). Through the analysis of this disease network, the researchers identified common genes associated with TB and other NCDs, establishing important gene–disease relationships.

To delve deeper into these diseases, the researchers constructed separate gene interaction networks for each disease by integrating carefully curated and experimentally validated human interactions. Additionally, they generated and analyzed a drug–target interactome network that encompassed clinically relevant drug–drug and drug–target interactions. This comprehensive network offered a more comprehensive understanding of the intricate landscape of drug–target interactions.

The primary objective of this study was to establish a comprehensive workflow that takes into account Tuberculosis (TB) and its overlapping Non-Communicable Diseases (NCDs), emphasizing the significance of reconsidering and redefining therapies and therapeutic management. The findings underscore the potential of exploring uncharted territories in disease research, particularly the shared gene sets that coexist among various diseases. By fostering collaboration, it becomes feasible to collectively impact the pathological phenotype at a physiological level (For more details, read the full article: https://doi.org/10.3389/fphar.2021.770762).

5.7 Conclusion

Multi-omics data will soon be regularly used in preclinical and clinical contexts, hence further constraints should apply standardized, rigorous bioinformatics techniques to process, normalize, and analyze the massive datasets from various study modalities. From this perspective, it should soon be possible to build networks of

networks to determine how various biological aspects are interconnected. Beyond intracellular molecular networks, we are now combining the connections.

between various cell-_types, organ-systems, hosts and microorganisms, hosts and environmental exposures, as well as other interconnections. It would also be essential to incorporate psychosocial components when defining disease networks in order to link all contributing factors to clinical results. As demonstrated, network controllability analysis continue to be useful for identifying important disease genes and prioritizing potential targets. Reticulocyte analysis is a recently developed idea in which a patient's unique integrative biological network (reticulome) and a set of molecular-mutants/variants are investigated. Each person has a unique set of biological networks. A biological network environment within an individual unquestionably influences the final result (phenotype) of a certain set of genetic variations (genotype) and therefore, should be an essential element of any patient-specific data assessment. Therefore, customized reticulotype-based network studies have the potential to strengthen the current genotype–phenotype correlation attempts and may make the search for customized targeted therapies.

References

Alam A et al (2019) Identification and classification of differentially expressed genes and network meta-analysis reveals potential molecular signatures associated with tuberculosis. Front Genet 10:932. https://doi.org/10.3389/fgene.2019.00932

Alam A, Imam N, Siddiqui MF, Ali MK, Ahmed MM, Ishrat R (2021) Human gene expression profiling identifies key therapeutic targets in tuberculosis infection: a systematic network meta-analysis. Infect Genet Evol 87:104649. https://doi.org/10.1016/j.meegid.2020.104649

Alam A et al (2022) An integrative network approach to identify common genes for the therapeutics in tuberculosis and its overlapping non-communicable diseases. Front Pharmacol 12:770762. https://doi.org/10.3389/fphar.2021.770762

Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH (2017) HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res 45(D1):D408–D414. https://doi.org/10.1093/nar/gkw985

Alonso-López D et al (2019) APID database: redefining protein–protein interaction experimental evidences and binary interactomes. Database 2019:baz005. https://doi.org/10.1093/database/baz005

Azuaje FJ, Zhang L, Devaux Y, Wagner DR (2011) Drug-target network in myocardial infarction reveals multiple side effects of unrelated drugs. Sci Rep 1(1):52. https://doi.org/10.1038/srep00052

Bader GD (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31(1):248–250. https://doi.org/10.1093/nar/gkg056

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4(1):2. https://doi.org/10.1186/1471-2105-4-2

Black JW, Leff P (1983) Operational models of pharmacological agonism. Proc R Soc Lond B Biol Sci 220(1219):141–162. https://doi.org/10.1098/rspb.1983.0093

Boran ADW, Iyengar R (2010) Systems pharmacology. Mt Sinai J Med 77(4):333–344. https://doi.org/10.1002/msj.20191

Calderone A, Castagnoli L, Cesareni G (2013) Mentha: a resource for browsing integrated protein-interaction networks. Nat Methods 10(8):690–691. https://doi.org/10.1038/nmeth.2561

Chatr-aryamontri A et al (2007) MINT: the molecular interaction database. Nucleic Acids Res 35, no. Database:D572–D574. https://doi.org/10.1093/nar/gkl950

Colquhoun D (2006) The quantitative analysis of drug-receptor interactions: a short history. Trends Pharmacol Sci 27(3):149–157. https://doi.org/10.1016/j.tips.2006.01.008

Costa AJ (1991) Potential drug interactions in an ambulatory geriatric population. Fam Pract 8(3):234–236. https://doi.org/10.1093/fampra/8.3.234

Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ (2022) Machine learning in drug discovery: a review. Artif Intell Rev 55(3):1947–1999. https://doi.org/10.1007/s10462-021-10058-4

Das J, Yu H (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol 6(1):92. https://doi.org/10.1186/1752-0509-6-92

Fahey ME et al (2011) GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics 12(1):298. https://doi.org/10.1186/1471-2105-12-298

Galletti C, Bota PM, Oliva B, Fernandez-Fuentes N (2021) Mining drug-target and drug-adverse drug reaction databases to identify target-adverse drug reaction relationships. Database (Oxford) 2021:baab068. https://doi.org/10.1093/database/baab068

Ghiassian SD, Menche J, Barabási A-L (2015) A disease module detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol 11(4):e1004120. https://doi.org/10.1371/journal.pcbi.1004120

Goel R, Harsha HC, Pandey A, Prasad TSK (2012) Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. Mol Bio Syst 8(2):453–463. https://doi.org/10.1039/C1MB05340J

Hermjakob H (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32(90001):452D–D455. https://doi.org/10.1093/nar/gkh052

Imam N, Alam A, Siddiqui MF, Veg A, Bay S, Khan MJI, Ishrat R (2023) Network-medicine approach for the identification of genetic association of parathyroid adenoma with cardiovascular disease and Type-2 diabetes. Brief Funct Genomics 22(3):250–262. https://doi.org/10.1093/bfgp/elac054

Kalathur RKR et al (2014) UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. Nucl Acids Res 42(D1):D408–D414. https://doi.org/10.1093/nar/gkt1100

Kim K-J, Moon S-J, Park K-S, Tagkopoulos I (2020) Network-based modeling of drug effects on disease module in systemic sclerosis. Sci Rep 10(1):13393. https://doi.org/10.1038/s41598-020-70280-y

Kotlyar M, Pastrello C, Sheahan N, Jurisica I (2016) Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res 44(D1):D536–D541. https://doi.org/10.1093/nar/gkv1115

Lee LY-H, Loscalzo J (2019) Network medicine in pathobiology. Am J Pathol 189(7):1311–1326. https://doi.org/10.1016/j.ajpath.2019.03.009

Luck K et al (2020) A reference map of the human binary protein interactome. Nature 580(7803):402–408. https://doi.org/10.1038/s41586-020-2188-x

Maehle A-H, Prüll C-R, Halliwell RF (2002) The emergence of the drug receptor theory. Nat Rev Drug Discov 1(8):637–641. https://doi.org/10.1038/nrd875

Morselli Gysi D et al (2021) Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proc Natl Acad Sci USA 118(19):e2025581118. https://doi.org/10.1073/pnas.2025581118

Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):036104. https://doi.org/10.1103/PhysRevE.74.036104

Oughtred R et al (2019) The biogrid interaction database: 2019 update. Nucleic Acids Res 47(D1):D529–D541. https://doi.org/10.1093/nar/gky1079

Patil A, Nakai K, Nakamura H (2011) HitPredict: a database of quality assessed protein–protein interactions in nine species. Nucleic Acids Res 39(suppl_1):D744–D749. https://doi.org/10.1093/nar/gkq897

Pinero J et al (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015(0):bav028. https://doi.org/10.1093/database/bav028

Sakle NS, More SA, Mokale SN (2020) A network pharmacology-based approach to explore potential targets of Caesalpinia pulcherima: an updated prototype in drug discovery. Sci Rep 10(1):17217. https://doi.org/10.1038/s41598-020-74251-1

Schoof H et al (2005) Munich information center for protein sequences plant genome resources. a framework for integrative and comparative analyses. Plant Physiol 138(3):1301–1309. https://doi.org/10.1104/pp.104.059188

Szklarczyk D et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1):D607–D613. https://doi.org/10.1093/nar/gky1131

Vella D, Marini S, Vitali F, Di Silvestre D, Mauri G, Bellazzi R (2018) MTGO: PPI network analysis via topological and functional module identification. Sci Rep 8(1):5499. https://doi.org/10.1038/s41598-018-23672-0

Wang Y, Qian X (2014) Functional module identification in protein interaction networks by interaction patterns. Bioinformatics 30(1):81–93. https://doi.org/10.1093/bioinformatics/btt569

Xenarios I (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28(1):289–291. https://doi.org/10.1093/nar/28.1.289

Xiong G et al (2022) DDInter: an online drug–drug interaction database towards improving clinical decision-making and patient safety. Nucleic Acids Res 50(D1):D1200–D1207. https://doi.org/10.1093/nar/gkab880

Role of R in Biological Network Analysis

6

Mohd Murshad Ahmed and Safia Tazyeen

Abstract

The life sciences are becoming increasingly data rich as a result of technical development in molecular biology (large-scale systems biology, genomics). Appropriate analysis of large-scale datasets (proteomics, genomes, multi-omics data, Htseq, RNA-seq, ChIPseq, and so on) is now a bottleneck. One of the most frequent ways of representing biological systems as complicated sets of binary interactions or relationships between diverse bio-entities is through networks. In fields like computational biology, finance, neurology, political science, and public health, network analysis is a technique that employs graph theory to investigate complicated real-world hurdles. Networks, which are mainly portrayed as graphs with hundreds of nodes and thousands of vertices, are the best way to describe the various components of a system and their interactions. Some of the potential applications of network analysis in biology and medicine include discovering drug targets, determining the function of a protein or gene, inventing successful strategies for treating various diseases, and providing early identification of abnormalities. Several software tools, such as Gephi and Cytoscape, are built for network analysis and the production of network graphs. R has evolved into a formidable tool for network analysis, despite not being specifically designed for it. R has three distinct advantages over standalone network analysis applications. The goal of this research/chapter is to provide a solid understanding of how to use R to perform advanced data analysis. Researchers will learn how to extract relevant information from multiple datasets using a variety of regularly

M. M. Ahmed (✉)
Singapore Institute of Clinical Sciences (SICS), Singapore, Singapore

S. Tazyeen
Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_6

used statistical and mathematical methodologies. The goal is also to demonstrate and highlight a few important methodologies related to network analysis, rather than to provide a fully-fledged analysis. In this chapter, we illustrate ideas, models, and methods from the graph theory world and explore how they might be utilized to uncover hidden aspects and features of a network.

Keywords
Packages · Co-expression network · Survival analysis · Modules analysis · Gene ontology

6.1 Introduction

In biology, network structures are everywhere because many biological systems depend on intricate interactions between the parts that make them up. Predator–prey partnerships are the most frequent kind of species interaction in ecosystems and are essential to preserving biodiversity (Oleskin 2014). Synapses in the human brain are used to transmit electrical and chemical signals between neurons. DNA, RNA, proteins, and other components play various roles in biochemical processes at the cellular level that control how cells function. These systems are best represented mathematically by networks, which have sets of elements, which we shall call vertices or nodes, with connections between them, called edges (Wang et al. 2021). Discovering the biological information from network concepts is of major significance since biological entities are involved in complicated and complex interactions (Liu et al. 2020). Experimental biology developments have increased the availability of large-scale biological network data (Vitkup 2004). We examine networks of protein–protein interactions (PPIs), in which proteins are represented as network nodes and interactions as network edges (Nabieva et al. 2005). Since proteins interact with one another to carry out nearly all biological functions, studying the structure of PPI networks may provide new information about disease and complex biological phenomena (Memisevic et al. 2010). Researchers have gained fresh insights into the challenge of modeling complex systems as a result of the growth of network data in a variety of fields, which has also sparked the development of various cutting-edge statistical approaches and computational tools/databases such as R/RStudio/GEO2R (Wang et al. 2021). R is a widely used open-source programming language for statistical computing and data analysis. R typically includes a command-line interface (Lovelace et al. 2019). R is accessible on popular operating systems like Windows, Linux, and macOS. The newest cutting-edge technology is the R programming language developed by Ross Ihaka and Robert Gentleman (Giorgi et al. 2022; Ihaka and Gentleman 1996). Much of the intricacy of cellular life is mediated by gene interaction networks, in which genes activate and repress the transcription of other genes, and their dysfunction can be fatal to an organism. So, systems biology has long sought to understand these networks (Saint-Antoine and Singh 2020). Exploring gene interactions experimentally by testing every possible gene pair is impractical due to the large number of genes present in the human

genome, which is approximately 20,000. However, modern molecular biology techniques, such as DNA microarrays and next-generation sequencing (NGS), have made it possible to obtain a quantitative understanding of the transcriptome profile of a single cell or a group of cells. RNA-sequencing (RNA-Seq), whether done on individual cells or as bulk sequencing, is a particularly valuable tool for this purpose. These methods have greatly facilitated the discovery of gene interactions and other molecular mechanisms underlying complex biological processes (Madsen et al. 2022). This chapter will demonstrate the numerous applications of the R language. Despite the fact that R is one of the most popular and effective programming languages in bioinformatics, R excels in the creation of graphs and figures of publication quality and in the use of a variety of statistical tools, such as RNA-Seq and genomics.

6.2 Installation of R Software and Packages

We will start by learning the statistical programming language R. R programming is a great option for processing and analyzing life science data since it has a large community of developers who are constantly creating new R programming packages. R software is deserving widespread acclaim, and it will continue to grow. A wide range of statistical techniques is supported by R software, including traditional statistical tests, modeling (both linear and nonlinear), classification, time series analysis, cluster analysis, and graphical data display. R software is a great option for manipulating big data and life science data because it is extremely flexible and simple to learn. We first need to get access to R, then we can begin learning it. The focus of this chapter will be on R implementation. An R installation is available at http://cran.r-project.org and RStudio at http://www.rstudio.com/. This is simple; the software package is not large, and I normally work on things offline. R console like below will appear, and the R program will get started, see Fig. 6.1.

6.2.1 Packages in R

R packages are a collection of preset functions that can be used as a library when deploying an R application to promote reuse and a low-code design. R packages are created outside of the R environment and can be imported to make use of the available functions they include. When a package is loaded into the R environment, it offers the necessary functions that can be used. A library in C, C++, or Java is analogous to a list of R Packages. Therefore, a package might essentially comprise a variety of functionality, such as functions, and constants, that we will then let the user use in the context of a certain situation. R comes with a number of packages that may be obtained from CRAN (Comprehensive R Archive Network) and GitHub.

Fig. 6.1 The diagram depicts the information about R. Sixty-four bit downloaded and installed in the window PC

6.2.2 Installing R Packages

Direct package installation is possible using either the IDE or commands. We use the function below and the package name to install packages:

Syntax:

```
install.packages()
```

Code:

```
install.packages("affy")
```

The above code installs the affy package in R.

6.2.3 Loading R Packages

To use the loaded packages in R, we must load them after installing the R package. To load the packages, we employ the functions below:

Syntax:

```
>Library (package name)
```

Code:

```
>Library (affy)
```

Now affy package is loaded in R Console. To overview the complete process of R installation and update refer to Fig. 6.2.

6.3 To Build Network Model Using R

Data must be in a specific form for the network analysis packages to build the unique type of object that is required by each program. Adjacency matrices, also known as sociomatrices, are the foundation for the object classes used by the programs network, igraph, and tidygraph (see Fig. 6.3).

A three-edged, undirected graph is produced by the code below. The edges are 1- > 2, 2- > 3, 3- > 1 since the integers are read as vertex IDs. M1 < − graph (edges = c(1,2, 2, 3, 1), *n* = 3, directed = F) plot(M1).

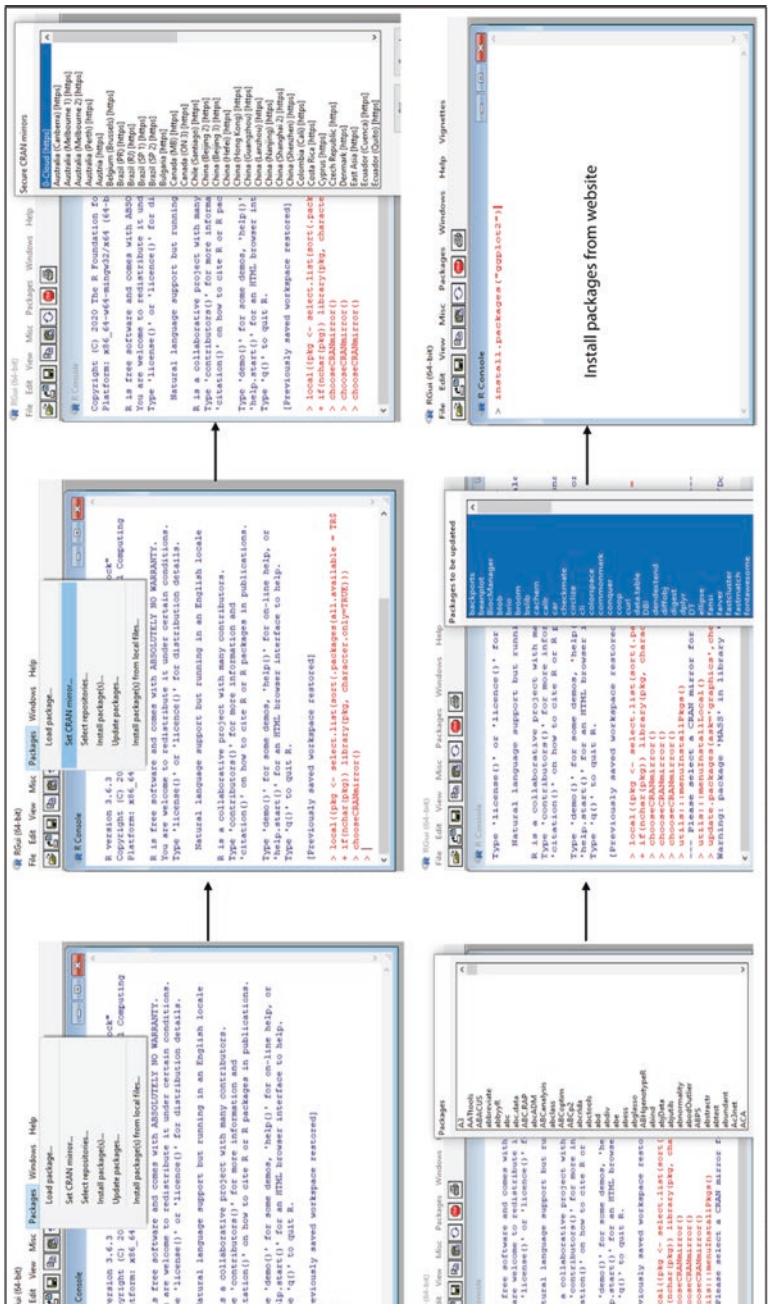

Fig. 6.2 The diagram depicts the packages of R. Direct installation of packages in R using a webserver see in the last corner image

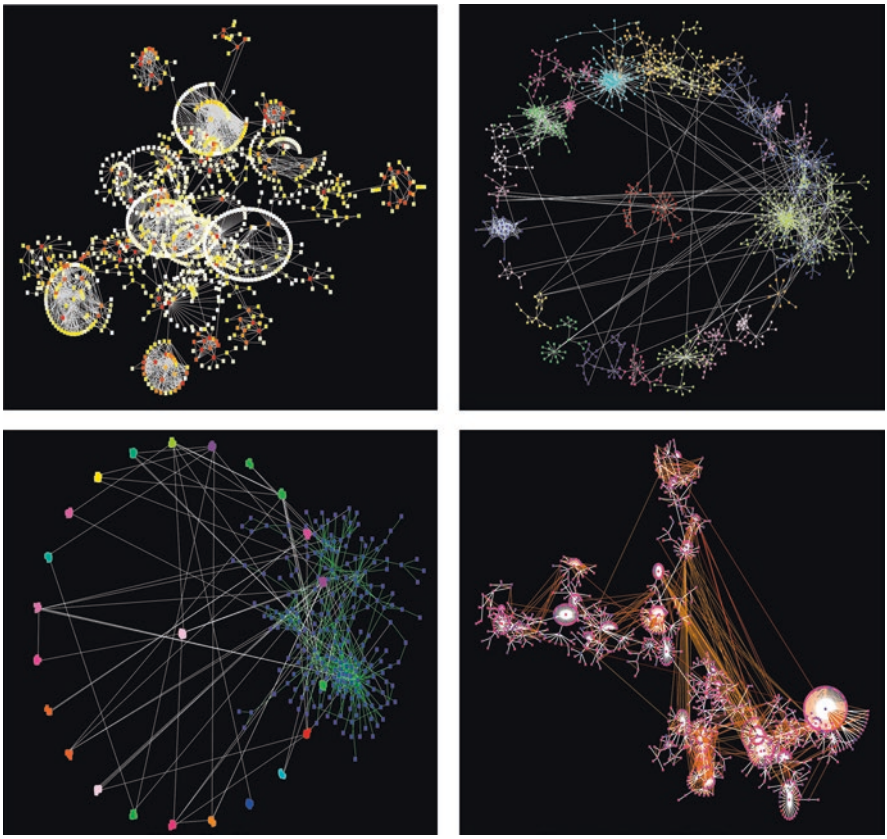

Fig. 6.3 The diagram depicts the network plot in R

6.4 Differential Gene Expression Analysis Microarray in R

Among all gene analyses, gene expression analysis has the potential to be the most important. Gene expression is the process by which genetic information is transferred from DNA to mRNA and subsequently translated into proteins. Although gene expression profiling has helped to better categorize some diseases (such as chronic myelogenous leukemia and breast cancer), it still presents a significant barrier to understanding the molecular mechanisms underlying disease pathogenesis. An overall picture of cellular activity can be obtained by studying the pattern of genes that are expressed at the transcriptional level in a particular cell or under particular circumstances. This process is known as gene expression profiling. Two techniques for doing so include DNA microarrays, which assess the relative activity of previously chosen target genes, or sequencing technologies, which enable profiling of all active genes. Our comprehension of the cellular processes, metabolic alterations, and transcriptional reprogramming of the diseased heart among inheritable forms of cardiovascular disease has been significantly impacted by recent

developments in single cell disorder. By changing the environment to which the cell is exposed and identifying which genes are expressed, gene expression profiling enables you to examine the effects of various situations on gene expression. Alternately, gene expression profiling enables you to ascertain whether a cell is performing a function for which you are already aware that a particular gene is important. For instance, some genes have been linked to cell division; if these genes are active in a cell, you can determine whether the cell is dividing or whether it is differentiating. The creation of hypotheses frequently uses gene expression profiling. Expression profiling under various circumstances can assist in developing a hypothesis when little is known about how and when a gene will be expressed.

If so, it might be determined through additional research. Investigating the impact of drug-like compounds on cellular response is another application of gene profiling. Determine if cells express genes known to be involved in responding to hazardous environments when exposed to the drug in order to find the genetic markers of drug metabolism.

Gene expression profiling is a potent method for locating DEGs and locating impacted pathways, and this method offers additional insight into the mechanisms underlying the progression of the disease. Suppose disease condition cells express higher or lower levels of certain genes, and these genes code for a protein receptor. In that case, this receptor may be involved in particular diseases, and targeting it with a drug might treat the disease. Gene expression profiling might then be a key diagnostic tool for people with complex diseases. In transcriptomics research, gene differential analysis is commonly utilized. When two groups of samples have distinct characteristics, their genes are almost certainly expressed differently. Differential analysis, in contrast to other analyses such as expression analysis or cluster analysis, focuses on the genes that are differentially expressed between two situations. The differentially expressed genes could explain why the two groups have different characteristics. Differential expression analysis aims to find genes whose expression varies depending on the situation. Correction for multiple testing is a significant factor in differential expression analysis. This is a statistical phenomenon that arises when a small number of samples are subjected to thousands of comparisons (for example, comparing the expression of numerous genes in different situations) (most microarray experiments have less than five biological replicates per condition). As a result, there is a higher chance of getting false-positive results. Advances in molecular biology and information technology have made it possible to explore a large portion of the genomes of numerous species in recent years. Because the amount of data created in the field of molecular biology is immense, current bioinformatics studies are focused on the structural and functional features of genes and proteins. The requirement to determine whether genes have different expression models by phenotype or experimental situation is the guiding premise in analyzing gene expression data. The “fold change” criteria is a straightforward method for choosing genes. Only if there are no or only a few repetitions is this conceivable. An analysis based solely on fold change, on the other hand, does not allow for the evaluation of the importance of expression differences in the presence of biological and experimental variables that may differ from gene to gene. This is the primary reason for evaluating differential expressions with statistical

tests. The fundamental tenet of biology explains how information is extracted from genes and used to make proteins. Proteins are created by RNA translation after RNA transcription. All living things use a process known as gene expression to produce the elements that makeup life from genetic information. A cell might read its genetic code differently since it only expresses a subset of the genes it possesses at any one time. The cell's ability to regulate its size, shape, and functions is accomplished by controlling which genes are expressed. The phenotype of an organism, such as the color of a mouse's hair or if it has any at all, depends on how the organism's cells express the genes that are present in them. Gene expression profiling counts the genes that are active at any particular time in a cell. With this technique, thousands of genes can be measured at once; in some experiments, the entire genome can be measured simultaneously. By measuring mRNA levels, gene expression profiling reveals the pattern of genes that are expressed by a cell at the transcription level. This sometimes entails measuring the relative levels of mRNA in two or more experimental circumstances, then determining which conditions led to the expression of particular genes. Many biomedical researchers, from molecular biologists to environmental toxicologists, use gene expression profiling. This method can support a wide range of experimental objectives by providing precise information on gene expression. The process of using a gene's information to create a functioning gene product, which may be a protein, is known as gene expression. Understanding the biological distinctions between healthy and sick situations requires knowledge of differential gene expression.

6.5 RNA-Seq Analysis in R

Microarray data has been the most prevalent sort of transcriptomics data available to scientists for a long time. This began to change in 2009 when technological breakthroughs made RNA-seq a viable alternative to microarray data. The identification and confirmation of biomarker signatures, as well as the importance of differential gene expression in normal biological and pathological processes, can all be aided by gene expression analysis. High-throughput sequencing tools are used in RNA sequencing (RNA-Seq), which provides data on a cell's transcriptome. Compared to prior Sanger sequencing and microarray-based techniques, RNA-Seq provides noticeably higher coverage and precision of the dynamic dynamics of the transcriptome. In addition to assessing gene expression, RNA-Seq data can be used to find novel transcripts, recognize alternatively spliced genes, and find allele-specific expression. Researchers can now better comprehend the functional complexity of the transcriptional machinery because of recent improvements in the RNA-Seq methodology, from sample preparation to library construction to data interpretation. R is used to analyze RNA-seq count data. With a focus on the DESeq, this will involve reading the data into R, doing quality control, differential expression analysis, and gene set testing.

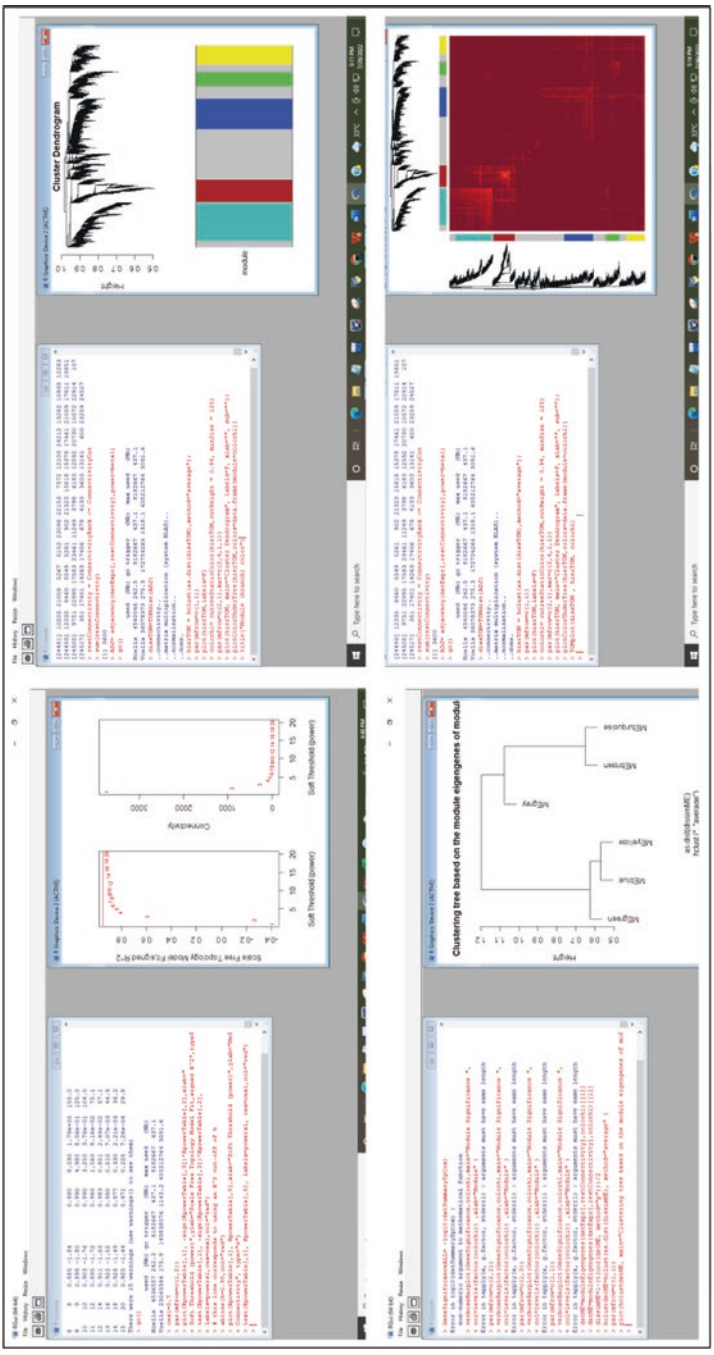

Fig. 6.4 The diagram depicts the various module plot in R

6.6 Weighted Correlation Network Analysis Using R

Applications in bioinformatics are increasingly using correlation networks. Weighted gene co-expression network analysis (WGCNA) is one systems biology method for summarizing the correlation patterns across genes across microarray sets. WGCNA can be used to locate clusters (modules) of highly correlated genes, summarize these clusters, connect modules to one another and to variables from external samples, and compute module membership measures. Network-based gene screening techniques for the discovery of potential biomarkers or therapeutic targets are made easier by correlation networks. These methods have been effectively applied in a wide range of scientific fields, including cancer research, mouse genetics, and yeast genetics. Steve Horvath and Peter Langfelder al did such type of analysis see link (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/)(see Fig. 6.4).

6.7 Network Component Analysis

A framework called Network Component Analysis (NCA) is used to infer the dynamics of regulatory signaling from network structure. NCA employs the connectivity structure of transcriptional regulatory networks to limit the decomposition to a single solution, in contrast to conventional approaches like principal component analysis or independent component analysis. The current version of NCA, however, is unable to take into account data from regulatory gene knockouts that limit the dynamics of regulatory signals, such as network topology data. By allowing for the incorporation of such data, NCA can be applied to systems that might not meet its identifiability requirements, resulting in more precise and self-consistent analysis across several experiments. High-throughput technologies like DNA microarrays commonly yield high-dimensional data sets that are the outputs of complex networked systems under the control of covert regulatory signals. Principal component analysis and independent component analysis are two examples of traditional statistical techniques for computing low-dimensional or hidden representations of these data sets. These techniques ignore the underlying network structures and provide decompositions based only on a priori statistical constraints on the computed component signals.

R Cluster Analysis Packages

There are many tools, databases, and plugins available in Cytoscape to find modules/communities/subnetworks. Modules from the network can be found using plugins like MCODE and Cytohubba. The web-based database centiserver, in contrast, has more than 20 centrality algorithms that are also used to find modules. In this chapter, we will discuss the igraph package for modules/community findings. The native network breaks into sets of subnetworks (sub-communities) using igraph coding in R. Finding the most influential nodes in Network using R packages is an advance level of centrality, see Figs. 6.5, and 6.6.

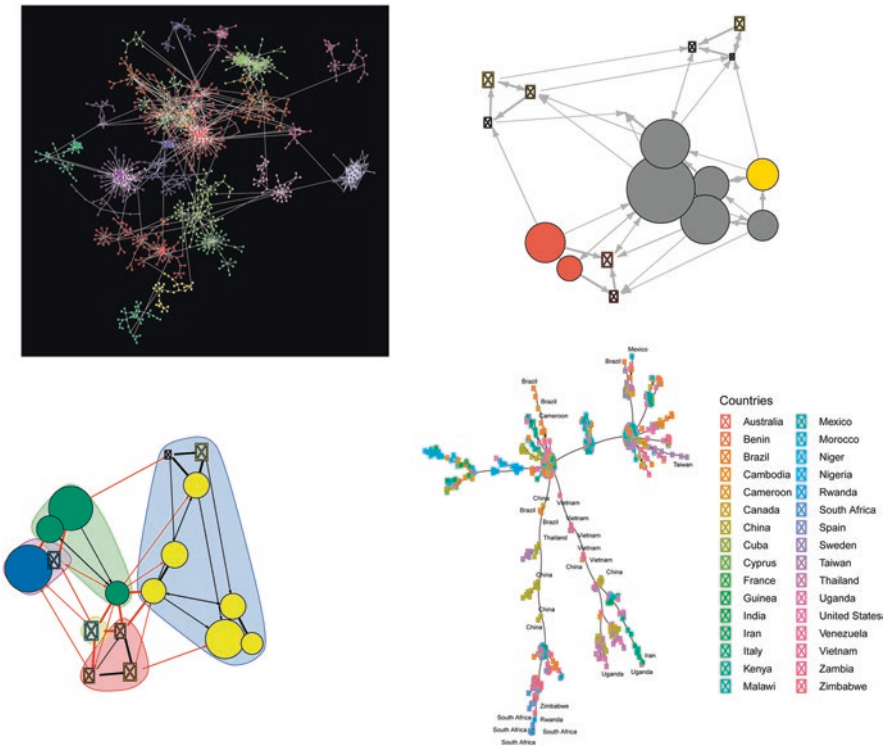

Fig. 6.5 The diagram depicts different modules in R. The left corner (upper side) shows different colors of network for modules. The right corner (upper side) shows the module based on degree, in which a higher degree means large-size nodes, whereas different color indicates different module nodes. The below-left corner indicates the modules with a color background. The below right corner shows the modules based on the country data

6.8 Preparing Network Data in R

The use of mathematical graph theory to represent, integrate, and analyze biological processes and data through networks is one of the fundamental ideas of systems biology. Depending on the type of data, biological networks such protein–protein interaction networks, gene regulatory networks, and metabolic networks can be created. Using network-based techniques as an integration and modeling tool, significant molecular connections can be uncovered. When used on specific patients, personalized network analysis can result in the identification of novel disease subtypes and therapeutic targets, enabling the development of novel drugs, the identification of novel biomarkers, and the repurposing of existing medications, as seen in the case of cancer. A biological system is represented as a network using a graph. It includes both biological components (such as cells, proteins, and genes) and the connections between them (like protein–protein interactions). In network biology, these are referred to as nodes and edges, respectively. Biological networks provide

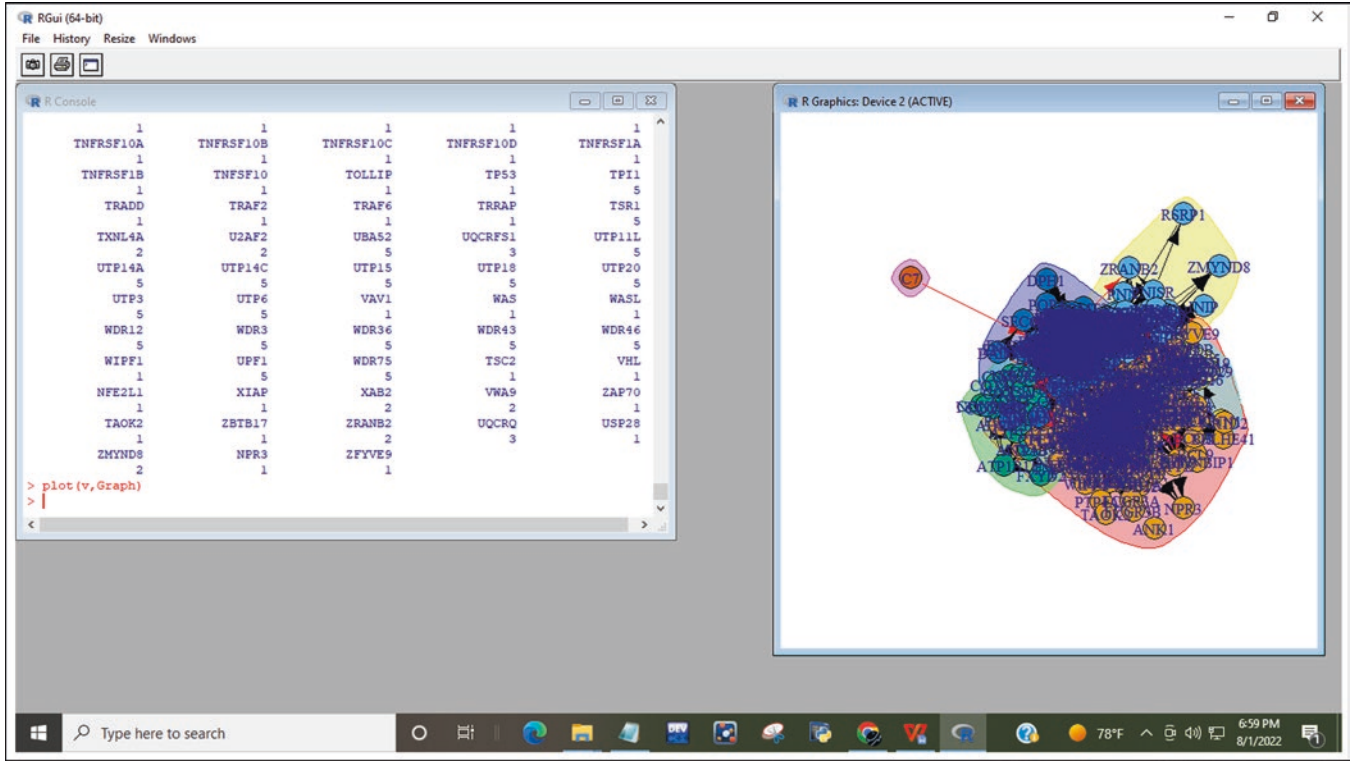

Fig. 6.6 The diagram depicts module of our data

a conceptual and accessible framework for studying, modeling, characterizing, and comprehending complicated interactions between various components in a biological system. In computer science and mathematics, network theory (which includes graph theory) has a wide range of applications, some of which are directly related to biological and disease networks. Examples of these applications include the Internet, social networks, particle physics, and other networks. Additionally, it has been discovered that network motifs, which are recurring and statistically significant sub-graphs in networks, have been preserved throughout evolution and are connected to particular biological functions. A system that consists of well-defined interaction elements, carrying certain topological features corresponding to physically or functionally related structures, can be represented by a graph (or network) G(n;m). The interacting elements are represented by nodes/vertices and their mutual interactions by edges/links. The system-specific relationship among the nodes and edges describes the topology of the network.

Based on edge attributes, there are three types of networks, directed, undirected, and weighted. In a directed network, the interaction among nodes has directional consequences, e.g., metabolic networks in which the nodes represent metabolites. The specific reaction pathways among the metabolites are represented by directional edges. Suppose the interactions can happen in both the direction with equal probability. In that case, the system is represented by an undirected network, e.g., a social friendship network, citation network, and large-scale functional biological network. Sometimes, edges can be attributed with weights to address a wide variance in the frequency of interaction among nodes, thus representing a weighted network.

6.9 Data Analysis and Visualization with R

The support for graphs in the R language is its most popular feature for producing various graphs and charts for visualizations. The R language includes a large selection of packages and features that may be used to generate the graphs using the input data set for data analytics. The most popular graphs in the R programming language are the scatter plot, box plot, line, pie, histogram, and bar graphs. R graphs enable both two- and three-dimensional visualizations for exploratory data analysis. Graphs are created using R methods like plot(), barplot(), hist(), and pie(). Advanced graph functions are supported by R packages such as ggplot2 (see Fig. 6.4).

6.9.1 Boxplot

Data can be represented graphically using boxes and whiskers using a boxplot. Variable value orders are sorted in ascending order before the data are divided into quarters. The box in the plot represents the IQR, or middle 50% of the data. The box's black line indicates the median.

See Code:

```
Boxplot(trees, col = c("yellow", "red", "cyan"), main = "bp
dataset")
```

6.9.2 Histogram

A histogram is a graphic instrument that analyses just one variable. A number of values known as the frequency are calculated when a number of variable values are categorized into bins. Following this calculation, frequency bars are plotted in the corresponding beans. Frequency is used to represent a bar's height. To create a histogram in R, we can use the hist() function, as seen below. Below is a straightforward histogram of tree heights:

Code:

```
hist(trees$Height, breaks = 10, col = "orange", main = "Histogram
of Tree heights", xlab = "Height Bin")
```

See Fig. 6.4.

6.9.3 Volcano Plot

In a volcano plot, the fold change is often found on the *x*-axis and the p-value is found on the *y*-axis. This form of scatter plot illustrates the differential expression of genes. It makes it possible to quickly visually identify genes that have significant statistical fold changes. These genes might be the ones with the most biological impact. In a volcano plot, the genes that are most upregulated are on the right, the genes that are most downregulated are on the left, and the genes that are most statistically significant are at the top, see Fig. 6.4.

6.9.4 Heatmap

The individual values found in a matrix are represented as colors in a heatmap, which is a graphical representation of data. R comes with the heatmap () function by default. It generates high-quality matrices and provides statistical tools for data normalization, grouping, and dendrogram visualization.

heatmap (data, scale="column")

R functions like plot(), barplot(), hist(), and pie() are used to create graphs in the R language. Advanced graph functions are supported by R packages such as ggplot2, see Fig. 6.7.

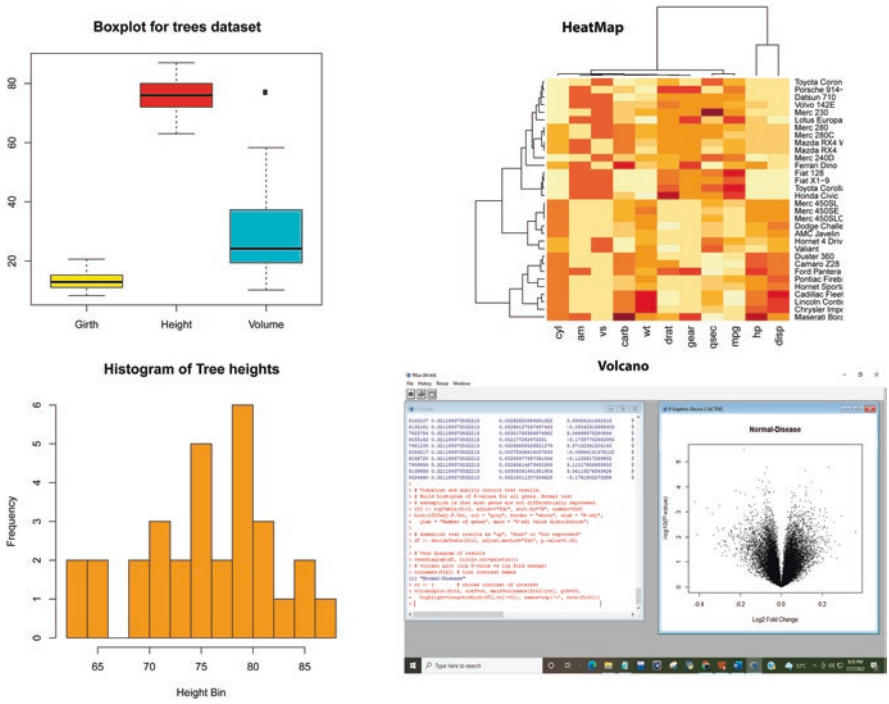

Fig. 6.7 The diagram depicts the visualizing plot in R

6.10 Case Study: Constructing a Protein–Protein Interaction Network from String Database to Find Out Influential Nodes Using R

One disadvantage of using the PSICQUIC web service to build an interaction network is that most biologists do not have easy access to it. Thankfully, there are some more user-friendly web-based tools available. We give a case study that demonstrates how to create and visualize a network of experimentally validated interactions from a gene list using a String database. In this case study, we used our published data on CVD and CKD (Ahmed et al. 2022). The mRNA expression patterns of CKD (GSE15072, GSE23609, GSE43484, GSE62792, GSE66494) and CVD (GSE26887, GSE42955, GSE67492, GSE71226, GSE141512, GSE48060) were collected from normal and treated samples. Using the R tool (version 3.6.0, 64-bit), gene expression microarray data from CVD and CKD were compared to find overlapped differentially expressed genes (DEGs). After that, all common/overlapping DEGs were analyzed using the online STRING version 9.1 tool, and the results were shown using Cytoscape (version 3.8.0). We have found 15 modules/clusters using Cystoscope's MCODE plugin, 10 of which contained genes of interest, and found that these were largely enriched in pathways. Nineteen important

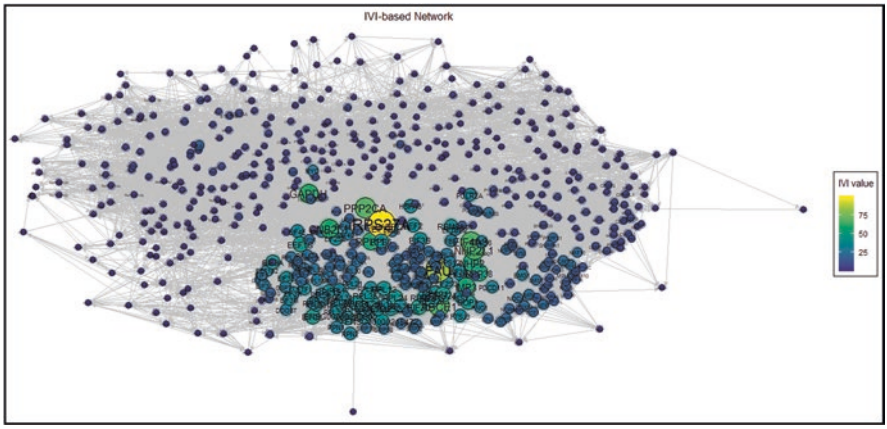

Fig. 6.8 There are 587 nodes and 13,887 edges in the network of CVD and CKD. The most significant nodes in the network are produced by R and significant packages are displayed. The nodes' color and interactions become more apparent as you zoom in on them. The original network's nodes are generally the same color, indicating that they have a score of less than 25. A color spectrum is used to indicate the nodes' values. Yellow indicates that the nodes are larger and have values higher than 75

genes (11 downregulated and 8 upregulated) were found in these 10 modules under study. The most important genes are found in modules 1 (RPL13 RPLP0 RPS24 RPS2) and 5 (MYC COX7B SOCS3). This study used IVI techniques to identify the most significant nodes in a gene–gene network, which may lead to the creation of new biomarkers. In this study, we combined the most important topological characteristics of the network to use a novel method called IVI (Integrated Value of Influence) to identify the network's significant players. The most important or influential nodes in a network are those with a high spreading potential (the spreader nodes are supposed to have the biggest impact on the information flow in the network) and a high hubness score. Top 20 nodes were extracted based on IVI values, hubness score, and spreading score, with RPS27A non-seed gene being the most significant node in the native network. However, among all the seed genes, RPS2 was the most significant node, see Fig. 6.8.

A constructed PPI data is available on the String database. The PSI-XML schema is a recent initiative by the Protein Standardization Initiative to formally standardize the way biologists should report molecular interaction data.

The code chunk below demonstrates how to download and parse the CVD and CKD merge PPI data from the String database using igraph package.

```
install.packages("igraph")
install.packages("influential")
library(influential)
library(igraph)
Interaction.Data<-read.delim("NETWORK STRING.txt" .sep ="\t")
```

```
MsData<- CVD.data
library(influential)
CVD.Data<-read.delim("NETWORK.txt" .sep ="\t")
MMData<- CVD.Data
Mss_graph<- graph_from_data_frame(MMData)
# Extracting the vertices
GraphVertices<- V(Mss_graph)
# Calculation of Spreading score
Spreading.score<- spreading.score(graph = Mss_graph,
                                  vertices = GraphVertices,
                                  weights = 0, directed = F, mode
= "all",
                                  loops = TRUE, d = 5, scaled = T)

# Calculation of Hubness score
Hubness.score<- hubness.score(graph = Mss_graph,
                                  vertices = GraphVertices,
                                  directed = 0, mode = "all",
                                  loops = TRUE, scaled = 1)
head(argument)
```

Upon reading the protein–protein interaction data into R environment, one intuitive approach to study it is to plot and query the statistics of the result String network. In a biological network, it is often important to know the number of direct interactions (degree) that possess a component. In graph theory, this corresponds to computing the degree (or valency)of a node of a graph. The degree of a graph is the number of edges incident to the node, with loops counted twice. Top ten genes extracted from the network based on degree see Table 6.1. Code for a degree such as below:

```
Mss_graph<- graph(MsData)
# Extracting the vertices
```

Table 6.1 Top ten genes based on degree are illustrated

Genes	Degree
“ABCE1”	159
“ACTB”	118
“ACTG1”	71
“ACTR5”	14
“AHSA2”	4
“AKT1”	174
“AKT3”	48
“ANK1”	2
“ANXA3”	4
“ABCE1”	159

```
GraphVertices<- V(Mss_graph)
# Calculating degree centrality
networkdegree<- degree(Mss_graph, v = GraphVertices, normal-
ized = 0)
head(networkdegree).
```

6.11 Future Direction

R programming has a bright future and is now popular since it is a straightforward language for those who are new to programming. The most common language used by statisticians and data scientists is R. The number of R users is thought to be around 2 million. It is regarded as a game-changer because R programming has proven to be the greatest tool for data analysis. Aside from the information technology sector, many other sectors use R programmers to make use of their data and solve their problems. Consider that the mumps vaccination took 5 years to develop, making it the second-fastest vaccine. The COVID-19 vaccination process took less than a year. AI, big data analytics, and bioinformatics all played a role in making it possible.

References

Ahmed MM, Tazyeen S, Haque S, Sulimani A, Ali R, Sajad M, Alam A, Ali S, Bagabir HA, Bagabir RA et al (2022) Network-based approach and IVI methodologies, a combined data investigation identified probable key genes in cardiovascular disease and chronic kidney disease. Front Cardiovasc Med 8:755321. https://doi.org/10.3389/fcvm.2021.755321

Giorgi FM, Ceraolo C, Mercatelli D (2022) The R language: an engine for bioinformatics and data science. Life 12:648. https://doi.org/10.3390/life12050648

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299. https://doi.org/10.2307/1390807

Liu C, Ma Y, Zhao J, Nussinov R, Zhang Y-C, Cheng F, Zhang Z-K (2020) Computational network biology: data, models, and applications. Phys Rep 846:1–66. https://doi.org/10.1016/j.physrep.2019.12.004

Lovelace R, Nowosad J, Muenchow J (2019) Geocomputation with R, 1st edn. Chapman and Hall/CRC. ISBN 978-0-203-73005-8

Madsen CD, Hein J, Workman CT (2022) Systematic inference of indirect transcriptional regulation by protein kinases and phosphatases. PLoS Comput Biol 18:e1009414. https://doi.org/10.1371/journal.pcbi.1009414

Memisevic V, Milenkovic T, Przulj N (2010) An integrative approach to modeling biological networks. J Integr Bioinform 7(3):120. https://doi.org/10.2390/BIECOLL-JIB-2010-120

Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21:i302–i310. https://doi.org/10.1093/bioinformatics/bti1054

Oleskin AV (2014) Network structures in biological systems. Biol Bull Rev 4:47–70. https://doi.org/10.1134/S2079086414010034

Saint-Antoine MM, Singh A (2020) Network inference in systems biology: recent developments, challenges, and applications. Curr Opin Biotechnol 63:89–98. https://doi.org/10.1016/j.copbio.2019.12.002

Vitkup D (2004) Biological networks: from physical principles to biological insights. Genome Biol 5:313. https://doi.org/10.1186/gb-2004-5-3-313

Wang YXR, Li L, Li JJ, Huang H (2021) Network modeling in biology: statistical methods for gene and brain networks. Stat Sci 36:89–108. https://doi.org/10.1214/20-STS792

Machine Learning in Biological Networks

7

Shahnawaz Ali

Abstract

In the past few years, AI has been a subject of discussion between general and also specific audiences. The countless articles published discussing the advancement and promises of AI can provide in medicine have been increasing. On the other hand, the concept and application of quantitative matrices for the networks in biology started by Neumann in 2010 has now grown as a respective field of its own, i.e., Network Biology. It is time to join these two fields of study and apply the rules with some modifications to biological networks. As a result of these integrations, the current concepts of precision medicine can be boosted from its core and benefited, like Network Medicine. So, let us tackle some of the pressing questions like what has been achieved, where are we heading? This chapter aims to elucidate the fundamental principles of machine learning and deep learning, particularly in the context of the medical field.

Keywords

Machine learning · Deep learning · Biological networks · Supervised · Unsupervised · Precision medicine

Abbreviations

Acc-motif	Accelerated motif
AI	Artificial intelligence
BIND	Biomolecular interaction network database
BioGRID	Biological general repository for interaction datasets

S. Ali (✉)
CGTRM, Kings College London, London, UK

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Ishrat (ed.), *Biological Networks in Human Health and Disease*,
https://doi.org/10.1007/978-981-99-4242-8_7

BN	Biological network
CNN	Convolutional neural network
CRAN	Comprehensive R archive network
DEDS	Discrete event dynamic systems
DIAMOnD	DIseAse module detection
DIP	Database of interacting proteins
DL	Deep learning
DNA	Deoxyribonucleic acid
DNN	Deep neural network
DQ	Data quality
DSI	Domain-specific interactome
GCN	Graph convolutional network
GNN	Graph neural networks
GRNs	Gene regulatory networks
GUI	Graphic user interface
HPRD	Human protein reference database
KEGG	Kyoto Encyclopedia of Genes and Genomes
LEV	Leading eigenvector
MIPS	Munich Information Center for Protein Sequences
ML	Machine learning
MSigDB	Molecular Signatures Database
nBN	Non-biological network
NCBI	National Center for Biotechnology Information
NetMODE	Network motif detection
NM	Network medicine
PM	Precision medicine
PPIN	Protein–protein interaction network
PPIs	Protein–protein interactions
RNA	Ribonucleic acid
RNN	Recurrent neural network
SSc	Systemic sclerosis
TFs	Transcription factors
WGCNA	Weighted gene co-expression network analysis

7.1 Machine Learning: Supervised, Unsupervised, and Deep Learning

Machine Learning (ML) as we know is a broader field of statistics and computer science. Think of ML as a compilation of sophisticated data analysis methods geared towards constructing models that have the ability to forecast probable results derived from complex and diverse data. In simple words, a classical ML model learns a 2D representation of that n-dimensional data that helps it to predict the outcome of a new dataset. Learning 2D representations as done in classical ML techniques or probabilistic modeling can only be done for problems that are well

structured like binary classification. For many real-world problems like text mining, recommendation system, and translation. We have to go deep (learn meaningful features successively), i.e., Deep Learning. As these are perceptual tasks involving skills that are intuitive to humans and has not been addressed with such a precision by classical ML. Some of these DL models have been used by social sites (aka on social networks) to build recommendation systems that are purely based on analyzing Real-World Networks (or Non-biological networks).

Over the last two decades, there has been an increase in the large datasets quantifying molecules, proteins, and their interactions within living system. Projects like Human proteome atlas, Human cell atlas, and many others of similar scope has generated petabytes of data. This volume of data and high dimensionality gives us the perfect case to apply ML techniques, as these techniques thrive on these two properties. Raw data presented in such high volume makes it imperative to apply techniques tailored to handle large data, i.e., ML. Application of ML is becoming more permeating and applied not only to genomic annotation but identification of Transcription factors, etc.

As discussed in Box 7.1, we have different ML techniques at hand that can be used for solving and understanding biological problems. Since, most of the biology questions either have constrained information or debatable data with high dimensionality, making it a perfect case of unsupervised learning. On the hand, with the advent of High throughput techniques, the volume of data generated has also increased exponentially. So, to take care of "*High dimensionality"* and "*volume"* of the data, the two most important requirement for applying ML has been fulfilling, and enough reason we can use the power provided by the advancement of ML techniques (i.e., DL) to get a perspective on any biological processes (genome, PPIs, GRNs, etc.) (Yip et al. 2013).

> "The core of ML technology depends on understanding the patterns from the raw data, while learning is based on its mathematical and statistical framework of rules and assumptions. It can simply be described in a conceptual way as in Fig. 7.1."

As it is clear from Fig. 7.1 (adapted from the book *Deep Learning in Python* by François Chollet), the shift in the overall perspective of solving problems is what makes ML or in general Artificial Intelligence (AI) interesting.

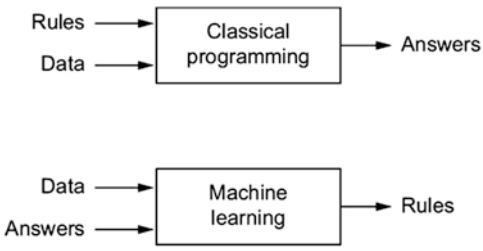

Fig. 7.1 The paradigm shifts from classical programming to Machine learning (Chollet 2021)

Definition "AI is an effort to automate tasks normally performed by humans. While ML is a subset of AI that generates trained model over tons of examples and find statistical structure that in the end allows to come up with rules for automating a particular task. On other hand Deep learning is a subfield of ML that means learning representations from data through successive layers of increasingly meaningful representations."

Box 7.1 Important

Classification of Machine Learning Techniques.

Machine learning implementations are classified into three major categories, depending on the nature of the learning "Input" or "Output."

1. *Categories based on Input*:
   (a) *Supervised:* When an algorithm learns from example data and associated target responses such as classes or tags, in order to later predict the correct response when posed with new examples. This approach is indeed similar to human learning under the supervision of a teacher. In biological context gene expression profiles to disease group, etc.
   (b) *Unsupervised*: When an algorithm learns from plain examples without any associated response, by determining the data patterns on its own. The algorithm in question possesses an ability of transforming data, almost mirroring a human's cognitive process of identifying objects by diligently assessing their degree of likeness.
   (c) *Semi-supervised learning:* Imagine a scenario where you are provided with an imperfect training input. This input consists of a training set wherein certain target outputs are mysteriously absent. Let's consider the specific example of protein structure classification. In this context, we are faced with the task of categorizing protein structures, but we are lacking crucial information regarding some of the target outputs.
2. *Categories based on Output:*
   (a) *Classification*: When inputs are divided into two or more classes, the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. An example of this type is Spam filtering that depends on the user to tag emails as spam or not spam thus supervised.
   (b) *Regression*: When the target output is continuous, not discrete, thus this type also comes under supervised learning.
   (c) *Clustering*: Being similar to classification and mostly misunderstood, it is the type where input can be divided into groups. Thus, making it a typical Unsupervised Learning. Sometimes we provide the expected no of clusters (as in k nearest neighbor) and other times we use other Dimensionality reduction methods to get to that number.

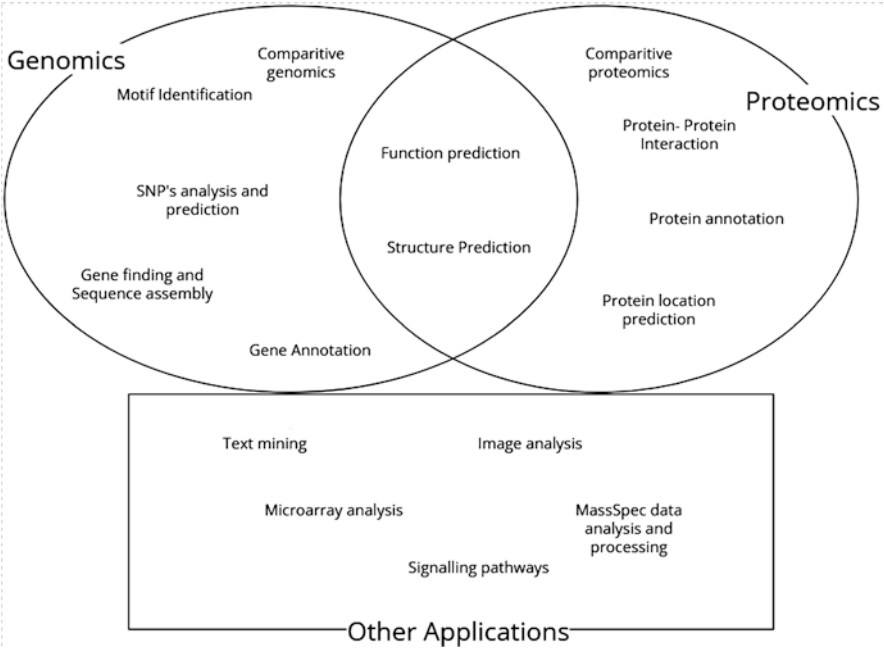

Fig. 7.2 Infographics for ML in bioinformatics

Now that we understand different types of ML techniques and a little in the context of Biology as discussed by Tarca et al. (2007) and how they are useful, we can now quickly talk about an interesting sub-field of ML, i.e., Deep Learning. As DL recently has given us a powerful tool like AlphaFold that can predict the structure of any protein with the outmost efficiency (Jumper et al. 2021).

7.2 Machine Learning Algorithms in Bioinformatics

The generation of large amounts of data has posed a challenge for computational biologist that is to efficiently extract useful information. The tools are required to analyze this heterogeneous data and give an insight in the form of testable models. Since, Biological data can fall under several domains to which ML techniques can be applied some of which are genomics, proteomics, and systemics (system biology and Network biology), see infographics Fig. 7.2 (adapted from Larranaga et al. 2006).

As evident from Fig. 7.2, the different domains where ML can be applied, now it is time to discuss some of the algorithms in this context.

7.2.1 Supervised Classification for Bioinformatics

In the supervised classification algorithm, we have a feature set (represented by $X \in \mathcal{R}^n$) and class vector ($C = f\left(X\right) \in \mathcal{R}$). The classification can be binary (for

example, predicting the splice site or sample type) or multiple (predicting the age group of patients, cell types). The main aim of supervised classification methods is to associate a new data point with the appropriate label based on the rules thus learned as a result of training.

The most important step of supervised classification is selection of Feature Subset (FSS). As FSS comes under searching the most appropriate subset of features thus search algorithms, to our disposal two methods can be used exhaustive search and heuristic search. Exhaustive search evaluates all possible subsets which can be impractical when we have large sample space, while heuristic searches involving deterministic and stochastic FSS algorithms have been proposed (Kuncheva 1993and Inza et al. 2000). The benefit of FSS in supervised classification is it reduces the dimension thus the cost of training and prediction. There are three important approaches proposed for FSS namely filter, wrapper, and hybrid see (Kohavi and John (1997), Inza et al. (2004), and Xing et al. (2001) for more.

Now, it is time to discuss some of the classifiers that can be trained using features selected using the FSS method. The application of a classifier depends on the type of problem that it can solve means no classifier is universal, see Table 7.1 for more information on the supervised classifiers models/algo (Table 7.2).

Table 7.1 Supervised classification algorithms

Classifier	Type	Description
Bayesian	Naive Baye, tree augmented naive Bayes, semi-naive Bayes, and k-dependence Bayesian (kDB)	Minimizes the total misclassification cost. Based on how the predictive variables given the class means $p(x \in X \mid c)$ is approximated. Different classifiers can be obtained
Logistic regression	–	It is represented by the $p(C=1\|x)=1/\left[1+e^{-\left(\beta_0+\sum_{i=1}^{n}\beta_i x_i\right)}\right]$, where x is an instance to be classified. The model preparation consists of *Wald* test with a likelihood test
Support vector machines	–	It maps inputs implicitly to a higher dimensional space so that they can be separated by hyperplane (Ivanciuc 2007)
Discriminative analysis (DA)	Fisher linear (FDA) and Linear (LDA)	FDA is where the ratio between—group to within—group sums of squares are minimized. While LDA constructs a hyperplane between two groups, such that LDA function is zero at hyperplane
Neural Nets	ANN	It is a set of multilayered neurons using activation function and backpropagation method to identify weights on each node (i.e., parameters)
Classification trees	Directed tree	It organizes the predictive variable in the form of a tree with branches split into mutually distinct and exhaustive link

Table 7.2 Domain-specific application of supervised classification algorithms

Domain (as in Fig. 7.1)	Implementation studies (some)
Genomics	Gene prediction methods (Le 2020), Protein-coding gene identification (Hoff et al. 2016, Numa and Itoh 2014 and Campbell et al. 2014)
Proteomics	Prediction of protein–protein interactions (Rodgers-Melnick et al. 2013; Zhu et al. 2016; Liu et al. 2017; Szklarczyk et al. 2019 and Ding and Kihara 2019)
Other applications	Gene Ontology category (Bradford et al. 2010, Kulmanov and Hoehndorf 2020, You et al. 2018, Fa et al. 2018 and You et al. 2019)

Table 7.3 *Clustering methods in Bioinformatics* (Kononenko and Kukar 2007)

Algorithms	Description
k-means, *k* medoid, and M-mean	Divides data into *k* clusters such that within group distance is minimized
Generalized Lloyd	It is a variant of k-means where vector is quantized using coder and encoder. Then these coded vectors are organized topologically using self-organized maps so that it captures the structure in low dimensions like 1 or 2D (Linde et al. 1980)
Agglomerative, divisive	Set of hierarchical: Agglomerative begins with N groups with one entity for N groups and then progressively merges until all points are contained in one group. While divisive starts with one group and successively divide it until N groups

7.2.2 Clustering Algorithms in Bioinformatics

Clustering also called segmentation are set of algorithms that deal with unlabeled data thus an Unsupervised way to naturally group data points into one or more clusters. These clustering algorithms can be divided into Partition based or Hierarchical based approaches. When diving into the world of Clustering, there are several pivotal steps to consider. First and foremost, it is crucial to weed out any outliers or noise from the dataset, ensuring that we are working with clean and relevant data. Next, selecting the right distance measure is of utmost importance. This could involve using methods such as Euclidean or Manhattan to accurately quantify the dissimilarity between data points. After that, we must turn our attention to selecting the proper criterion for our clustering process. This entails optimizing the cost function or establishing a set of rules that align with our objectives. Once the criterion has been determined, we can proceed to choose the most suitable algorithms for our analysis. And lastly, it is imperative to validate the outcomes of our clustering efforts, ensuring their reliability and accuracy. By meticulously following these steps, we can maximize the effectiveness of our clustering (Table 7.3).

7.3 Integrating Machine Learning in Biological Networks

All interactions in a living organism consisting of either of the four molecular entities of DNA, RNA, Protein, and small molecules (ions, etc.) can be represented as respective as well as mixed multilayer (aka complex) networks called Biological Networks. These networks can be divided broadly into Signaling and Non-signaling. So, in order to understand the properties of interaction between these molecular entities, we need methods that can be applied to networks of these entities like GNNs. There are subsets of GNNs like GCNs that have adapted CNN, which is highly successful spatial based method. Table 7.4 describes a list of methods and their implementations related to BN domain (Muzio et al. 2021).

Thus, using these BNs, we can predict and classify various features that can be subdivided in terms of task involved like Node classification, Prediction of links, Graph Embedding, classification, and regression (Bhagat et al. 2011; Lü and Zhou 2011; Tsuda and Saigo 2010). When it comes to protein networks and the prediction of unknown protein functions, algorithms that classify nodes (i.e proteins) are particularly valuable. This allows us to effectively categorize and analyze the various components within these networks. While link prediction algorithms are useful when we want to know the regulation between genes in a gene regulation networks; thus predicting the edges of network. Graph embedding set of methods are useful when the goal is to find a lower dimension representation of a protein in a PPI. Graph embedding is used before applying any DL algorithm. As summarized in Table 7.4, these techniques and methods fulfil the need to extract useful information from the large amount of data given that can be represented in the form of graphs/networks.

Table 7.4 Some important methods related to BN and their source code

Method (some)	Implementations
Defferrard et al. (2016)	https://github.com/mdeff/cnn_graph
Duvenaud et al. (2015)	http://github.com/HIPS/neural-fingerprint
Grover and Leskovec (2016)	https://github.com/aditya-grover/node2vec
Hamilton et al. (2017)	https://github.com/williamleif/GraphSAGE
Kipf and Welling (2017)	https://github.com/tkipf/gcn
Perozzi et al. (2014)	https://github.com/phanein/deepwalk
Tang et al. (2015)	https://github.com/tangjianpku/LINE
Baranwal et al. (2020)	https://github.com/baranwa2/MetabolicPathwayPrediction
Senior et al. (2020)	https://github.com/deepmind/deepmind-research/tree/master/alphafold_casp13
Yue et al. (2020)	https://github.com/xiangyue9607/BioNEV
Zeng et al. (2019)	https://github.com/CSUBioGroup/DeepEP

7.4 Machine Learning in Disease Dynamics

Understanding disease and its dynamics are two different yet interconnected research topics. One is being defining the molecular basis and later being its progression in terms of contagion rate and other biological and non-biological factors. Since ML techniques are data hungry they can be applied in several ways to fulfill different goals let us explore some of these applications and discuss some of the state-of-the-art algorithms that are being used for this purpose.

One of the main applications is predicting the spread of infectious diseases like in case of viral infections Influenza, COVID-19, and Ebola. This has been useful for public health agencies as they plan their response to outbreaks (Golumbeanu et al. 2022; Venkatramanan et al. 2021).

Another important application is that it can be used to identify the risk factors for the disease. This involves analyzing large datasets to identify patterns and trends that may be associated with an increased risk of developing a particular disease. Factors like diet, lifestyle, or genetic predisposition may increase a person's risk of developing a particular disease. This information can be used to develop targeted prevention strategies or to identify individuals at substantial risk who may benefit from early intervention (Salathé and Khandelwal 2011).

In addition to predicting spread and identifying risk factors, ML can also be used to enhance the diagnosis with increased accuracy and develop personalized treatments for individual patients. Finally, machine learning can also be used to improve public health interferences. ML algorithms can be used to identify the most effective strategies for preventing the spread of diseases or improving the health of populations. Thus, researchers may use ML to analyze data on the effectiveness of different public health interventions, like vaccination campaigns or health education programs, to identify the most effective strategies for improving the health in general (Palaniappan and David 2022; Francisco et al. 2021).

There are different types of machine learning algorithms that can be used to study disease dynamics, including supervised learning algorithms, unsupervised learning algorithms, reinforcement learning algorithms, and deep learning algorithms. Supervised learning algorithms are trained on labeled data and can be used to predict the likelihood of a particular outcome, such as the likelihood that a patient will develop a particular disease. Unsupervised learning algorithms are not given labeled data and must instead identify patterns and relationships in the data on their own and can be used to identify clusters or groups within a dataset useful for identifying risk factors. Reinforcement learning algorithms learn by interacting with their environment and associated rewards or punishments based on their actions. These can be used to identify the most effective strategies for preventing the spread of diseases or improving the health of populations. Last yet more effective are Deep learning algorithms are a type of neural network that can process substantial amounts of data (say big data) and identify complex patterns and relationships. They are particularly suited for analyzing medical images and other high-dimensional data and have been used to improve the accuracy of diagnoses and identify risk factors for various diseases.

7.5 Superlative Features of Machine Learning over Probabilistic Models

Machine learning algorithms and probabilistic models are both used to make predictions or decisions based on data. However, there are several key differences between the two approaches, and machine learning algorithms have several advantages over probabilistic models in certain situations (Jaynes 2003; Robert 2014).

One key advantage of machine learning algorithms is their ability to learn from large amounts of data. Machine learning algorithms can analyze vast amounts of data and identify patterns and trends that may not be apparent to human analysts. This allows machine learning algorithms to make more accurate predictions or decisions than probabilistic models, which may not be able to capture all of the relevant patterns and trends in the data.

Another advantage of machine learning algorithms is their flexibility. Machine learning algorithms can be applied to a wide range of tasks and can be fine-tuned to specific problem domains. In contrast, probabilistic models are typically designed for specific types of problems and may not be as flexible or adaptable.

Another advantage of machine learning algorithms is their ability to improve over time. Machine learning algorithms can continue to learn and improve as they are exposed to more data, allowing them to make increasingly accurate predictions or decisions. Probabilistic models, on the other hand, are typically fixed and do not improve as they are exposed to more data.

Finally, machine learning algorithms are generally more efficient than probabilistic models at making predictions or decisions. Machine learning algorithms can process large amounts of data quickly and make predictions or decisions in real time, while probabilistic models may require more time to make predictions or decisions.

Overall, machine learning algorithms offer a number of advantages over probabilistic models, including their ability to learn from large amounts of data, their flexibility and adaptability, their ability to improve over time, and their efficiency at making predictions or decisions. As a result, machine learning algorithms are becoming increasingly popular for a wide range of applications, including healthcare, finance, marketing, and many others.

Despite the many advantages of machine learning algorithms, probabilistic models still have their place in certain situations. For example, probabilistic models may be more suitable in cases where the underlying relationships between variables are well-understood and can be accurately modeled. Probabilistic models may also be more suitable in cases where the goal is to understand the underlying causal relationships between variables, rather than just making predictions or decisions based on the data.

In general, machine learning algorithms are more suitable for tasks where the goal is to make accurate predictions or decisions based on large amounts of data, while probabilistic models are more suitable for tasks where the goal is to understand the underlying causal relationships between variables. Choosing the

appropriate approach will depend on the specific goals of the project and the nature of the data and problem domain.

Thus, machine learning algorithms offer several advantages over probabilistic models, including their ability to learn from large amounts of data, their flexibility and adaptability, their ability to improve over time, and their efficiency at making predictions or decisions. However, probabilistic models still have their place in certain situations, and the appropriate approach will depend on the specific goals of the project and the nature of the data and problem domain (Heath et al. 2008).

7.6 Machine Learning-Based Big Data Processing in Cancer

In an era of big data, which in biology consists of domains like omics, imaging, and signal processing. The main challenge is to get important information out of large amount of data. With the advent and increased popularity of Deep Learning from early 2000, it has been discussed and emphasized in both academic and industrial settings. The main advantages of this branch of ML are its ability to use power of parallel computing, being data driven rather than hand-designed features and learning the representation of data in lower dimensions. The key elements of DL are its network architecture (for example, RNN, CNN, and DNN) and model training approaches (say optimization, choice of cost function, etc.). As seen in Table 7.5, some of the categories of DL are applied in a particular domain of biology.

On the other hand, cancer is a complex and heterogeneous disease, and large amounts of data are generated from a variety of sources like biobanks, including genomic, clinical, and imaging data (Huppertz and Holzinger 2014). ML algorithms can be used to analyze these large datasets and identify patterns and trends that may not be apparent to human analysts, providing insights that can inform the development of new diagnostic and treatment approaches. Thus, the advanced understanding of risk factors associated with cancer will reduce the burden on healthcare (Leatherdale and Lee 2019). Overall, the use of machine learning in the processing of big data in cancer has the potential to improve the accuracy and speed of diagnosis, treatment planning, and disease monitoring, and to help identify new strategies for preventing and treating cancer (Dash et al. 2019; Iqbal et al. 2021).

Table 7.5 *Categories of DL in bioinformatics* (Min et al. 2017)

Category	Biological domain	Implementations (some)
DNN	Omics	Protein structure, protein and anomaly classification, and gene expression regulation
	Imaging	Segmentation, recognition, and decoding of brain
	Signal processing	Anomaly classification and decoding of brain
CNN	Omics	All of the above
RNN	Omics	All of the above

7.7 Future Prospects of AI in the Field of Medicine

AI, in general, is capable of performing tasks that normally require human intelligence, is slowly permeating modern day-to-day life. Healthcare has lagged behind in adopting AI, but the pace of implementation is picking up. Computer-based decision support systems based on machine learning can take over the complex tasks currently assigned to experts. This has the potential to revolutionize healthcare by increasing diagnostic accuracy, improving clinical workflow, reducing labor costs, and improving treatment modalities. The growing interest in AI and machine learning in various industries, including healthcare, is largely due to the rise of deep learning. This is the process by which AI uses various forms of neural networks similar to the human brain to recognize patterns, and the process is based on: Availability based on big data repositories. The promise of AI in healthcare is to improve the quality and safety of care and democratize expertise using mobile devices such as smartphones. Mobile devices can be deployed with algorithms, potentially universally and inexpensively accessed in a low-cost, essential way anywhere in the world. Diagnosis provides care. AI is ripe for AI because healthcare has large datasets (big data). This is ideal for AI as it requires large datasets for computers to learn. As such, AI is rapidly becoming a key component of the healthcare environment. AI algorithms will play a key role in predicting cancer and helping cancer patients make treatment decisions in the near future. Already digitized fields like radiology are already undergoing an AI revolution. Deep neural networks will be able to provide a synergistic combination of disciplines such as radiology, nuclear medicine, and surgical pathology which will hopefully allow the achievement of a medical paradigm which recognizes that every human being is unique. Although pathology especially surgical pathology was late to adopt AI, mainly due to practical and financial obstacles, and will require resources for additional workflows, personnel, equipment, storage of data, the time is now ripe, with rapid development of new and better AI technology at lower cost (reduced costs of digital data and availability of digital images) for AI to succeed in surgical pathology. Various studies cited above demonstrate the increasingly effective role of AI in medicine. By increasing speed and accuracy of diagnosis and by improving prognostication, use of AI is translating into better patient care. AI will, in the near future, not replace humans but by performing routine repetitive tasks quickly and accurately, allowing us to give time to more complex cognitive tasks. Therefore, we need to adopt and train AI to take its advantage. This integration takes a lot of time as AI methods need to be integrated into training programs and we (professionals, researchers, etc.) need to be familiar with these data using computer algorithms in our daily work. Synergistic collaboration between disciplines such as will play a major role. Financial hurdles need to be overcome, especially for poorer developing countries, so they can benefit from improved applications of AI in medicine.

References

Baranwal M, Magner A, Elvati P, Saldinger J, Violi A, Hero AO (2020) A deep learning architecture for metabolic pathway prediction. Bioinformatics 36(8):2547–2553

Bhagat S, Cormode G, Muthukrishnan S (2011) Node classification in social networks. In: Aggarwal CC (ed) Social network data analytics. Springer, New York, pp 115–148

Bradford JR, Needham CJ, Tedder P, Care MA, Bulpitt AJ, Westhead DR (2010) GO-at: in silico prediction of gene function in Arabidopsis thaliana by combining heterogeneous data. Plant J 61:713–721

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J et al (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol 164:513–524

Chollet F (2021) Deep learning with python. Simon and Schuster

Dash S, Shakyawar SK, Sharma M, Kaushik S (2019) Big data in healthcare: management, analysis and future prospects. Journal of Big Data 6(1):1–25

Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of the 29th International conference on neural information processing systems, pp 3844–3852

Ding Z, Kihara D (2019) Computational identification of protein-protein interactions in model plant proteomes. Sci Rep 9:1–13

Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. Adv Neural Inf Proces Syst 28:2224–2232

Fa R, Cozzetto D, Wan C, Jones DT (2018) Predicting human protein function with multi-task deep neural networks. PLoS One 13:e0198216

Francisco ME, Carvajal TM, Ryo M, Nukazawa K, Amalin DM, Watanabe K (2021) Dengue disease dynamics are modulated by the combined influences of precipitation and landscape: a machine learning approach. Sci Total Environ 792:148406

Golumbeanu M, Yang G, Camponovo F, Stuckey EM, Hamon N, Mondy M et al (2022) Leveraging mathematical models of disease dynamics and machine learning to improve development of novel malaria interventions. MedRxiv 2021(11):61

Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining, New York: Association for Computing Machinery, 2016, pp. 855–64

Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: proceedings of the 30th International conference on neural information processing systems, 2017, pp. 1024–34

Heath J, Kwiatkowska M, Norman G, Parker D, Tymchyshyn O (2008) Probabilistic model checking of complex biological pathways. Theor Comput Sci 391(3):239–257

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769

Huppertz B, Holzinger A (2014) Biobanks–a source of large biological data sets: open problems and future challenges. In: Interactive knowledge discovery and data mining in biomedical informatics. Springer, Berlin, Heidelberg, pp 317–330

Inza I, Larranaga P, Etxeberria R et al (2000) Feature subset selection by Bayesian network-based optimization. Artif Intell 123:157–184

Inza I, Larranaga P, Blanco R et al (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 31(2):91–103

Iqbal MJ, Javed Z, Sadia H, Qureshi IA, Irshad A, Ahmed R et al (2021) Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int 21(1):1–11

Ivanciuc O (2007) Applications of support vector Machines in Chemistry. Rev Comput Chem 23:291–400
Jaynes ET (2003) Probability theory: the logic of science. Cambridge university press
Jumper J, Evans R, Pritzel A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589
Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: proceedings from the 5th International conference on learning representations (ICLR), 2017
Kohavi R, John G (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
Kononenko I, Kukar M (2007) Chapter 12-cluster analysis. Machine Learning and Data Mining:321–358
Kulmanov M, Hoehndorf R (2020) DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36:422–429
Kuncheva L (1993) Genetic algorithms for feature selection for parallel classifiers. Inf Process Lett 46:163–168
Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I et al (2006) Machine learning in bioinformatics. Brief Bioinform 7(1):86–112
Le D-H (2020) Machine learning-based approaches for disease gene prediction. Brief Funct Genomics 19(5–6):350–363
Leatherdale ST, Lee J (2019) Artificial intelligence (AI) and cancer prevention: the potential application of AI in cancer control programming needs to be explored in population laboratories such as Compass. Cancer Causes Control 30(7):671–675
Linde Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Commun 28(1):84–95
Liu S, Liu Y, Zhao J, Cai S, Qian H, Zuo K, Zhao L, Zhang L (2017) A computational interactome for prioritizing genes associated with complex agronomic traits in rice (Oryza sativa). Plant J 90:177–188
Lü L, Zhou T (2011) Link prediction in complex networks: a survey. Physica A: Statistical Mechanics and its Applications 390(6):1150–1170
Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851–869
Muzio G, O'Bray L, Borgwardt K (2021) Biological network analysis with deep learning. Brief Bioinform 22(2):1515–1530
Numa H, Itoh T (2014) MEGANTE: a web-based system for integrated plant genome annotation. Plant Cell Physiol 55:e2. https://doi.org/10.1093/pcp/pct157
Palaniappan S, David B (2022) Prediction of epidemic disease dynamics on the infection risk using machine learning algorithms. SN computer science 3(1):1–3
Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: proceedings of the 20th ACM SIGKDD International conference on knowledge discovery and data mining, KDD'14, 2014, pp. 701–710
Robert C (2014) Machine learning: a probabilistic perspective. MIT Press, Cambridge
Rodgers-Melnick E, Culp M, DiFazio SP (2013) Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. BMC Genomics 14:608
Salathé M, Khandelwal S (2011) Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 7(10):e1002199
Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577(7792):706–710
Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M et al (2019) String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607–D613
Tang J, Qu M, Wang M, et al.. LINE: large-scale information network embedding. In: proceedings of the 24th International conference on World Wide Web, New York, United States: Association for Computing Machinery, 2015
Tarca AL, Carey VJ, Chen XW, Romero R, Drăghici S (2007) Machine learning and its applications to biology. PLoS Comput Biol 3(6):e116

Tsuda K, Saigo H (2010) Graph classification. In: Managing and Mining Graph Data. Springer, New York, pp 337–363

Venkatramanan S, Sadilek A, Fadikar A, Barrett CL, Biggerstaff M, Chen J et al (2021) Forecasting influenza activity using machine-learned mobility map. Nat Commun 12(1):1–12

Xing EP, Jordan MI, Karp RM. Feature selection for high-dimensional genomic microarray data. In: proceedings of the eighteenth International conference in machine learning. ICML, 2001: pp. 601–8

Yip KY, Cheng C, Gerstein M (2013) Machine learning and genome annotation: a match meant to be? Genome Biol 14(5):205

You R, Huang X, Zhu S (2018) DeepText2GO: improving large-scale protein function prediction with deep semantic text representation. Methods 145:82–90

You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S (2019) NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res 47:W379–W387

Yue X, Wang Z, Huang J, Parthasarathy S, Moosavinasab S, Huang Y et al (2020) Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics 36(4):1241–1251

Zeng M, Li M, Wu FX, Li Y, Pan Y (2019) DeepEP: a deep learning framework for identifying essential proteins. BMC bioinformatics 20(16):1–10

Zhu G, Wu A, Xu X-J, Xiao P-P, Lu L, Liu J, Cao Y et al (2016) PPIM: a protein-protein interaction database for maize. Plant Physiol 170:618–626