Chapter 12
Biological Networks: Tools, Methods, and Analysis

Basharat Ahmad Bhat, Garima Singh, Rinku Sharma, Mifftha Yaseen, and Nazir Ahmad Ganai

Contents

B. A. Bhat (✉)
Department of Life Sciences, School of Natural Sciences, Shiv Nadar University, Greater Noida, UP, India
e-mail: bb284@snu.edu.in

G. Singh · R. Sharma
Department of Life Sciences, Shiv Nadar University, Greater Noida, UP, India

M. Yaseen
School of Interdisciplinary Sciences and Technology, Jamia Hamdard University, New Delhi, New Delhi, India

N. A. Ganai
Animal Genetics and Breeding, Sher-e-Kashmir University of Agricultural Sciences and Technology – Kashmir, Srinagar, Jammu and Kashmir, India

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_12

12.1 Introduction to Biological Networks

The biology of organisms is complex and driven by the interplay of genes, proteins, small molecules, metabolites, and nucleic acids. To understand the biological system, it is important to interpret these interactions. As the genetic code suggests, DNA is transcribed to RNA, and then RNA is translated to proteins (Fig. 12.1), depending on the coding potential of mRNAs. The fundamental objective of systems biology is to comprehend the complete biological system by elucidating the behavior of all components and their interactions.

Over the years, the huge volume of data has been generated by various high-throughput techniques like next-generation sequencing, microarrays, and mass spectrometry to understand the molecular mechanism behind specific diseased state. These techniques provide the expression profiles of proteins and other genomic information for a biological system in one or the other format. However, interpretation of this complex and multidimensional data is a great challenge. In this chapter, we tried to elaborate on the data types from such high-throughput technologies, giving details about the methodologies and software to extract valuable and legible information from such complex data. Network analysis can be one of the promising approaches to address this issue and understand the biology behind the myriad of mechanisms and biological processes.

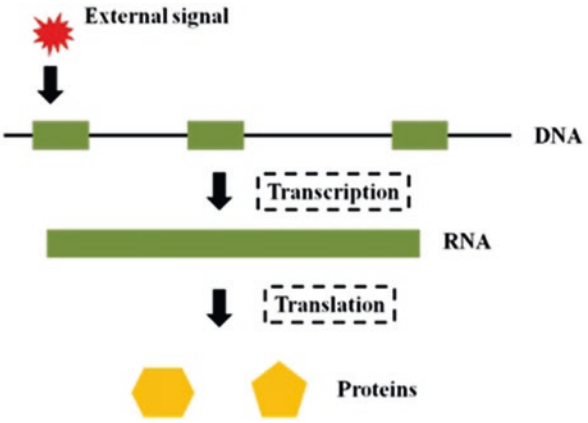


Fig. 12.1 The central dogma of biology. DNA is transcribed to RNA, and RNA is translated to proteins, which are the protagonist in biological systems

12.2 Types of Biological Networks

Biological networks are the mathematical representation of interactions between different types of molecules in a biological system. There are different types of biological networks as described below.

12.2.1 Protein-Protein Interaction Networks (PPIN)

The most important biochemical molecule in the organism is DNA, which stores the genetic information. The central dogma quotes that information from DNA is transferred to RNA and then from RNA to proteins (Fig. 12.1). However, the theory quoted by Beadle and Tatum (Beadle and Tatum 1941) about one gene-one enzyme-one function theory has come a long way. Now the biological processes are more complex, where proteins serve as the major molecule guiding a specific biological pathway. Proteins are long chains of amino acids, which are folded in a particular configuration. It is this specific configuration that enables a protein to physically interact with other proteins to form protein complexes and serve in downstream processes. Since proteins play a principal role in determining the molecular mechanisms and cellular responses, understanding the protein interaction networks is becoming a salient subject in research. Compiling the dense omics data from high-throughput techniques into meaningful biological networks is important to understand the cellular functions in a normal and diseased condition of the organism. This knowledge can further be translated into effective diagnostic strategies.

The reason behind the formation of protein complexes is still enigmatic. Proteins are folded in a specific configuration, which allows them to interact with other proteins via domains. Protein domains are the small conserved sequence of amino acids. These domains can function independently of the chain of protein and interact with other proteins to trigger biochemical processes like posttranslational modification, e.g., phosphorylation, glycosylation, etc. In one way, functional domains bind to other domains via protein interfaces to initiate a cellular response, e.g., interaction between Ras and its GTPase activating protein Ras-GAP, leading to a signaling process (Bader et al. 2008). Such type of interaction has high binding affinity and stability in lower volumes. In another way, domains bind to a stretch of amino acid sequence (3–10 in length) called motifs, present in the disordered region of a protein. For example, PDZ domain binds to C-terminus motifs of interacting proteins. The folds in the protein tertiary structure create active sites or catalytic domains, which interact with other proteins having similar conformations to initiate an enzymatic reaction (an induced-fit model). This model was proposed to be a lock-and-key model (Alberts et al. 2002), where the enzyme and substrate physically interact with each other to stimulate a biochemical reaction. Further, protein interactions in cell signaling pathways help in understanding cellular transports and interconnected modules in a biological process, e.g., p53 pathway.

12.2.1.1 Structure of Protein-Protein Interaction (PPI) Networks

PPI network is an organization of functional modules that comprises of a set of proteins having similar functions. The biological process can be interpreted as a modular network where proteins in a module are densely connected with each other sharing a similar function. Proteins are represented as "nodes" in the PPIN. Some proteins in the network have more interactions than other proteins, and these are called hubs. These nodes have very few interactions outside the module (Yook et al. 2004). PPIN are scale-free networks (Albert 2005). Hubs play a centralized role in scale-free networks and are classified as "party hubs" and "date hubs" (Han et al. 2004). Party hubs function inside the module and bind to interacting partners simultaneously, while date hubs bridge different modules and bind to interacting partners in different time and locations.

Network topology includes modularity and hub-oriented structure. There are four elements which define network topologies: (i) average degree (K) which can be calculated as degree distribution $P(k)$, (ii) clustering coefficient (C) calculated as degree distribution of cluster coefficients $C(k)$, (iii) average path length (L) calculated as shortest path distribution SP(i), and (iv) diameter (D) calculated as topological coefficient distribution TC(k). This concept is further explained in the chapter.

To understand the biochemical networks in a particular species, condition, or biological state, scientists are trying to merge the expression data from the myriad of experimental and computational techniques with the existing networks. For example, when expression data of each phase of yeast cell cycle was merged with PPIN in yeast cell cycle, most proteins were expressed continuously and found in the PPIN in each cell cycle, but there were some proteins which are expressed in a specific cell cycle phase and thus present in a PPIN of that phase (Batada et al. 2006). This is how computational algorithms are making the understanding of biological systems in different conditions (species, diseases, drug treatments) much easier than in earlier times. We can translate these results into therapeutic advancements in biomedical science.

12.2.2 Disease-Gene Interaction Networks

A disease is caused by the malfunctioning of any crucial biomolecule of an organism which can be a gene, protein, metabolite, or some unwanted genetic interactions, leading to the structural and functional aberration in the organisms. The genes, proteins, and other cellular components carry out their biological function in a complex network. With the advent of genomic sequencing and large-scale proteomics techniques, abundant genetic information is now available to build interactomes (biological networks). These biological networks help in understanding the pathophysiology of a specific disease and lead to a better understanding of the disease pharmacokinetics. Also, new disease-specific genes are identified which play an important role in disease prognosis.

12.2.2.1 Structure of Disease-Gene Interaction Networks

The important property of molecular networks is that they are dynamic. These networks change with space and time to adapt to different biological conditions. Hence this property of networks can be used to identify disease progression and also prognostic pathways specific for that disease.

Infection or disease progression occurs mainly due to molecular interactions. During host-pathogen interactions, host proteins interact with pathogen's proteins to initiate aberrated pathways. Such networks help researchers in understanding the mechanisms by which pathogens can attack the hosts. These networks are scale-free following the power law.

Recently, a research on human disease network (Goh et al. 2007) has given insight on how diseases are connected to each other through genes associated with them. The diseases are connected to their genes in which the associated mutation was found. This network is called "diseasome." One genetic mutation can be associated with several diseases. This resulted in a bipartite graph.

Diseases are also connected to each other if they have a common linked gene with a mutation, thus leading to human disease network (HDN). Genes are also connected to each other if they are found in the same disorder, thus resulting in disease-gene network (DGN) (Fig. 12.2).

12.2.3 Metabolic Networks

Metabolism is a complex association of metabolic reactions involving substrate, products, molecules, compounds, and cofactors. In general, metabolic reactions are reversible reactions, and they interact with each other, i.e., a product of one reaction can be the reactant of other reaction. The network of these metabolic reactions is called a metabolic network. An example of the metabolic network is the glycolysis process in humans.

12.2.3.1 Structure of Metabolic Networks

Metabolic pathways consist of enzymes, main substances, and co-substances. Main substances are metabolites, and co-substances are molecules like ATP, NADPH, etc., which help in transferring electrons. Metabolic networks have unique properties different from other networks because of (a) conservation constraints at each node and (b) the representation where nodes are metabolites and links are reactions catalyzed by specific gene products. This representation is very different from PPIN, where nodes are gene products and links are interactions. Also, a node in the metabolic network cannot be deleted by genetic techniques but links. A node in PPIN can be deleted using different molecular techniques, but it can result in a lethal phenotype. Metabolic networks have flux distribution with average path length longer, and their functional state does not have scale-free characteristics (Arita 2004).

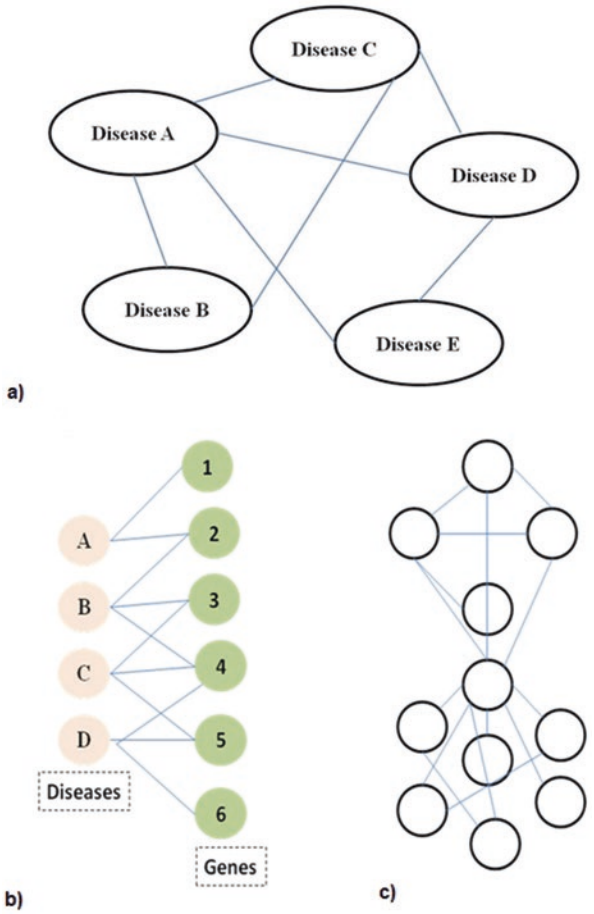


Fig. 12.2 (**a**) Human disease network (HDN): Different types of disease nodes are connected to each other if they share a common mutated gene. (**b**) The diseasome: The set of diseases are connected to the associated mutation in a gene. Genes are green in color while disease nodes are in orange color. (**c**) The disease gene network (DGN): The genes are connected to each other if they are associated with the same disorder

The metabolic networks can be of three types:

(a) *Simplified metabolic network:* A network of enzymes, reactions, and main substances but not co-substances (Fig. 12.3).
(b) *Simplified metabolite network:* A network of metabolites only. This kind of network is not always directed, and the metabolites are not directly connected to each other, but such type of interaction can be obtained from correlation analysis (Fig. 12.4).
(c) *Enzyme network:* A network of enzymes only. This kind of network can be obtained from PPIN (Fig. 12.5).

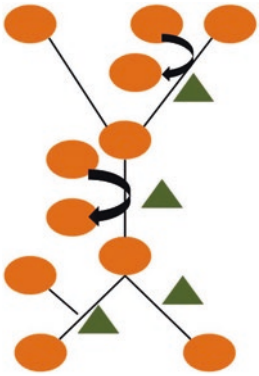


Fig. 12.3 Simplified metabolic network. The circles represent metabolites and the triangles are enzymes

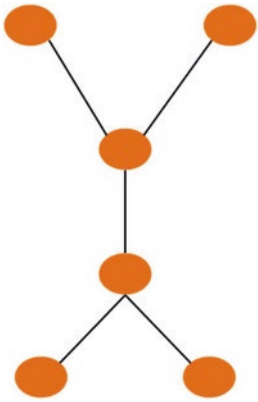


Fig. 12.4 Simplified metabolite network. The circles represent the metabolites

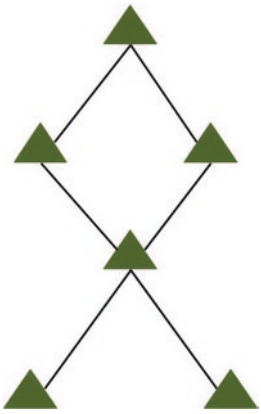


Fig. 12.5 Enzyme network. The triangles represent the enzymes

12.2.4 Gene Regulatory Networks

Gene regulation is the control of gene expression and thus the synthesis of proteins at transcription as well as translational level. The biological system is hardwired by the explicitly defined gene regulatory codes that control transcription as well as translation of the gene in a spatial and temporal manner. These control systems consist of transcription factors (TFs), signaling molecules, microRNAs, long noncoding RNAs, and epigenetic modulators. The molecules like TFs are cis-regulatory modules, which control the expression of the neighboring gene. Small RNAs like miRNAs control protein synthesis at the translation levels. Epigenetic modulators control the protein activity. Such kind of association of genes with its regulatory elements forms a gene regulatory network (GRN). GRNs include feedback, feed-forward, and cross-regulatory loops which define the regulation of gene at various levels.

12.2.4.1 Structure of Gene Regulatory Network

GRNs consist of many sub-circuits like signal transduction sub-circuit, metabolic reaction sub-circuit, and protein-protein interaction sub-circuits. Also, there can be a sub-circuit where TFs can regulate the expression of regulatory molecules like miRNAs. These sub-circuits connect to each other along with gene regulatory molecules to design a GRN.

GRNs are used to study the rationale behind the differential expressed genes in various diseased states and also help drug designing. An example of GRN is depicted in Fig. 12.6 where TFs are regulating the genes, which are in turn regulated by miRNAs.

12.2.5 Gene Co-expression Networks

A gene co-expression network is a kind of undirected graph where nodes (genes) are linked to each other on the basis of similarity in expression patterns (co-expression) under various experimental conditions. Gene co-expression network

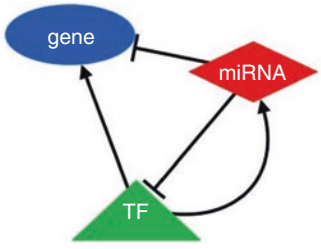


Fig. 12.6 Gene regulatory network

analysis helps in the simultaneous identification and grouping of genes with similar expression profiles. This analysis is of biological importance because co-expressed genes are regulated by the same transcription factors, functionally related or involved in same biological pathway(s). This kind of networks is built using expression data generated from high-throughput techniques such as microarray and RNA-Seq.

The co-expression network construction involves two steps:

1. Co-expression/expression relatedness measure calculation
2. Significant threshold selection

12.2.5.1 Co-expression Measure Calculation

The expression values of a gene for different samples are generally $\log_2$ transformed before co-expression measure calculation in order to scale the values to the same dynamic range. The following are four measures used for co-expression measure (Weirauch 2011) calculation between genes:

- *Pearson's correlation coefficient*: This measure is widely used for calculating expression similarity among genes for gene co-expression network construction. It gauges the inclination of two vectors to increment or abatement together, rendering a measure of their general relationship. Its value varies from -1 to 1 where absolute values near to 1 represent strong correlation. The positive values represent positive correlation, i.e., activation mechanism where a gene expression value is directly proportional to the expression value of other co-expressed gene and vice versa. When the relation between expression values of co-expressed genes is inverse, it represents the inhibition mechanism, and they will have negative correlation value. Assuming linear correlation, normally distributed values and being sensitive to outliers are some of the drawbacks of the Pearson correlation measure.
- *Mutual Information*: It describes nonlinear relations between genes, which measure the similarity between two genes based on their relations with other genes.
- *Spearman's rank correlation coefficient*: It is the nonparametric version of Pearson's correlation which is calculated for the ranks of gene expression values in a gene expression matrix.
- *Euclidean distance*: To calculate the geometric distance between gene pairs, both positive and negative expression values are considered. It is not suitable when the absolute expression values of related genes are highly varying.

12.2.5.2 Threshold Selection

After calculating co-expression measures between all pairs of genes, a cutoff is imposed for selecting the gene pairs that should be connected in the network. Several methods can be used for selecting a threshold for gene co-expression

network construction, for example, weighted gene co-expression network analysis (WGCNA) package which follows a power-law distribution approach for threshold selection.

12.2.5.3 WGCNA (Weighted Gene Co-expression Network Analysis)

It is a systems biology approach, which illustrates the correlation gene patterns across a series of microarray samples. It has been widely used in the genomic applications. It can be used to define modules of highly correlated genes, for summarizing such modules based on intra-modular hub genes and for calculating module membership for network nodes, i.e., genes, to study the relationships between co-expressed genes and external sample traits. It can also be used to compare the network topology of different networks. WGCNA (Langfelder and Horvath 2008) can be used as:

1. Data reduction technique
2. Clustering method
3. Feature selection method
4. Framework for integrating genomic data based on expression value.

The WGCNA pipeline is shown in Fig. 12.7.

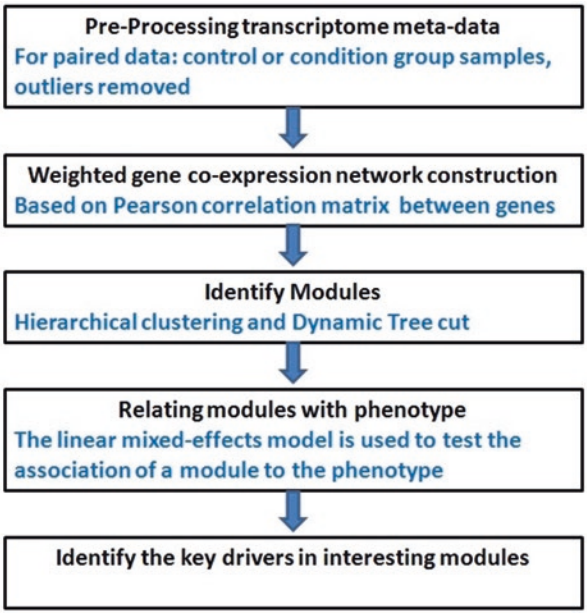


Fig. 12.7 WGCNA pipeline

12.3 Network Measures

A complex biological system can be considered as networks wherein components within a complex system can be represented as nodes and are connected through their interactions, also known as edges. It enables analysis of the network's topology, which gives insight into molecular mechanism operating within a cell under given condition. Network topology considers knowledge about the global and local properties of the network. Graph-theoretic network analysis can be used to measure the topological properties quantitatively (Ma'ayan 2011). Centrality indices are one of the measures which tell about the important nodes or edges, for the connectivity or the information flow within the network. The following are some of the centrality measures which can be calculated to define local properties of a network:

1. *Degree centrality*: It tells about the number of links for each node. The nodes with the highest degree may act as a hub, regulating multiple other nodes in the network.
2. *Node betweenness centrality*: It tells about the number of shortest paths between all possible pairs of nodes. The nodes with high betweenness centrality lie on communication paths and can control information flow.
3. *Closeness centrality*: It is the average shortest path from one node to all other nodes. It estimates how fast the flow of information would be through a given node to other nodes.
4. *Eigenvector centrality*: It accesses the closeness to highly connected nodes.
5. *Edge betweenness centrality*: It is the number of shortest paths that go through an edge among all possible shortest paths between all the pairs of nodes.

The following are some of the global properties of a network:

1. *Degree distribution:* It is the probability distribution of degrees over the whole network. For most of the biological networks, this distribution follows power-law, giving scale-free architecture to the network. It makes network stable and robust to random failures.
2. *Characteristic path length*: It represents the average shortest path between all pairs of nodes.
3. *Clustering coefficient*: It is the local density of interactions by measuring the connectivity of neighbors for each node averaged over the entire network. It demonstrates the tendency of the nodes to cluster together. High clustering coefficient means the presence of communities in a network. The communities are very important in the biological network as they represent functional modules or protein complexes working together to achieve a biological process.

12.4 Gene Ontology

The gene ontology is a cooperative attempt to bring together a consolidated description of gene and gene product for all organisms. It can be a promising approach to decipher key components from complex biological networks and helps in organizing the biological networks in a meaningful way to improve performance and biological interpretability.

Comparative genomics has apparently shown that a vast portion of the genes specifying the major biological functions are common to all organisms. Information of the biological role of such common proteins in one organism can often be exchanged with other organisms. The objective of the Gene Ontology Consortium is to deliver a dynamic, controlled vocabulary that can be connected to all organisms even as information of gene and protein roles in cells is gathering and evolving. The undertaking started in 1998 as a coordinated effort between three model organism databases, the FlyBase (Drosophila), the Saccharomyces Genome Database (SGD), and the Mouse Genome Database (MGD). The GO Consortium (GOC) has since developed to join numerous databases, including a few of the world's significant vaults for the plant, animals, and microbial genomes (Reference Genome Group of the Gene Ontology Consortium 2009).

There are three separate aspects to this effort:

1. The development and maintenance of the ontologies themselves
2. The annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases
3. The development of tools that facilitate the creation, maintenance and use of ontologies

The GO project has created three organized ontologies that associate gene products with their biological processes, cellular components, and molecular functions in a species-independent manner (Botstein et al. 2000).

- Cellular component: The location in the cell where a gene product is functional. In most of the situation, annotations connecting gene product with cellular component types are made on the basis of a direct observation of an instance of the cellular component in a microscope. Cellular component incorporates terms like "ribosome" or "proteasome," specifying where different gene products would be found.
- Molecular function: A molecular function term is an enduring potential of a gene product to act in a certain way or in other words the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition likewise applies to the ability that a gene product conveys as a potential. It portrays just what is done without indicating where or when the occasion really happens. For example, glucose transport, regulation of cell death, etc.

- Biological process: It defines what the gene or gene product contributes. A process is defined by means of at least one requested gathering of molecular functions for example, “cell growth and maintenance”, “signal transduction”, “cAMP biosynthesis”, etc.

Gene ontology (GO) has a graph-like structure where GO terms are nodes and relationships among them are links between nodes. The structure is loosely hierarchical having a parent-child relationship between nodes. Child node terms are more specialized than their parent node terms, but a child may have more than one parent term. For example, “integral component of external side of plasma membrane” is a child of the “integral component of plasma membrane” and “intrinsic component of external side of plasma membrane” (Fig. 12.8).

GO terms are designed with a unique identifier and term name, for example, GO:0015758~ glucose transport. The unique identifier is a zero-padded seven-digit identifier prefixed by “GO:”. The link between two nodes represents the relationship between them. For example, in Fig. 12.9, GO term “GO:1900117” has two types of relationship with parent nodes, i.e., “is a” and “regulates” which means GO:1900117 *is a* kind of regulation of apoptotic process (GO:0042981) and it *regulates* execution phase of apoptosis (GO:0097194).

The ontologies are dynamic, as in they exist as a network that is changed as more data gathers yet have adequate uniqueness and accuracy with the goal that databases in light of the ontologies can consequently be refreshed as the ontologies develop. The ontologies are adaptable in another way, so they can reflect the numerous distinctions in the biology of the assorted life forms, such as the breakdown of the nucleus during mitosis. The GO vocabulary is intended to be species-impartial and incorporates terms relevant to prokaryotes and eukaryotes and single and multicellular organisms.

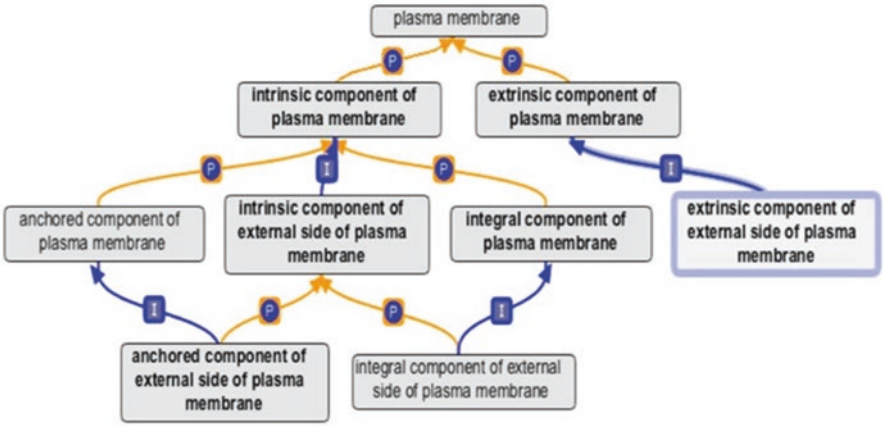


Fig. 12.8 Relationship between GO terms

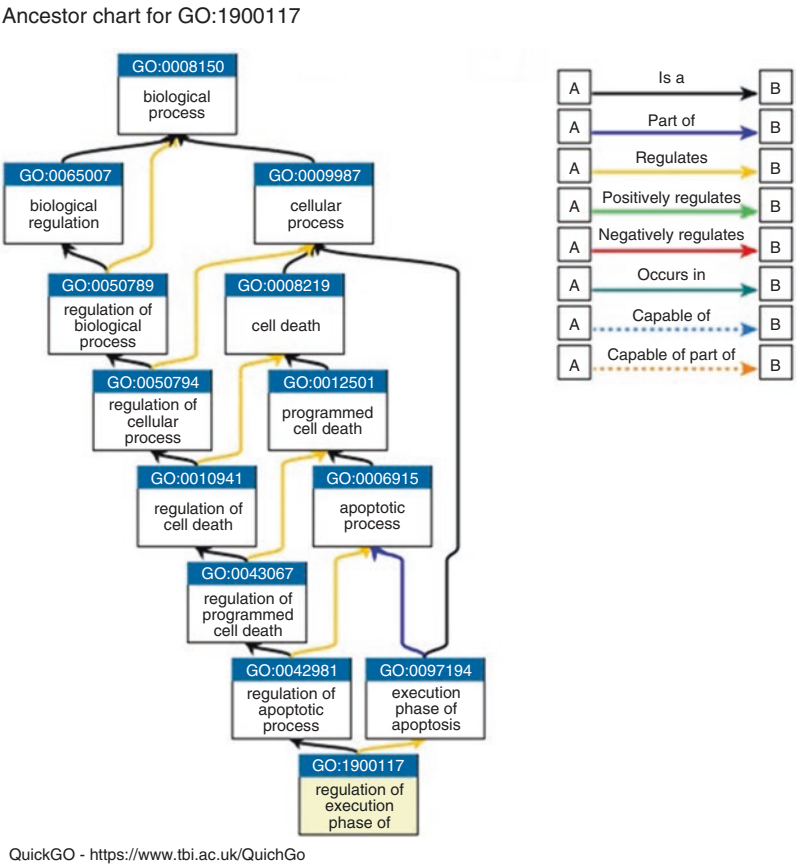


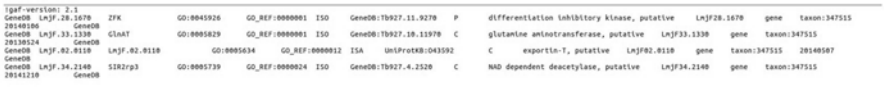
Fig. 12.9 GO ancestor chart

12.4.1 Applications of Gene Ontology

The gene ontology annotation is most widely utilized for deciphering large-scale "omics" data. Gene ontology enrichment analysis is one of the uses of GO annotation which helps in finding the significant clusters of genes associated with biological processes and thus reduce the bulk amount of data to the much smaller number of biological function getting altered under different experimental conditions.

12.5 GO Annotation

GO annotation is a link between the gene product and what that gene product can do, which molecular and biological processes it adds to, and where in the cell it is functioning. The GO annotation focuses on the identification of functional activities



S.No	Data	Description
1	DB (GeneDB)	Database providing the gene association list
2	DB_Object_ID (LmjF.28.1670)	a primary identifier in the database for the object being annotated
3	DB_Object_Symbol (ZFK)	Contain gene name, used to identify gene annotation.
4	Qualifier	Rarely used; changes interpretation of GO annotation
5	GO ID (GO:0045926)	The GO identifier for the term attributed to the DB_Object_Symbol
6	DB:Reference (GO_REF:0000001)	Literature evidence.
7	Evidence (ISO)	The evidence code is associated with a specific GO annotation term to describe what type of evidence was present in that reference to make the annotation.
8	With (GeneDB:Tb927.11.9270)	Identifier connecting evidence code with annotation
9	Aspect (P)	One of the three ontology classes F: molecular function P: biological process C: cellular component
10	DB_Object_Name	Gene name or gene product name
11	DB_Object_Name_Synonym	Gene symbol aliases
12	DB_Object_Type (gene)	The entity being annotated (gene, protein, exon etc).
13	Taxon_ID (taxon:347515)	Identifier for the species being annotated.
14	Date (YYYY:MM:DD)	Annotation date
15	Assigned_By	Database contributing to the annotation.

Fig. 12.10 Annotation format provided by the GO consortium

of a gene or a protein. GO annotation is principally divided into two parts: first, a link between the gene product and a representative GO term and second is an evidence to establish that link (Weirauch 2011). The annotation data in the GO database is contributed by members of the GO Consortium (GOC); more than 15 major contributing groups are actively working for GOC (Blake 2013). GOC is a dynamic ontology-based resource that contains the most updated and exhaustive set of annotations available in the literature. Keen utilization of GO annotation assures the best result in advancing biological research. GO annotation process follows a basic three-step paradigm in which:

1. Relevant experimental data is being identified from the biomedical literature.
2. Correlation of gene product with GO terms.
3. Finally, annotation quality control and refinement process are employed to ensure that the annotation has a correct formal structure.

GO annotation data file provided to GOC consists of 15 columns (Fig. 12.10). To fully comprehend the GO annotation file, a few important terms are worth to discuss:

An *annotation* is a process of assigning GO terms to the gene product. These assignments are made based on the conclusion drawn from experiments.

A *gene product* is an output generated from RNA or protein molecule that has some defined role in the biology of an organism.

A *molecular function* encompasses activities of a gene product such as catalytic or binding activities, influencing at the molecular level.

A *biological process* is a recognized sequence of molecular events performed by one or more ordered assemblies of molecular functions. For example, the progression of the brain development over time would be an instance of the biological function *brain development*.

A *cellular component* is a part of a cell where a gene product is active.

Curation is the formulation of annotation on the basis of the gene and gene product information from experimental observations.

An *evidence code* is a three-letter code that specifies the type of evidence identified from literature to support the association between gene and gene product. There are 21 (Hill et al. 2008) evidence (Table 12.1) codes classified broadly into five groups.

Table 12.1 Evidence codes classification

Category	Evidence codes
Experimental Evidence codes: literature cited indicates that there is evidence from an experiment directly supporting an association between gene and gene product	Inferred from Experiment (EXP) Inferred from Direct Assay (IDA) Inferred from Physical Interaction (IPI) Inferred from Mutant Phenotype (IMP) Inferred from Genetic Interaction (IGI) Inferred from Expression Pattern (IEP)
Computational Analysis evidence codes: literature cited contains observations from in silico analysis	Inferred from Sequence or structural Similarity (ISS) Inferred from Sequence Orthology (ISO) Inferred from Sequence Alignment (ISA) Inferred from Sequence Model (ISM) Inferred from Genomic Context (IGC) Inferred from Biological aspect of Ancestor (IBA) Inferred from Biological aspect of Descendant (IBD) Inferred from Key Residues (IKR) Inferred from Rapid Divergence(IRD) Inferred from Reviewed Computational Analysis (RCA)
Author statement evidence codes: annotation was made on the basis of declarations made by the author(s) in the literature	Traceable Author Statement (TAS) Non-traceable Author Statement (NAS)
Curator statement evidence codes: when annotation does not support any direct evidence	Inferred by Curator (IC) No biological Data available (ND) evidence code
Electronic Annotation evidence code: specifies that annotation was assigned by automated methods, without curator	Inferred from Electronic Annotation (IEA)

12.5.1 Utilities for GO Annotation

The gene ontology (GO) provides core biological knowledge representation for modern biologists, whether computationally or experimentally based. It has become an extremely useful tool for the analysis of OMICS data and structuring of biological knowledge. With the aim of high-quality annotation and easy access to GO annotation database, a number of online tools are available, such as *QuickGO* (Binns et al. 2009), which have been developed at the EBI, and *AmiGO* (Carbon et al. 2008), which is developed by the GO Consortium.

12.5.1.1 Viewing GO Terms Using QuickGO

A responsive web-based tool that allows easy access to GO annotation. QuickGO can be queried online at https://www.ebi.ac.uk/QuickGO/ or can be downloaded freely from http://www.ebi.ac.uk/QuickGO/installation.html.

The *QuickGO* home page (Fig. 12.11) provides a query box (Fig. 12.11 (A)) to start searching for GO annotation. *QuickGO* takes a wide range of gene identifiers and symbol for annotation retrieval, for example, NCBI Gene IDs, RefSeq accessions, Ensembl Ids, UniProtKB accessions, InterPro IDs, Enzyme Commission (EC) numbers, and GO IDs.

A search for the keyword "apoptosis" retrieves all terms where "apoptosis" is present in the term name and gene product (Fig. 12.12). Here search term "apoptosis" is underlined in red color, and matched terms are shown in green color.

Fig. 12.11 *QuickGO* home page

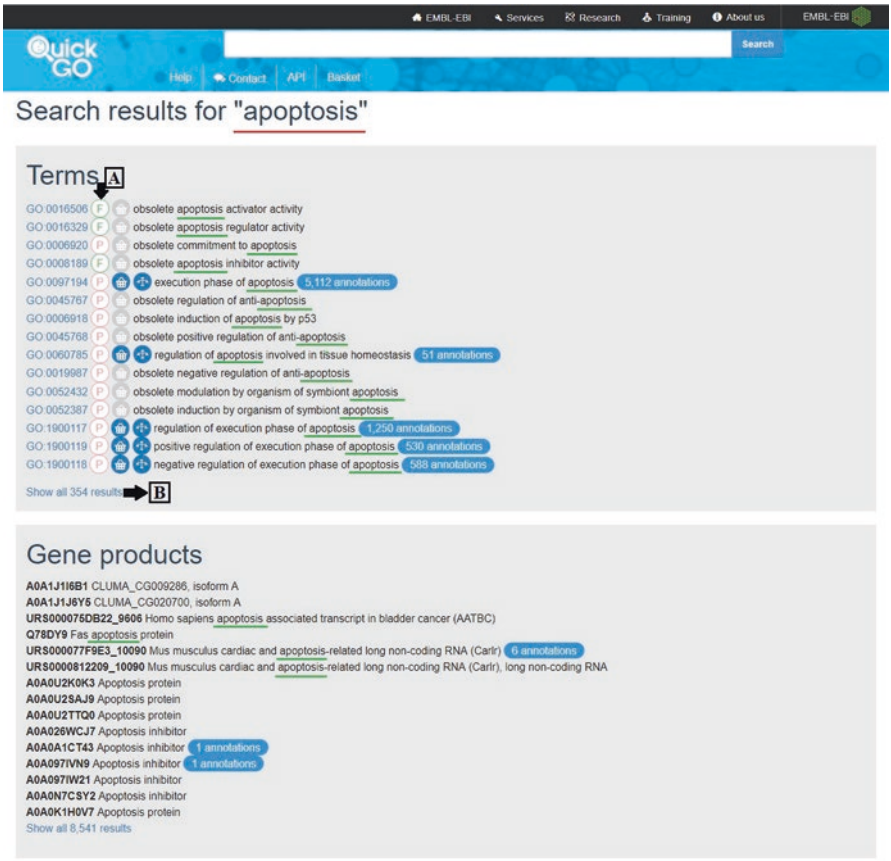
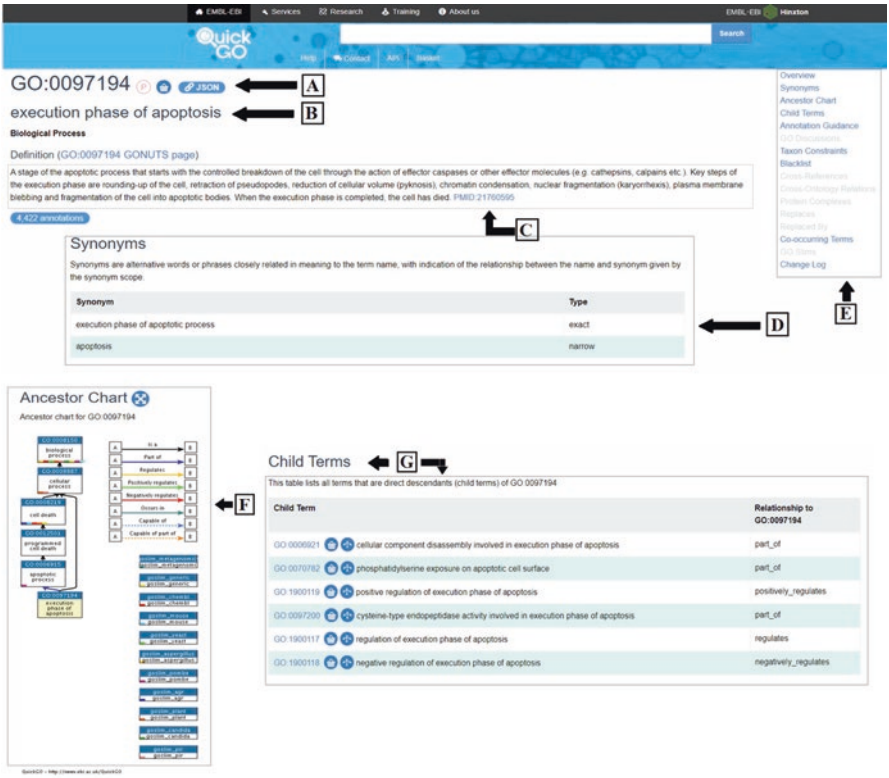


Fig. 12.12 QuickGO: search for keyword "apoptosis"

Clicking on the GO term (e.g., GO:0097194) redirect user to *Term Information Page* (Fig. 12.13), providing complete information for the selected GO term.

12.5.1.2 Viewing GO Terms Using AmiGO

AmiGO is another web-based application provided by the Gene Ontology Consortium for identification and visualization of GO terms related to genes. AmiGO can be accessed from GOC (http://amigo.geneontology.org) or can be downloaded (http://sourceforge.net/projects/geneontology/) to use as the stand-alone application.



[A] → A unique, stable identifier for the GO term
[B] → The primary GO term name
[C] → The term definition and evidence information
[D] → Term synonyms
[E] → Output columns
[F] → Ancestor terms to the selected GO term alongwith their relationship
[G] → Terms that are direct descendants of selected GO term.

Fig. 12.13 QuickGO: GO term information page view

The *AmiGO* home page (Fig. 12.14a) provides a search box (Fig. 12.14a (A)) to start searching for GO annotation. *AmiGO* takes a wide range of gene identifiers and symbol for GO annotation retrieval. Search keyword "apoptosis" is used to retrieve all terms where "apoptosis" is present in the GO terms, GO annotation, and gene products (Fig. 12.14b).

Clicking on "*Ontology*" will return all GO IDs containing "apoptosis" keyword in gene ontology term, synonym, or GO definition (Fig. 12.15).

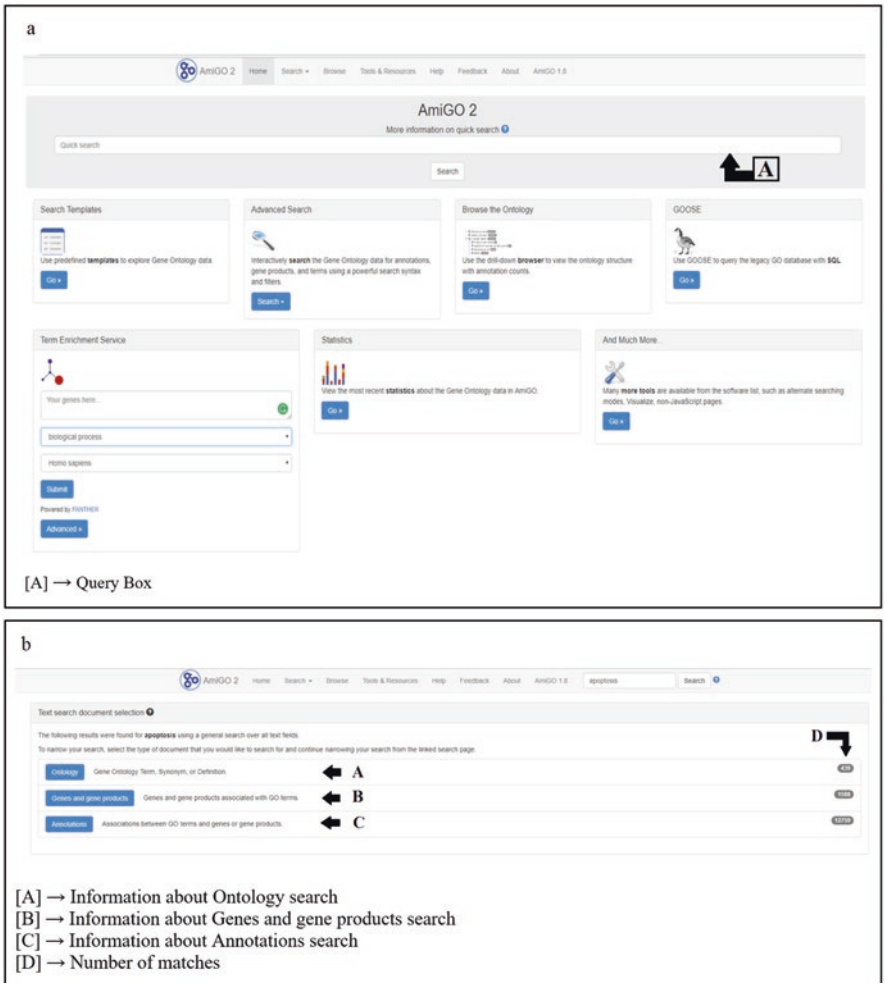


Fig. 12.14 AmiGO home page and "apoptosis" keyword search page

12.5.1.3 The Database for Annotation, Visualization, and Integrated Discovery (DAVID)

DAVID (Huang et al. 2008) provides a comprehensive set of functional annotation tools for investigators to comprehend the biological meaning behind large list of gene/protein lists generated from a variety of high-throughput genomic experiments. In this tutorial, given a list of differentially expressed genes, we will use DAVID to identify the enriched GO terms, such that we can have a clue on the role of genes played in the underlying biological processes.

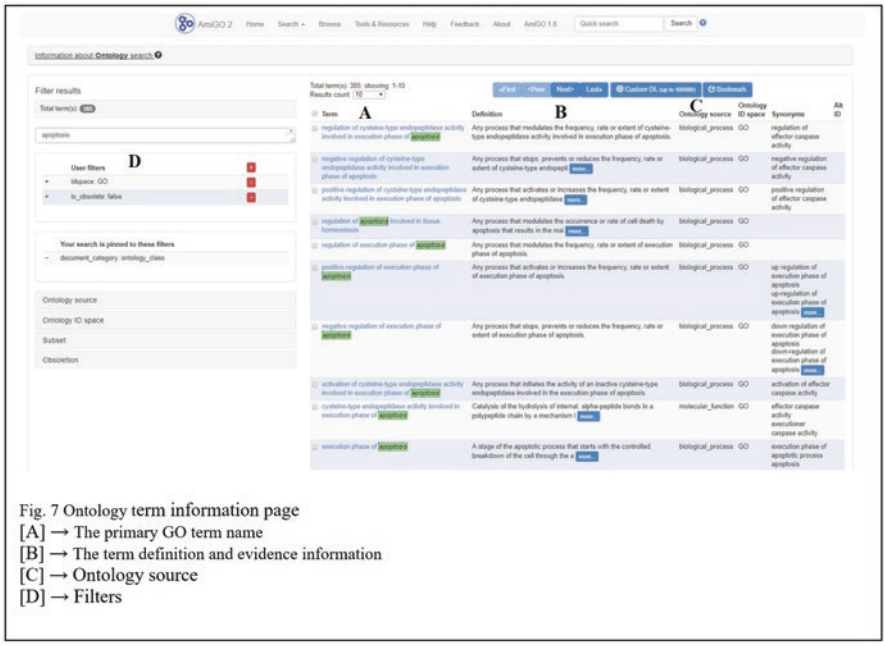


Fig. 12.15 AmiGO: Ontology term information page

Perform Function Annotation Test

(a) Open the server DAVID 6.8 (https://david.ncifcrf.gov/).
(b) Click "Start Analysis" tab (A) as shown in Fig. 12.16.
(c) Submit a gene list to DAVID using input interface (Fig. 12.17). Paste the Affymetrix_geneID list from (A) to the text box (B), or load a text file containing gene IDs using browse option (C). Select the appropriate gene identifier type for input gene IDs using (D). User can also convert gene IDs to other formats using DAVID "Gene ID conversion" tool (E). Specify input IDs as gene list (i.e., genes to be analyzed) or as background genes (i.e., gene population background) at (F). Finally, click "Submit" (G).
(d) After job submission, the progress bar at the top shows job progress. If >20% of gene_identifiers are ambiguous or unrecognized, user will be redirected automatically to "DAVID Gene ID Conversion Tool" Fig. 12.18 (D). Implicitly, the background is set up to the species that contain majority of genes in the user's input list (Fig. 12.18 (B)). User can change background using "Background" section as in Fig. 12.18 (A). Run "Functional Annotation chart" (Fig. 12.18 (C)) for functional enrichment analysis and biological knowledge base selection.
(e) Now user needs to input what type of functional annotations are required. For this purpose, the user needs to deselect the "Check Defaults" tab in Fig. 12.19 (A). Then select the GOTERM_BP_FAT (Fig. 12.19 (C)), which is the summarized version of Biological Processes in the GO, by clicking (+) sign as in

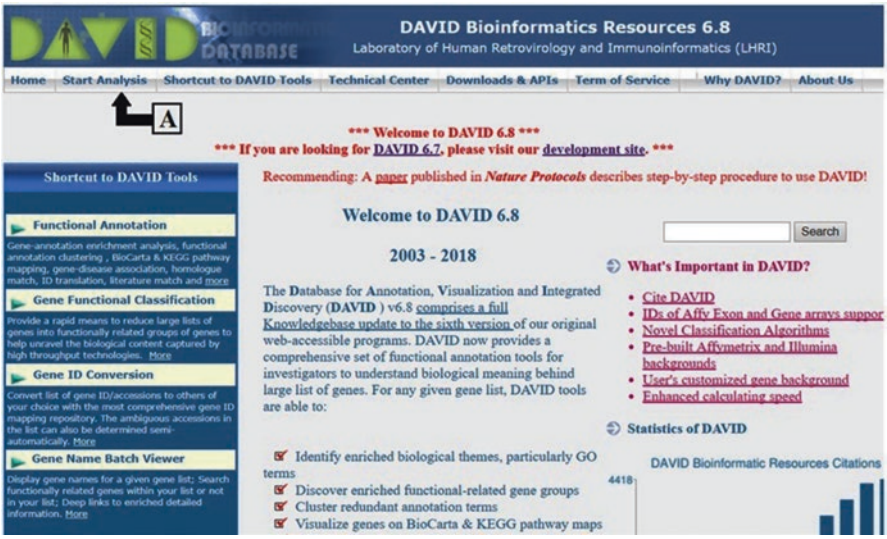


Fig. 12.16 The DAVID 6.8 home page

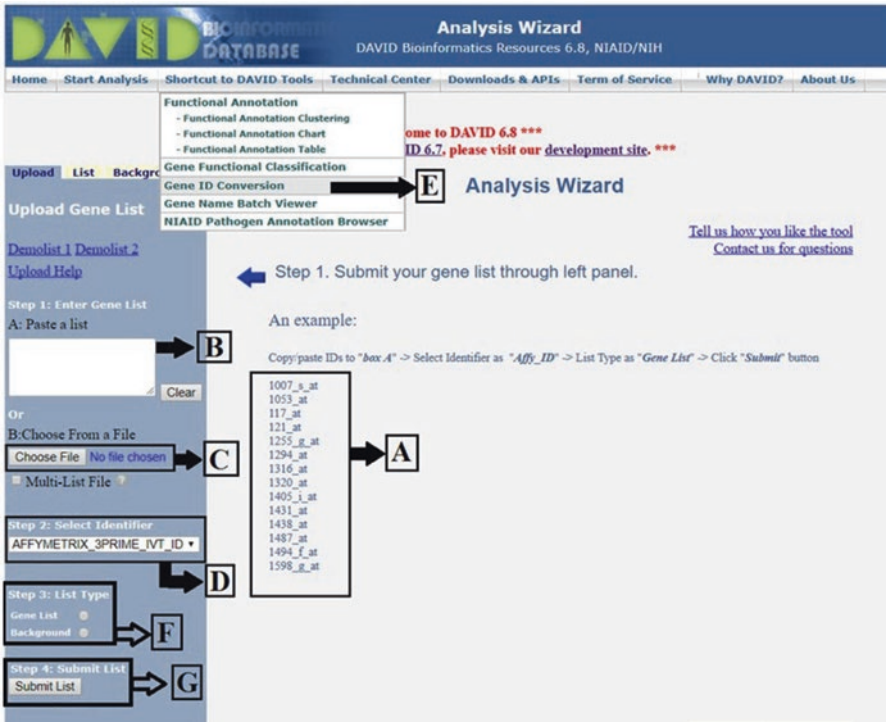


Fig. 12.17 Gene list submission page to DAVID

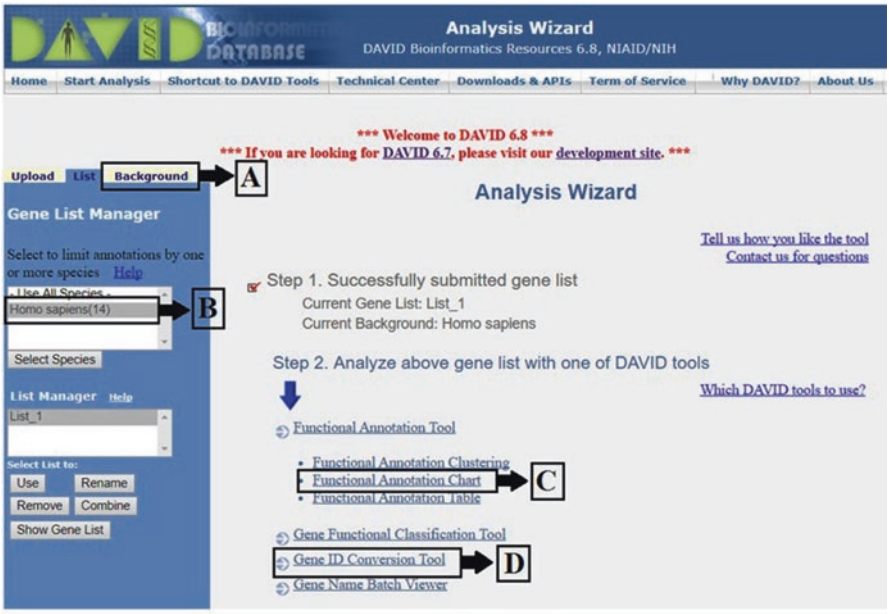


Fig. 12.18 Webpage to access various analytic tools/modules available in DAVID

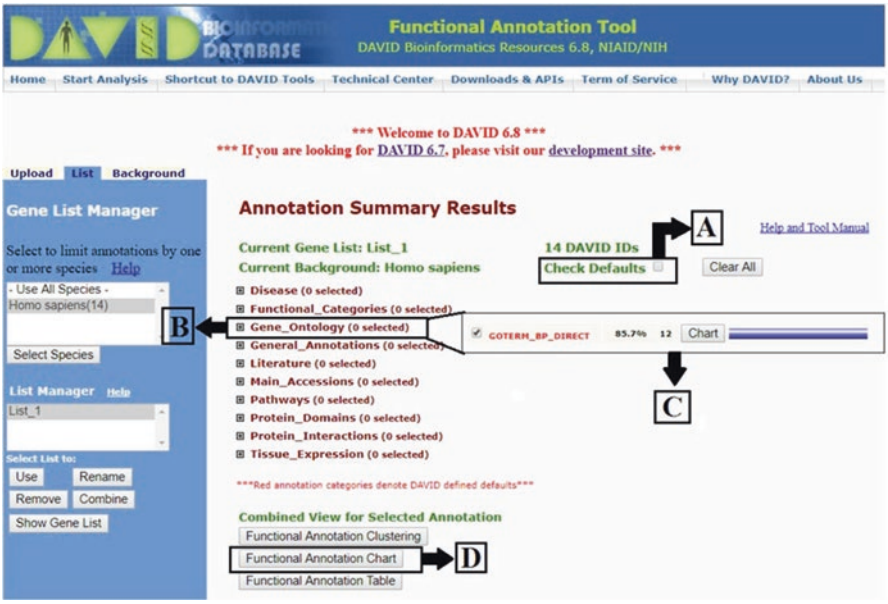


Fig. 12.19 Layout of DAVID "Functional Annotation Chart"

Fig. 12.19 (B). User can try other annotation categories, for example, classifying genes based on pathways using KEGG database, gene-gene interactions identification using BIOGRID database, domain identification, etc.

(f) Click on "Functional annotation chart" button (Fig. 12.19 (D)); a window will be prompted to show the results of functional enrichment test. This statistical test identifies the significantly enriched terms in GOTERM_BP_FAT knowledgebase (Fig. 12.20 (B)). Each row represents an enriched functional term (Fig. 12.20 (C)) and is ordered by their significance level; the smaller the score (Fig. 12.20 (D)), the better is the result. User can download the complete annotation file from Fig. 12.20 (A).

When to and Why Use DAVID?

High-throughput techniques like next-generation sequencing and mass spectrometry generate a huge amount of data, which finally yield gene identifiers.

The gene identifier table can be of various types:

- If data is generated from RNA sequencing or MS experiments, these gene identifiers are linked to respective expression values in a particular condition. These expression values can be as FPKM or RPKM units.

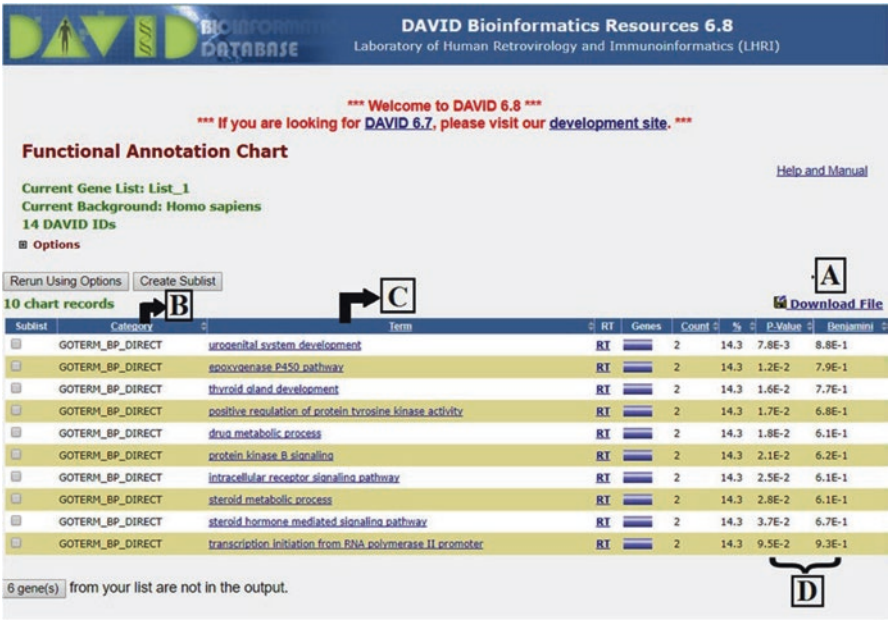


Fig. 12.20 DAVID: Functional annotation chart

- These genes need to be classified according to their molecular functions, biological processes, and cellular locations to identify the major pathways operating in a particular biological condition (e.g., diseased state in which sequencing was performed). Such classification or grouping of genes is called gene enrichment. Genes are also clustered based on their functional annotation. Such functional clustering is essential to identify genes having similar functions. Such kind of functional annotation and clustering can be performed using DAVID.
- Data generated from exome sequencing have gene identifiers linked to respective variant information (e.g., in a diseased state).
- This gene set has to functionally annotate to predict the role of respective variants associated. Also, clustering of genes will recognize the genes with polymorphisms, belonging to similar molecular functions. This will give new leads toward building hypothesis on disease pathogenesis.

12.5.1.4 STRING

STRING (Szklarczyk et al. 2016) is a web-based tool for making protein-protein interaction networks.

Create a PPIN Using STRING

The tutorial is for the set of proteins you have.

Step 1: You can search interaction network by clicking on "Multiple proteins" (Fig. 12.21 (A)) and paste a list of gene IDs into text box provided (Fig. 12.21 (B)) or load a text file containing gene IDs using "Browse" option (Fig. 12.21 (C)). In the organism field, you can specify organism name explicitly (e.g., Homo sapiens) or leave it to default as "auto-detect" (Fig. 12.21 (D)). Then click the search button (Fig. 12.21).

Step 2: You will be redirected to the page listing the gene symbols you have entered with their alias and function (Fig. 12.22). The user needs to ensure that specific protein of interest being queried. Then click on "Continue" button (Fig. 12.22 (A)).

Step 3: You will be redirected to a network page (Fig. 12.23). In the protein-protein interaction network (Fig. 12.23 (A)), the circles represent the nodes or proteins. The edges represent the associations between nodes. The legend (Fig. 12.23 (B)) section gives information about nodes and interacting partners or edges.

Step 4: User can change the research parameters from "Setting" section (Fig. 12.24 (A–D)).

Step 5: Visualize the Analysis section (Fig. 12.25). The Analysis section provides network statistics (Fig. 12.25 (A)). The functional enrichment analysis of the input gene set is provided in Fig. 12.25 (B). The information about the statistical background used for functional enrichment is also given in Fig. 12.25 (C).

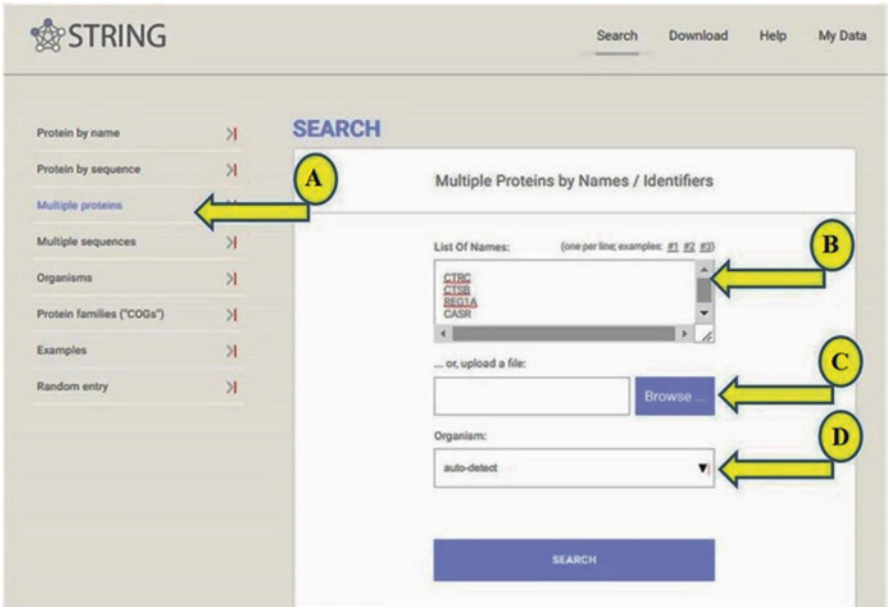


Fig. 12.21 STRING: Use multiple protein identifiers input for PPIN construction

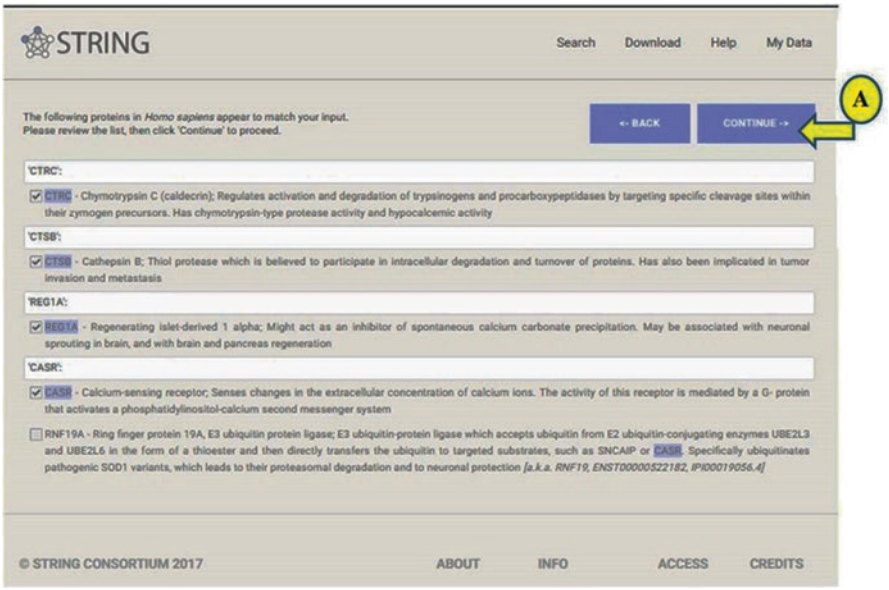


Fig. 12.22 STRING: Ensuring the correct protein identifiers are being used for PPIN construction

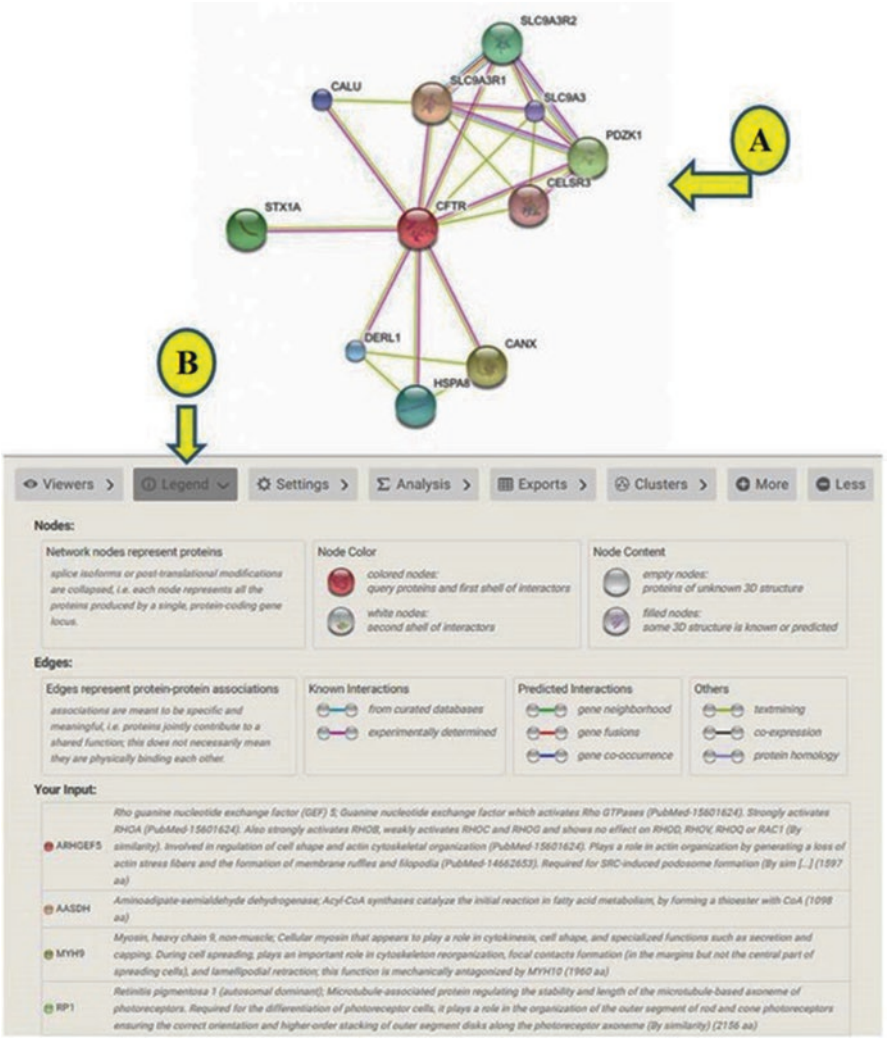


Fig. 12.23 STRING: Network visualization

Step 6: Finally you can export the network files (Fig. 12.26 (A)) in different formats (Fig. 12.26 (B)) to analyze it further using Cytoscape or any other network visualization tool(s).

12.5.1.5 Cytoscape

Cytoscape (Shannon et al. 2003) is an open source tool for visualizing biomolecular interaction networks, integrating functional annotations and high-throughput gene expression profiles into a unified conceptual framework, and identifying their

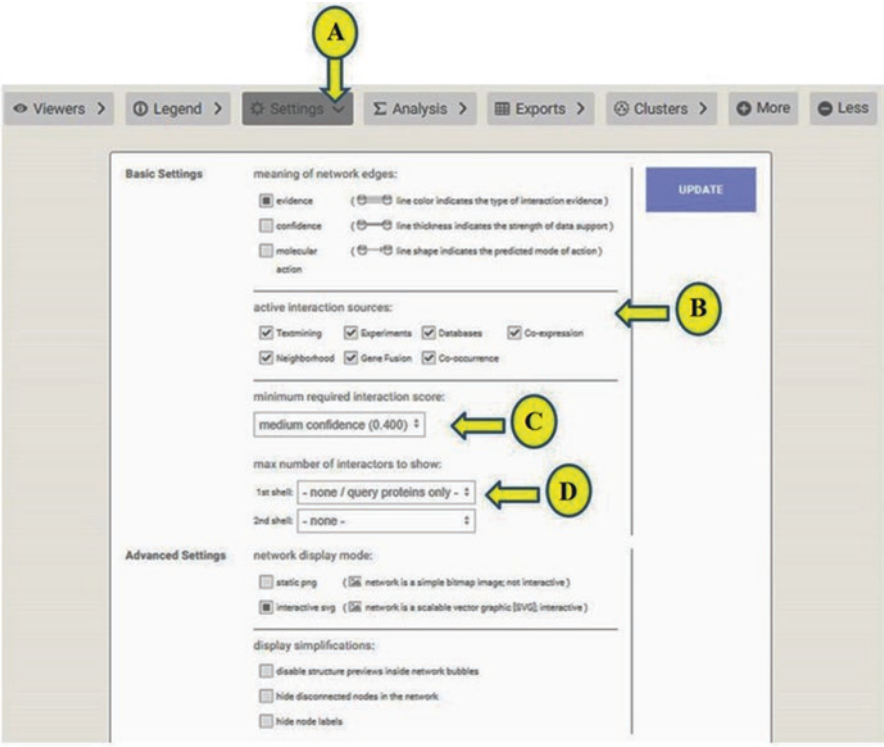


Fig. 12.24 STRING: Change research parameters for PPIN construction

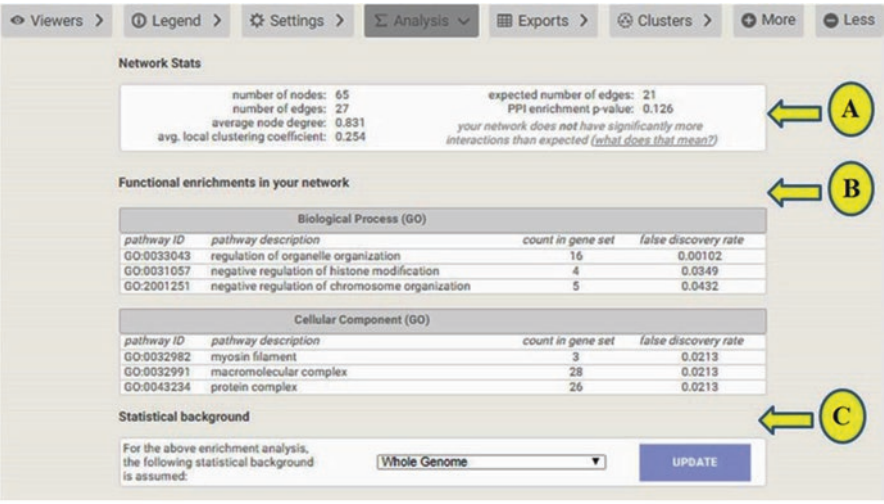


Fig. 12.25 STRING: Analysis section providing network statistics and functional enrichment analysis of input protein identifiers

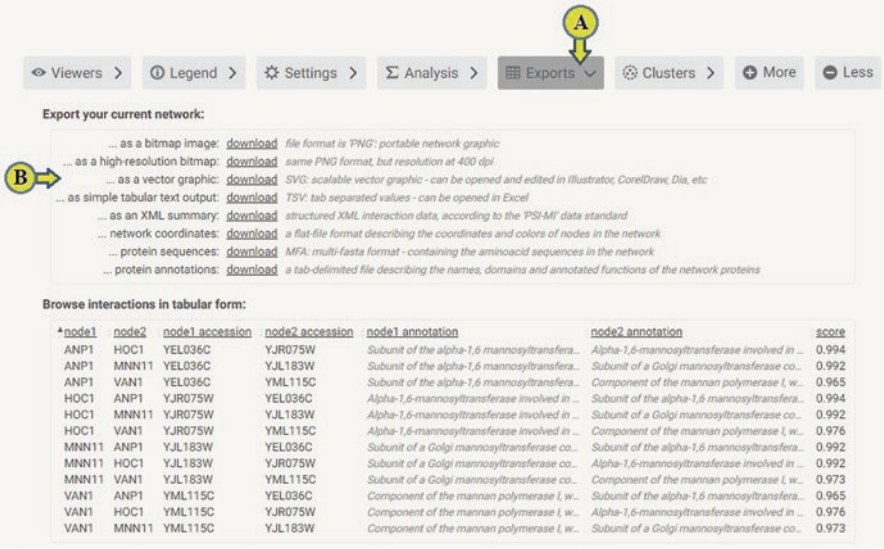


Fig. 12.26 STRING: Export required network files

properties. Additional utilities are available in the form of plugins. Plugins are available for network properties and molecular profiling analyses, various layouts for better visualization, additional file format support, and connection with databases and searching in large networks. Cytoscape additionally has a JavaScript-driven sister venture named *Cytoscape.js* that can be utilized to dissect and visualize networks in JavaScript environments through a web browser.

Examples of Uses

Gene function prediction – examining genes (proteins) in a network context shows connections to sets of genes/proteins involved in the same biological process that is likely to function in that process (plugin for analysis: jActiveModules, PiNGO, etc.).

Detection of protein complexes/other modular structures – protein complexes are groups of associated polypeptide chains whose malfunctions play a vital role in disease development. Complexes can perform various functions in the cell, including dynamic signaling, and can serve as cellular machines, rigid structures, and posttranslational modification systems. Many disorders are consequences of changes in a single protein and, thus, in its set of associated partners and functionality (plugin for analysis: Motif Discovery, Mclique, MCODE, PEWCC, etc.).

Identification of disease sub-networks and potential biomarkers – identification of disease network sub-networks that are transcriptionally active in the disease and also provide a rich source of biomarkers for disease classification. These suggest key pathway components in disease progression and provide leads for further study and potential therapeutic targets (plugin for analysis: PhenomeScape, PSFC, etc.).

Dynamics of a biological network – the molecular interactions in a cell vary with time and surrounding environmental conditions. The construction and analysis of dynamic molecular networks can elucidate dynamic cellular mechanisms of different biological functions and provide a chance to understand complex diseases at the system level (plugin for analysis: DyNetViewer, DynNetwork, DynNet, etc.).

INPUT Type

Cytoscape can read network/pathway files written in the following formats:

- Simple interaction file (SIF or .sif format)
- Nested network format (NNF or .nnf format)
- Graph Markup Language (GML or .gml format)
- XGMML (eXtensible Graph Markup and Modeling Language)
- SBML
- BioPAX
- PSI-MI Level 1 and 2.5
- GraphML
- Delimited text
- Excel Workbook (.xls, .xlsx)
- Cytoscape.js JSON
- Cytoscape CX

The SIF format specifies nodes and interactions only, while other formats store additional information about network layout and allow network data exchange with a variety of other network programs and data sources.

Visualization

Substantial progress has been made in the field of "omics" research (e.g., genomics, transcriptomics, proteomics, and metabolomics), leading to a vast amount of biological data generation. In order to represent large biological data sets in an easily interpretable manner, this information is frequently visualized as graphs, i.e., a set of nodes and edges. Cytoscape assists in visual exploration and analysis of biological network in several ways:

- Provides customize network data display using powerful visual styles.
- Helps in integrating gene expression values with the network. This can be done by mapping expression values to network nodes which represent the gene as

color, label, border thickness, etc. according to the user-defined mapping file and provide several layout options in two as well as three dimensions for network visualization, for example, edge-weighted spring-embedded layout, attribute circle layout, etc.
- The network manager can be utilized to manage multiple networks in a single session file. Easily navigate large networks through an efficient rendering engine.

Analysis

- Filter the network to select subsets of nodes and/or interactions based on the current data. For instance, users may select nodes involved in a threshold number of interactions, nodes that share a particular GO annotation, or nodes whose gene expression levels change significantly in one or more conditions according to p-values loaded with the gene expression data.
- Find active sub-networks/pathway modules. The network is screened against gene expression data to identify connected sets of interactions, i.e., interaction sub-networks, whose genes show particularly high levels of differential expression. The interactions contained in each sub-network provide hypotheses for the regulatory and signaling interactions in control of the observed expression changes.
- Find clusters (highly interconnected regions) in any network loaded into Cytoscape. Depending on the type of network, clusters may mean different things. For instance, clusters in a protein-protein interaction network have been shown to be protein complexes and parts of pathways. Clusters in a protein similarity network represent protein families.
- Plugins available for network and molecular profile analysis.

12.6 Conclusion

Complex biological networks are the reservoir for the plethora of biological information about pathways and cellular mechanisms. This chapter summarized different types of biological networks, methodologies to analyze such networks and biological relevance. These networks can provide researchers with critical information about the pathogenesis of diseases (disease-gene networks), identification of drug targets (protein-protein networks, protein-ligand interaction), and biological pathways. Functional and pathway analysis of genes (gene ontology) determine significant gene clusters associated with a specific biological process, molecular function or pathway. This chapter succinctly provides relevant information about the applications of biological networks in the molecular biology field. Our hope is that the tutorials provided in this chapter will guide researchers to annotate genes on gene products and enrich GO annotation both qualitatively and quantitatively on the available tools.

References

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118:4947–4957
Alberts B et al (2002) Molecular biology of the cell, 4th edn. Garland Science, New York
Arita M (2004) The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A 101:1543–1547
Bader S, Kühner S, Gavin AC (2008) Interaction networks for systems biology. FEBS Lett 582(8):1220–1224
Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD et al (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol 4:e317
Beadle GW, Tatum EL (1941) Genetic control of biochemical reactions in Neurospora. Proc Natl Acad Sci U S A 27:499–506
Binns D et al (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25(22):3045–3046
Blake JA (2013) Ten quick tips for using the gene ontology. PLoS Comput Biol 9(11):e1003343
Botstein D et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25(1):25–29
Carbon S et al (2008) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2):288–289
Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. Proc Natl Acad Sci 104(21):8685–8690
Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430(6995):88–93
Hill DP et al (2008) Gene Ontology annotations: what they mean and where they come from. BMC Bioinf 9(5):S2; BioMed Central
Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44
Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 9(1):559
Ma'ayan A (2011) Introduction to network analysis in systems biology. Sci Signal 4(190):tr5
Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 5(7):e1000431
Shannon P et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
Szklarczyk D et al (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw937
Weirauch MT (2011) Gene coexpression networks for the analysis of DNA microarray data. In: Applied statistics for network biology: methods in systems biology. Wiley-Blackwell, Weinheim, pp 215–250
Yook SH, Oltvai ZN, Barabási AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4(4):928–942