The study of biological networks, their analysis and modeling are important tasks in life sciences today. Most biological networks are still far from being complete and they are often difficult to interpret due to the complexity of relationships and the peculiarities of the data. This worksheet describes major types of biological networks and useful public databases that contain biological networks.

Types of Biological Networks

Many important biological networks are defined on molecules such as DNA, RNA, proteins and metabolites, and the networks describe interactions between these molecules. **Gene co-expression networks** are constructed by looking for pairs of genes which show similar expression patterns across biological conditions, where the activation levels of two co-expressed genes rise and fall together across conditions. **Signal transduction and gene regulatory networks** describe how genes can be activated or repressed, and therefore contain information about which proteins are produced in a cell at a particular time. **Protein-protein interaction networks** represent interactions between proteins such as the building of protein complexes and the activation of one protein by another protein. **Metabolic networks** show how metabolites are transformed, for example to produce energy or to synthesize specific substances. Other types of biological networks include **phylogenetic trees, special networks and hierarchies** which are often built based on information from molecular biology such as DNA and protein sequences. Phylogenetic trees represent the ancestral relationships between different organisms, i.e., their origins, how they survive or become extinct.

Gene regulatory, signal transduction, protein-protein interaction and metabolic networks interact with each other and build a complex biological network. Furthermore, these networks are not universal but are organism-specific and environment-specific, i.e. the same network differs between different organisms and environments in which these organisms live.

Biology is often more complicated than what appears in a network. For example, protein-protein interactions can be dependent on the location within the cell. Another such complexity level is the time dimension. For example, one protein can be at one time bound to another protein, suppressing its activity, and at other time this protein can be bound to a third protein, in which case it cannot bind to the second protein. Both interactions will appear in the protein-protein interaction network, although they do not occur simultaneously.
Gene Co-expression Networks

Gene co-expression is the process by which a set of genes are expressed in coordination to produce proteins. A gene co-expression network captures information on the correlation of gene expression in different biological conditions, such as during the time when cells are activity dividing, or when cells are reacting to a particular drug treatment.

A gene co-expression network is a weighted undirected network $G = (V, E, \delta)$, where the set of nodes $V$ represent genes, the set of edges $E$ represent pairs of genes that are significantly co-expressed, and edge weights $\delta : E \rightarrow [-1, 1]$ represent correlation of pairs of genes. A pair of nodes is connected with an edge if the corresponding genes have significantly similar expression patterns, meaning that the genes are active under the same biological conditions.

**Major public databases:** The Cancer Genome Atlas [1], NCBI Gene Expression Omnibus [2], GeneMANIA [3], EBI Array Express [4], GTEx Data Portal [5], MGI-Mouse Gene Expression Database [6], STRING [7], Bgee [8].

Signal Transduction and Gene Regulatory Networks

Signal transduction is a communication process within a cell to coordinate its responses to an environmental change. The response is a reaction of the cell, e.g., the activation of a gene or the production of energy. A signal transduction pathway is a directed network of chemical reactions in a cell from a stimulus (an external molecule which binds to a receptor on the cell membrane) to the response (e.g., a gene whose activity is changed due to the binding of external molecule). The signal transduction network of a cell is the complete network of all signal transduction pathways.

Gene regulation can also be seen as the response of a cell to an internal stimulus. Often one gene is regulated by another gene via the corresponding protein that is called a transcription factor. Gene regulation is thus coordinated in a gene regulatory network. A gene regulatory network is a directed network where nodes represent genes and directed edges represent regulatory interactions, such as binding of a transcription factor (i.e., source of an edge) to a gene (i.e., target of an edge). Compared to a gene co-expression network, a gene regulatory network attempts to represent the causal (direct) relationships between genes. Ideally, a directed edge in a gene regulatory network from node $v_i$ to node $v_j$ is present if and only if a causal effect runs from node $v_i$ to $v_j$ and there exist no nodes or subsets of nodes that are intermediating the causal influence.
**Protein-Protein Interaction Networks**

Protein-protein interaction networks are networks where nodes represent proteins and edges represent interactions, that is, two proteins are connected if they interact with each other. A protein can interact with another protein, e.g., to build a protein complex or to activate it.

A protein-protein interaction network is an undirected graph $G = (V, E, \tau)$ where $V$ is the set of proteins, $E$ the set of interactions, and $\tau : E \rightarrow T$ defines the type of each edge (interaction type). Often only the existence of an interaction between two proteins is known, but the interaction type $T$, such as “activation”, “binding to”, or “phosphorylation”, remains unknown. However, for the understanding of biological processes, information about the interaction type is crucial, although up to now databases contain little information about that. It is also possible to represent a protein-protein interaction network with a directed graph $G$, in this case, $E$ denotes a set of directed interactions where a protein initiating the interaction defines the source of an edge. Protein-protein interaction networks can be derived from databases such as BioGRID [15] and STRING [16].

**Metabolic Networks**

Metabolic networks are directed networks where each node represents a metabolite (a molecule) and an edge represents a metabolic reaction. A metabolic reaction is a chemical process that transforms chemical substances or metabolites (i.e., reactants) into other substances (i.e., products) usually catalyzed by enzymes.

Metabolic reactions interact with each other, i.e., the product of one reaction is usually a reactant of another reaction. A metabolic path $P = (R_1, \ldots, R_n)$ is a sequence of metabolic reactions $R_i$ where for all $1 \leq i \leq n$ at least one product of reaction $R_i$ is a reactant of reaction $R_{i+1}$. The metabolic network or metabolism of an organism is then the complete network of metabolic reactions of this organism. A metabolic pathway is a connected sub-network of the metabolic network either representing specific processes or defined by functional boundaries.
e.g., the network between an initial and a final chemical substance.

Formally, metabolic pathway is a hyper-graph. The nodes represent the substances and the hyper-edges represent the reactions. A hyper-edge connects all substances of a reaction, is directed from reactants to products and is labeled with the enzymes that catalyze the reaction. Additionally to the nodes representing substances, metabolic reactions are nodes and edges are binary relations connecting the substances of a reaction with the corresponding reaction node. A metabolic pathway is modeled as a directed bipartite graph $G = (V_s, V_r, E)$ with nodes $u_1, \ldots, u_n, w_1, \ldots, w_m \in V_s$ representing substances, nodes $v_1, \ldots, v_k \in V_r$ representing metabolic reactions and directed edges $(u_1, v_1), \ldots, (u_n, v_1), (v_1, w_1), \ldots, (v, w_m) \in E$ representing the transformation of substances $u_1, \ldots, u_n$ to substances $w_1, \ldots, w_m$ by the reaction $v_1$.

Major public databases: BRENDA [23], KEGG PATHWAY Database [24], MANET [25], Reactome [26], Small Molecule Pathway Database [27], MetaNetX [28].

Phylogenetic Trees, Special Networks and Hierarchies

One of the fundamental principles in biology is the hierarchical organization of organisms in an evolutionary context, i.e., reconstruction of ancestral relationships between different species, genes, or DNA sequences. The common representation for analyzing such relationships is a phylogenetic tree. A phylogenetic tree (in literature also called evolutionary tree) $T = (V, E, \delta)$ is a tree consisting of nodes $V$ (i.e., taxons) and edges $E$ (i.e., links). Leaf nodes, i.e., nodes with exactly one link, represent species, sequences, or similar entities. Internal nodes represent (hypothetical) ancestors generated based on phylogenetic analysis. The edge weights $\delta : E \to \mathbb{R}_0^+$ quantify biological divergence between incident nodes, e.g., biological time or genetic distance.

One example of special networks are gene-phenotype networks, such as [29, 30, 31]. A gene-phenotype network is a bipartite network with two sets of nodes. One set represent phenotypes, e.g., diseases, of a particular organism (e.g., humans) and the other set represents organism’s genes. A gene and a phenotype are connected by an edge if the gene is involved in the disease (e.g., causal disease genes).

A very prominent example of biological data hierarchies is the Gene Ontology (http://geneontology.org) [32]. The Gene Ontology is a major organism-agnostic bioinformatics initiative that develops a computational representation of our evolving knowledge of how genes encode biological functions at the molecular, cellular and tissue systems levels. The Gene Ontology provides controlled vocabularies of terms representing gene properties. The
vocabularies cover three domains: cellular components, the parts of a cell and its extracellular environment; molecular functions, the activities of genes at the molecular level, such as binding and catalysis; and biological processes, operations or sets of molecular events with a defined beginning and end that are relevant for the functioning of living systems. The Gene Ontology is structured as a directed acyclic graph where each term has defined relationships to one or more other terms. Additionally, the Gene Ontology provides gene annotations, which assign Gene Ontology terms to genes. Gene annotations are lists of gene-Gene Ontology term pairs that are often used as gold standard when analyzing data (e.g., [33]).

Major public databases: Gene Ontology [32], Disease Ontology [34], OMIM Inherited Diseases [29], The Comparative Toxicogenomics Database [30], DisGeNET [31], DrugBank [35], STITCH [36], STRING [16], MSigDB [37], UniProt [38], ENCODE [39], NCBI Taxonomy [40].

References


