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 $\delta : E \to [-1, 1]$ 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.

Major public databases: Netpath [9], Pathway Commons [10], WikiPathways [11], NCI-Nature Pathway Interaction Database [12], RegulonDB [13], TRANSFAC [14].

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].

Major public databases: BioGRID [15], HPRD [17], MIntAct [18], STRING [16], GeneMANIA [19], CCSB Interactome [20], DIP [21], MINT [22].

Metabolic Networks

Metabolic networks are directed networks where each node represents a metabolite (a molecule) and and 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 $(u_1, v_1), \ldots, (u_{n_1}, v_1), (v_1, w_1), \ldots, (v, w_{m_1}) \in E$ representing the transformation of substances $u_1, \ldots, u_{n_1}$ to substances $w_1, \ldots, w_{m_1}$ 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

[1] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[2] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[3] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl 2):W214–W220, 2010.

[4] Gabriella Rustici, Nikolay Kolesnikov, Marco Brandizi, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Jon Ison, Maria Keays, et al. ArrayExpress updatetrends in database growth and links to data analysis tools. *Nucleic Acids Research*, 41(D1):D987–D990, 2013.

[5] GTEx Consortium et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[6] Constance M Smith, Jacqueline H Finger, Terry F Hayamizu, Ingeborg J McCright, Janan T Eppig, James A Kadin, Joel E Richardson, and Martin Ringwald. The mouse gene expression database (GXD): 2007 update. *Nucleic Acids Research*, 35(suppl 1):D618–D623, 2007.

[7] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, 2013.

[8] Frederic Bastian, Gilles Parmentier, Julien Roux, Sebastien Moretti, Vincent Laudet, and Marc Robinson-Rechavi. Bgee: integrating and comparing heterogeneous transcriptome data among species. In *International Workshop on Data Integration in the Life Sciences*, pages 124–131. Springer, 2008.

[9] Kumaran Kandasamy, S Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S Sameer Kumar, Abhilash K Venugopal, Deepthi Telikicherla, J Daniel Navarro, Suresh Mathivanan, Christian Pecquet, et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biology*, 11(1):1, 2010.

[10] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690, 2011.

[11] Thomas Kelder, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, 2012.

[12] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. PID: the pathway interaction database. *Nucleic Acids Research*, 37(suppl 1):D674–D679, 2009.

[13] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(D1):D203–D213, 2013.

[14] Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(suppl 1):D108–D110, 2006.

[15] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'Donnell, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43(D1):D470–D478, 2015.

[16] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, page gku1003, 2014.

[17] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database2009 update. *Nucleic Acids Research*, 37(suppl 1):D767–D772, 2009.

[18] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The MIntAct projectIntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, page gkt1115, 2013.

[19] Khalid Zuberi, Max Franz, Harold Rodriguez, Jason Montojo, Christian Tannus Lopes, Gary D Bader, and Quaid Morris. GeneMANIA prediction server 2013 update. *Nucleic Acids Research*, 41(W1):W115–W122, 2013.

[20] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

[21] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002.

[22] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861, 2012.

[23] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, 39(suppl 1):D507–D513, 2011.

[24] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360, 2010.

[25] Hee Shin Kim, Jay E Mittenthal, and Gustavo Caetano-Anollés. MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics*, 7(1):1, 2006.

[26] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al.

The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2014.

[27] Timothy Jewison, Yilu Su, Fatemeh Miri Disfany, Yongjie Liang, Craig Knox, Adam Maciejewski, Jenna Poelzer, Jessica Huynh, You Zhou, David Arndt, et al. SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Research*, page gkt1067, 2013.

[28] Sébastien Moretti, Olivier Martin, T Van Du Tran, Alan Bridge, Anne Morgat, and Marco Pagni. MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Research*, page gkv1117, 2015.

[29] Joanna Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. McKusick's online Mendelian inheritance in man (OMIM). *Nucleic Acids Research*, 37(suppl 1):D793–D796, 2009.

[30] Allan Peter Davis, Benjamin L King, Susan Mockus, Cynthia G Murphy, Cynthia Saraceni-Richards, Michael Rosenstein, Thomas Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2011. *Nucleic Acids Research*, 39(suppl 1):D1067–D1072, 2011.

[31] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028, 2015.

[32] Gene Ontology Consortium et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(suppl 1):D258–D261, 2004.

[33] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.

[34] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 2012.

[35] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42(D1):D1091–D1097, 2014.

[36] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Christian Von Mering, Lars Juhl Jensen, and Peer Bork. STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Research*, 40(D1):D876–D880, 2012.

[37] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[38] UniProt Consortium et al. The universal protein resource (UniProt). *Nucleic acids research*, 36(suppl 1):D190–D195, 2008.

[39] Jie Wang, Jiali Zhuang, Sowmya Iyer, Xin-Ying Lin, Melissa C Greven, Bong-Hyun Kim, Jill Moore, Brian G Pierce, Xianjun Dong, Daniel Virgil, et al. Factorbook. org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Research*, 41(D1):D171–D176, 2013.

[40] Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, 2012.