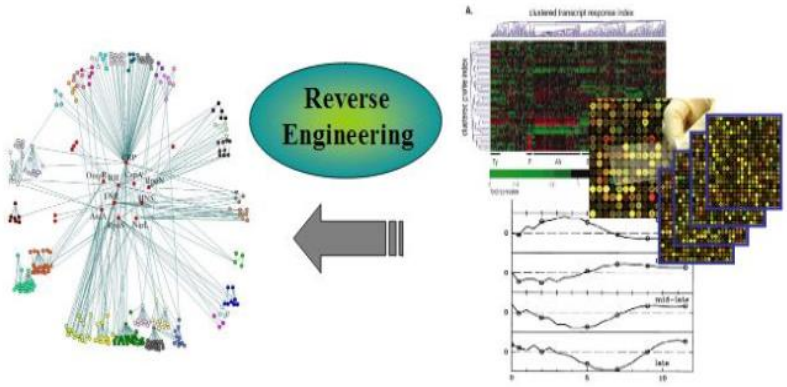
UNIT V

REVERSE ENGINEERING OF BIOLOGICAL NETWORKS

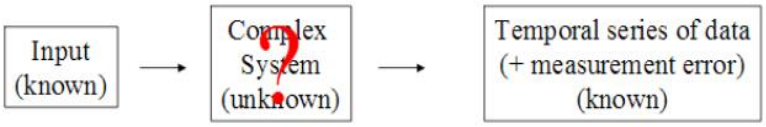
Reverse Engineering:
means building a network structure from the observed gene expression patterns.



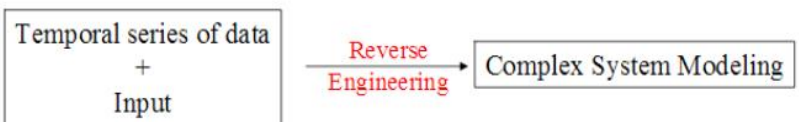
Network structure Observed data

Reverse engineering of Biological networks

The interplay of mathematical modelling with experiments is one of the central elements in systems biology. The aim of reverse engineering is to infer, analyse and understand, through this interplay, the functional and regulatory mechanisms of biological systems. Reverse engineering is not exclusive of systems biology and has been studied in different areas, such as inverse problem theory, machine learning, nonlinear physics, (bio)chemical kinetics, control theory and optimization, among others



Temporal dynamics, between one state of the system and another, are necessary to infer the structure of the system.



Biological networks:

Biological processes are often represented in the form of networks such as protein-protein interaction networks and metabolic pathways. The study of biological networks, their modeling, analysis, and visualization are important tasks in life science today.An understanding of these networks is essential to make biological sense of much of the complex data that is now being generated. This increasing importance of biological networks is also evidenced by the rapid increase in publications about network-related topics and the growing number of research groups dealing with this area. Most biological networks are still far from being complete and they are usually difficult to interpret due to the complexity of the relationships and the peculiarities of the data. Network visualization is a fundamental method that helps scientists in understanding biological networks and in uncovering important properties of the underlying biochemical processes

Several highly important biological networks are related to molecules such as DNA, RNA,proteins and metabolites and to interactions between them.
Gene regulatory and signal transduction networks describe how genes can be activated or repressed and therefore which proteins are produced in a cell at a particular time. Such regulation can be caused by regulatory proteins or external signals.

Protein-protein interaction networks represent the interaction 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 synthesize specific substance phylogenetic trees, special networks or hierarchies which are often built on information from molecular biology such as DNA or protein sequences. Phylogenetic trees represent the ancestral relationships between different species. They are used to study evolution, which describes and explains the history of species, i.e., their origins, how they change, survive, or become extinct. Finally, signal transduction, gene regulatory, protein-protein interaction and metabolic networks interact with each other and build a complex network of interactions; furthermore these networks are not universal but species-specific, i.e., the same network differs between different species.

Signal Transduction and Gene Regulatory Networks:

Signal transduction is a communication process within a cell to coordinate its re sponses to an environmental change. The stimulus comes from the cell's environment, e.g., molecules such as hormones. 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., the activation of a gene). Here we focus on signal transduction pathways that aim at transcription factors and thus alter the expression of genes in a cell. The signal transduction network of a cell is the complete network of all signal transduction pathways. A signaling cascade is a process where signal transduction involves an increasing number of molecules in the steps from the stimulus to the response.

Gene regulation is a general term for cellular control of the synthesis of proteins at thetranscription step. 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 (called transcription factor), thus gene regulation is coordinated in a gene regulatory network. Thisnetwork directs the level of expression for each gene in the cell by controlling whether and how often that gene will be transcribed into RNA. Similar to signaling cascades in signal transduction networks a gene can activate more genes in turn and an initial stimulus can trigger the expression of large sets of genes.

Events of signal transduction and gene regulatory processes occur in different parts of a cell (cellular compartments). To represent compartments these networks can be modeled as clustered graphs. A clustered graph C= (G, T) consists of a directed graph G= (V, E) and a rooted tree T, such that the leaves of Tare exactly the nodes of G. The nodes $v \in V$ of the graph are chemical and biochemical compounds (ranging from ions, to small molecules,macromolecules and genes) and the edges $e \in E$ are biochemical events (e.g., binding, transportation and reaction). The occurrence of signal transduction and gene regulatory events in different cellular compartments can be modeled be the treeT. Each node $t \in T$ represents a cluster of nodes of G consisting of the leaves of the subtree rooted at t. The modeling of such networks based on clustered graphs can be used for cluster-preserving layout algorithms. However, as it is only partly known in which compartment an event occurs, signal transduction and gene regulatory processes are usually modeled by graphs. The pathways and networks can be derived from databases such as KEGG and TransPath.

Visualization Requirements

Important goals of the visualizations of signal transduction and gene regulatory pathways are the understanding of the regulation of cellular processes by external and internal signals,the flow of information through the pathways and networks, the interconnection of genes,the discovering of master-genes responsible for the regulation of larger sets of genes, and the identification of main and alternative regulatory paths.

The main visualization requirements are:

•**Pathways**: The main direction of the processes (e.g., from top to bottom) should be clearly visible to express the temporal order of the events.

•**Compartments:** Events of signal transduction and gene regulation occur in different cellular compartments and this information should be visually represented.
•**Complexes:** Especially during signal transduction one event occurring frequently is the building of molecular complexes. Their structure and how they are built by interacting molecules should be displayed.

Layout Methods

There are two established approaches to visualize signal transduction and gene regulatory pathways and networks: force-directed and hierarchical layout methods. It should be noted that some visualizations of gene regulatory networks in books and articles also use orthogonal or grid-based drawing styles.

There are some systems supporting force-directed layouts for the visualization of signal transduction and gene regulatory pathways and networks. These tools are either based on re-implementations of well-known algorithms or on existing layout libraries. Usually the visualizations do not meet the main requirements, especially the main direction and the consideration of compartments. There are a few approaches to improve the general force directed method. Examples are the PATIKA system where the force directed layout has been extended to deal with several application specific requirements, e.g., cellular compartments, and the approach presented in where placement, directional, compartmental and other constraints are considered.

Another common approach for the visualization of signal transduction and gene regulatory networks are graph drawing solutions based on hierarchical layout methods,
There exist several systems which use hierarchical layouts for the visualization of these networks, e.g., TransPath . Most are based on existing layout libraries such as dot and Pajek . These approaches meet some visualization requirements such as the main direction of pathways.

Protein-Protein Interaction Networks

A protein can interact with another protein, e.g., to build a protein complex or to activate it.Protein-protein interactions form large networks. Their visualization aids biologists in pinpointing the role of proteins and in gaining new insights about the processes within and across cellular processes and compartments, e.g., for form ulating and experimentally testing specific hypotheses about gene function.

Often only the existence of an interaction between two proteins is known, but the interaction type, 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. Therefore we define a protein-protein interaction network as a directed g

Raph G= (V, E, τ) whereV is the set of proteins,E the set of directed interactions (the initiator defines the source),and τ:E→T defines the type of each edge (interaction type). Protein-protein interaction networks can be derived from databases such as BIND and DIP.

Metabolic Networks

A metabolic reaction R is a transformation of chemical substances or metabolites (reactants) into other substances (products) usually catalyzed by enzymes. In general metabolic reactions are reversible, that is, they occur in both directions. Such reactions are characterized by a steady state, i.e., if occurring isolated they reach a state where the amount of change in both directions is equal. A cell is in a constant exchange of substances with its environment. Furthermore, many reactions are regulated, i.e., they are suppressed or enhanced by other factors (allosteric control). This shifts the steady state and together with the steady supply of substances from outside and their final use, e.g., by exporting them from the cell, one can consider a main direction of a reaction. This is also expressed by the differentiation of substances into reactants and products. As already seen, metabolic reactions interact with each other, i.e., the product of one reaction is usually a reactant of another reaction. A metabolic path P= (R1, . . . , Rn) is a sequence of metabolic reactionswhere for all 1≤ i < nat least one product of reaction Ri is a reactant of reaction R I+1.

The metabolic network or metabolism of a particular cell or an organism is the complete network of metabolic reactions of this cell or 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 substance.

From a formal point of view a 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. Hyper-graphs can be represented by bipartite graphs.

Additionally to the nodes representing substances, the reactions are nodes (either labeled with the enzymes or with further nodes for enzymes) and edgesare binary relations connecting the substances of a reaction with the corresponding reaction node. This is a common modeling of metabolic pathways, e.g., for their simulation using Petri-nets .

For the analysis and visualization of metabolic pathways substances are often divided into two types :main substances and cosubstances. Co-substances are usually small or current metabolites, e.g., ATP, ADP, H2 O, NH3 and NADH. These substances normally transfer electrons or functional groups such as phosphate and amino groups]. Main substances are all other metabolites. However, this is not a global property but is given according to the reaction, and a

small metabolite such as ATP may be considered as main substance in a particular reaction. For visualization purposes this distinction is important as main substances and co-substances are often differently visually represented.

Here a metabolic pathway is modeled as directed bipartite graph G= (VS , V R, E) with Nodes u 1 , . . . , u N , w 1, . . . , w m∈V S representing substances, nodes v∈V R representing reactions (including the enzyme(s) catalyzing the reaction) and directed edges (u 1, v), . . . ,(un, v),(v, w1), . . . ,(v, w m) ∈ E representing the transformation of substances u 1, . . . , u N to substances w 1, . . . , w M by the reaction v. A reversible reaction does not contain backward edges as in some models for simulation purposes, instead this property of an reaction isrepresented by an attribute.

Another attribute is used to mark main and co-substances
There are several networks which are closely related to metabolic pathways or networks:
•Simplified metabolic network: A network which contains reactions, enzymes and main substances, but no co-substances.
•Metabolite network and simplified metabolite network: A network which consists only of substances (metabolites); in the simplified case only of main substances.
•Enzyme network: A network which consists only of the enzymes catalyzing the reactions

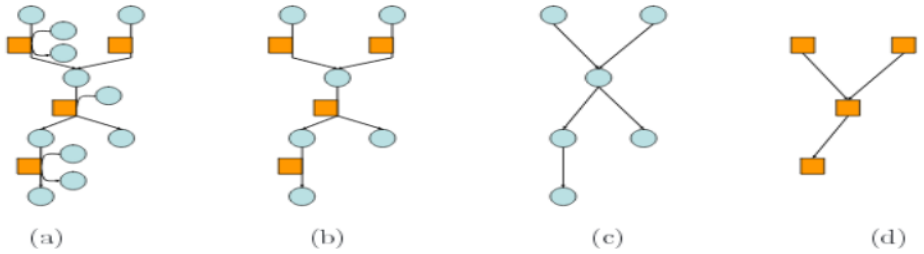


Figure :A metabolic network (a) and corresponding networks: (b) the Simplified metabolic network, (c) the simplified metabolite network and (d) the enzyme network.Circles denote metabolites and rectangles represent enzymes

These networks are not always directly associated with a metabolic network. For example,the metabolites in a metabolite network are not necessarily connected according to the reactions of a metabolic network, but can be established by correlation analysis of metabolite profiles . An enzyme network can be derived from a protein-protein interaction network. Again for relations in such a network a corresponding (connecting) substance cannot always be found within the metabolic network and protein-protein interaction networks may be undirected

Metabolic pathways can be derived from several databases such as EcoCyc,UM-BBD , and MetaCrop. For an overview and comparison between different databases.

Simplified metabolic networks are widely used, a popular example is the KEGG/LIGAND database.

1.Parts of reactions: The display of substances and enzymes is application and user-specific. Usually for main substances their name, structural formula or both should be shown. Co-substances should be displayed using their name or abbreviation and enzymes should be represented by their name or EC -number .

2.Reactions: The reaction arrow(s) should be shown from the reactants to the products with enzymes placed on one side of the reaction arrow and co-substances on the opposite side. The reversibility of a specific reaction should be clearly visible. For co-substances their temporal order, which depends on the reaction mechanism, is important, and they should be placed according to this order.

3.Pathways: The main direction of reactions (e.g., from top to bottom) should beclearly visible to express the temporal order of reactions.

There are important exceptions to the main direction used for the visualization of specific pathways,e.g., the citrate acid cycle or the fatty acid synthesis. The structure of these cyclic reaction chains should be emphasized. Such pathways are characterized by the continuous repetition of a reaction sequence in which the product of the sequence re-enters in the next loop as a reactant. There are two mechanisms. First, the reactant and the product of the reaction sequence are identical from loop to loop (e.g., citrate acid cycle)— a mechanism called a closed cycle. Second,the reactant of the reaction sequence varies slightly from the product (e.g., fattyacid cycle) - this is called an open cycle There are two established approaches to visualizing metabolic pathways and networks:force-directed and hierarchical layout methods.Force-directed methods are often used and several pathway analysis tools support such layout. Frequently they visualize not only metabolic and metabolite pathways, but different types of biochemical pathways and networks. Examples are PathwayAssist,PathDB and pathSCOUT. These tools use either their own implementations of well-known algorithms or are based on existing layout libraries. For example, VisANT contains an algorithm based on the layout method of Eades, and the method described by Rojdestvenski is based on the force-directed method of Kamada and Kawai . On the other hand Cytoscape uses the yFiles li-632.

Approaches for inference of biological networks – Boolean networks, Bayesian networks:

In order to understand complex biological networks and pathways, we need to investigate global structures instead of individual behaviors since there are interactions and associations between genes.

A Bayesian network is a directed acyclic graph (DAG) comprised of two components. The first component is comprised of nodes that correspond to a set of variables and a set of directed edges between variables with Markov properties. The second component describes a conditional distribution for each variable given its parents in the graph. Recently, Bayesian network models have been applied to analyze microarray expression and biological data.

A Bayesian model is a directed acyclic graph that explicitly establishes probabilistic relationships between network nodes. It describes the probability distribution governing a set of variables by specifying a set of conditional independence assumptions in addition to a set of conditional probabilities. Thus, the arcs between nodes not only indicate the regulatory relationships, but also describe the conditional dependencies (i.e. a family of joint probability distributions of all nodes) between them. In this way, the joint probability for any desired assignment of values to the network variables can be computed. More details of Bayesian networks can be found in relevant volumes. The variables in a Bayesian network can also be continuous, though Bayesian networks are classified as discrete variable models here.

Bayesian methods can be classified as either static or dynamic, depending on whether temporal expression profiles are used for the consideration of dynamics. As the directed network graphs are acyclic by definition, there can be no auto-regulation and no time-course regulation. With these limits, the static Bayesian methods cannot be used to infer regulatory networks with feedback loops. To consider the dynamic processes of networks, dynamic Bayesian methods were thus developed, and these can yield more accurate models. It should be noted, however, that the computational complexity increases significantly.

Modeling Bayesian networks involves two steps: model selection (or structure learning) and parameter learning. Model selection involves creation of the network structure, and parameter learning involves estimation of the probability values in the tables associated with each network node. The network structure can be inferred by employing a Bayesian scoring metric for model evaluation. For each possible model, the metric defines the score as the logarithm of the probability that the model correctly describes a given set of data. To avoid the overfitting problem, this likelihood is averaged over all parameter values that could possibly define the conditional probability distribution for each model. It should be noted that though Bayesian models have rich statistics and probability semantics, the learning network structure for such models is computationally expensive. For this purpose, some supplementary methods, such as

network decomposition (for dimension reduction), and Monte Carlo strategies using random sampling have been developed to enhance performance.

Boolean network models

The first type of GRN model assumes that genes only exist in discrete states. This approximation is usually implemented by Boolean variables in which the gene is in either on (active or expressed) or off (inactive or unexpressed). Boolean networks are easy to simulate and are therefore less computationally taxing, but it has been proven that Boolean networks are not able to capture certain system behaviors that can be captured by continuous models.

To construct a Boolean network, many computational methods can be adopted. If there is only qualitative knowledge available, literature-based methods are useful. In these methods, sentences in different documents are analyzed and compared to extract the relationships and links between genes. Alternatively, if experimental data are available, the Boolean network can be inferred from time course data. Two classes of methods are often used to infer Boolean networks. One is based on correlation measurement, where different methods are employed to extract information about gene relationships, and this information is then used to model the topological connections between genes. For example, the information-theoretic approach is commonly used to calculate the mutual information between genes, which can be used as a correlation measurement. The other class to infer Boolean networks is machine learning-based, in which genetic algorithm (GA) is the most common method for network modeling . The regulatory functions of network nodes and the relationships between nodes are encoded in the string representations often used in GA, and adaptive operators (such as crossover and mutation) can be used to create new solutions. In addition to using a linear encoding scheme in GA, a graph structure (generally described as a tree) can also be used to represent a Boolean network, with subsequent genetic programming (GP) used to infer the network structure directly.

Because traditional evolutionary algorithms are global search methods that mainly concentrate on exploring the solution space without considering local information, they cannot be optimized via local fine-tuning. Therefore, many enhanced methods have been proposed that combine GA with different local search techniques. These include taboo search, hill-climbing, simulated annealing and the simplex method; all exploit local information to determine promising directions in the search space. More recently, a new suite of intelligent population-based optimization techniques, known as swarm intelligence methods (including ant colony system and particle swarm optimization was proposed as an alternative to the traditional evolutionary algorithms. Some hybrid methods have now been proposed to effectively exploit the qualities to each of these two types of methods. It is now commonly agreed that a hybrid model comprising both evolutionary algorithm and swarm intelligent methods can lead to further improvements in performance.

Probabilistic Boolean network models

Though Boolean networks are easy to simulate at a reduced computational cost, they are generally considered to be unable to capture many important system behaviors. The dynamics of a Boolean network are deterministic, and they depend on the initial node states. These unexpected features make the Boolean network model lack realism. To overcome this problem, the probabilistic Boolean network (PBN), a similar but revised model, was proposed. Uncertainty is introduced into this newly developed model by providing each network node with multiple regulatory functions, each with a predefined probability. The function at each node is then determined probabilistically. At each time step, the regulatory function of each node is randomly selected from this pool of functions according to their given probabilities. With this randomizing effect, PBNs are stochastic, and the dynamics of the network are no longer deterministic. Any given set of initial node states can result in multiple subsequent network states.

The first step for generating a PBN model is to identify some candidate Boolean networks by the aforementioned methods. Once the candidates have been identified, the next step is to compile the functions that the different candidate networks ascribe to each node into n sets of predictor functions with certain designated probabilities. The details of this compilation process can be found in the literature. The main disadvantage of PBN, perhaps not surprisingly, is its increased computational complexity. The methods used in Boolean network inferencing can be revised and applied to PBN, but much more computation time is required to calculate the predicator probabilities. It thus becomes difficult to scale this approach to large networks. Some heuristic methods have been proposed to reduce the amount of computation. For example, Ivanov and Doherty developed two methods (mapping reduction and projection) and considered their effect on the original probability structure of a given PBN. Marshall *et al*. also proposed a method that separates the data sequence into sub-sequences, infers a Boolean network for each given sub-sequence, and then infers the probabilities of perturbation as well as the selection probabilities governing which network is to be selected.

Stoichiometric analysis:

One of the most important challenges in systems biology is to understand the functionality of metabolic networks that can be reconstructed from genomic and biochemical data for a wide variety of organisms. Current theories have different strengths and shortcomings in providing an integrated, predictive description of complex metabolic networks. For dynamic mathematical modeling of large-scale systems, often the necessary mechanistic detail and kinetic parameters are not available. In contrast, structure-oriented analyses only require the usually well-characterized network topology.

For this reason, Stoichiometric Network analysis (SNA) of biochemical reaction systems has become an important approach for understanding the functionality of metabolic networks.

In brief, at a very abstract level, cellular metabolism can be thought of as a complex network in which substances (nodes) are linked to each other via reactions (links) .

For analysis of these networks, two major classes of approaches can be distinguished:
In the narrower field of systems biology, most analysis methods rely on network stoichiometry, reaction reversibilities and potentially other constraints such as maximal pathway capacities . Applications of graph theoretical methods on metabolism only use the scheme of network connections as a starting point. The underlying principles of both classes of approaches will be described in the next sections after an introduction to the basic terms and concepts used in stoichiometric network analysis in general, and an overview of applications of this type of network analysis.

For a comparison of strengths and limitations of the four approaches to network analysis discussed herein, we will address the following issues:

Detection of functional pathways and cycles:

The original purpose of one class of approaches in SNA, namely of metabolic pathway analysis, is to provide a formalism for the detection of functional pathways even in large networks. Thus identified pathways include 'classical' biochemical pathways such as glycolysis. Pathway analysis may also lead to the identification of new hypothetical routes that only emerge in the context of the large, complex network, and are of importance, for instance, if one is interested in finding a route from a specific substrate or metabolite to a product or biomass, respectively. The identification of 'futile cycles', those chains of reactions that involve only a net consumption of energy, can help to recognize potential energy wasting routes. Cycles without any net energy consumption enable the identification of thermodynamic inconsistencies. Intuitively, it is clear that a route (or pathway) should be a set of connected reactions. A more crucial task is to establish a notion of 'meaningful' pathways that covers physiological as well as biotechnological aspects and, hence, has to be handled by the theoretical approaches.

Identification of optimal / sub-optimal operating modes:
The evaluation of, for instance, maximal product yields in terms of the moles of product generated per mole of substrate has clear relevance for biotechnological applications. Stoichiometrically derived yields may give indicators of the maximal efficiency of engineered organisms. The identification of alternative optimal pathways, or of sub-optimal pathways can, however, be of equal importance with regard to the feasibility of genetic engineering approaches
.

Assessing the importance of single reactions:

A prominent application of network analysis is to determine the importance of single reactions for the overall systems performance. This might include estimates of the relative importance of a reaction under different growth conditions. A particular application concerns the study of knockout mutations. For instance, it is of medical relevance to predict the effects of enzyme deficiencies, which may be causative for human diseases. Furthermore, this type of investigation should allow for a differentiation between essential and non-essential genes. The reliability of the predictions of gene deletions that are generated by the various theoretical approaches will therefore be a central evaluation criterion.

Analysis of correlated reactions:

Reactions that always have to operate together, for instance when being involved in an un-branched linear pathway are likely to be coregulated. Typical examples are many biosynthetic pathways, for example, for the production of amino acids. Similarly, if reactions never appear together in metabolic pathways at the same time, this is indicative of differential regulations, for instance, to establish qualitatively different operation modes of the network depending on the environmental conditions. The identification of such groups of reactions therefore gives hints on the regulation of metabolic networks that can be exploited to understand, and possibly predict, features of the corresponding regulatory networks.

Determination of pathway lengths:

The length of a metabolic pathway is the number of reaction it comprises. This quantity might be of interest, because it gives an indication of the amount of cellular resources that is needed to establish a pathway, e.g. to provide for the necessary enzymes. Moreover, the distribution of pathway lengths can be used to assess the complexity of a given network, or to characterize seemingly similar networks for different organisms.

Analysis of network functionality:

This application refers, for instance, to the study of the effects of adding reactions to, or deleting reactions from a given network. As such, it is closely related to the analysis of the importance of single reactions. In addition to such investigations, one may be interested in investigating how (additional) constraints on the reversibility of reactions influence the set of possible pathways in the network. Moreover, it can be important to assess the effect of newly introduced genes with respect to the functional capabilities - including potential, un-anticipated side-effects – of the modified metabolic network. In practice, such techniques could be applied to find suitable targets for the addition or removal of metabolic genes.

Investigations into network flexibility and robustness:

Robustness is generally defined as the (relative) insensitivity of a system to changes in its parameters. Flexibility stresses the capability of a system to switch between different states or functional modes. In the case of structural analysis of metabolic networks, both concepts can be regarded as being equivalent, because a metabolic system should be able to tolerate changes, for instance in the set of enzymes present, once it provides for alternative pathways that can operate even when a specific reaction is not functional.

This, in turn, implies that for the analysis of robustness in metabolic systems, it will be necessary to investigate the set of all possible behaviors of the system . It is not sufficient to find a single (or no) functional pathway, as is the case for a judgment on whether a metabolic gene is essential or not.

Stoichiometric analysis (also termed structural analysis) can be applied to biochemical reaction networks where the stoichiometries of the reactions are known. The stoichiometry matrix N is then used to derive conservation relationships, enzyme subsets (or reaction correlation coefficient φ), elementary modes, and so on. An advantage of structural analysis is that it requires no information on concentrations of species or on volumes of compartments, nor any knowledge of kinetic parameters of enzymatic or nonenzymatic reactions. It can be performed on large-scale metabolic networks and even on the genome-scale networks mentioned

Every model that consists of a list of biochemical reactions is also represented through its stoichiometry matrix N. The elements of N, the stoichiometric coefficients of the reactions, relate the rate of change of the concentrations of each network component to the reaction rates of the reactions that produce or consume the component:

$$\frac{dX}{dt} = Nv(X, p)$$

Knowledge about the structure of a metabolic[2] network, reflected by N and details of its reactions' reversibility are sufficient to perform a stoichiometric analysis on this reaction network. It is important to realize that the steady state assumption $Nv = 0$ is the premise for most of the concepts in structural modeling but not for all (e.g., the conservation relationships hold true at every point in time).

By analyzing N, one can determine a variety of model properties that could not be found by any other means:

1. Conserved moieties: Sets of internal metabolites with a fixed total concentration metabolites that contribute to such a moiety are not free to take on every concentration but are dependent on the concentrations of the other metabolites contributing.
2. Enzyme subsets: Groups of enzymes that operate jointly in fixed flux proportions at steady state
3. Elementary modes: Minimal sets of reactions that can operate at steady state with all irreversible reactions proceeding in the appropriate direction ; the concept of elementary modes provides a mathematical tool to define and comprehensively describe all metabolic routes that are stoichiometrically feasible for a certain reaction network; Schuster, et al. gave an overview, a calculation algorithm, and an example for this concept.

Graph Theory and Analysis of Biological Data in Computational Biology

The theory of complex networks plays an important role in a wide variety of disciplines, ranging from communications to molecular and population biology. The focus of this article is on graph theory methods for computational biology. We'll survey methods and approaches in graph theory, along with current applications in biomedical informatics. Within the fields of Biology and Medicine, potential applications of network analysis by using graph theory include identifying drug targets, determining the role of proteins or genes of unknown function. There are several biological domains where graph theory techniques are applied for knowledge extraction from data. We have classified these problems into several different domains, which are described as follows.

- Modeling of bio-molecular networks. It presents modeling methods of bio-molecular networks, such as protein interaction networks, metabolic networks, as well as transcriptional regulatory networks.
- Measurement of centrality and importance in bio-molecular networks. To identify the most important nodes in a large complex network is of fundamental importance in computational biology. We'll introduce several researches that applied centrality measures to identify structurally important genes or proteins in interaction networks and investigated the biological significance of the genes or proteins identified in this way.
- Identifying motifs or functional modules in biological networks. Most important biological processes such as signal transduction, cell-fate regulation, transcription, and translation involve more than four but much fewer than hundreds of proteins or genes. Most relevant processes in biological networks correspond to the motifs or functional modules. This suggests that certain functional modules occur with very high frequency in biological networks and be used to categories them.
- Mining novel pathways from bio-molecular networks. Biological pathways provide significant insights on the interaction mechanisms of molecules. Experimental validation of identification of pathways in different organisms in a wet-lab environment requires monumental amounts of time and effort. Thus, there is a need for graph theory tools that help scientists predict pathways in bio-molecular networks.

The concept of a graph

The concept of a graph is fundamental to the material to be discussed in this chapter. A graph G consists of a set of vertices $V(G)$ and a set of edges $E(G)$. In a simple graph, two of the vertices in G are linked if there exists an edge $(v_i , v_j)E(G)$ connecting the vertices v_i and v_j in graph G such that v_i $V(G)$ and v_j $V(G)$. The number of vertices will be denoted by $V(G)$, and the set of

vertices adjacent to a vertex v_i is referred to as the neighbors of v_i , $N(v_i)$. The degree of a vertex v_i is the number of edges with which it is incident, symbolized by $d(v_i)$. Two graphs, G_1 and G_2 , are said to be isomorphic (G_1 G_2) if a one-to-one transformation of V_1 onto V_2 effects a one-to-one transformation of E_1 onto E_2 . A subgraph G' of a graph G is a graph whose set of vertices and set of edges satisfy the relations: $V(G')V(G)$ and $E(G')$ $E(G)$, and if G' is a subgraph of G, then G is said to be a supergraph of G'. The line graph $L(G)$ of an undirected graph G is a graph such that each vertex in $L(G)$ indicates an edge in G and any pairs of vertices of $L(G)$ are adjacent if and only if their corresponding edges share a common endpoint in G.

Directed and undirected graphs

A graph may be *undirected*, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be *directed* from one vertex to another. Formally, a finite directed graph, G, consists of a set of vertices or nodes, $V(G) = \{v_1 ,...,v_n \}$, together with an edge set, $E(G)$ $V(G)V(G)$. Intuitively, each edge (u, v) $E(G)$ can be thought of as connecting

the starting node u to the terminal node v. An undirected graph, G, also consists of a vertex set, $V(G)$, and an edge set $E(G)$. However, there is no direction associated with the edges in this case. Hence, the elements of $E(G)$ are simply two element subsets of $V(G)$, rather than ordered pairs as directed graphs. As with directed graphs, we shall use the notation uv (or vu as direction is unimportant) to denote the edge $\{u, v\}$ in an undirected graph. For two vertices, u, v, of an undirected graph, uv is an edge if and only if vu is also an edge. We are not dealing with multi-graphs, so there can be at most one edge between any pair of vertices in an undirected graph. That is, we are discussing the simple graph. A simple graph is an undirected graph that has no loops and no more than one edge between any two different vertices. In a simple graph the edges of the graph form a set and each edge is a pair of *distinct* vertices. The number of vertices n in a directed or undirected graph is the size or order of the graph.

Node-degree and the adjacency matrix

For an undirected graph G, we shall write $d(u)$ for the degree of a node u in $V(G)$. This is simply the total number of edges at u. For the graphs we shall consider, this is equal to the number of neighbors of u, $d(u) = |N(u)|$. In a directed graph G, the *in-degree*, $d^+(u)$ (out-degree, $d^-(u)$) of a vertex u is given by the number of edges that terminate (or start) at u. Suppose that the vertices of a graph (directed or undirected) G are ordered as $v_1,..., v_n$. Then the adjacency matrix, A, of G is given by

$$a_{ij} = \begin{cases} 1 if u_i v_j \in E(G) \\ 0 if u_i v_j \notin E(G) \end{cases}$$

Path, path length and connected graph

Let u, v be two vertices in a graph G. Then a sequence of vertices $u = v_1 , v_2 ,..., v_k = v$, such that for $i = 1,..., k$-1, is said to be a path of length k-1 from u to v. The geodesic distance, or simply distance, $d(u, v)$, from u to v is the length of the shortest path from u to v in G. If no such path exists, then we set $d(u, v) = 1$. If for every pair of vertices, (u, v), in graph G, there is some path from u to v, then we say that G is connected.

Modeling of bio-molecular networks

Several classes of bio-molecular networks have been studied: Transcriptional regulatory networks, protein interaction network, and metabolic networks. In Biology, transcriptional regulatory networks and metabolic networks would usually be modeled as directed graphs. For instance, in a transcriptional regulatory network, nodes would represent genes with edges denoting the transcriptional relationships between them. This would be a directed graph because,

if gene A regulates gene B, then there is a natural direction associated with the edge between the corresponding nodes, starting at A and terminating at B. In recent years, attentions have been focused on the protein-protein interaction networks of various simple organisms. These networks describe the direct physical interactions between the proteins in an organism's proteome and there is no direction associated with the interactions in such networks. Hence, PPI networks are typically modeled as undirected graphs, in which nodes represent proteins and edges represent interactions. In next sections, we individually introduce these bio-molecular networks.

Transcriptional regulatory networks

Transcriptional regulatory networks describe the regulatory interactions between genes. Here, nodes correspond to individual genes and a directed edge is drawn from gene A to gene B if A positively or negatively regulates gene B. Networks have been constructed for the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* and are maintained in databases such as RegulonDB and EcoCyc .Such networks are usually constructed through a combination of high-throughput genome location experiments and literature searches. Many types of gene transcriptional regulatory related approaches have been reported in the past. Their nature and composition are categorized by several factors: considering gene expression values, the causal

relationship between genes, e.g. with Bayesian analysis or Dynamic Bayesian Networks, and the time domain e.g. discrete or continuous time. One of the limitations of graph theory applications in analyzing biochemical networks is the static quality of graphs. Biochemical networks are dynamical, and the abstraction to graphs can mask temporal aspects of information flow. The nodes and links of biochemical networks change with time. Static graph representation of a system is, however, a prerequisite for building detailed dynamical models. Most dynamical modeling approaches can be used to simulate network dynamics while using the graph representation as the skeleton of the model. Modeling the dynamics of biochemical networks provides closer to reality recapitulation of the system's behavior *in silico*, which can be useful for developing more quantitative hypotheses.

3.3. Protein interaction networks

Understanding protein interactions is one of the important problems of computational biology. These protein-protein interactions (PPIs) networks are commonly represented by undirected graph format, with nodes corresponding to proteins and edges corresponding to protein-protein interactions. The volume of experimental data on protein-protein interactions is rapidly increasing by high-throughput techniques improvements which are able to produce large batches of PPIs. For example, yeast contains over 6,000 proteins, and currently over 78,000 PPIs have been identified between the yeast proteins, with hundreds of labs around the world adding to this list constantly. Humans are expected to have around 120000 proteins and around 10^6 PPIs. The relationships between the structure of a PPI network and a cellular function are waited to be explored. Large-scale PPI networks have been constructed recently using high-throughput

approaches such as yeast-2-hybrid screens or mass spectrometry techniques to identify protein interactions.

Vast amounts of PPI related data that are constantly being generated around the world are being deposited in numerous databases. Data on protein interactions are also stored in databases such as the database of interacting proteins (DIP). We briefly mention the main databases, including nucleotide sequence, protein sequence, and PPI databases. The largest nucleotide sequence databases are EMBL, DDBJ, and GenBank. They contain sequences from the literature as well as those submitted directly by individual laboratories. These databases store information in a general manner for all organisms. Organism specific databases exist for many organisms. For example, the complete genome of yeast and related yeast strains can be found in *Saccharomyces* Genome Database (SGD). FlyBase contains the complete genome of the fruit fly *Drosophila melanogaster*. It is one of the earliest model organism databases. Ensembl contains the draft

human genome sequence along with its gene prediction and large scale annotation. SwissProt and Protein Information Resource (PIR) are two major protein sequence databases. SwissProt maintains a high level of annotations for each protein including its function, domain structure, and post-translational modification information.

Understanding interactions between proteins in a cell may benefit from a model of a PPIs network. A full description of protein interaction networks requires a complex model that would encompass the undirected physical protein-protein interactions, other types of interactions, interaction confidence level, or method and multiplicity of an interaction, directional pathway information, temporal information on the presence or absence of PPIs, and information on the strength of the interactions. This may be achieved by designing a scoring function and assigning weights to nodes and edges of a PPIs network.

3.4. Metabolic networks

Metabolic networks describe the bio-chemical interactions within a cell through which substrates are transformed into products through reactions catalysed by enzymes. Metabolic networks generally require more complex representations, such as hyper-graphs, as reactions in metabolic networks generally convert multiple inputs into and multiple outputs with the help of other components. An alternative is a weighted bipartite graph to reduce representation for a metabolic network. In such graphs, two types of nodes are used to represent reactions and compounds, respectively. The edges in a weighted bipartite graph connect nodes of different types, representing either substrate or product relationships. These networks can represent the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. Metabolic networks are complex. There are many kinds of nodes (proteins, particles, molecules) and many connections (interactions) in such networks. Even if one can define sub-networks that can be meaningfully described in relative isolation, there are always connections from it to other networks. As with protein interaction networks, genome-scale

metabolic networks have been constructed for a variety of simple organisms including *S. cerevisiae* and *E. coli*, and are stored in databases such as the KEGG or BioCyc databases. A common approach to the construction of such networks is to first use the annotated genome of an organism to identify the enzymes in the network and then to combine bio-chemical and genetic information to obtain their associated reactions (Kauffman et al., 2000;. While efforts have been made to automate certain aspects of this process, there is still a need to validate the networks generated automatically manually against experimental biochemical results. For metabolic networks, significant advances have also been made in modelling the reactions that take place on

such networks. The overall structure of a network can be described by several different parameters. For example, the average number of connections a node has in a network, or the probability that a node has a given number of connections. Theoretical work has shown that different models for how a network has been created will give different values for these parameters. The classical random network theory (Erdös & Renyi, 1960) states that given a set of nodes, the connections are made randomly between the nodes. This gives a network where most nodes have the same number of connections. Recent research has shown that this model does not fit the structure found in several important networks. Instead, these complex networks are better described by a so-called scale-free model where most nodes have only a few connections, but a few nodes (called hubs) have a very large number of connections. Recent work indicates that metabolic networks are examples of such scale-free networks. This result is important, and will probably lead to new insights into the function of metabolic and signaling networks, and into the evolutionary history of the networks. Robustness is another important property of metabolic networks. This is the ability of the network to produce essentially the same behavior even when the various parameters controlling its components vary within considerable ranges. For example, recent work indicates the segment polarity network in the *Drosophila* embryo can function satisfactorily with a surprisingly large number of randomly chosen parameter sets (von Dassow *et a.l*, 2000). The parameters do not have to be carefully tuned or optimized. This makes biological sense, which means a metabolic network should be tolerant with respect to mutations or large environmental changes.

Another important emerging research topic is to understand metabolic networks in term of their function in the organism and in relation to the data we already have. This requires combining information from a large number of sources, such as classical biochemistry, genomics, functional genomics, microarray experiments, network analysis, and simulation. A theory of the cell must combine the descriptions of the structures in it with a theoretical and computational description of the dynamics of the life processes. One of the most important challenges in the future is how to make all this information comprehensible in biological terms. This is necessary in order facilitate the use of the information for predictive purposes to predict what will happen after given some specific set of circumstances. This kind of predictive power will only be reached if the complexity of biological processes can be handled computationally.

4. Measurement of centrality and importance in bio-molecular networks

Biological function is an extremely complicated consequence of the action of a large number of different molecules that interact in many different ways. Genomic associations between genes reflect functional associations between their products (proteins). Furthermore, the strength of the genomic associations correlates with the strength of the functional associations. Genes that frequently co-occur in the same operon in a diverse set of species are more likely to physically

interact than genes that occur together in an operon in only two species, and proteins linked by gene fusion or conservation of gene order are more likely to be subunits of a complex than are proteins that are merely encoded in the same genomes.

Other types of associations have been used for network studies, but these focus on certain specific types of functional interactions, like subsequent enzymatic steps in metabolic pathways, or physical interactions. Elucidating the contribution of each molecule to a particular function would seem hopeless, had evolution not shaped the interaction of molecules in such a way that they participate in functional units, or building blocks, of the organism's function. These building blocks can be called modules, whose interactions, interconnections, and fault-tolerance can be investigated from a higher-level point of view, thus allowing for a synthetic rather than analytic view of biological systems. The recognition of modules as discrete entities whose function is separable from those of other modules introduces a critical level of biological organization that enables *in silico* studies.

Intuitively, modularity must be a consequence of the evolutionary process. Modularity implies the possibility of change with minimal disruption of function, a feature that is directly selected for. However, if a module is essential, its independence from other modules is irrelevant unless, when disrupted, its function can be restored either by a redundant gene or by an alternative pathway or module.

Furthermore, modularity must affect the evolutionary mechanisms themselves, therefore both robustness and evolvability can be optimized simultaneously. The analysis of these concepts requires both understanding of what constitutes a module in biological systems and tools to recognize modules among groups of genes. In particular, a systems view of biological function requires the development of a vocabulary that not only classifies modules according to the role they play within a network of modules and motifs, but also how these modules and their interconnections are changed by evolution, for example, how they constitute units of evolution targeted directly by the selection process. The identification of biological modules is usually based either on functional or topological criteria.

For example, genes that are co-expressed or coregulated can be classified into modules by identifying their common transcription factors, while genes that are highly connected by edges in a network form clusters that are only weakly connected to other clusters. From viewpoint of

evolutionary, genes that are inherited together but not with others often form modules. However, the concept of modularity is not at all well defined. For example, the fraction of proteins that constitutes the core of a module and that is inherited together is small, implying that modules are fuzzy but also flexible so that they can be rewired quickly, allowing an organism to adapt to novel circumstances.

A set of data is provided by genetic interactions, such as synthetic lethal pairs of genes or dosage rescue pairs, in which a knockout or mutation of a gene is suppressed by over-expressing another gene. Such pairs are interesting because they provide a window on cellular robustness and modularity brought about by the conditional expression of genes. Indeed, the interaction between genes epistasis has been used to successfully identify modules in yeast metabolic genes. However, often interacting pairs of genes lie in alternate pathways rather than cluster in functional modules. These genes do not interact directly and thus are expected to straddle modules more often than lie within one .

In silico evolution is a powerful tool, if complex networks can be generated that share the pervasive characteristics of biological networks, such as error tolerance, small-world connectivity, and scale-free degree distribution. If furthermore each node in the network represents a simulated chemical or a protein catalyzing reactions involving these molecules, then it is possible to conduct a detailed functional analysis of the network by simulating knockdown or over-expression experiments.

This functional datum can then be combined with evolutionary and topological information to arrive at a more sharpened concept of modularity that can be tested in vitro when more genetic data become available. Previous work on the in silico evolution of metabolic, signaling, biochemical, regulatory, as well as Boolean , electronic, and neural networks has begun to reveal how network properties such as hubness, scaling, mutational robustness as well as short pathway length can emerge in a purely Darwinian setting. In particular, *in silico* experiments testing the evolution of modularity both in abstract and in simulated electronic networks suggest that environmental variation is key to a modular organization of function. These networks are complex, topologically interesting, and function within simulated environments with different variability that can be arbitrarily controlled.

Identifying motifs or functional modules in biological networks

Biological systems viewed as networks can readily be compared with engineering systems, which are traditionally described by networks such as flow charts. Remarkably, when such a comparison is made, biological networks and engineered networks are seen to share structural principles such as modularity and recurrence of circuit elements .

Both biological systems function and engineering are organized with modularity. Engineering systems can be decomposed into functional modules at different levels, subroutines in software and replaceable parts in machines. In the case of biological networks, although there is no consensus on the precise groups of genes and interactions that form modules, it is clear that they possess a modular structure. Alon proposed a working definition of a module based on comparison with engineering.

A *module* in a network is a set of nodes that have strong interactions and a common function. A module has defined input nodes and output nodes that control the interactions with the rest of the network.

Various basic functional modules are frequently reused in engineering and biological systems. For example, a digital circuit may include many occurrences of basic functional modules such as multiplexers and so on. Biology displays the same principle, using key wiring patterns again and again throughout a network. For instance, metabolic networks use regulatory circuits such as feedback inhibition in many different pathways. Besides basic functional modules, recently a small set of recurring circuit elements termed *motifs* have been discovered in a wide range of biological and engineering networks. Motifs are small (about 3 or 4 nodes) sub-graphs that occur significantly more frequently in real networks than expected by chance alone, and are detected purely by topological analysis. This discover kindled a lot of interest on organization and function of motifs, and many related papers were published in recent years. The observed over-representation of motifs has been interpreted as a manifestation of functional constraints and design principles that have shaped network architecture at the local level. Some researchers believe that motifs are basic building blocks that may have specific functions as elementary computational circuits. Although motifs seem closely related to conventional building blocks, their relation lacks adequate and precise analysis, and their method of integration into full networks has not been fully examined. Further, it is not clear what determines the particular frequencies of all possible network motifs in a specific network. Mining novel pathways from bio-molecular networks

In the studying organisms at a systems level, biologists recently mentioned the following questions:

(1) Is there a minimal set of pathways that are required by all organisms?

(2) To what extent are the genomic pathways conserved among different species?

(3) How are organisms related in terms of the distance between pathways rather than at the level of DNA sequence similarity?

At the core of such questions lies the identification of pathways in different organisms. However, experimental validation of an enormous number of possible candidates in a wet-lab environment requires monumental amounts of time and effort. Thus, there is a need for comparative genomics tools that help scientists predict pathways in an organism's biological network. Due to the complex and incomplete nature of biological data, at the present time, fully automated computational pathway prediction is excessively ambitious. A metabolic pathway is a

set of biological reactions where each reaction consumes a set of metabolites, called substrates, and produces another set of metabolites, called products. A reaction is catalyzed by an enzyme (or a protein) or a set of enzymes. There are many web resources that provide access to curated as well as predicted collections of pathways, e.g., KEGG, EcoCyc, Reactome, and PathCase.

Work to date on discovering biological networks can be organized under two main titles:

(i) Pathway Inference, and
(ii) Whole-Network Detection.

Even with the availability genomic blueprint for a living system and functional annotations for its putative genes, the experimental elucidation of its biochemical processes is still a daunting task. Though it is possible to organize genes by broad functional roles, piecing them together manually into consistent biochemical pathways quickly becomes intractable.

A number of metabolic pathway reconstruction tools have been developed since the availability of the first microbial genome, Haemophilus influenza. These include PathoLogic, MAGPIE and WIT and PathFinder. The goal of most pathway inference methods has generally been to match putatively identified enzymes with known or reference pathways. Although reconstruction is an important starting point for elucidating the metabolic capabilities of an organism based upon prior pathway knowledge, reconstructed pathways often have many missing enzymes, even in essential pathways.

The issue of redefining microbial biochemical pathways based on missing proteins is important since there are many examples of alternatives to standard pathways in a variety of organisms. Moreover, engineering a new pathway into an organism through heterologous enzymes also requires the ability to infer new biochemical routes. With more genomic sequencing projects underway and confident functional characterizations absent for many of the genes, automated strategies for predicting biochemical pathways can aid biologists inunraveling the complex processes in living systems. At the same time, pathway inference approaches can also help in designing synthetic processes using the repertoire biocatalysts available in nature.