Chapter 2

Introduction to Metabolic Network Reconstruction
Cells consist of different types of biomolecules and these molecules are discovered with the help of modern molecular biology techniques[20]. The process of studying the properties and roles of individual molecules leads to produce a large amount of useful knowledge. The modern cellular biology is usually understood and studied based on central dogma, which describes how information encoded in genes (genotype) is reflected at the level of cellular function and state (phenotype).

\[ \text{flux} = f(M', M', \text{Enzyme}^k) \]

![Figure 2.1.](a) Central dogma in molecular biology. DNA replication can be thought of as information flow from genome to genome. Information coded in genes is transformed into proteins via transcription and translation. (b) Enzymes catalyze chemical transformation of metabolites. The rate of enzyme catalyzed reaction is not only a function of enzyme availability and properties, but also concentration of substrates and products. Several of such enzymatic steps constitute a metabolic network where products of some reactions serve as substrates for other reactions, thus creating an interconnected reaction network. The overall function...
of a metabolic network can then be viewed as utilizing environmentally available nutrients to generate energy and building molecules for growth and maintenance of the cell.

The genotypic information is stored in the genes in the form of sequence order comprised of four nucleotides (Adenine, Thymine, Guanine and Cytosine) and triplet codes for one of the 20 amino-acids. The several of the amino acids joined together constitute a protein. Thus gene codes generate a specific protein via transcription (DNA to mRNA) and translation (mRNA to protein). Apart from the manifestation of the classical central dogma, of DNA to mRNA and mRNA to protein pathway, it is an important to note that the genes in the DNA are themselves regulated by the presence or absence of certain proteins [21]. Furthermore, many of the interactions going on in the cell occur entirely at the protein level, which can cause significant discrepancies between protein and mRNA levels. Protein which plays an important role as catalyst for chemical and physical transformation of various chemical substances in cellular process is usually referred to as enzymes [22]. Enzymes create a network of reactions where substrates available in the environment are broken down to generate energy and building block molecules[23]. There are many networks present in cell like reaction networks and understanding such networks gives many ideas of working mechanisms of cells.

In the post genomic era, the key aim for researchers is to understand the interactions between various biological processes within the living cell. Biological networks can be broken down into genetic, protein and metabolic networks [24]. Various types of interaction networks such as protein–protein interaction, metabolic, signalling and transcription regulatory networks are emerging from the various bimolecular interactions. None of these networks is independent, instead they form a 'network of networks' that is responsible for the behaviour of the cell. A major challenge of contemporary biology is to embark on an integrated theoretical and experimental programme to map out, understand and model in quantifiable terms the topological and dynamic properties of the various networks that control the behavior of the cell[7].

Currently, most studies trying to infer expression mechanisms from cell state data use mRNA levels, since they are easiest to measure because of the availability of large-scale gene expression technologies[25]. Large-scale protein measurements are found to be incomplete, when there is a requirement for measuring the highest abundant proteins. However, it can be supplemented with more exact measurements of individual proteins, which are known to play an important role in cells. For measuring gene expression data in a process involving cellular metabolism an effort should be
made to quantify metabolites with the help of important metabolite and nutrient levels at particular condition. For example, in the diauxic shift, yeast switch from anaerobic fermenting glucose to form ethanol to aerobically consume this ethanol when the glucose is exhausted. Analysis of the diauxic shift using DNA microarrays revealed that many genes change their expression during this switch in nutrient utilization. The interrelationship between specific gene expression changes and metabolic changes however remains unclear primarily because the temporal course of changes in metabolite concentrations and enzymatic steps has not been studied adequately.

It is possible to uncover the patterns in these networks that follow a common transcriptional response by using such information on the network topology from genome-scale protein & metabolic reconstruction. The gene expression pattern is a strong indicator of perturbations induced in specific parts of the metabolic network. Any changes due to the perturbation are then propagated through the biological network because of the highly connected nature of metabolism [26]. The biological networks can be organized into three broad groups (Fig. 4.2), (1) gene regulatory (2) protein networks and (3) metabolic networks.

![Network overview](image)

*Figure 2.2: Network overview. The figure illustrates the three main network layers, metabolic, protein and gene (TF = Transcription Factors)*
Chapter 2 Introduction to Metabolic Network Reconstruction

The metabolic, protein and gene regulatory networks each have a characteristic mode of operation and differ by the employed molecular mechanisms and the operating time scales. In general metabolic networks operate on the smallest time scale, followed by protein networks and gene regulatory networks. Getting idea of protein networks, gene regulatory networks and metabolic network gives more understanding of biological network.

2.1.1 Gene Regulatory Networks

The gene network inference techniques are sophisticated and sensitive with generation of large amount of data. Measuring gene expression time series has the fastidious feature of yielding lots of data. However, all the data points in a single time series tend to be about a single dynamic process in the cells. The data are related to the surrounding of particular time points. A data set of ten expression measurements under different environmental conditions will actually contain more information than a time series of ten data points on a single phenomenon. The advantage of the time series is that it can provide crucial insights into the dynamics of the process for gene networks.

**Figure 2.3:** A simplified scheme of intracellular regulation circuits with gene expression (box I), signal transduction (box II), and metabolic (box III) processes. It has been viewed as a bipartite graph consisting of molecular entities (ellipses) and regulatory events (rectangles). Molecular entities are genes, proteins, modified proteins, their complexes, peptides, and metabolites. Regulatory events are gene expression (GE), protein modification and complex formation (PM, CF), protein degradation (PD) into short peptides, and metabolic reactions (MR). Mass flow is shown with fat arrows. Dashed arrows represent the catalytic action of
molecular entities on the corresponding regulatory event; catalysts are themselves not consumed in the corresponding process.

In a genetic network, genes are represented by nodes and their interaction is given by edges. Most of the genetic networks are boolean in nature, thereby the weights on the edges can be assigned a weight of either 0 or 1. Multiple data sets of each type of data are likely to be needed to open up the regulatory interactions of the genes. Indeed, to correctly infer the regulation of a single gene, we need to observe the expression of that gene under many different combinations of expression levels of its regulatory inputs. This implies a wide variety of different environmental conditions and perturbations. Gene regulation is a general name for a number of sequential processes, the most well-known and understood being transcription and translation, which controls the level of a gene's expression, and ultimately results in specific quantity of a target protein[27].

A gene regulation system consists of genes, cis-elements, and regulators. The regulators are most often proteins, called transcription factors, but small molecules, like RNAs and metabolites, sometimes also participate in the overall regulation. The interactions and binding of regulators to cis-elements in the cis-region of genes control the level of gene expression during transcription. The cis-regions serve to aggregate the input signals, mediated by the regulators, and thereby affect a very specific gene expression signal. The genes, regulators, and the regulatory connections between them, together with an interpretation scheme form gene networks[28].

Gene regulation networks control gene expression in cells. The expression of one gene can be controlled by the gene product of another gene. Thus, a directed graph in which the vertices are genes and the directed edges represent control can be used to model these networks. Only fragments of these networks have been modeled quantitatively and with the help of assigning rate laws to every step. Recent advances in data collection and analysis made it possible to elucidate large scale gene regulation networks. It has been found that in this network type, certain motifs such as feed-forward loops or single input modules are overrepresented when compared with randomly generated networks. Through these investigations it was possible to define the "basic computational elements" of biological networks.

The collection of such elements information in a computer-readable form is a prerequisite for making use of these data for analysis. A suitable database structure as well as the quality and consistence of the stored data are of great importance in research community. Presently, many databases on different aspects of gene regulation are available. Updated information on them
appears every year in the Nucleic Acids Research database issue [10]. Representative examples of such databases and knowledge bases containing information about regulatory interactions are aMAZE, EcoCyc, GeneNet, KEGG, RegulonDB, Reactom, TRANSPATH, and TRANSFAC.

The aMAZE database focuses on information about genetic regulation, biochemical pathways, signal transductions, and aims on modeling the systems of catalyzed chemical reactions by means of simulation software package such as GEPASI (http://www.gepasi.org/) and databases like KEGG and EcoCyc which provide rich information on metabolism, metabolic pathways, as well as on signal transduction, gene regulation, and cellular processes [29]. GeneNet database provides information on structure and functional organizations of gene networks and metabolic and signal transduction pathways. Reactome is a rich resource on pathway information and reactions in human biology. TRANSPATH provides encyclopedic information about the intracellular signal transduction pathways and offers molecular details of the signal flow from the cell surface into the nucleus. TRANSPATH and accompanying tools can be used for data visualization and modeling, as well as for the analysis of gene expression data[30]. TRANSFAC is the database on many aspects of transcription regulation in eukaryotes[31] [32]. It presents the largest archive of eukaryotic transcription factors, their genomic binding sites, and DNA-binding profiles. The gene regulatory network contains thousands of vertices and edges and generates unipartite and bipartite networks. These networks can be analyzed with a wide range of algorithms for network analysis.

Recently, a new database EndoNet on the endocrine cell-cell signaling in human has been developed, which enables the analysis of intercellular regulatory pathways. The data model includes two classes of components and can be viewed as a bipartite directed graph. One class represents the signaling molecules (hormones, cytokines, growth factors, or other messengers), which are secreted by defined donor cells. The other class represents the acceptor or target cells expressing the corresponding receptors. The identity and anatomical environment of cell types, tissues, and organs are provided through references to the CYTOMER ontology. CYTOMER is a relational database on tissues, cell types, physiological systems, and developmental stages in a human organism. In EndoNet, all entries for signaling molecules and receptors are provided with links to external information resources that include TRANSPATH database on intracellular signal transduction. That promises to greatly increase the scale of molecular regulatory circuits available for analyses. That might help to bridge the gap between known genotypes and their molecular and clinical phenotypes in the area of medical research and its applications[33].
Chapter 2 Introduction to Metabolic Network Reconstruction

Figure 2.4: The gene regulatory network in Hepatocellular carcinoma (HCC). Different colors represent nodes in different network modules. Size of nodes is proportional to the out-degree of nodes. Black edges represent regulations in the core GRN, and the width of the edges in the core GRN are proportional to the edge-betweenness values calculated from the global GRN.

2.1.2 Protein Interaction Networks

Proteins control and mediate the vast majority of biological processes in a living cell. They act as catalysts, transport or store other molecules, provide mechanical strength, confer immunity, transmit signals, and control growth and development. Proteins are polymers of amino acids, covalently linked through peptide bonds into a chain. The function of a protein is determined by its three
dimensional structure, which in turn is defined by its amino acid sequence. Monitoring the alterations in expression of specific proteins in response to changing environments or across different developmental stages of a given organism has provided substantial insight into the complex regulatory networks controlling life. Proteins operate entirely on the basis of interactions with other molecules such as low molecular weight compounds, lipids, nucleic acids, or other proteins. The functionally active form of a protein is rarely its monomer, that is, the single molecule. Rather close association with partner proteins, or assembly into larger protein complexes is necessary for biological activity. Besides the obvious role of protein–protein interactions in the assembly of the cell's structural components, such as the cytoskeleton, they are also crucial for processes ranging from transcription, splicing, and translation to cell cycle control, secretion, and the assembly of enzymatic complexes. Prominent examples for the latter are the organization of enzymes catalyzing sequential steps in a metabolic pathway, such as glycolysis or fatty acid biosynthesis, into multienzyme complexes. A major advantage of such spatial organization is the transfer of biosynthetic intermediates between catalytic sites without diffusion into the enzyme's surrounding[35].

Proteins perform distinct and very well-defined tasks, but little is known about how interactions among them are structured at the cellular level. In a protein interaction network each link in the network is assigned a length of 1. In principle, if a protein interacts with its partner, the link is designed as one. If however, a protein does not interact with other proteins, zero is given in the link. For constructing the network, the basic principle, therefore, follows adjacent matrix that is a matrix with rows and columns labeled by graph nodes with a 1 or 0 in position \((i, j)\) according to whether \(i\) and \(j\) are adjacent or not. Protein complexes and modules are derived from clustering the protein interaction network.

A study of protein interaction network in yeast cells found that they are not random, but well organized. In another study by Schwikowski et al a global analysis of 2,709 published interactions between proteins of the yeast *Saccharomyces cerevisiae* was performed, enabling the establishment of a single large network of 2,358 interactions among 1,548 proteins[36]. They found that proteins of known function and cellular location tend to cluster together, with 63% of the interactions occurring between proteins with a common functional assignment and 76% occurring between proteins found in the same subcellular compartment. They suggested that possible functions can be assigned to a protein based on the known functions of its interacting partners. This approach was used to
correctly predict a functional category for 72% of the 1,393 characterized proteins with at least one partner of known function, and 364 previously uncharacterized proteins.

Apparently in scale-free protein–protein interaction networks, or 'interactome' networks, most proteins interact with few partners, whereas a small but significant proportion of proteins, the 'hubs', interact with many partners. Both biological and non-biological scale-free networks are particularly resistant to random node removal but are extremely sensitive to the targeted removal of hubs. A link between the potential scale-free topology of interactome networks and genetic robustness seems to exist, because knockouts of genes encoding for hubs are approximately three times more likely to confer lethality than those of non-hubs. Han et al investigated how hubs might contribute to robustness and other cellular properties for protein–protein interactions dynamically regulated both in time and in space. They uncovered two types of hub: 'party' hubs, which interact with most of their partners simultaneously, and 'date' hubs, which bind their different partners at different times or locations. Both, in silico studies of network connectivity and genetic interactions described in vivo support a model of organized modularity in which date hubs organize the proteome, connecting biological processes or modules, to each other, whereas party hubs function inside modules[37].

Besides of hubs, some percent of proteins can be expected to work in relative isolation, the majority operate in coordination with other proteins in PPI networks to arrange the processes revolving around cellular structure and function. These processes include cell cycle control, differentiation, protein folding, signaling, transcription, translation, post-translational modification, and transport. Protein interaction networks are generated out of different types of large-scale experimental and computational approaches. The different methods are resulting in significantly different networks, so that we can speak only of a network for a certain organism determined by using a certain method. The protein–protein network analysis has demonstrated that the most highly connected proteins in the cell are the most important for its survival. In the network, this corresponds to the vertices with the highest number of connections. In the same network, it has been shown that certain motifs are overrepresented. Protein interactions play key roles in these processes. For instance, signals from the exterior of a cell to the inside of that cell are conveyed by protein–protein interactions of the signaling molecules. A protein may modify another protein via interaction. For example, a protein kinase will add a phosphate to a target protein. Such modification of proteins can itself change protein–protein interactions virtually every process in a living cell; information about these
Chapter 2 Introduction to Metabolic Network Reconstruction

interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches. Thus, to gain a thorough understanding of biological function, it is important to look at a protein in the context of other interacting proteins [38]. Given protein-protein interactions are of central importance for. Moreover, disease is often related to alterations in certain protein–protein interactions. Hence, the manipulation of protein–protein interactions that contribute to certain diseases provides a potential therapeutic strategy. The complete sequence of genomes of many bacteria, viruses, and small and large eukaryotes has provided a vast new resource to define gene function at the morphological, biochemical, and physiological level. Sequence information by itself, however, does not lead to a clear insight into the underlying principles of cellular systems. This is mainly because the biological function of the plethora of predicted genes remains experimentally uncharacterized. Thus, an understanding of biological mechanisms and disease processes demands a "systems" approach that goes beyond the one-at-a-time studies of single components to more global analyses of the structure, function, and dynamics of the networks in which proteins function. Recent technological advancements using highly parallelized and automated approaches has opened the possibility to assess protein–protein interactions on a genome wide scale. This has led to the establishment of complete protein-linkage maps, called the interactome, which can be regarded as "framework" information. Protein networks aid functional annotation of unknown proteins by opening the possibility to group unknowns into a known biological context. Moreover, increasingly detailed and reliable biological models can be generated by integrating other functional genomic and proteomic data sets into interaction maps[39].

To effectively exploit the large amounts of protein–protein interaction data generated by current experimental and computational methods for biological research, it is necessary for these data to be stored in a consistent and reliable way. A number of protein–protein interaction databases have been developed recently and are now publicly available. These databases greatly differ by their coverage and contents and only a selection of databases is described here (Table 2.1).

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbr.</th>
<th>Entries</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database of interacting proteins [40]</td>
<td>DIP</td>
<td>55,732</td>
<td><a href="http://dip.doe-mbi.ucla.edu/">http://dip.doe-mbi.ucla.edu/</a></td>
</tr>
<tr>
<td>Biomolecular Interaction Network Database [41]</td>
<td>BIND</td>
<td>201,732</td>
<td><a href="http://binddb.org">http://binddb.org</a></td>
</tr>
<tr>
<td>Molecular Interactions Database [42]</td>
<td>MINT</td>
<td>61,330</td>
<td><a href="http://mint.bio.uniroma2.it">http://mint.bio.uniroma2.it</a></td>
</tr>
</tbody>
</table>

Table 2.1: Comprehensive Interaction Databases
For a more comprehensive description of protein interaction databases, including more specialized ones, the reader is referred to some excellent recent reviews and the listings at http://www.pathguide.org. The databases listed in Table 2.1 combine protein interaction data from a number of sources such as from high- and low-throughput yeast two-hybrid and AP-MS analyses, respectively, as well as from data mining of the literature. Therefore, these databases can be considered meta servers of protein interaction data. The Database of Interacting Proteins (DIP) developed at the University of California is a relational database that contains experimentally determined protein–protein, protein–nucleotide, and protein–ligand interactions. Interactions in DIP are curated both manually (by expert curators) and automatically (by text-mining approaches). This database is useful in identifying interacting partners of a protein of interest and visualizing the interactions between proteins. The interaction diagrams also provide researchers with a confidence level of every interaction, as indicated by the width of edges connecting interacting proteins. Access to DIP is free to members of the academic community upon registration. The top-level search page of DIP allows different search strings to be applied on the data set. Proteins can be found by protein name (vertex), sequence motif, BLAST search, or article. Once a protein has been found, a new window can be opened showing the interactions, and information concerning the interaction partners can be browsed and displayed in several ways. The information in the DIP is available for download. The Biomolecular Interaction Network Database (BIND) contains by far the most comprehensive interaction dataset and is designed to store full descriptions of interactions, molecular complexes, and pathways. All interactions contained in BIND have been experimentally verified and published in at least one peer-reviewed publication and are collected by literature mining or investigator submission, with review by curators before incorporation. Incorporated visualization tools help visualizing complex multiprotein interactions. The basic search mode uses text entries to query BIND [43]. The results page from a search shows interaction between what was queried and what is in the database in a binary set. Querying the database with a protein of interest, either using the text search or the BLAST search option using sequence information, generates a results page with interaction listed in sets of two protein of interest with its potential partners. The results page also lists comprehensive additional information concerning the interaction partners such as a general description, molecular function, cellular component, biological process, experiments, and links. The Molecular Interaction Database (MINT), developed at the University of Rome Tor Vergata, focuses primarily on experimentally verified protein–protein interactions from mammalian genomics. The MINT data set was generated by mining scientific literature using a respective algorithm and then
Chapter 2 Introduction to Metabolic Network Reconstruction

reviewed by expert curators. MINT combines basic protein and gene information with data on binary interactions. The database can be searched for individual proteins by name, keyword, structure, or accession number. Such a search results in a list of entries matching the search criteria, and clicking on an individual result leads to a page describing a particular entry, containing a variety of annotation information, as well as a link to all binary interactions involving this protein. Additionally, interactions from MINT can be viewed graphically as an interaction network using the interactive JAVA-based MINT viewer. Each protein in such an interaction network is linked back to the corresponding information page, and interacting proteins are linked to the corresponding binary interaction page. A long list of interacting proteins or a table of protein pairs falls short of capturing what actually happens in a cell, which is a dynamic process that occurs in at least four dimensions (including time). Instead, the use of graphics suits the human preference of visual perception over every other sensory system. Commonly, protein interaction networks are built from experimental data sets using network visualization tools, which in the simplest case represent protein interaction networks as a graph composed of vertices (proteins) connected by edges (interactions). A number of purely automatic and general algorithms have been developed for visualizing biological networks. These tools rely on a layout algorithm to organize a graph of vertices and edges into an easily navigable layout that is usually featured by minimizing the number of edges that cross each other, and grouping groups of vertices that are highly connected to each other. Two commonly used open source visualization tools for protein networks are BioLayout and Cytoscape, which both are based on JAVA and thus readily portable between a variety of computer environments. They also allow the interactive editing of graphs, including movement of vertices, vertex labeling, and changing graph appearance. BioLayout utilizes the weighted Fruchterman–Rheingold algorithm and has a number of options for graph customization, data overlay, export, and graph analysis. Cytoscape provides a number of different layout algorithms for producing useful visualizations such as circular, hierarchical, organic, embedded, and random layouts. Circular and hierarchical algorithms try to lay out networks as their name suggests. Organic and embedded are two versions of a force-directed layout algorithm[44].
Figure 2.5: Map of protein-protein interactions in yeast. Each point represents a different protein and each line indicates that the two proteins are capable of binding to one another. Only the largest cluster, which contains ~78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein [45] (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown).

2.1.3 Metabolic Networks

Metabolic network refers to the network composed of metabolites and their interconversions biochemical reactions in an organism. In the metabolic group, small molecules are chemically transformed by enzymes. Generated energy is then used to assemble these building blocks towards creating new cells and for maintaining the existing cells. The term metabolite is generally used to refer to only relatively “low” molecular weight compounds and excludes all cellular substances that are genetically encoded. These molecules serve either as energy sources or as building blocks for more complex molecules, particularly polymers such as polysaccharides, nucleic acids and proteins. Metabolites are usually small molecules such as glucose and amino acids but can also be
macromolecules such as polysaccharides and glycans. The interconversions are usually catalyzed by enzymes (proteins). Only a few reactions in the cell are spontaneous and thus non-enzymatic. Metabolic pathway, an important concept in biochemistry, is a series of successive or tightly associated biochemical reactions for a specific metabolic function, for instance, glycolysis, lysine degradation, and penicillin biosynthesis. A metabolic pathway can be considered as a small local area of a metabolic network, whereas a metabolic network gives a better and more complete view of the cellular metabolism[46].

Figure 2.6: Overview of the genome-scale metabolic network of Escherichia coli. Adapted from Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Nodes represent metabolites and edges represent metabolic reactions. Cofactors and some small molecules such as water are omitted for simplicity.
The basic architecture of metabolic networks is largely conserved across several different species ranging from microscopic bacteria to plants and humans. Thus the cellular machinery fueling distinct functionality and phenotypes is founded on identical metabolic processes. Understanding of these general organizational principles of metabolic networks can be facilitated by graph-theoretical representation of metabolic networks (Fig. 2.6). Enzymes and metabolites can be viewed as nodes in this network, while interactions between them form edges. The number of neighbors for a node is referred to as its degree. The study of network properties based on its connectivity is called topological analysis. Metabolic networks generally form a fully connected network, i.e., it is possible to travel from any node to all other nodes in the network. The number of edges traversed in such path is referred to as the distance between two nodes.

**Figure 2.7:** Graph theoretical representation of metabolic networks. A group of reactions (a) are generally represented as pathways (b) where co-factors and other highly connected metabolites are depicted only at individual reaction level. (c) Graphical representation where enzymes/reactions and metabolites form nodes and interaction among them as edges. Such graph is essentially bi-partite, as neither enzyme nor metabolite nodes are directly connected to other nodes in the same category.

Organization and operation of metabolic networks is traditionally viewed and understood in terms of ensemble of pathways. Pathways are comprised of groups of enzymes acting towards
production/breakdown of certain metabolites. Glycolysis and TCA cycle are familiar examples of pathways. Although the concept of a pathway is widely used and very useful for pictorial representation, the definition of a pathway is very vague from stoichiometric point of view. An alternative and more comprehensive way of understanding the operation of metabolism is through the enumeration of all possible combinations of reactions that can sustain a balanced flow from substrates to final products. Each such combination can then be seen as a definition of a stoichiometrically "complete" pathway. An important fact is that the number of such possible routes is of the order of millions, even in a simple microbial metabolic network. Any active metabolism at a steady state can be represented as a linear combination of these elementary flux modes. What are the factors/mechanisms responsible for particular phenotype under given conditions? It appears that this choice is achieved via coordinated regulation of enzymes. Thus, not only a single enzyme is regulated for its optimal operation, but also the whole network is subject to regulation for optimal network functionality[47].

Regulation of enzymes can occur at the level of transcription, translation or post-translational modifications (e.g. phosphorylation). Furthermore, enzyme activity may be regulated by small effector molecules. The flux through a reaction is dependent not only on the availability and activity of the enzyme, but also on the concentrations of substrates, products and effectors. Such relationships are usually non-linear. Additionally, due to the interconnected nature of the metabolic network, all steps in the metabolism can in principle influence all other steps. Consequently, understanding, simulation and prediction of both dynamic and steady state operations of metabolic network are challenging tasks [48].

A complete metabolic network should show all the possible modes of material flows in the cell, therefore indicating all the metabolic potential and capacity of the cell. In other words, metabolic network is the material processing center for a functioning cell. The cell relies on this network to uptake and digest substrates from the environment, to generate energy (e.g., in form of ATP) and to synthesize components that are necessary for its growth and survival [49].
2.2 Metabolic Network Reconstruction

It is quite interesting to study metabolic network for its fundamental importance in biology and particular way because many applications are directly built on the use of cellular metabolism. Mapping metabolic pathways has been a focus of significant scientific efforts dating from the emergence of biochemistry as a distinct scientific field in the late 19th century. This endeavor remains an important effort for at least two compelling reasons. First, cataloguing and characterizing the full range of metabolic processes across species is a fundamentally important step towards a complete understanding of our ecological environment. Second, mapping metabolic pathways in organisms, many of which can be found with specialized properties shaped by their environment, facilitates metabolic engineering to advance nascent industrial biotechnology efforts ranging from replacing petroleum-derived chemical precursors or fuels to biopharmaceutical production. However, despite laudable efforts to enable high-throughput 'genomic enzymology', the traditional biochemical approaches of enzyme expression, purification, and characterization remain time-intensive, capital-intensive, and labor-intensive, and have not expanded in scale like our ability to identify and characterize life genomically. Characterizing new metabolic function is further hampered by the challenge of cultivating environmental isolates in laboratory conditions. Fortunately, recent efforts to leverage genome functional annotation and established knowledge of biochemistry have enabled the computational assembly of 'draft metabolic reconstructions', which are parts list of metabolic network components. In this context, a reconstruction is not just the information embodied in the stoichiometric matrix describing metabolic network structure, but also the associated metadata and annotation that entails an organism-specific knowledge base. Such a reconstruction can serve as the basis for making functional models amenable to mathematical simulation. Thus, a reconstruction is a bottom-up assembly of biochemical information, and a model can serve as a framework for integrating top-down information. This information is used for statistically inferred gene regulatory networks from model. Such computational approaches are significantly faster and less expensive than biochemical characterization. They also provide new resources which facilitate cultivation of novel environmental reconstruction model, and the scope of draft metabolic network coverage across the community has increased much faster than wet lab characterization. If the distinction between reconstruction and model formulation can be strengthened and supported through software implementation, there is great opportunity for using both tasks to further advance rapid discovery of biological function[50].
Chapter 2 Introduction to Metabolic Network Reconstruction

Metabolic networks across different species are different in many aspects. Differences are more evident at the level sequences of individual enzymes and the regulation of enzymes in response to environmental and genetic challenges. Biotechnologists modify the cells and use them as cellular factories to produce bulky or fine chemicals, antibiotics, industrial enzymes, antibodies, and so on. In biomedicine, people cure metabolic diseases of human beings through better understanding of the metabolic mechanism, and control infections of pathogens by making use of the metabolic differences between human beings and pathogens.

The earliest metabolic network reconstructions, including *Clostridium acetobutylicum, Bacillus subtilis* and *Escherichia coli* all belong to this category. Also, it is important to know that this reconstruction model may be revised from time to time. One example is the genome-scale metabolic network of *Escherichia coli*, which was created and refined several times in the past ten years[51].

These manually reconstructed metabolic networks provide valuable datasets about cellular metabolism, such as collections of metabolic reactions that are being used when reconstructing metabolic networks for other organisms. Beyond the metabolic reaction sets, researchers also accumulated experiences in reconstructing metabolic networks and summarized them into standard procedures. After great amounts of annotated genome sequences became available, the reconstruction process mainly relies on these genomic data. Despite some recent developments, the reconstruction process is still labor intensive and time consuming. In a suggested protocol, there are 96 steps involved to reconstruct a high-quality metabolic network for the genome data (Figure 2.8), which might cost years for a well studied, medium sized bacterial genome. The situations will become even worse if the organism is not well studied and even not cultivable, for which only limited experimental data are available. Another limitation that constrains the application of manual reconstruction is the inconsistency between different databases and datasets. An important task in the manual reconstruction process is reconciliation of model predictions with experimental data. The inconsistency between different databases and datasets collected manually makes this task much more difficult[52].

The iterative steps should continue until the model predictions are close to the experimental phenotypic characteristics of the organism. To overcome the limitations of manual reconstruction, automated or semi-automated procedures that directly generate metabolic reconstructions from annotated genome data are of great interest. A number of user-friendly resources have been developed to facilitate this demanding process.
Chapter 2 Introduction to Metabolic Network Reconstruction

1. Pratt Reconstruction

1. Obtain genome annotation.
2. Identify candidate metabolic functions.
3. Obtain candidate metabolic reactions.
4. Assemble of draft reconstruction.
5. Collect of experimental data.

II Obtain genome annotation.

2. Refinement of reconstruction

6. Determine and verify substrate and cofactor usage.
7. Determine reaction directionality.
8. Add information for gene and reaction localization.
9. Add subsystems information.
11. Determine and add confidence score.
12. Add subsystems information.
14. Calculate reaction stoichiometry.
15. Determine substrate and cofactor usage.
16. Obtain neutral formula for each metabolite.
17. Obtain charged formula.
18. Calculate reaction stoichiometry.
19. Determine reaction directionality.
20. Add additional reactions to the reconstruction.
22. Add intracellular transport reactions.
23. Draw metabolic map (optional).
25. Add biomass reactions.
26. Add ATP maintenance reaction (ATPM).
27. Add demand reactions.
28. Add sink reactions.
29. Determine growth medium requirements.

3. Conversion of reconstruction into computable format

38. Initialize the COBRA toolbox.
39. Load reconstruction into Matlab.
40. Add exchange constraints to the reconstruction.
41. Set objective function.
42. Set simulation constraints.

Data assembly and Dissemination

95. Print Matlab model content.
96. Add gap information to the reconstruction output.

Network evaluation

43-44. Test if network is mass- and charge balanced.
45. Identify metabolic dead-ends.
49. Add missing exchange reactions to the model.
50. Set exchange constraints for simulation condition.
51-58. Test for stoichiometrically balanced cycles.
59. Recompute gap list.
60-65. Test if biomass precursors can be produced in standard medium.
66-71. Test if biomass precursors can be produced in other growth media.
72-78. Check for blocked reactions.
81-82. Test for known incapacities of the organism.
83. Compare predicted physiological properties with known properties.
84-87. Test if the model can grow fast enough.
88-94. Test if the model grows too fast.

Some of these tools are designed for helping manual metabolic network reconstruction. For example, MetaFluxNet provides an interface for users to manually input information of metabolic networks and then carries out metabolic flux analysis (MFA). Similarly, MetNetMaker allows users to select metabolic reactions from reaction databases by EC number or other information and then generate a metabolic network. METANNOGEN and rBioNet are effective tools for data management for metabolic network reconstruction process. YANAsquare provides a interface that connect to KEGG database and allow user to select metabolic reactions in each pathways that are associated with annotated genes. By combining those selected reactions, the tool can generate draft genome-scale metabolic networks[53, 54].

A number of methods and tools have been developed to facilitate metabolic network curation and gap filling. In 2004, Green and Karp designed a Bayesian based method for identifying missing reactions in database [55]. In the same year, Kharchenko et al. published an algorithm that is able to fill metabolic gaps in a metabolic network using expression information [56]. Along this line, Kumar et al. developed an optimization based procedure that can find and fill metabolic gaps by searching all downstream no-production metabolites and minimizing the number.
of added reactions. COBRA Toolbox also provides the metabolic gap-filling functions for metabolic network reconstruction. Recently, rBioNet, a COBRA toolbox extension for metabolic network reconstruction, was developed by Thorleifsson and Thiele. rBioNet enables users to combine metabolites and reaction database from different sources during the curation process. This function is important when the metabolic gaps are filled by reactions in another database or model. In 2009, Kumar and Maranas developed GrowMatch algorithm that can reconcile in silico and in vivo growth predictions and revise the metabolic reconstruction at the same time. GrowMatch has been implemented in COBRA toolbox v2.0, which provides a more power platform for both metabolic network reconstruction and network analysis[57].

Several tools have been developed for network analysis and curation rather than creating a new metabolic network. BioMet, a web-based resource for stoichiometric analysis and integration of transcriptome and interactome data, has been developed. BioMet contains a tool that can convert metabolic networks written in system biology makeup language (SBML) into its own data framework. Acorn is another web tool providing constraint based modeling and visualization for existing genome wide metabolic networks written in SBML. In addition to BioMet and Acorn, OptFlux and SBRT both provide comprehensive software platform for in silico metabolic modeling and engineering[58, 59].

All these tools support increasingly sophisticated network analyses, but rely largely on existing network models and have very limited capabilities for creating new networks. To our best knowledge, there are two tools available that can reconstruct metabolic network automatically from genome sequence or annotation, including Model SEED and GEMSiRV which is based on MrBac. The Model SEED, based on an automated genome annotation tool RAST, first compares the annotated genome with a self-maintained database containing both metabolic reactions and associated genes to generate a draft metabolic reconstruction. This draft metabolic network is then refined with minimal modifications to meet the growth requirements preassumed in the model. However, users cannot specify the growth condition and biomass compositions during the automated curation, which limits its application. The gene candidates of the filled metabolic gaps are not provided by Model SEED, which is important information in metabolic network reconstruction process[60-62].

Different form SEED, GEMSiRV, which is based on MrBac, can generate a new metabolic network by comparing the genome sequences to a known organism with reconstructed metabolic network.
Chapter 2 Introduction to Metabolic Network Reconstruction

Therefore, those metabolic reactions in the reference metabolic network will be added into the genes that are associated with the metabolic reactions in the reference genome. Sequence alignment is applied to identify these genes by setting a threshold value or E value. This method can only apply to those organisms that are close to model organisms and the reconstructed metabolic network model. Otherwise, this method cannot provide reliable metabolic networks due to the diversity of cellular metabolism. Another limitation of this method is it cannot make use the efforts spending in the manual curation of gene annotation because only sequence alignment results are utilized to identify reactions[63].

2.3 Genome annotation databases

Cells consist of different types of biomolecule. New technological approach of molecular biology deals with considering the properties and roles of individual molecules and produces large amount of useful knowledge. A discrete biological function cannot be attributed to an individual molecule, but to a complex web of interactions between a set of molecules. Therefore, describing biological systems requires explaining how they arise from interactions among components in the cell. The availability of complete genome sequences of several organisms has opened doors for development of high-throughput omics technologies. The sequencing of genome and bioinformatics and functional genomics studies of sequences make it now possible to reconstruct the metabolic network of a specific organism at genome level and thus opens up new horizons in many areas of biotechnology and life science. According to NCBI and Genome online database (GOLD) (Status April, 2014), 6653 genomes are completed, including human-beings and many clinically or industrially important organisms, while other 1887 genome sequencing projects are ongoing [64].

Reconstructions rely on many types of biological evidence, including genetic, biochemical, and physiological data. Many of these networks are available on online databases: Kyoto Encyclopedia of Genes and Genomes, EcoCycn and BioCyc. EcoCyc is an extensively human-curated database specialized for Escherichia coli K-12 whereas KEGG and BioCyc maintain databases for many organisms but with little human curation effort. Biological information can be readily obtained from online available databases with detailed organism-specific data. This data must usually be extracted from the annotation pipeline for eukaryotic genome annotation, which are discussed in this chapter later part. We consider reconstruction and network generation with specific four main different eukaryotic algae (1. Chlamydomonas reinhardtii, 2. Ostreococcus lucimarinus, 3. Ostreococcus tauri and 4. Volvox carteri). Getting data for above specific organisms and
Chapter 2 Introduction to Metabolic Network Reconstruction

annotation are described here; also, other relevant databases for microbial reconstructions have been discussed [65-67].

The figure 2.9 provides an overview of some omics experimental techniques for measuring different biological data types, which are elaborated in this chapter. Gene expression microarrays and Serial Analysis of Gene Expression (SAGE), which enable us to measure the abundances of thousands of gene transcripts simultaneously, have empowered the first omics discipline known as transcriptomics. The omics technologies have contributed to rapid accumulation of knowledge such as gene and gene product annotations and biomolecular interactions. Simultaneously, the availability of genome sequences also enables development of computational algorithms for sequence analyses which also help to rapidly annotate new sequence data, and predict the structure and interactions. Finally, text mining also helps to retrieve important molecular interaction information, and careful manual literature curation leading to higher quality information as compared to high-throughput data. In the spirit of genomic data sharing, many types of the biological data have been made accessible through world wide web.

These web-accessible databases and their associated search and mining tools are primary resources serving thousands of biology researchers worldwide. These tools allow researchers to effectively
Chapter 2 Introduction to Metabolic Network Reconstruction

Figure 2.9: Various types of data in molecular biology and the experimental techniques employed for obtaining the data. High-throughput techniques capable of measuring all or at least a large number of components (several hundreds to thousands) simultaneously are known as omics techniques.

mine the databases and answer one's biological questions. The databases cover a wide range of information including literature, sequences and annotations of genes and gene products, and a variety of molecular interactions such as biochemical reactions, transcriptional regulatory interactions, signal transduction pathways, to name a few. This chapter is a backbone of research theme, where primary data collection is required with publicly available database and special database created for green algae species. This chapter introduces some of the commonly used databases for microarray gene expression profiles, protein-protein interactions, metabolic interactions, transcriptional regulatory interactions, and signal transduction networks.

Genome annotation databases are comprehensive, consisting of information about gene centric resources that provide an abundance of information on gene identifiers (e.g., abbreviations, names, and synonyms), gene-protein relationships (e.g., alternative transcripts, isozymes, and protein complexes), and protein localization. They also typically contain links to primary research articles in PubMed, making them an ideal starting point for manual network curation. One of the largest and most widely used annotation databases is Entrez Gene, which consists of over 3,600 taxa to date. Most model organisms have species-specific genome annotation databases that are actively updated by their research community, such as the Cynoref Genome Database and Comprehensive ALAGE Genome Database. Several human genome annotation databases have been generated computationally using data mining algorithms (GeneCards) and high-throughput data (Kyoto Encyclopedia of Genes and Genomes (KEGG) and HumanCyc).

In systems biology, networks are commonly reconstructed at the cellular level and are considered ‘genome-scale’ if they include all of the components encoded by an organism’s DNA. The interactions between genes, proteins, and chemical compounds constitute the two-dimensional annotation of a genome, and there are many parallels between their hierarchy and the one-dimensional, sequence-based annotation.

Networks are comprehensive reconstructions of a system’s components and their interactions. There are two general strategies for assembling a reconstruction: top-down and bottom-up. Top-down approaches rely on inference methods to identify and formulate relationships between network components. They are typically implemented in a computer, enabling rapid assembly of
large, comprehensive networks. In contrast, bottom-up networks are manually assembled in a component-by-component manner based on direct physical evidence from multiple data sources. Manual reconstruction can be a time-consuming and laborious process, but is oftentimes favored for modeling applications because it produces self-consistent networks. Thus, top-down and bottom-up methods each have distinct advantages and disadvantages, and the best choice for reconstructing a network of interest can depend on many factors, including time, data availability, and the number of components[49].

![Diagram of Genome Annotation and Network Annotation](image)


**Figure 2.10**: The hierarchical relationship of network components. There are many parallels between levels of detail in genome and network annotations. Chromosomes are analogous to cellular biochemical networks, which are described in terms of reactions. Contigs are delineated by sequence reads, which describe individual base pairs, the primary components of a sequence annotation. Similarly, reactions are catalyzed by enzymes, which are derived from genes and their transcripts, and act on compounds, the primary components in a network annotation.

The **Sequence Read Archive** is a database that provides a public repository for DNA sequencing data, especially the "short reads" generated by High-throughput sequencing machine like Roche 454,
Chapter 2 Introduction to Metabolic Network Reconstruction

HiSeq X Ten, Illumina GA2, Ion Torrent, MiSeq, MinION, NextSeq, PacBio, Polonator Proton and SOLiD, which are typically less than 1,000 base pairs in length. The archive is a part of the International Nucleotide Sequence Database Collaboration (INSDC), and runs as a collaboration between the NCBI, the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). These high-throughput sequencing technologies make deep transcriptome sequencing and transcript quantification, whole genome sequencing and resequencing available to many more researchers and projects.

GenBank is a sequence database that stores all known DNA sequences gathered by direct submission of sequence data from individual laboratories and from large-scale sequencing projects. There are two other major DNA sequence databases. Namely EMBL nucleotide sequence database, and DDBJ and the data among these three databases are synchronized. The Ensembl project offers an integrated source of genome sequences and annotations for a comprehensive set of chordate genomes with a particular focus on human, mouse, rat, zebrafish etc. The University of California Santa Cruz (UCSC) Genome Browser Database is a source for genome sequence and annotation data. UCSC Genome Browser, is a tool associated with the UCSC database that provides rapid visualization and querying of the data. The annotations provided by these genome databases include mRNA and expressed sequence tag (EST) alignments, gene predictions, cross-species homologies, high-level maps, single nucleotide polymorphisms (SNPs) and so on. Besides these general genomic databases, there are organism specific genomic databases for model organisms. The Saccharomyces Genome Database (SGD) is a database for the molecular biology and genetics of the yeast Saccharomyces cerevisiae that provides functional annotations, mapping and sequence information, protein domains and structure, expression data, mutant phenotypes, physical and genetic interactions and the primary literature from which these data are derived. FlyBase is a database of genetic and genomic data concerning fruit flies of which Drosophila melanogaster is an extensively studied model organism. FlyBase is populated with information from a variety of sources ranging from large-scale genome projects to the primary research literature. FlyBase provides access to information on gene models, molecular classification of gene product functions, mutant phenotypes, mutant lesions and chromosome aberrations, gene expression patterns, transgene insertions, and anatomical images. WormBase is a central data repository for nematodes of which Caenorhabditis elegans is an extensively studied model organism. WormBase includes genomic sequences, gene predictions and orthology assignments from a range of
Chapter 2 Introduction to Metabolic Network Reconstruction

related nematodes and relies on manual curation of information from the corpus of C. elegans literature[70].

Transcription factors are proteins that are vital for the transcriptional regulation of gene expression. A transcription factor has a DNA binding domain which can bind to a particular region in the DNA sequence of a gene, called the binding site, and helps in enhancing or inhibiting the expression of the gene. TRANSFAC database primarily provides information about entities involved in the transcriptional regulation such as transcription factors, binding sites and genes among a variety of other related information [31].

Living cells interact with their environment by exchanging a variety of signals. Signaling pathways of the receiver cells forward the signals to the nucleus through cascades of interactions and trigger the appropriate adaptation of the genetic program. The TRANSPATH database provides information about signal transduction pathways involved in the transcriptional regulation of gene expression via regulating the activity of the transcription factors[30].

The Universal Protein Resource (UniProt) provides information about protein sequences and functional information. The central database in Uniprot, termed UniProt Knowledgebase, provides accurate, consistent and rich sequence and functional annotations and consists of two sections: UniProt/Swiss-Prot and UniProt/TrEMBL. UniProt/Swiss-Prot consists of manually curated protein functional information, resulting from literature information extraction and curator-evaluated computational analysis. UniProt/TrEMBL consists of protein sequences translated from EMBL gene sequences and annotated with computational annotation tools, pending manual curation[71].

BRENDA (BRaunschweig ENzyme DAtabase) represents a comprehensive collection of enzyme and metabolic information, based on primary literature. The database contains data from at least 83,000 different enzymes from 9800 different organisms. The database contains more than 40 data fields with enzyme-specific information on more than 4800 EC numbers that are classified according to the IUBMB. The different data fields cover information on the enzyme's nomenclature, reaction and specificity, enzyme structure, isolation and preparation, enzyme stability, kinetic parameters such as Km value and turnover number, occurrence and localization, mutants and engineered enzymes, application of enzymes and ligand-related data. Currently, BRENDA contains manually annotated data from over 130,000 different scientific articles. Each enzyme entry is clearly linked to at least one literature reference, to its source organism, and, where available, to the protein
sequence of the enzyme. An important part of BRENDA represents the more than 107,000 enzyme ligands, which are available on their names, synonyms or via the chemical structure. The term "ligand" is used in this context to all low molecular weight compounds which interact with enzymes. These include not only metabolites of primary metabolism, co-substrates or cofactors but also enzyme inhibitors or metal ions. The origin of these molecules ranges from naturally occurring antibiotics to synthetic compounds that have been synthesized for the development of drugs or pesticides. Furthermore, cross-references to external information resources such as sequence and 3D-structure databases, as well as biomedical ontologies, are provided. The database is accessible free of charge to the academic community[72].

Database of Interacting Proteins (DIP) is a database of manually curated protein-protein interactions. A curator enters each interaction entry into the database after manually reading the publication reporting an experimentally verified interaction. This is intended to be a comprehensive and integrated tool for browsing and efficiently extracting information about protein interactions and interaction networks in biological processes. DIP provides access to combined information from multiple observations and experimental techniques, from multiple organisms, as well as to networks of interacting proteins. Each interaction entry in the DIP database contains information about the protein domains and range of amino acids involved in the interaction, and the corresponding experiments. The inter factors are identified by Swissprot, PIR or GenBank accession numbers and each inter factor entry contains information about the organism, function, superfamily, cellular location and so on[40].

The Biomolecular Interaction Database (BIND) stores pairwise interactions between biological ‘objects’ which could be protein, RNA, DNA, molecular complex, small molecule, photon (light) or gene. Moreover, it contains higher level functional structures called molecular complexes and pathways which are collections of the pairwise interactions with some additional data. The minimum amount of information required to define an interaction is a description of the interacting objects and a publication reference to PubMed. Data in BIND is primarily obtained via submissions of individual contributors across the world. However, it also incorporates interaction data imported from other databases such as PDB, and a number of large-scale cell mapping studies using yeast two hybrid, mass spectrometry, genetic interactions and phase display. SeqHound is a data integration system that provides extensive C, C++, and Perl application programming interfaces (API) for data
Chapter 2 Introduction to Metabolic Network Reconstruction

in BIND. SeqHound system provides also functions to link the biological objects with other biological databases in public domain[43].

The Molecular Interaction database (MINT) stores information about experimentally verified molecular interactions extracted from publications from peer-reviewed journals. The main focus is on physical interactions between proteins. Genetic or computationally inferred interactions are not included in MINT. MINT includes an additional database called HomoMINT, which is a database of interactions between human proteins inferred from interactions between orthologous proteins in model organisms. A large number of MINT data comes from large scale, genome wide experiments, although curating data from low-throughput published experiments is given emphasis. Each interaction entry contains reference to Swiss-Prot/TrEMBL protein accession number for the interacting factor and contains the experimental information and pubmed reference for describing the experimental conditions and other properties of the interaction[73].

BioGrid is a database of protein and genetic interactions. It is aimed to be a generic repository providing comprehensive information on molecular interactions in several organisms such as Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans, and Homo sapiens. It currently hosts protein-protein interaction data from high-throughput experiments such as yeast two-hybrid (Y2H) method and mass spectrometry analysis of purified protein complexes. Additionally, the BioGrid team also compiles interaction data by extensive manual curation of literature. Literature curated data for Saccharomyces cerevisiae and Schizosaccharomyces pombe have already been added to BioGrid and curation efforts for other organisms are underway[74].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) system consists of three main components: the genomic space (KEGG GENES), the chemical space (KEGG LIGAND), and network space (KEGG PATHWAY). The KEGG GENES is a collection of gene catalogues for completely or partially sequenced genomes, compiled by automatically extracting information from databases such as NCBI GenBank, and RefSeq. The KEGG LIGAND is further divided into several components including COMPOUND, GLYCAN, REACTION, ENZYME, and so on. The COMPOUND database contains manually entered and computationally verified chemical structures of known metabolic compounds, and some pharmaceutical and environmental compounds. The GLYCAN database consists of carbohydrate structures, a few hundreds of which were manually entered and the rest derived from CarbBank project. The REACTION database contains reaction formulae for enzymatic reactions, the reactants of which are represented in COMPOUND or

Page | 31
GLYCAN databases. The ENZYME database contains enzyme nomenclature. Each enzyme is identified by an Enzyme Commission (EC) number, which can be linked to other public databases such as UniProt. The KEGG PATHWAY database is a collection of manually drawn diagrams, called KEGG reference pathway diagrams (maps), each of which corresponds to a known network of functional significance. Moreover, PATHWAY database are also contains organism specific pathways, which are automatically generated by superimposing genes in given organisms. The KEGG pathways are provided in an XML-based markup language called KGML. Each metabolic reaction in a KEGG metabolic pathway is linked with one entry in the REACTION database, and the enzymes in the enzymatic reactions can be linked to the databases in the genome space as well as to other public databases via EC numbers [75, 76].

PANTHER (Protein ANalysis THrough Evolutionary Relationships) classification system is a large curated biological database of gene/protein families and their functionally related subfamilies that can be used to classify and identify the function of gene products. PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function (ontology terms and pathways), as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. The latest version, 5.0, contains 6683 protein families, divided into 31,705 subfamilies, covering approximately 90% of mammalian protein-coding genes [77].

Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. The Pfam database contains information about protein domains and families. Pfam-A is the manually curated portion of the database that contains over 10,000 entries. It has generated 1182 new families and maintained sequence coverage of the UniProt Knowledgebase (UniProtKB) at nearly 80%, despite a 50% increase in the size of the underlying sequence database. For each entry a protein sequence alignment and a hidden Markov model is stored. These hidden Markov models can be used to search sequence databases with the HMMER package written by Sean Eddy. Because the entries in Pfam-A do not cover all known proteins, an automatically generated supplement is provided called Pfam-B. Pfam-B contains a large number of small families derived from clusters produced by an algorithm called ADDA. Although of lower quality, Pfam-B families can be useful when no Pfam-A families are found [78].
Saccharomyces cerevisiae iND750 is a manually reconstructed genome-scale metabolic model describing Saccharomyces cerevisiae metabolism with 750 genes, their transcripts, proteins and reactions. Manual reconstruction process involves curating reaction lists based on information from genome annotations, biochemical pathway databases, biochemistry textbooks, and publications. All reactions in iND750 model are elementally and charge balanced, and compartmentalised to eight cellular locations: extracellular space, cytosol, mitochondrion, peroxisome, nucleus, endoplasmic reticulum, Golgi apparatus, and vacuole. Similar semi-automated manual curation has been employed to construct the first consensus metabolic network for yeast global human metabolic network, and so on[79].

Gene Expression Omnibus (GEO) stores a variety of high-throughput molecular abundance data of which microarray gene expression data is a major data type. The data in GEO is organized into GEO Platforms (GPL), GEO Samples (GSM), GEO Series (GSE) and GEO Data sets (GDS). A Platform describes the set of elements that can be detected and quantified in the experiment. A sample describes a single hybridization or experimental condition. A Series is a group of related samples that make up one single study. A data set is an assembly of biologically meaningful samples that are statistically comparable. Of these, GPL, GSM and GSE are direct submissions of contributors, while GDS is a curated collection[80].

Gene Ontology (GO) consists of three independent ontologies: Biological process, Molecular function, and Cellular component to describe the roles of genes and proteins in eucaryotes. GO is built on the premise that a large fraction of the genes specifying core biological functions are shared by all eukaryotes. It is aimed to be a dynamic controlled vocabulary applicable to all eucaryotes even as our knowledge of gene and protein roles in cells continuously evolves[81].

Although the high-throughput experimental techniques of modern molecular biology empower us to measure multiple components of a biological system simultaneously, they often produce data that is inferior in quality to low-throughput techniques. On the one hand, high-throughput techniques may produce large number of false positives, meaning that not all findings are necessarily correct. On the other hand, they may also exhibit large number of false negatives or detection biases, meaning that they may miss some true phenomena, leading to the lack of coverage or comprehensiveness in the findings. von Mering et al. performed a detailed comparison of multiple high-throughput techniques as well as a computational approach commonly employed for the study of protein-protein interactions. They estimated that more than half of all high-throughput protein-protein
interaction data are false positives. While the estimated lower-bound to the number of protein-protein interactions is 30,000, the number of interactions supported by more than one method is only approximately 2,400, which demonstrates the sparing coverage of the methods employed. For these reasons, manual curation of published literature has, despite the practical difficulties, received significant interest. In order to compile high-quality data, manual literature curation is expected to consider high-confidence data coming from low-throughput techniques, and perhaps also additionally supported by multiple independent studies. However, a recent study that has systematically compared the quality of a literature curated data has speculated that the quality of the literature curated yeast protein interaction data from BioGrid is at the most as good as or even inferior to high-throughput yeast two-hybrid (Y2H) data. On the other hand, high-throughput techniques such as Y2H method are also improving, and it is increasingly believed that the little overlap among different high-throughput protein interaction studies is due not to the false positives, but to the false negatives. However, still the coverage of such data is quite little. For instance, while reporting a newly produced high quality comprehensive binary Y2H protein interaction map for Saccharomyces cerevisiae, Yu et al. indicated that three proteome-level Y2H studies taken together, only account for approximately 20% of the empirically estimated protein binary interactions in Saccharomyces cerevisiae. Y2H interaction maps have also been generated for other model organisms and humans, and their quality and coverage are similar to those of S. cerevisiae interaction maps. Global metabolic network reconstruction approaches described earlier presumably produce high quality metabolic networks but they are not complete and continuously keep growing. Comprehensive system-level data of high quality is a key ingredient of systems biology. Similarly, efficient computational tools to effectively handle current and future high-throughput data and turn them into knowledge are equally important [82].