Integration of Pathway Data as Prior Knowledge into Methods for Network Reconstruction

Dissertation

zur Erlangung
des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

vorgelegt von
Frank Kramer
aus Erlangen

Göttingen 2014

D7

Referent:	Prof. Dr. Tim Beißbarth
Korreferent:	Prof. Dr. Stephan Waack

Tag der mündlichen Prüfung: 16.09.2014

Abstract

Hundreds of databases offer vast amounts of literature knowledge about biological signaling networks. However, this knowledge is rarely integrated into current bioinformatic analyzes due to challenges in the programmatic access and transformation of this data. This thesis focuses on the integration of prior knowledge into methods for network reconstruction. The motivation is to improve the performance of bioinformatic algorithms and methods by facilitating the integration of available pathway data as prior knowledge.

First, the fundamentals of biological networks and pathways, their encoding using ontologies, methods for network reconstruction, and high-throughput gene expression technologies are introduced.

Three central results are presented in this work: First, the novel software package `rBiopaxParser`, which enables the generic import of BioPAX-encoded pathway databases into the R Project for Statistical Computing. An overview of the functionality, the internal data model and visulization options is given. Second, a proof-of-concept implementation of the transformation and merging of pathway data to be used as prior knowledge for methods for network reconstruction is presented. The interactomes, the entirety of interactions, of three databases, Reactome, Pathway Interaction Database, and BioCarta, are generated and merged as a basis for prior pathway knowledge. Third, network reconstruction using Nested Effects Models is performed based on the generated prior knowledge networks and experimental high-throughput data of 16 gene knockdowns in breast cancer cell lines.

Finally, this thesis compares the implemented software to similar concurrent developments and discusses the generated prior knowledge and the results of network reconstruction.

Zusammenfassung

Über 300 Datenbanken bietet Zugang zu dem unüberschaubaren Literaturwissen über biologische Signalnetze. Derzeit wird dieses Vorwissen, aufgrund von Hindernissen beim programmatischen Zugriff und der weiteren Verarbeitung, nur selten in bioinformatischen Analysen eingesetzt. Der Fokus dieser Arbeit liegt in der Integration von Vorwissen in Methoden zur Netzwerkrekonstruktion. Das Ziel hierbei ist, die Ergebnisse von bioinformatischen Algorithmen und Methoden zu verbessern, indem die Integration von verfügbarem Vorwissen vereinfacht wird.

Zuerst werden in dieser Arbeit die Grundlagen von biologischen Netzwerken und Signalwegen, sowie ihre Kodierung mittels Ontologien, eingeführt. Desweiteren werden Methoden zur Netzwerkrekonstruktion und Hochdurchsatz-Technologien zur Messung von Genexpressionsdaten beschrieben.

Drei zentrale Ergebnisse werden in dieser Arbeit beschrieben: Das erste Ergebnis ist die Implementierung des Open Source Softwarepakets `rBiopaxParser` für das R Project for Statistical Computing. Es wird ein Überblick über das R-Paket, welches den Import von BioPAX-kodierten Pathwaydatenbanken erlaubt, das interne Datenmodell und die Visualisierungsoptionen gegeben. Das zweite Ergebnis ist die beispielhafte Implementierung eines Workflows für das Einlesen, die Transformation und das Zusammenführen von Pathwaydatenbanken, welches für die Erstellung von Vorwissen für Netzwerkrekonstruktionsverfahren benötigt wird. Hierbei werden die Interaktome, die Gesamtheit aller Interaktionen, der drei Pathwaydatenbanken Reactome, Pathway Interaction Database und BioCarta, konstruiert und als Basis für Vorwissen zusammengeführt. Das dritte Ergebnis ist schließlich die Anwendung von Nested Effects Models zur Netzwerkrekonstruktion basierend auf den generierten Vorwissennetzwerken und experimentellen Daten von 16 Gen-Knockdowns in Brustkrebs-Zelllinien.

Anschließend werden in dieser Arbeit dem implementierten Softwarepaket ähnliche Entwicklungen gegenübergestellt. Desweiteren werden der Workflow, das generierte Vorwissen, sowie die Ergebnisse der Netzwerkrekonstruktion diskutiert.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"Cells are the intrinsic center of health and disease."* (Virchow, 1855)

This insight was announced by Rudolph Virchow, generally acknowledged as the father of modern pathology, as early as 1855. Decoding the interactions within a cell therefore leads to a new understanding of diseases. Deciphering the inner workings of living cells fascinates researchers all over the world. However, the processes within each cell are highly complex, with countless participants constantly interacting via biochemical reactions, signaling cascades and feedback loops.

Knowledge about these processes can be organized into so-called "pathways" by grouping sets of interactions which share a common goal or function (Alberts, 2008). Two examples are the apoptosis pathway (Kerr et al., 1972), which includes the cell signaling cascade that leads to programmed cell death, and the glycolysis pathway (Meyerhof, 1927), a metabolic process in which glucose is degraded and leads to a gain in energy-rich molecules within the cell. In fact wall charts, huge poster prints with detailed data on metabolic processes within the cell, cover many laboratory walls across the world (Miura and Duncan, 1973). Due to the directed nature of signaling and catalytic processes, pathways are often depicted computationally in the manner of directed graphs (Kohn, 1999).

Methods for network reconstruction are approaches to infer the graph structure of pathways from experimental data (Tresch and Markowetz, 2008),

enabling researchers to extend the current pathway knowledge. One possible approach to network reconstruction derives the interactions of genes by comparing expression profiles between perturbed and untreated samples (Fröhlich et al., 2009). Furthermore, a number of methods for network reconstruction are able to integrate prior knowledge into their computations and thus improve the power or robustness of their predictions (Fröhlich et al., 2007a; Mukherjee and Speed, 2008).

Over the course of the last decades an enormous amount of knowledge on molecular interactions within cells has been accumulated. These insights range from the assembly of molecular complexes from single proteins, to the catalysis of biochemical reactions and the signaling cascades triggering certain functions within the cell. A meta-database on pathway databases, pathguide.org (Bader et al., 2006), currently contains links to over 300 databases which collect and curate knowledge on biological pathways.

Methods for network reconstruction can be used to infer the topology of a cellular network from biological experiments, which are measured using high-throughput technology (Markowetz and Spang, 2007). Literature knowledge of molecular interactions might overlap with the reconstructed network. Integrating relevant parts of this literature knowledge as a prior knowledge network can enhance the performance of network reconstruction (Fröhlich et al., 2007a).

The motivation of this thesis is to facilitate the integration of multiple pathway data sources as prior knowledge for methods for network reconstruction. Furthermore, the influence of these computationally merged data sources are evaluated.

1.1 Aims and Organization of the Thesis

The specific aims of this thesis, in order to integrate multiple pathway data sources as prior knowledge for network reconstruction, are:

A1

First, to enable the access to and the interoperability of pathway data from different data sources. This warrants the integration of biological knowledge from pathway databases encoded in an ontology into the R Project for Statistical Computing. This aim is accomplished by the implementation of a new software package `rBiopaxParser`.

A2

Second, the computational transformation and merging of available pathway data. Here, a proof-of-concept for the transformation and merging of pathway data from different sources is provided. This aim is reached by applying the newly-developed software to existing pathway databases and compiling a consensus network.

A3

Third, to implement a workflow for the integration of pathway knowledge into methods for network reconstruction. An exemplary reconstruction of a gene network is performed, integrating the merged consensus network into methods for network reconstruction.

A4

Fourth, to evaluate the results of network reconstruction with and without integrated prior knowledge. This evaluation of the performance of methods for network reconstruction is assessed based on the results of the exemplary reconstruction with and without integrated prior knowledge.

Figure 1.1 depicts the underlying workflow of the methods used within this thesis: Gene perturbations, i.e. knockdowns of genes in cell line samples, are measured using microarrays. This data is analyzed and used as input for the network reconstruction algorithm. Literature knowledge, stored in pathway databases and encoded using an ontology, is parsed and transformed into a

directed graph, representing the interactions between the perturbed genes. The experimental data and the generated prior knowledge network are used as input for Nested Effects Models (NEMs) to reconstruct the network topology of the perturbed genes.

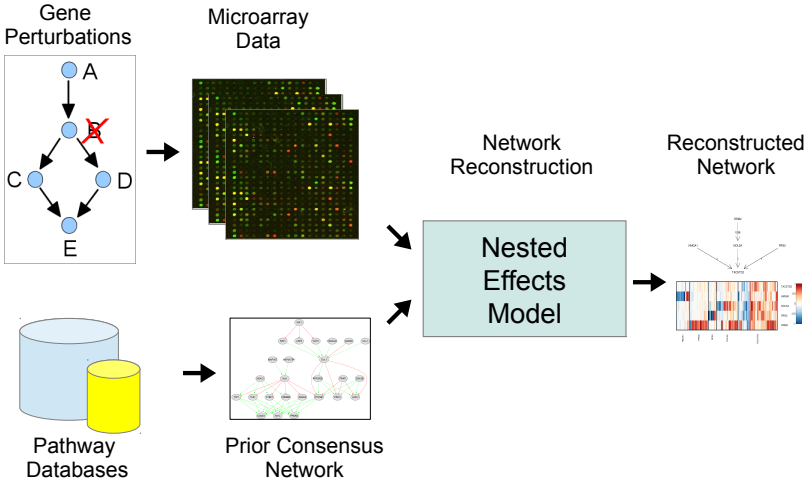

FIGURE 1.1 *Detailed workflow of integrating prior knowledge into methods for network reconstruction.*

This thesis touches upon different areas of computer science, statistics, bioinformatics and computational biology. Understanding the underlying mechanisms, for example modeling knowledge via ontologies, or measuring gene expression via microarrays, is a prerequisite.

The current Chapter 1, *Introduction*, covers the most relevant aspects of biology, computer science and statistical bioinformatics for this thesis. In Section 1.2, biological pathways and their organization and structure are introduced. Section 1.3 describes high-throughput technologies used to measure experimental gene expression data. Afterwards, Section 1.4 covers ontologies as a way to model knowledge of a specific domain. Section 1.5 presents the workings of methods for network reconstruction along with a general overview of published methods.

Chapter 2 *Materials and Methods* covers the methods, software and modeling approaches used within this thesis as well as the experimental data used to conduct network reconstruction. First, Section 2.1 presents BioPAX, a widely used ontology to model pathway knowledge. Here, an overview on pathway modeling approaches is given. Furthermore, a number of pathway databases

which collect and curate biological pathway knowledge are presented. Second, Nested Effects Models (NEMs), a framework of methods for network reconstruction, are covered in Section 2.2. The section explains in detail how a pathway topology is reconstructed by analyzing gene expression data. Section 2.3 details the experimental data and reveals the setup of the perturbation experiments used within this thesis. Finally, Section 2.4 introduces the R Project for Statistical Computing, a language and programming environment. This section also describes several R software packages implementing NEMs as well as functions to perform statistical bioinformatic analyzes.

These methods act as a foundation for my own work, presented in Chapter 3 *Results*. These results describe in detail how the aims defined for this thesis were reached. The three central and novel results are described in the following sections: In the first section of Chapter 3, the new R software package `rBiopaxParser` is introduced in detail. The focus of this section lies on the workflow, how BioPAX pathway data is parsed, the internal data model and how this data can be accessed and visualized. This section fulfills the first aim **A1**, to integrate biological knowledge into the R Project for Statistical Computing. In Section 3.2 *Prior Knowledge Generation*, the merging of several BioPAX databases and their transformations into suitable prior knowledge input is described. This section offers a solution for the second aim **A2**, as a proof-of-concept for the merging of pathway data from different databases using the newly implemented R package. *Network Reconstruction*, the last section of Chapter 3, applies NEMs to reconstruct networks from experimental data integrating prior knowledge parsed from different pathway databases, which fulfills the third aim **A3**.

The achieved results are assessed in Chapter 4 *Discussion*, weighing the pros and cons of the used methods, the workflow implementation and the results of network reconstruction. Section 4.1 discusses the data modeling format BioPAX and compares the R package `rBiopaxParser` to similar available approaches. In Section 4.2 *Prior Knowledge Generation*, the used data sources and the steps towards merging a consensus prior network from literature knowledge are analyzed. The last section of this chapter, Section 4.3 *Network Reconstruction*, evaluates the reconstructed network with respect to differences in the results for network reconstruction with and without integrated prior knowledge. The

evaluation of the results in Chapter 4 *Discussion*, accomplishes the fourth aim **A4** of the thesis.

Finally, Chapter 5 *Conclusion* rounds off the work described within this thesis and mentions (con-)current developments in the fields of standardization of pathway modeling formats, pathway databases and computational pathway generation.

1.2 Biological Networks and Pathways

The mechanisms of the inner cell are commonly described using the pathway representation. In biological terms a "pathway" is used to describe a collection of processes within a cell that lead to one or more actions. The graphical representation of these processes enables the reader to understand complex relationships and interactions much more easily compared to free-text descriptions (Kohn, 1999). Pathways are a way of organizing the multitude of cellular processes and events into modules responsible for a certain process of a higher abstraction level (Novère et al., 2009), for example cell proliferation or cell death. While there is usually agreement on the existence and function of these high-level processes, the specific molecules and their interactions are often disputed and a matter of current research. The following sections aim to give the reader an idea of the organization of common pathways as well as to illustrate examplary pathways.

While the nomenclature in literature often differs, usually pathways are divided into three subgroups: Metabolic pathways, signaling pathways and regulation of gene expression. Within this introduction of the biological fundaments the general nomenclature of Karp (2010) and Alberts (2008) is used, which define metabolic pathways as series of chemical reactions with educts and products, while signaling pathways are defined as cascades of molecular interactions and cellular processes.

Graphical representations of pathways often contain not only processes subject to only one of these pathway groups, but incorporate signaling events as well as regulatory events and biochemical processes. The graphical representation of pathways commonly includes a multitude of biological processes, for example: Biochemical reactions of metabolites, the assembly of complex

molecules, cell signaling, phosphorylation or the transport of proteins within the cell. Section 1.2.4 *Visualization of Pathway Knowledge* illustrates different possibilities and standards for visualizing pathway knowledge at different levels of detail.

Furthermore, pathways may be represented as graphs, allowing a broad variety of mathematical and bioinformatical operations. This makes it possible to use pathway information in a multitude of different algorithms. Due to the directed nature of signaling and catalytic processes, pathways are often depicted computationally in a manner of directed graphs. A more pronounced definition of the participants and interactions within pathways, as utilized in a computational manner within this thesis, is given in Section 2.1 *Modeling Pathway Knowledge*.

1.2.1 Metabolic Pathways

Although the continuously running metabolic pathways are a fundament of cellular activity, this thesis focuses on the more abstract regulatory events of signaling pathways and gene regulation. However, for the sake of completeness metabolic pathways are shortly described.

A metabolic pathway is characterized by a series of chemical reactions catalyzed by enzymes. Enzymes may use organic as well as inorganic co-factors for their catalysis. A number of distinct major metabolic pathways are known and form the so-called metabolic network of the cell. The metabolic network is a central aspect to sustain homeostasis of the cells, a balance between educts and products for the various processes. The fact that metabolic processes are fundamental biochemical reactions, catalyzed by enzymes, has sparked a strong industrial research interest. The enzymatic nature of the reactions means that genetically modified yeast or bacteria may be used to increase the amount of product or lower the energy costs for reactions. Additionally, the cross-species similarity of the metabolism means that new findings can be easily validated and adopted (Pace, 2001). New findings as well as suggestions for techniques to extend and validate metabolic pathways have been published for a long time (Stanier, 1947). Due to the extent of available data, metabolic pathway curators have been early adopters of database infrastructure (Ochs and Conrow, 1991). Nowadays, many pathway databases detailing literature

knowledge of metabolic pathways are available, for example the well-known Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Ogata et al., 1999), the Human Metabolome Database (Wishart et al., 2007) or MetaCyc (Karp et al., 2002).

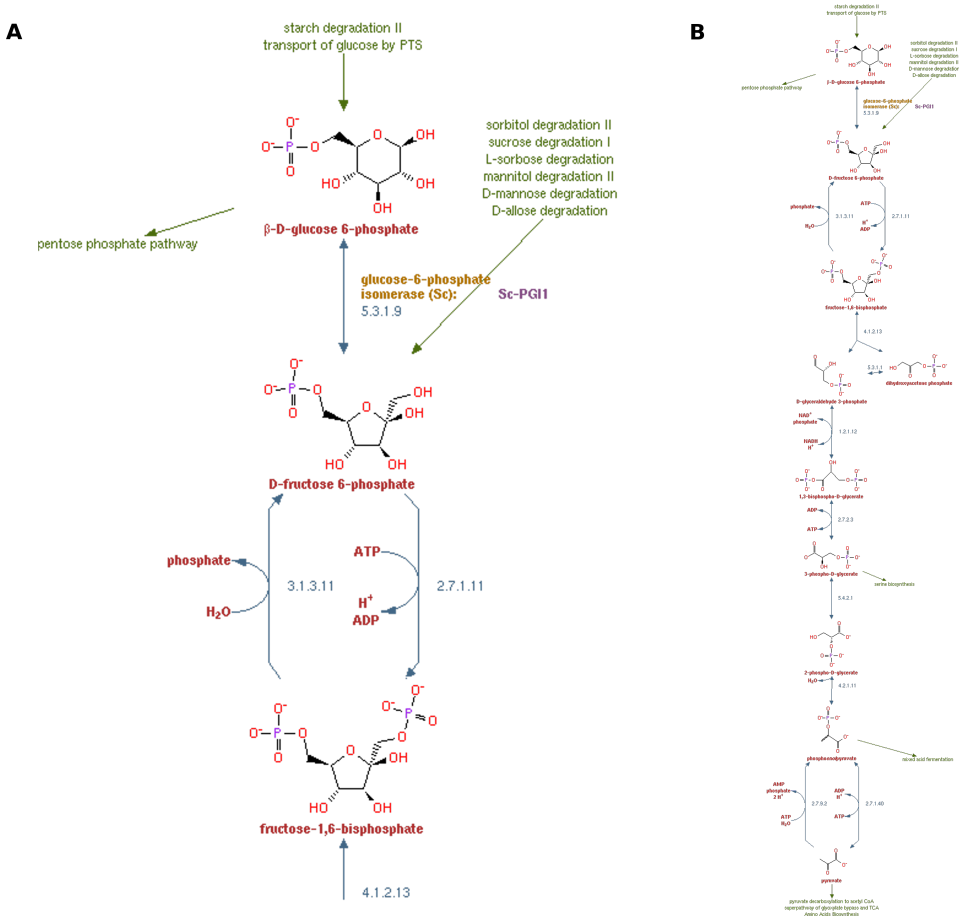

FIGURE 1.2 *Representation of the glycolysis pathway in yeast (Saccharomyces cerevisiae). Part A shows a portion of the whole pathway with detailed biochemical reactions. Part B shows the complete pathway. Courtesy of MetaCyc. (Karp et al., 2002).*

In Figure 1.2 the glycolysis pathway as shown in MetaCyc is displayed as an example of a metabolic pathway. Its main task is the conversion of glucose into pyruvate at a gain of energy, in order to generate energy-rich adenosine triphosphate (ATP) (Meyerhof, 1927). The first part of Figure 1.2 (A) shows a portion of the pathway, depicting the chemical reactions as edges and the chemical compounds in their structural and molecular formulas. Green edges represent molecule transports and blue edges represent biochemical reactions. The source or destination of a transport is written in green text. Chemical

compounds are stated in red text and enzymes catalyzing reactions are encoded in blue text, stating their Enzyme Commission number (e.g. 2.7.1.11). The second part (B) of Figure 1.2 displays the complete glycolysis pathway.

1.2.2 Signaling Pathways

Signaling pathways are chains of molecular interactions and cellular processes which let a cell respond to changes in its microenvironment. This communication can appear in a variety of settings: The signaling may occur between different organisms, like mating yeast cells or early embryos of mammals, it may occur between different cells of the same organism, or the source and target of cellular signaling can be within the same cell.

When compared with metabolic networks, which have been published as early as 1927 (Meyerhof, 1927), the process of signal transduction has only recently been discovered. In 1994 Martin Rodbell recieved the Nobel Prize in Medicine for his discovery of the G-protein, a major protein family involved in transmitting a signal from outside the cell to its inside, in 1971 (Rodbell et al., 1971; Coles, 1994).

Problems with cellular signaling events may coincide with cancer development, autoimmune diseases and metabolic diseases like diabetes (Karp, 2010). However, the complexity of the signaling networks makes good treatment very hard to achieve. Due to the complexity of the signaling network, pathway boundaries are often arbitrarily chosen or different pathways might overlap and share the same molecular interactions (Schaefer et al., 2009). Examples of signaling pathways are cell proliferation and cell death, apoptosis, as well as tissue repair and immune responses.

For example, apoptosis, the programmed cell death, is a central process in embryonal development, in cancer suppression and immune response (Kerr et al., 1972). Furthermore, apoptosis is also a normal deconstruction process for cells that are no longer needed. In an average human about 50 billion cells undergo apoptosis daily (Alberts, 2008). Indeed apoptosis is a major antagonist in the fight of the human body against cancer development (Karp, 2010). Cells which have sustained serious DNA damage might become cancerous and proliferate further. Apoptosis hinders cancer development by triggering on cells

with serious DNA damage. One of the best-researched parts of the apoptosis pathway is the signaling of the tumor necrosis factor (TNF).

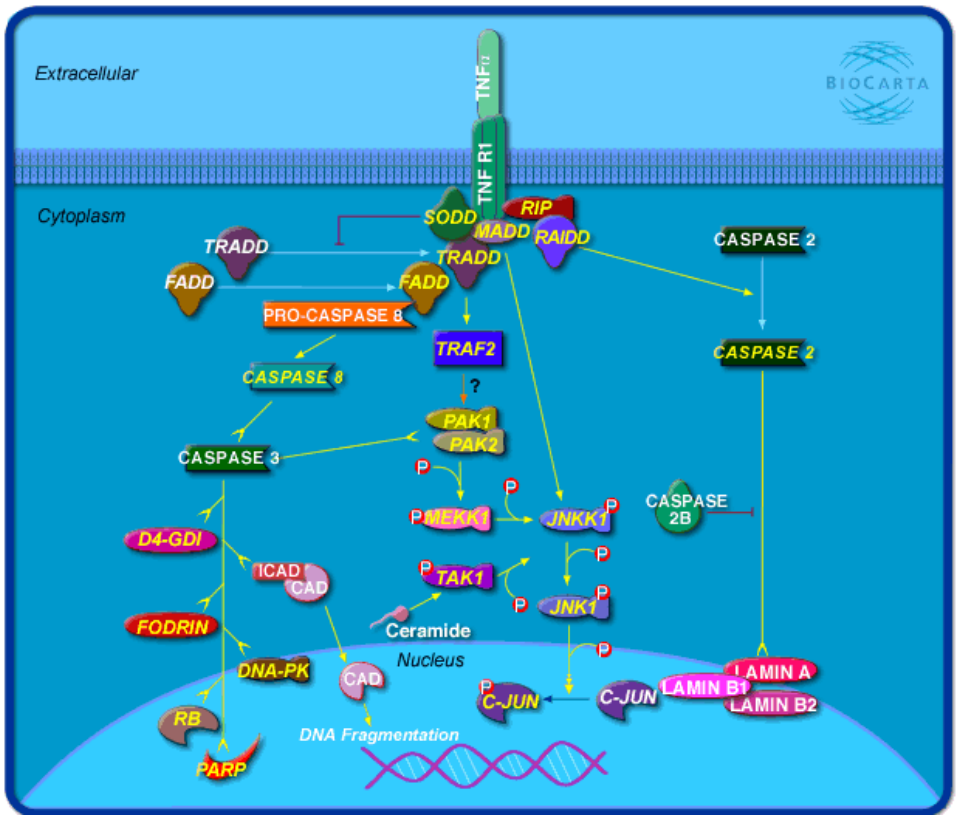

FIGURE 1.3 *The apoptosis signaling pathway of the tumor necrosis factor R1 in homo sapiens. (TNFR1 Signaling Pathway, Courtesy of BioCarta) (Nishimura, 2001).*

In Figure 1.3 the apoptosis pathway downstream of the TNF receptor 1 is illustrated (Nishimura, 2001). Here, apoptosis is induced by binding of the TNF protein to the transmembrane TNF receptor. This leads to a complex assembly by binding the proteins TRADD and FADD. Further downstream, this complex binds two procaspase-8 molecules, which leads to an activation of caspase-8 and the initiation of programmed cell death (Karp, 2010).

The signaling processes within the cell are usually very complex, with possible feedback loops and self-regulation, and might induce a number of metabolic pathways downstream. Different pathway collections and databases are available, often including not only signaling but also metabolic and regulatory information (Bader et al., 2006; Schacherer et al., 2001; Krull et al., 2006).

1.2.3 Regulation of Gene Expression

Genes, defined by sequences on the DNA within the nucleus, are continuously read from the DNA and assembled within the cell to take part in almost all cellular activities (Alberts, 2008). In general the term "gene expression" describes the process of transcription of genomic DNA into messanger RNA (mRNA) molecules, which are later translated into polypeptides and assembled into proteins. In a nutshell, gene expression is a two-step process. The first step transcribes a gene from the DNA to RNA and the second step translates this RNA into a protein (Karp, 2010). The process of gene expression is used by eukaryotes and prokaryotes alike. Gene expression can be measured on the mRNA level, i.e. transcriptomics, and on the protein level, i.e. proteomics. The expression levels of the transcriptome and the proteome depict the current state of the cell and influence responses to cellular signaling as well as control of the metabolic processes. Among many other aspects, gene expression regulation is responsible for cellular differentiation in adult stem cells, leading to daughter cells which differ vastly in size, shape and function.

The regulation of gene expression includes various mechanisms which can be used to adapt the production of proteins or RNA within a cell. An overview of these mechanisms is shown in Figure 1.4.

Proteins can be regulated for short durations of time by phosphorylation, and on DNA level the transcription of genes can be regulated for longer periods via processes such as methylation. So-called transcription factors play a major role in the up- and downregulation of gene expression. Transcription factors are proteins, which can bind to the DNA in the nucleus and therefore regulate gene expression by making the transcription of corresponding genes more or less likely. This is called transcriptional regulation. On the other hand, post-transcriptional regulation is the control of gene expression at the RNA level via processes like RNA capping or alternative splicing. Transcribed RNA has to use the nuclear export mechanism to leave the nucleus towards the cytoplasm via a nuclear pore. Finally, translational regulation controls the abundance of protein synthesis from exported RNA. Following the expression of a gene, translated proteins can be regulated via post-translational modifications, protein binding or self-regulation (Alberts, 2008).

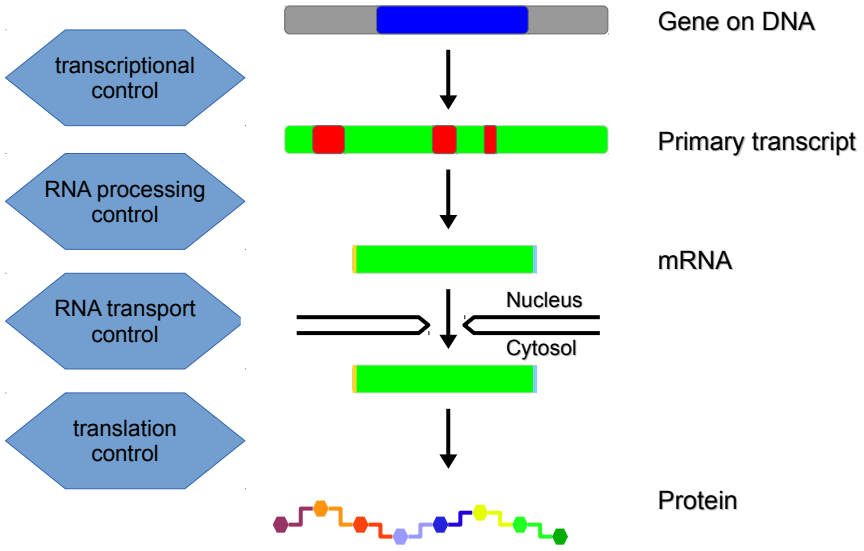

FIGURE 1.4 *The process of gene expression and possibilities for its regulation. Adapted from Wikimedia Commons (Arnelh, 2009).*

1.2.4 Visualization of Pathway Knowledge

With his remark, "A good sketch is better than a thousand words", Napoleon Bonaparte probably did not have biologists and bioinformaticians in mind. However, visualization of pathways has been performed long before personal computers or databases were commonly used (Meyerhof, 1927; Stanier, 1947; Hendricks, 1953). Visualizing pathways helps readers to understand complex molecular interactions and relationships more easily. A single pathway sketch can contain dozens of molecules or chemicals, a huge number of interactions and can still be perceived by a human. However, this information would be tedious to read and difficult to understand in text form. A standardized computational representation of biological networks has become desirable, especially with the recent surge in new knowledge generation in biology and medicine due to the advancements in bioinformatics and computational biology.

Before the start of the millennium, Kohn and colleagues (Kohn, 1999) had already begun first attempts to standardize pathway representation. The proposed "Molecular Interaction Map" (MIM) was intended as a diagram convention aiming at unambiguous representation of pathways. By defining a

fixed set of glyphs and a mapping convention, pathway sketches became less ambiguous and easier to understand. The focus of MIMs was on modeling reactions of molecules and their interactions. In 2005 Kitano and colleagues undertook another approach, which allowed graphs to have a much finer granularity, for example depicting all possible states of tyrosine and theronine phosphorylation sites of an molecule (Kitano et al., 2005). Although both approaches aimed at standardization, their scope was too limited and lacked the support for computationally encoding and handling diagrams.

Finally, in 2009 a joint work of Kohn and Kitano was published (Novère et al., 2009), proposing the Systems Biology Graphical Notation (SBGN), which consists of three different diagram types, shown in Figure 1.5.

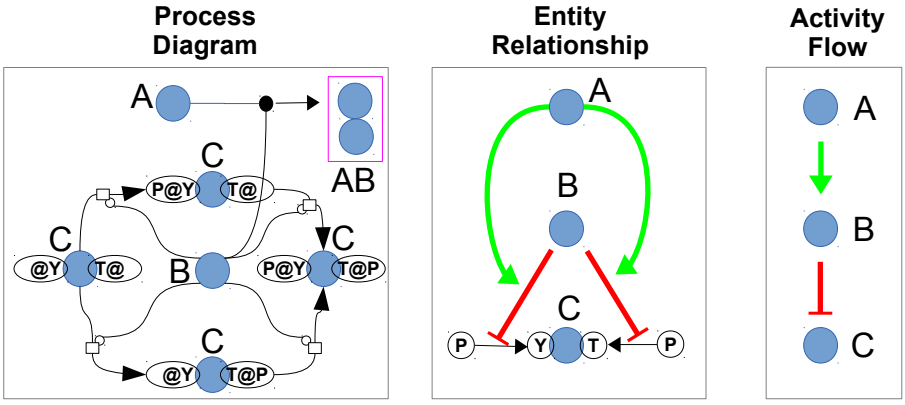

FIGURE 1.5 *Diagrams of the same biological process visualized using the three different diagram types available in SBGN. Reproduced according to (Novère et al., 2009).*

The process diagram (left) resembles Kitano's notation to represent all possible states, educts and products of a biological pathway at finest granularity. In this figure molecule C, with all possible states of its two phosphorylation sites (T@ and Y@), and its interaction with molecule B, which can bind to molecule A, are illustrated. In contrast, the entity relationship diagram (center) is quite similar to Kohn's Molecular Interaction Maps and mainly focuses on describing the interactions of entities and their influence upon each other, but leaves out exact variable states. Finally, the activity flow diagram (right) is the coarsest diagram, depicting only activating and inhibiting relationships between molecules. SBGN graphs, or rather the information contained in them, can be represented and exchanged using SBGN-ML, a markup language to

encode SBGN entities and interactions. This is also the reason why SBGN itself does not dictate shape, color or layout of graphs. These details are part of layout styles which can be applied to any SBGN-ML encoded graphs.

There are a number of possibilities available to generate pathway sketches programmatically. The library *libSBGN* is provided as a Java and a C++ library and allows programs to visualize graphs in SBGN notation using the SBGN-ML schema[1] (Iersel et al., 2012). *Cytoscape* is a Java-based modular software for generating, editing and visualizing networks and graphs (Shannon et al., 2003). A large number of plugins are available and offer extended functionality like pathway analyzes, interfacing with the R Project for Statistical Computing as well as importing SBGN-ML diagrams (Lotia et al., 2013). *Graphviz*, short for Graph Visualization Software, is a collection of open-source tools initially developed by the AT&T Bell Labs for drawing graphs (Ellson et al., 2002). Graphviz is available for many operating systems, and its main focus is to offer layouting functionality for common graph types.

A large number of further tools to visually explore and map biological networks are available (Suderman and Hallett, 2007), for example *VisANT* (Hu et al., 2008), *CellDesigner* (Funahashi et al., 2003) and *PathVisio* (Iersel et al., 2008).

The pathway databases used within this thesis and the corresponding data models are further detailed in Chapter 2 *Materials and Methods*, Section 2.1 *Modeling Pathway Knowledge*.

1.3 Omics-Technologies

Measuring the abundance of proteins, metabolites and expressed genes within cells is a requirement in order to pursue further insight into biological pathways. Traditional techniques measure single protein or RNA expression levels. Nowadays the so-called "omics" in biology, for example genomics and proteomics, cover the complete genome or proteome and measure all parts of the field. This section introduces methods to measure the abundance of gene expression within cells. Furthermore, the last part of this section explains the experimental design and possible ways to analyze microarray experiments.

[1] The libSBGN project: http://www.sbgn.org/LibSBGN

1.3.1 Measuring Gene Expression

Measuring gene expression levels (i.e. mRNAs) within cells enables the researcher to trace the change within the cells, for example after drug treatment or due to immune response. Several traditional methods are available to measure the current level of gene expression (Alberts, 2008). Northern and western blotting are methods to measure mRNA and protein levels, respectively, by using gel electrophoresis. For northern blotting the sample is hybridized to a complementary target mRNA sequence and for western blotting the sample is probed with a matching protein antibody. A drawback for both methods is the relatively high consumption of material, which might be very valuable and hard to come by, for example biopsies of human cancer tissue. Another approach for measuring the mRNA level of cells is the reverse transcriptase real-time quantitative polymerase chain reaction (RT-qPCR), where qPCR is used to amplify and measure a DNA sequence which was previously acquired by generating the complementary DNA (cDNA) using reverse transcriptase (Karp, 2010). Although recent development brought plates for hundreds of parallel runs of RT-qPCRs, the sheer amount of known genes, roughly 25,000 for Homo sapiens, makes these methods more convenient for validation purposes of smaller gene sets, but less useful for exploratory research of the entire transcriptome.

On the other hand, "omics" technologies like microarrays and RNA sequencing allow expression profiling of the whole human genome in a single run (Alberts, 2008). These methods enable fast and reproducible expression profiling on a whole-genome scale.

1.3.2 Gene Expression Profiling using Microarray Technology

Microarrays are chips with an array of thousands of oligonucleotide probes attached to their surface. These oligonucleotide sequences bind specific DNA or RNA targets, and labeling techniques are used to quantify the abundance of these targets.

Using microarray scanners, the intensity of light emitted by the labels allows comparative quantification of target expression. Originally microarrays evolved from parallelized southern blotting, a method similar to northern blotting, where DNA is fragmentized and fixated and then probed using a

single complementary DNA sequence (Augenlicht and Kobrin, 1982). The 1990s saw the introduction of commercially available microarrays and computer-aided scanning devices (Maskos and Southern, 1992), and a development from the first custom spottable cDNA arrays with comparatively few probes to the first whole genome chip for Saccharomyces cerevisiae (Lashkari et al., 1997).

Currently most microarrays come pre-spotted and enable whole genome expression profiling in many different settings. These include, for example, different species, like human, rat or mouse genomes, and different types of targets, for example mRNA, miRNA and single nucleotide polymorphisms.

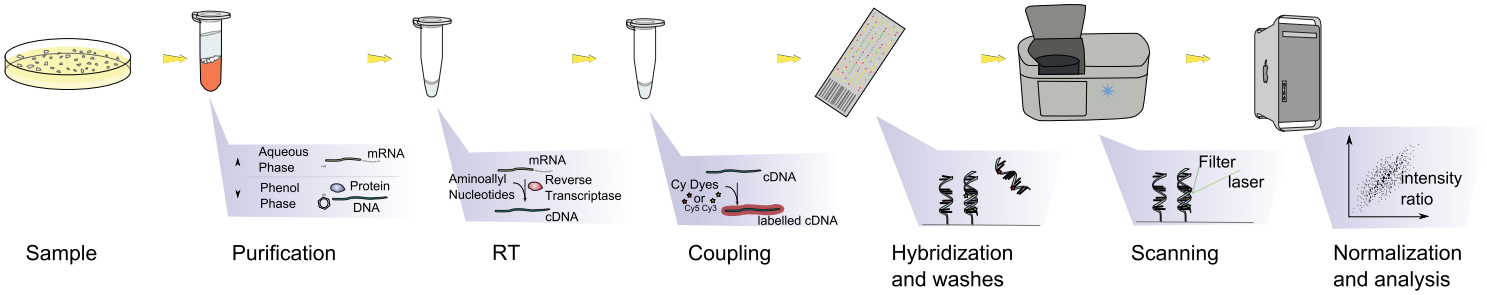

FIGURE 1.6 *This figure illustrates the single steps in the workflow of microarray experiments. (Courtesy of Wikimedia Commons, Public Domain) (Squidonius, 2008).*

Figure 1.6 shows the workflow of mRNA microarray experiments. In the first step the cells' mRNA is retrieved by purification of the samples, for example from tissues or cell lines. Then cDNA is created by applying reverse transcriptase (RT) and in the coupling-step the cDNA is labeled with fluorescent markers. In the next step labeled cDNA is then hybridized onto the microarray and non-binding fragments are washed off. Finally, the last step of wet lab work is reached: The microarray chip is inserted into the scanning device and a picture of the light intensities of all probes on the chip is scanned (for an example, see Figure 1.7).

Single-channel and two-channel microarray chips exist. Formerly two-channel chips were very popular, allowing two samples, for example control and treatment, to be hybridized to the same chip. However, experiment design proved to be more complex and was not easily adopted for large cohort studies in patients (Smyth, 2004). Drastically reduced prices per microarray chip as well as application in clinical practice has led to a dominance of single-channel chips nowadays.

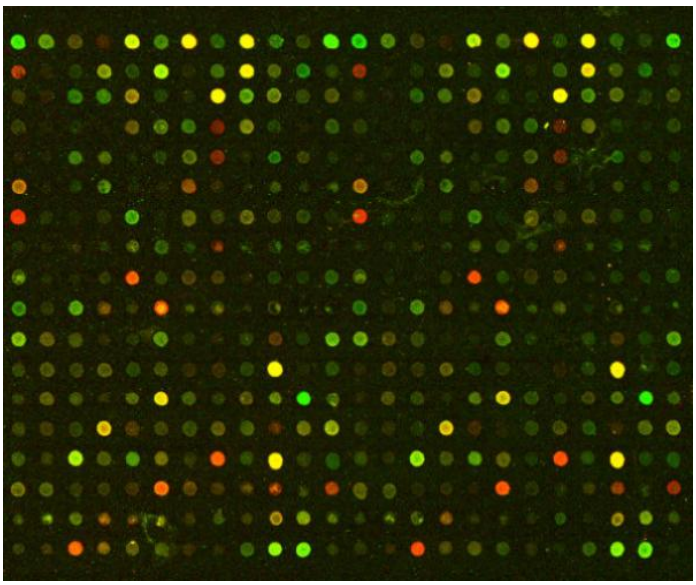

FIGURE 1.7 *This image shows a portion of a scanned two-color microarray. The individual probes and their fluorescent green and red coloring are clearly visible. (Courtesy of the Transkriptomanalyselabor at the University Medical Center Göttingen).*

1.3.3 Experimental Design of Microarray Experiments

Several mechanisms to measure gene expression have been introduced in the previous sections. In order to measure whole genome mRNA levels, RNA-sequencing or mRNA microarrays are available. The statistical design of microarray experiments is essential to correctly measure and analyze the effects of biological interest. The basic idea for many analyzes is the measurement and comparison of expression levels of a single gene between two or more conditions (Smyth, 2004). The type of analysis depends on several factors, a main aspect being the chosen end-point of an experiment. In general most microarray experiments belong to one of two categories:

The first category are cohort studies, where measurements from samples across a specific population are correlated with time-to-event data. Cohort studies use microarrays to measure whole genome expression profiles of patient samples from a study cohort and try to correlate their expression levels with clinical parameters, for example tumor progression or survival time.

The second category are group-wise comparisons, where measurments of samples from different groups are compared. In group-wise comparisons microarrays are used to compare two or more groups of samples on a gene-by-gene

basis. Statistical tests are used to determine significant differential expressions. Examples suited for group-wise comparisons are the analysis of different types of the same cancer, or the testing of samples treated with drugs or irradiation against untreated controls. Figure 1.8 illustrates the different approaches in a basic sketch.

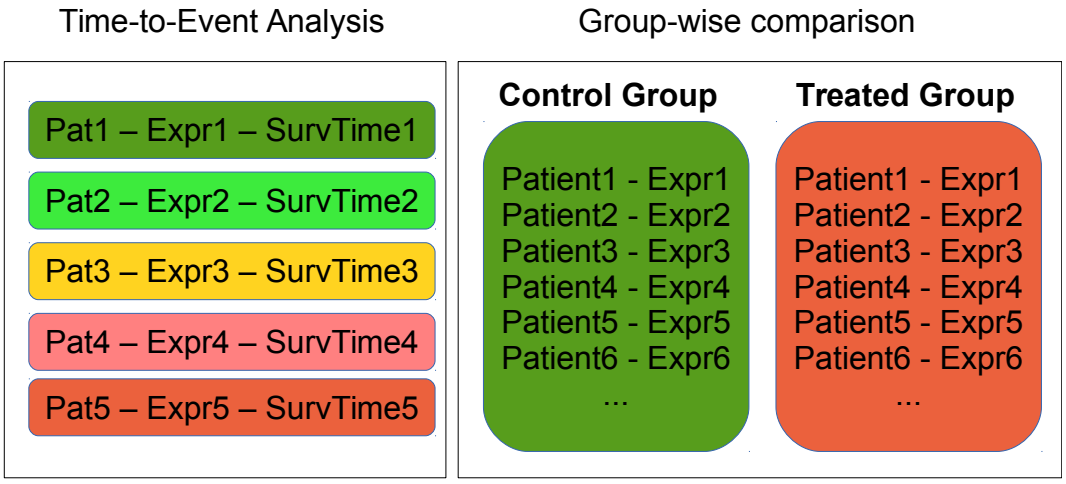

FIGURE 1.8 *In time-to-event analyzes the objective is to model the occurance of an event, for example death, as a function of time and other variables, for example the expression level of a specific gene. On the other hand, group comparisons try to evaluate whether there is a significant difference of the mean expression levels between the groups.*

Gene perturbation experiments belong to the category of group-wise comparisons. A common setting is that within samples of a specific cell line a gene is perturbed and subsequently compared to control samples of this cell line. Various approaches to perturb genes are available. Overexpression of gene products can be achieved by injecting corresponding gene and a promoter into the target cell via transfection of a viral vector. Furthermore, genes can be down-regulated by knockout and knockdown protocols (Alberts, 2008). In a knockout approach, the DNA corresponding to the gene is rendered unusable and can subsequently no longer be transcribed. Consequently, this leads to a complete lack of corresponding gene product. In gene knockdowns, also called RNA interference, the mRNA product of a gene is targeted by introducing short hairpin RNA (shRNA) or small interfering RNA (siRNA) into the cell.

These gene knockdowns do not entirely remove all corresponding gene products from within the cell but constantly degrade newly transcribed mRNA.

While the assignment of samples to single color microarrays is trivial, experimental design for two-color microarray experiments poses a bigger challenge. This is due to the fact that there is a bias between the colors, which leads to a shift of expression values measured by red compared to green. In order to handle this bias the dye-swap design was commonly used. Although several different approaches were published (Yang and Speed, 2002), the basic idea usually remains the same: By design, the amount of replicates on green and red channels are identical and the expression ratios between green/red-channel are used for analysis between groups.

The experimental data used within this thesis is described in Section 2.3 of Chapter 2 *Material and Methods*. The results of the statistical analysis of the experimental data can be found in Section 3.3.1 of Chapter 3 *Results*.

1.4 Modeling Knowledge using Ontologies

A vast amount of knowledge about biological processes and molecular interactions has been accumulated over the past decades. In order to make use of this complex data, it has to be archived in an accessible and well-documented way.

Modeling knowledge or data for storage and usage in computer systems is a difficult task. Usually, once the architecture of data storage has been decided upon, the users have to cope with the design decisions for a long time. This poses a special challenge for biological knowledge: On the one hand biological entities and their interactions are highly complex. Relationships exist between DNA, RNA, proteins and small molecules, and interactions as well as feedback between them is possible, as illustrated in the different examples of Section 1.2 *Biological Pathways*. On the other hand, the underlying assumptions on the data structure might change or might be extended with new entities or relationships. Although a fundamental change of underlying assumptions may not be addressable, the advent of ontologies in computer science has offered a flexible, extensible way for modeling specific domains of knowledge (Gruber, 1995; Berners-Lee et al., 2001).

The term "Ontology" originates from philosophy, where it denotes the studies of existance and reality, known as a branch of metaphysics, founded on the work of the philosopher Aristotele (Burkhardt and Smith, 1991). In computer science an ontology can be defined as following:

> *"A specification of a representational vocabulary for a shared domain of discourse – definitions of classes, relations, functions and other objects – is called an ontology."* (Gruber, 1993)

An ontology is always based on a conceptualization, i.e. an abstract, simplified view of the domain which is to be modeled. An ontology is a specific implementation of this conceptualization, it defines existing classes of objects, as well as the relationships between them (Gruber, 1995).

The main goals for developing an ontology are to formalize the structure of domain-specific information, to separate knowledge about the data structure and the data itself, and to enable the reuse and sharing of the structure and knowledge (Noy et al., 2001). Furthermore, it is possible to model description logics, which enables automated reasoning and inference based on the knowledge base and logical operations (Hitzler et al., 2011).

Every ontology is made up of a number of core components: Classes define types of objects or things, properties define the respective attributes and features of these classes. Restrictions on these properties allow the modeling of assertions and pre-determined values. Classes can be instantiated for specific objects and are called instances. Properties of objects can either reference objects or consist of numeric or textual facts, for example a name property (Noy et al., 2001). Furthermore, rules in an if-then form and axioms can be used to infer statements about a domain of knowledge.

In practice ontologies are often used to add a layer of abstraction when the underlying reality is very complex and the available knowledge can be detailed in very different granularity. An example of this would be a full-length research paper about Gene A activating Gene B compared to the simple statement "Gene A activates Gene B". On a very high abstraction level these statements would be identical, however this conclusion cannot be drawn by comparing the free text format of a research paper and the short statement (Plessis et al., 2011).

Another notable development in knowledge encoding using ontologies is the concept of so-called nanopublications. Starting with so-called microattributions for genomic findings (NatGenEditorial, 2008; Giardine et al., 2011), nanopublications were introduced as the idea of being the smallest publishable scientific knowledge facts (Groth et al., 2010; Mons et al., 2011). The concept has received considerable attention and aims at offering a standardized modeling framework for scientific knowledge, with the goal in mind to interconnent findings and infer new findings automatically in the near future (Beck et al., 2012; Patrinos et al., 2012). Lately, the OpenPhacts website has been opened to support the publication of nanopublications in biosciences (Sansone et al., 2012).

Ontologies have been defined to model knowledge domains within biology and medicine, for example to encode the knowledge about the biological pathways introduced in Section 1.2.

1.4.1 Overview of Published Biomedical Ontologies

A large number of ontologies have been suggested, defined and published in the last decade. Several web sites are available which list and categorize biomedical ontologies (Noy et al., 2009; Rubin et al., 2008), even a search machine for these ontologies exists (Orchard et al., 2011). Examples of notable developments in the biomedical community are the ontologies Chemical Entities of Biological Interest (ChEBI, Degtyarenko et al., 2008), Gene Ontology (GO, Ashburner et al., 2000), as well as the ontology for Biological Pathways Exchange (BioPAX, Demir et al., 2010).

The first two are part of the Open Biomedical Ontologies Foundry (OBO, (Smith et al., 2007)), a collaboration to standardize the way biomedical ontologies are developed and to allow cross-ontology referencing between members of the OBO Foundry. ChEBI is a dictionary of small chemical molecules and molecular entities commonly used in metabolic processes, as well as pharmaceuticals, laboratory reagents, and subatomic particles. However, more complex macromolecules like proteins are generally excluded. The idea behind ChEBI is to provide an extensive, cross-referencing dictionary of basic biochemical entities, their machine-readable structural information, their biological role

(e.g. antibiotic or hormone) and their applications (e.g. pesticide or drug) (Degtyarenko et al., 2008).

The Gene Ontology emerged from a cooperation of three model organism databases: FlyBase, Mouse Genome Informatics (MGI) and the Saccharomyces Genome Database (SGD). A major goal of GO arose from the discovery that there are large amounts of DNA sequences which are identical between species, as well as functional conservation within these genes (Ashburner et al., 2000). The desire for a common site of annotation for genes is a consequence of this finding. The idea of GO is to model the knowledge about genes and gene products across species and to provide access to this information. GO consists of three independent ontologies, each modeling a different domain: biological process, molecular function and cellular component (Ashburner et al., 2000). Aiming for a generalizing model, the cellular component ontology models the parts and pieces of eukaryotic cells and their microenvironments. The biological process ontology contains all processes and events which take place within cells and organisms. Finally, the molecular function ontology describes the functional activities of proteins within a cell. GO is constructed in a manner that the ontologies can be understood as a directed acyclic graph. Each node in this graph represents one GO term, its name, annotations and references to other databases or GO domains. In this graph every GO term is connected via edges to its parents and children, representing the ancestry between these GO terms. This hierarchical modeling enables GO to provide an open controlled vocabulary where the user is able to retrieve knowledge about a certain item, as well as more generalized or detailed knowledge about the GO term. GO is not static, but continuously developed and curated as the biological knowledge increases (Consortium, 2008). Being widely used and hierarchical in structure, GO has sparked numerous new approaches in bioinformatics. Statistical testing procedures (Beißbarth and Speed, 2004; Beißbarth, 2006) can be used to find significantly overrepresented GO terms within a group of genes. Furthermore, semantic similarity measures have been proposed to assess functional similarity of genes (Fröhlich et al., 2007b; Pesquita et al., 2008) and pathways (Guo et al., 2006). Based on these measures a large number of methods have been proposed, ranging from disease gene identification (Jiang et al., 2011) to drug repurposing (Andronis et al., 2011).

The ontology Biological Pathways Exchange (BioPAX) (Demir et al., 2010) aims at easing the sharing of pathway knowledge by offering a standardized knowledge model for the pathway domain. Research groups and database providers can use this common model to make their information easily accessible and sharable by users. The main classes of BioPAX are physical entities, interactions and pathways. Physical entities are defined as all physically existing objects, for example proteins, small molecules, as well as RNA and DNA fragments. The interaction class and its subclasses define all biological processes and events within pathways, e.g. complex assembly, cell transport and regulatory events. Depending on the interaction, its participants are physical entities, interactions and whole pathways. The pathway class models pathways which are made up of a number of interaction instances. A more detailed account of the BioPAX ontology is given in Section 2.1.2 *BioPAX Format for Encoding Knowledge of Biological Pathways* of Chapter 2 *Materials and Methods*. A large number of pathway databases are available in BioPAX format (Bader et al., 2006) and several well-known sources for BioPAX-encoded data are described in Section 2.1.3 *Pathway Databases*.

1.5 Network Reconstruction

In bioinformatics and systems biology the term *network reconstruction* denotes methods which aim at inferring biological networks from experimental data. The predominant goal of these methods is to infer new insights into the processes within cells (Markowetz and Spang, 2007). Methods for network reconstruction either perform de-novo reconstruction of a new network from scratch or extend previously known pathways by further nodes or edges. The central challenge for these methods is that complex interactions involving a multitude of genes have to be inferred from sparse and noisy high-dimensional data (Werhli and Husmeier, 2007). This challenge has attracted many researchers from the fields of statistics and computer science alike. Depending on the specific aims, the experimental data and the availability of prior knowledge, different approaches for network reconstruction have been pursued (Markowetz and Spang, 2007). The following section offers an overview of commonly chosen

aims and approaches for network reconstruction as well as an overview of published methods.

1.5.1 Aims and Approaches for Network Reconstruction

The general idea of network reconstruction in bioinformatics is to derive knowledge about biological interactions of molecules from experimental data. The result of methods for network reconstruction is usually a graph representing the inferred biological interactions. These resulting graphs can have directed or undirected edges, depending on the chosen algorithm. On the one hand, the data required can differ from algorithm to algorithm. On the other hand, the choice of measured tissue and measuring technology can restrict the possible algorithms for network reconstruction.

Network reconstruction has been conducted on a wide range of different experiments using various different statistical inference or machine learning approaches. A plethora of methods have been proposed in the statistics as well as in the bioinformatics community. Several extensive reviews of popular methods (Markowetz and Spang, 2007; Ideker and Lauffenburger, 2003; Hecker et al., 2009; Werhli et al., 2006) offer an overview of the field. Network inference challenges, like the Dialogue on Reverse Engineering Assessment and Methods (DREAM) challenges (Marbach et al., 2010; Prill et al., 2010; Marbach et al., 2012), enable researchers to contest their implementations with other methods.

The reasons for the heterogeneity of the field are mainly two-fold: First, the biological complexity of different interacting processes of metabolites, signaling receptors and regulatory activities, and second, the varying biological questions or aims behind network reconstruction. Both reasons can be illustrated using a model adapted from Brazhnik and colleagues (Brazhnik et al., 2002) by splitting up biological processes into three layers: gene space, protein space and metabolite space (see Figure 1.9).

In Figure 1.9 the biological entities are grouped into their corresponding stages of gene expression, genes and DNA fragments are depicted in the gene space layer, proteins and mRNA transcripts are nodes within the protein layer and chemicals and their reactions take place within the metabolite space. Regulations and interactions can occur within one layer of entities as well as

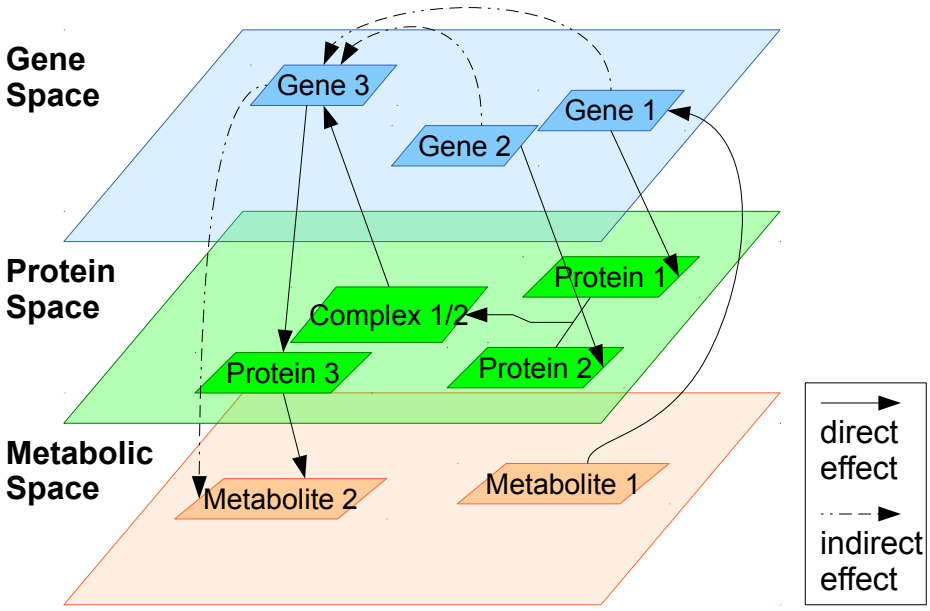

FIGURE 1.9 *A schematic view of biological interactions, layered into gene space, protein space and metabolite space, illustrating possible interactions within and between these layers. Adapted from Brazhnik et al. (2002) and Penfold and Wild (2011).*

span across different layers, including the biological processes within pathways introduced in Section 1.2. Genes in the gene layer can encode for transcription factors, i.e. proteins which can regulate the transcription of genes by binding upstream of their target promoter regions, leading to edges between gene and protein space. Complex assembly and regulatory processes like phosphorylation can lead to regulating edges within protein space. Enzymes can catalyze biochemical reactions, while metabolites are able to degrade enzymes, leading to regulations between protein and metbolite space.

Figure 1.9 illustrates that observed correlations between data might in fact be due to indirect interactions, covering different layers of different pathway types. This implies that network reconstruction is highly dependent on the type of available data, which are also further detailed in the next section, dividing available methods for network reconstruction into two groups, based either on correlating expression profiles or based on intervention experiments.

1.5.2 Overview of Published Methods

A number of facts determine which network reconstruction approaches are viable: the type of data, e.g. whether there is mRNA or protein expression data available, or if intervention or time-course measurements were conducted. However, the heterogenity and extent in methods and applications has also led to a wide range of differing definitions and nomenclature (Aittokallio and Schwikowski, 2006; Markowetz and Spang, 2007; Kaderali and Radde, 2008; Hecker et al., 2009). Following the nomenclature of Markowetz and Spang (2007), methods for network reconstruction can be divided into two groups: models of conditional independence, which are based on clustering co-expressed molecules, and intervention models, which are based on observing cause-effect relationships of perturbation experiments.

1.5.2.1 Conditional Independence Models

Conditional independence models derive the network structure from the correlation structure of the measured molecules. In the most basic approach, a so-called coexpression network is built from the similarity of measured expression profiles.

Coexpression networks are built following the guilt-by-association principle: if two genes are co-expressed, i.e. they share a similar expression profile, they are assumed to participate in the same biological processes. First uses of this approach have already been made in the last century and have helped to identify genes participating in the cell cycle (Eisen et al., 1998; Spellman et al., 1998). The most basic approach to building a network from coexpression profiles simply treats genes, or clusters of genes, as independent if their correlation is zero and connects dependent genes and gene clusters (Stuart et al., 2003). This approach has been extended in several ways: to account for time lag in expression profiles of time-course data (Bickel, 2005), to account for "differential coexpression" between different sample groups (Kostka and Spang, 2004), to include different data source weighting, and to account for non-linear correlations (Yamanishi et al., 2004).

Different models of conditional independence have been proposed for network reconstruction: *full conditional models*, *first order conditional independence*

models and *Bayesian networks*. The central difference between these models are the number of tests performed to assure that a correlated pair of genes is indeed indepedent of the remaining genes.

Full conditional models are implemented as Gaussian graphical models and infer correlations between two genes, depending whether this correlation can be explained by the set of all other remaining genes (Heckerman et al., 2001). A big advantage of this model is the small number of tests performed: one test per gene pair. However, the drawback of full conditional models is that in comparison to the number of genes, a large number of samples is needed to compute the model. Unfortunatly, this setting is very rarely found in -omics data. However, different model estimation strategies like bootstrapping and linear shrinkage approaches have been proposed to increase modeling performance (Schäfer and Strimmer, 2005a,b).

Unlike the strategies to improve model estimation, the idea behind *first order conditional independence models* is to tackle the problem of $p \gg n$ by restricting the model conditions. Full conditional models account for conditional independence of two genes with the set of all other genes. In contrast, first order conditional independence models assure conditional independence of two correlated genes with any single third gene (Markowetz and Spang, 2007). Wille and colleagues (Wille et al., 2004) applied their implementation of sparse Gaussian graphical models to identify gene clusters and cross-talk between pathways in the Isoprenoid gene network in Arabidopsis thaliana and perform further simulation studies. Another notable representative of lower order conditional independence models is ARACNE (Margolin et al., 2006), which has been published and applied in several settings, for example the reverse engineering of regulatory networks in human B cells (Basso et al., 2005).

The assumed independence of coexpression clusters in full conditional models (the correlation of two genes cannot be explained by all other genes) and first order conditional independence (the correlation of two genes cannot be explained by any single other gene) can be further extended. An even higher resolution of network knowledge is provided by networks for which the correlation of two genes cannot be explained by any other subset of the remaining genes. It can be shown that the knowledge of all orders of independence of gene subsets

implies the joint probability distribution of all variables and results in a directed *Bayesian network* (Markowetz and Spang, 2007).

Bayesian networks are probabilistic graphical models, represented as directed acyclic graphs (DAGs), which connect variables via their probabilistic relationships and dependencies. One advantage of using a DAG as representation is that it formally contains the joint probability distribution of the variables, and still remains informative for a human reader. In a DAG, nodes represent random variables and the edges represent the conditional probabilities between the variables. A vast number of different network reconstruction methods based on Bayesian networks have been proposed in order to tackle various problems. The first problem arises from the fact that in Bayesian networks for every pair of genes independence tests for every possible subset of all other nodes have to be assessed, while for full conditional and low order independence only a few statistical tests, in the order of magnitude of the number of graph nodes, have to be conducted. Unfortunately, the extensive amount of tests required for Bayesian Networks are computationally not feasible for networks with more than half a dozen genes (Pearl, 2000; Markowetz and Spang, 2007). In order to avoid this problem, networks are scored on how well the measured data fits a specific network. This poses the problems of network selection and network scoring. In order to tackle the first problem, different approaches for selecting networks from a huge network space have been used to smartly traverse through the network space, for example greedy hillclimbing or sampling strategies like Markov Chain Monte Carlo (Hastings, 1970; Husmeier, 2003). The second problem is the scoring of networks, i.e. computing a score for the network to define how well the measured data fits a selected network. Maximum likelihood as well as Bayesian scores are often applied to rate the goodness of fit between network and data (Pearl, 2000).

Although good results have been obtained and verified, reviews and benchmarks have shown that conditional independence models exhibit severe limitations in many areas. A major problem of these basic approaches lies in the failure to reveal more information about cliques of a graph, i.e. fully connected clusters of genes (Markowetz and Spang, 2007): For a clique of genes $X - Y - Z$, basic coexpression networks are not able to distinguish if the underlying biological regulation is $X \rightarrow Y \rightarrow Z$ or $X \leftarrow Y \rightarrow Z$ or if in fact a

hidden fourth regulator is triggering all genes independently. Furthermore, Husmeier and colleagues found that network inference performance varies greatly based on prior knowledge, experimental sampling strategy and training set size (Husmeier, 2003). Wimburly et al. demonstrate that reconstruction is unreliable and quickly degrades with added noise and small sample size (Wimburly et al., 2003). However, one factor has been shown to greatly improve network reconstruction performance: The use of interventions on biological networks to experimentally generate perturbation data (Werhli et al., 2006; Zak et al., 2003).

1.5.2.2 Intervention Models

In gene intervention experiments external stimuli or inhibitions, which either enhance or reduce the gene expression of a particular gene, are provided to cells. The idea of intervention models is that the observed effects of these interventions can then be used to infer knowledge about the network (Markowetz, 2010). Various approaches for network reconstruction using intervention data have been published, notably *Boolean networks*, *correlation networks*, *ideal interventions* and *Nested Effects Models*.

Boolean networks are directed, however not necessarily acyclic, graphs that are defined by one Boolean function per node. This Boolean function derives the state of the node from the state of its parents nodes. Boolean networks are deterministic in the way that a regulatory edge within a regulatory network either exists or not. Due to noisy data and other influences, models which account for uncertainties are usually preferred for intervention models (Ideker et al., 2000; Akutsu et al., 1998).

Correlation has been used to model intervention data similarily to the conditional independence models (Rice et al., 2005). In these *correlation networks* the expression levels for perturbed genes, both in perturbation and control samples, are correlated with the expression levels of all other genes. Two nodes within the model are connected if a high correlation for these genes is computed. Although the model is accurate in reconstructing relationships between genes, the number of needed perturbation experiments and replicates is prohibitive for bigger networks (Markowetz and Spang, 2007).

Ideal interventions have been proposed by Pearl and colleagues (Pearl, 2000) to model interventions in Bayesian networks. Ideal interventions assume perfect perturbation of a knocked-out gene and fix its state, making it independent of all parent nodes. This model has been integrated for network reconstruction using full conditional independence models (Rogers and Girolami, 2005) as well as Bayesian networks (Pe'er et al., 2001; Markowetz and Spang, 2003). Simulation studies have shown that intervention data strongly increases the performance of network reconstruction algorithms (Werhli et al., 2006; Zak et al., 2003).

Nested Effects Models (NEMs) are a family of graphical models which try to further tackle a central problem of network reconstruction: the fact that observed effects are often only indirect effects, nested below a number of upstream regulators. The general idea of NEMs is that the observed effects of interventions on a pathway are nested into each other. The regulator at the very top of the pathway affects a very large number of targets. However, a perturbation further downstream in the pathway affects only a subset of these genes.

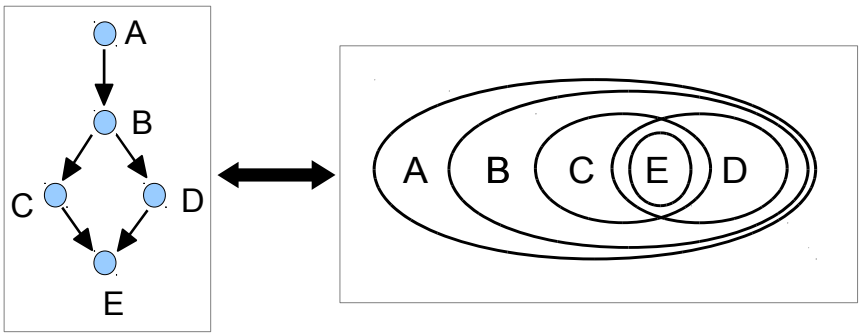

FIGURE 1.10 *NEMs are a probabilistic model to infer network topology from the nesting of observed perturbation effects. Figure adapted from Markowetz and Spang (2007).*

Figure 1.10 visualizes the concept that perturbations at different steps of a pathway result in a number of sets of effected genes, which indirectly reflect the original network topology. The framework for NEMs has been proposed by Markowetz (2005) and has been extended over time by Tresch and Markowetz

(2008), Fröhlich et al. (2007a, 2009, 2011), Anchang et al. (2009) and Failmezger et al. (2013).

Nested Effects Models are used for the purpose of network reconstruction within this thesis. A more in-depth description of NEMs can be found in Section 2.2 of Chapter 2 *Material and Methods*.

Chapter 2

Material and Methods

This thesis combines statistical bioinformatics, concepts from computer science and requires knowledge about certain aspects of biology. This chapter describes the methods used within the scope of this thesis. The idea is to give the reader insight into the required methods to understand the workflow and the results of this thesis. The first section describes the modeling of pathway knowledge and the implementation of the ontology for Biological Pathway Exchange (BioPAX). Furthermore, several renowned pathway databases as well as methods for visualizing pathway data are introduced. The second section describes Nested Effects Models (NEMs), a method for network reconstruction in detail. The third section describes the setup of the intervention experiments and the data used for network reconstruction. In the last section, a description of the R Project for Statistical Computing and of the R packages used within this thesis are given. In order to increase readability, ontology classes and R software packages are printed in *italics* and R functions are printed in `monospaced` font.

2.1 Modeling Pathway Knowledge

Modeling pathway knowledge facilitates new opportunities of data exchange between researchers and asserts a common vocabulary and understanding of underlying principles. Evolving from home-grown databases to ontologies ensures that knowledge models follow a standardized encoding and therefore are easier to document and understand. Several different approaches to model pathway knowledge have been proposed. The next section introduces a common

ground of minimal requirements to model the different types of pathways introduced in the previous chaper. Section 2.1.2 introduces the BioPAX ontology for modeling pathway knowledge. Finally, in Section 2.1.3 pathway databases in general and three databases used within this thesis are described.

2.1.1 Modeling the Structure and Composition of Biological Pathways

Formally representing pathway knowledge requires the definition of atomic entities as well as relationships between these entities, which take part in the composition of a pathway. Section 1.2 *Biological Pathways* introduced the different types of pathways, this section focuses on signaling pathways, which are used within this thesis. Modeling signaling pathways induces certain requirements on the amount of biological entities and interaction that need to be encoded.

Signaling pathways represent the communication within and between cells. The key players in signaling pathways are receptors and ligands. Receptors are proteins which are embedded either in the nucleus, the cytoplasm or the plasma membrane of the cell. Ligands are molecules which can bind to a receptor and form complexes with the receptor. This change of the receptor leads to a change in the functional state of the receptor and a set of changes downstream of the receptor. Finally, the binding of a ligand to a receptor triggers a cellular response according to the associated pathway. The processes that have to be modeled in signaling pathways are binding (association) and its reverse reaction (dissociation). Furthermore, the cross-talk between different signaling pathways can be modeled to account for overlapping pathways, feedback and feedforward signaling. Finally, several interactions can interfere with signal transduction, for example phosphorylation, ubiquitylation or methylation. These interactions can activate or inhibit receptors and regulate signal transduction.

In essence, a signaling pathway consists of participating molecules as nodes and two different types of edges. The first type of edge, the biochemical reaction, connects educt molecules and product molecules. The second type of edge, an interaction, connects a controlling molecule and a controlled biochemical reaction edge.

Different approaches to standardizing the encoding for one or several types of pathways have been published, for example the Biological Pathway Exchange (BioPAX, Demir et al., 2010), the Systems Biology Markup Language (SBML, Hucka et al., 2003) and the Human Proteome Organizations Proteomics Standards Initiative's Molecular Interaction format (PSI-MI, Hermjakob et al., 2004).

An overview of the capabilities of these standards was published by Strömbäck and Lambrix (2005) and by Cary et al. (2005), and a short comparison is performed in Chapter 4 *Discussion*. The following sections give an overview of visualization options for pathways and an introduction to the BioPAX standard, which is used within this thesis for modeling signaling and regulatory interactions.

2.1.2 The BioPAX Format for Encoding Knowledge about Biological Pathways

A central element to integrate pathway knowledge from different sources within this thesis is the ontology for Biological Pathway Exchange (BioPAX) (Demir et al., 2010).

An ontology is a formal system to model knowledge about a specific domain. This ontology defines entities, like a protein, their properties, e.g. the name and sequence of a protein, and their relationships to other entities, by using predefined vocabulary. A strong advantage of encoding knowledge using an ontology is the fixed modeling space which eases the exchange and portability of knowledge by ensuring compatibility. Links to external resources, i.e. other ontologies or databases, help standardization and the reuse of knowledge.

Three specifications are relevant for parsing ontology-encoded data within the scope of this thesis: The defintions of classes and properties that make up an ontology can be defined via the Web Ontology Language (OWL), a World Wide Web Consortium (W3C) standard (McGuinness et al., 2004). These OWL definitions can be encoded in an XML/RDF file format (Beckett and McBride, 2004) based on the extensible markup language (XML, Bray et al., 1997) and the Resource Description Framework (RDF, Klyne et al., 2004). In short, XML is a markup language, which encodes data using tags (‘<>‘) for annotation, and

RDF defines so-called triples in form of subject-predicate-object expressions to specifiy statements.

The ontology Biological Pathways Exchange (BioPAX, Demir et al., 2010) is defined using OWL and the XML/RDF encoding. In this ontology the domain of pathway knowledge is modeled. The ontology is under active development and currently contains a total of 70 classes including utility classes for links to open vocabularies and external resources. Figure 2.1 shows a simplified class tree for the BioPAX ontology.

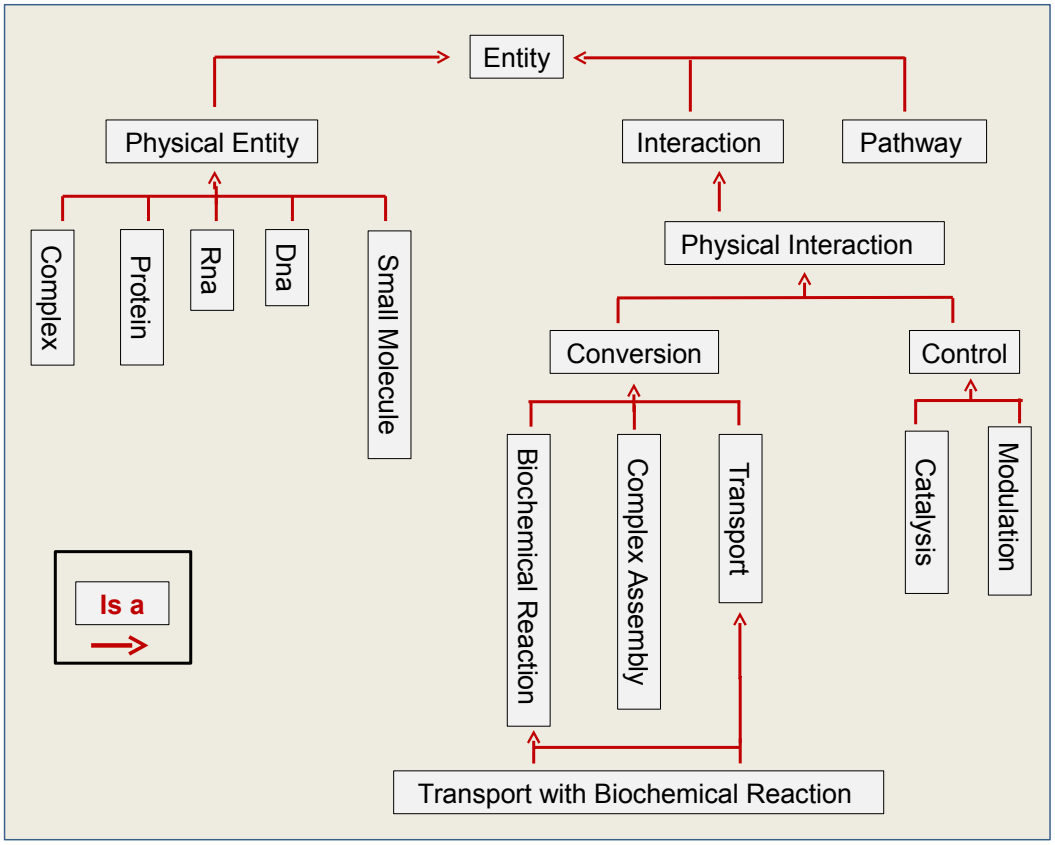

FIGURE 2.1 *This diagram shows the central classes and their inheritance relationships, (Demir et al., 2010). Reproduced according to the BioPAX specification*[(1)].

There are four distinguished central classes: *physical entities*, *interactions*, *pathways* and *support classes*. All classes inherit the name and comment properties from root class Entity.

[(1)] BioPAX Ontology Specification: http://www.biopax.org

Physical entities are all physically existing objects, which are a part of pathways, i.e. proteins, complexes, RNA, DNA and small molecules. Apart from name and comment, these classes have further properties, for example the participants of a complex molecule or a RNA or DNA sequence. Physical entities take part in all kinds of interactions. These entities can be further described and annotated by references to support classes, for example by using external database identifiers like UniProt (Bairoch et al., 2005) or Entrez Gene IDs (Maglott et al., 2005).

Interactions are split up into two different sub-classes, conversions and controls. All interactions share the property participants, a term which references physical entities. Conversions include the properties left, right and direction, in contrast to controls, which have the properties controller and controlled as sub-properties of the participants property. Conversions describe interactions like complex assemblies and biochemical reactions, for example. The property direction specifies whether the conversion occurs from left to right or vice versa. Each conversion can have any number of physical entities referenced via left or right properties. Controls are interactions with one controller property referencing the controlling physical entity and any number of controlled properties referencing interactions.

The *pathway class* has the properties: name, comment, organism, and any number of pathway components referencing interactions.

Support classes include internally defined open vocabulary terms to describe interactions, external references to publications or protein databases and references from DNA sequence to mRNA or protein products. Furthermore, references to other ontologies likes the Gene Ontology (Ashburner et al., 2000) are possible.

The summary of interaction classes above already indicates that the BioPAX ontology models pathways similarly to the ER diagram of the SBGN: An interaction is represented by an edge going from one physical entity to another edge. A biochemical reaction is an edge from one or more entities to one or more converted entities.

Section 3.1 of Chapter 3 *Results* introduces the rBiopaxParser, an R package to parse and work with BioPAX-encoded data within R. The following section

introduces a number of pathway databases which are available in BioPAX encoding.

2.1.3 Pathway Databases

A plethora of pathway databases exists and many of them offer free access to their encoded knowledge. Pathguide.org, a website listing all types of pathway databases, currently contains links to over 300 different pathway data resources (Bader et al., 2006). The different types of databases include protein-protein interactions, metabolic pathways, signaling pathways and transcription factor networks (Matys et al., 2003) for example, as well as collections of pathway sketches. However, some of these require a paid subscription or do not offer a data export in a standardized encoding.

Many notable pathway databases have been developed and are actively curated. Probably, the best known is the Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa et al., 2004), which includes metabolic and signaling pathways. On the other hand, WikiPathways is a community approach to pathway editing (Kelder et al., 2011). It allows everyone to join and share new pathways or curate existing ones. Pathway Commons is a meta-database aiming at providing a single point of access to publicly available pathway knowledge (Cerami et al., 2011). It is a collection of pathway databases covering many aspects and common model organisms trying to ease access to a large number of different sources.

Within the scope of this thesis, the focus lies on renowned and freely accessible databases offering BioPAX exports. Three exemplary databases have been picked in order to demonstrate the parsing as well as the transformation and merging of pathway databases to provide prior knowledge for methods for network reconstruction: first, the Pathway Interaction Database (PID, Schaefer et al., 2009) of Nature and the National Cancer Institute (NCI); second, the BioCarta pathway database (Nishimura, 2001), which is available via the NCI as well; and third, the Reactome database (Croft et al., 2011), an open source database featuring a peer-review process.

2.1.3.1 Pathway Interaction Database

The Pathway Interaction Database was launched as a collaborative project between the NCI and the Nature Publishing Group in 2006 (Schaefer et al., 2009). Three main data sources are available via the NCI website: Reactome, BioCarta and the PID database curated by NCI and Nature. All these data sources focus exclusively on the human as their model system. Initially, two external databases were integrated in order to be able to offer data right from the beginning. A partially annotated version of BioCarta export was integrated without any peer-review. Although this data covers large parts of known signaling pathways, only molecules are annotated using the Entrez Gene ID (Maglott et al., 2005). Neither references nor evidence of interactions nor post-translational molecule states are annotated. The second database to be integrated was Reactome version 22 released in 2007. Within this import molecules are annotated using UniProt identifiers (Bairoch et al., 2005) and post-translational modifications are included. Finally, the most important data source of PID is the NCI-Nature curated data, which was peer-reviewed and curated by Nature editors. This data includes molecules annotated using UniProt identifiers and post-translational modifications (Bauer-Mehren et al., 2009). Evidence codes and references are used to annotate interactions. This NCI-Nature curated data is very extensive and well curated and are referenced within this thesis as PID data. New pathways are curated by NCI editors in order of biological relevance and un-disputedness. Relevant interactions are identified within peer-reviewed literature and added to existing pathways.

The PID knowledge is encoded using a proprietary data model based on XML. The PID data model resembles the SBGN ER-diagram style: Molecular reactions transform input molecules to output molecules. These reactions are targeted by regulatory interactions, which can inhibit or promote the specific reaction. Valid types of molecules are proteins, complexes, RNA and small molecules. Unlike the BioPAX-model the PID model does not include DNA. Molecules can be tagged, for example as active, inactive or phosphorylated.

All PID data sources are also accessible in different encodings, i.e. BioPAX Level 2 and BioPAX Level 3. Furthermore, the PID website offers browsing capabilities for pathways as well as querying algorithms to search for molecules

or merge different pathways. As of June 2013, the available PID export contains about 2000 pathways including nested sub-pathways.

2.1.3.2 BioCarta

The BioCarta database originated from a collection of pathway maps (Nishimura, 2001). BioCarta focuses on human and mouse as model systems but includes selected plant pathways as well. Pathways are curated via templates for drawing software tools and each pathway has one or more curators to update and extend existing pathways. Pathways are available for browsing via the website and can be queried for participating molecules. Furthermore, pathways can be ordered as printed posters.

Although BioCarta is mainly a collection of pathway sketches, the BioCarta knowledge has been transferred into a standardized encoding within the PID project. BioCarta exports are available in BioPAX Level 2 and Level 3 encoding via the PID website. As of June 2013, the available BioCarta export contains about 350 pathways.

2.1.3.3 Reactome

Reactome is an open-source, manually curated, and peer-reviewed pathway database including an interactive website for querying and visualizing data (Vastrik et al., 2007). Reactome is a joint effort of the European Bioinformatics Institute, the New York University Medical Center and the Ontario Institute for Cancer Research. The database is focused on pathways in homo sapiens, however, equivalent processes in 22 other species are inferred from human data (Vastrik et al., 2007). Reactome includes signaling pathways, information on regulatory interactions as well as metabolic pathways. The data model of Reactome is based on a relational database schema. Its central model defines entities and events, the conversion of input entities to output entities. Pathways are grouped events which can also include further sub-pathways in a hierarchical manner (Matthews et al., 2009). Reactome is currently available in version 45 and contains about 1500 pathways and sub-pathways, as of June 2013. Data exports are available as a MySQL database dump using the internal database model and in the SBML and BioPAX, Level 2 and Level 3, encodings.

2.2 Methods for Network Reconstruction

Methods for network reconstruction aim at inferring the topology of a network given experimental data. An overview of different approaches and techniques is given in Section 1.5. Nested Effects Models have been thoroughly described and used in a number of publications, e.g. by Markowetz (2005); Markowetz et al. (2007), Fröhlich et al. (2007a, 2008b,a, 2009) and Tresch and Markowetz (2008). This section describes Nested Effects Models (NEMs) as means to reconstruct networks within this thesis and explains how prior knowledge can be integrated.

2.2.1 Nested Effects Models

Nested Effects Models are graphical models, which reconstruct networks based on the nested structure of intervention effects generated by perturbation experiments, for example gene knockdowns. The perturbed genes, which constitute the nodes of the reconstructed network, are selected in such a way that they are known or suspected to interact or interdepend, knockdowns of genes of the same signaling pathway, for example. These experiments can be measured using omics technologies introduced in Section 1.3, for example microarrays. The data measured for each of these experiments can then be statistically evaluated. Effected genes are commonly detected by testing genes which are significantly differentially expressed between the control and the knockdown experiment. This step yields a list of differentially expressed genes for each knockdown experiment. Usually, all measured genes which show no differential expression between any two comparisons of control versus a treated group are filtered out.

A Nested Effects Model can be described in form of a matrix product of two matrices representing two directed graphs: the network topology Φ and the bipartite graph Θ attaching effected genes to perturbation experiments. Figure 2.2 shows the definition of Nested Effects Models.

Another way of visualizing a NEM is shown in Figure 2.3 which depicts the NEM F as the product of Φ and Θ using the adjacency matrix representation of Φ and a dichotomized effect graph Θ.

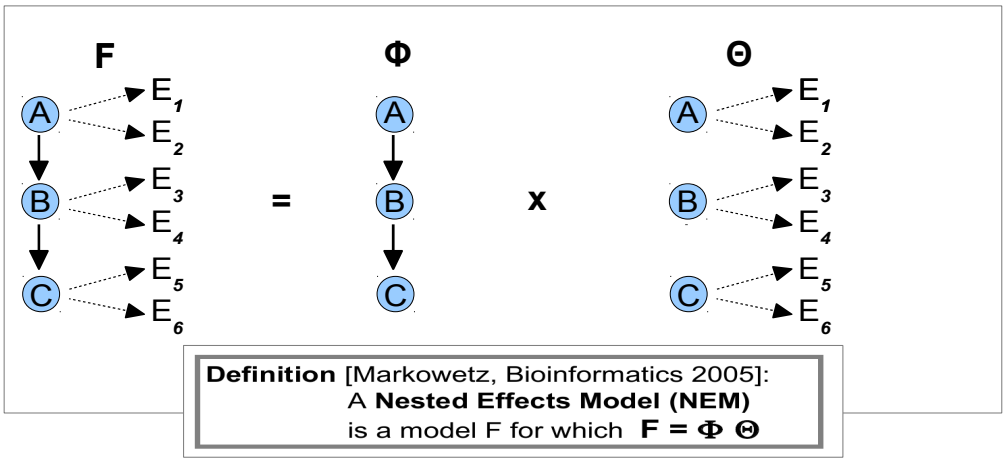

FIGURE 2.2 *The definition of Nested Effects Models. (According to Markowetz et al. (2005))*

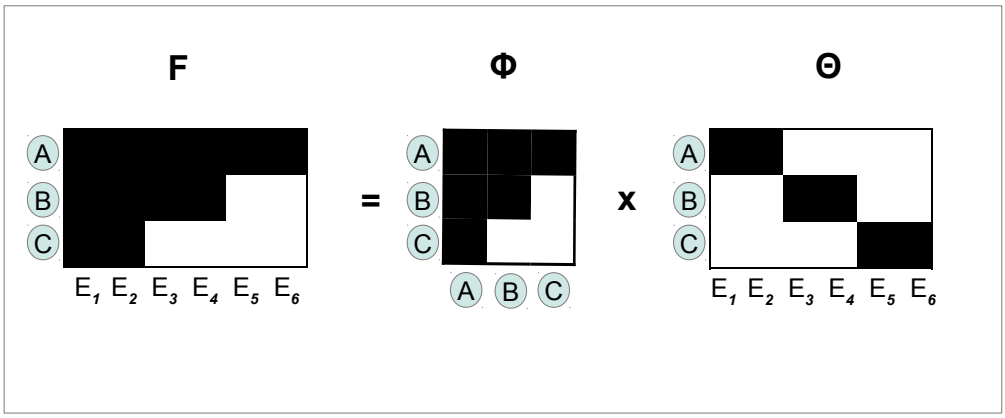

FIGURE 2.3 *Visualization of a NEM: matrix representation of* Φ *(black = edge present, white = edge absent); dichotomized representation of the effect graph* Θ *(black = effect, white = no effect). (According to Markowetz et al. (2007))*

Nested Effects Models reconstruct the network of perturbed genes and the effects attached to each perturbation by optimizing F given the observed data. Based on generated network and effect graph hypotheses, the resulting NEMs can be scored according to their fit to the experimental data (see Figure 2.4). The NEM fitting the experimental data best is selected.

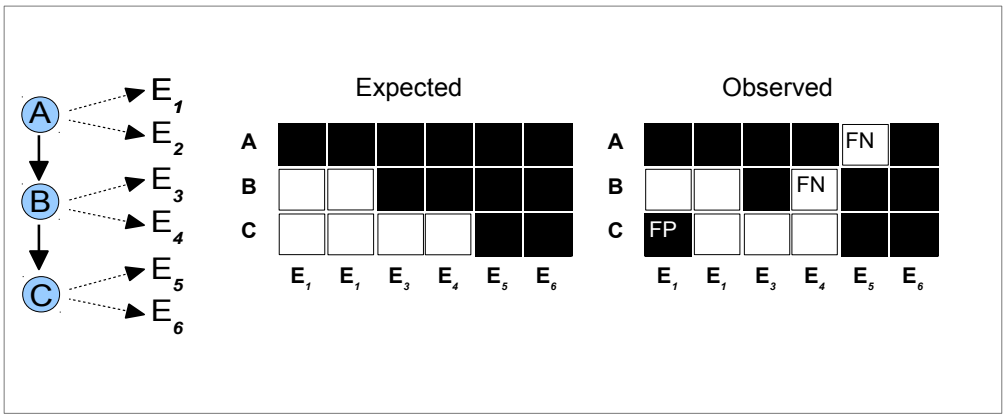

FIGURE 2.4 *A NEM can be scored by comparing the expected effects based on the model with the observed effects. (According to Markowetz et al. (2007))*

Markowetz et al. (2007) and Fröhlich et al. (2007a, 2008b) state that the main goal of NEMs is the inference of the signaling graph Φ, hence statistical independence of effect positions is assumed and Θ is integrated out following a Bayesian point of view (Fröhlich et al., 2009). An alternative approach was proposed by Tresch and Markowetz (2008), the maximization of the NEM score by using maximum a posteriori / maximum likelihood (MAP/ML) probability estimate in an alternating fashion for Φ and Θ.

From an algorithmic point of view, analyzing the nesting structure of the experimental data and selecting the best fitting NEM are the crucial points. It might not be feasible for larger networks to search the complete space of network topologies for the best model and inference mechanisms have to be used, for example greedy hillclimbing. A number of heuristics has been proposed in order to select Φ from the space of possible network topologies. Furthermore, several likelihood models were introduced to compare the network hypotheses to the observed experimental data.

The most simple way of finding the optimum for this problem is traversing the complete network topology space of Φ via an exhaustive search algorithm. However, this is not feasible for larger networks due to the exponential growth of possible network hypotheses with the number of nodes. Several strategies were proposed to deal with this problem: The *greedy hillclimbing* algorithm and

the divide-and-conquer algorithms *pairwise heuristic* (Markowetz et al., 2007), *triplets inference* (Markowetz et al., 2007) and *module networks* (Fröhlich et al., 2007a).

Greedy hillclimbing is a search strategy for finding local optima from a given starting position. In NEMs an empty network topology, without any edges, is used as starting position and during each iteration the edge improving the network score the most is added to the graph. The algorithm terminates when no edge remains which improves the network score.

The pairwise heuristic divides the network into the smallest possible subsets of all pairs of genes. For each of these pairs the most likely of one of four models is inferred, either $X \rightarrow Y$, $X \leftarrow Y$, $X \leftrightarrow Y$ or XY. The inferred network topology is the set of all pairwise relationships (Markowetz et al., 2007). *Triplets inference* further extends the scoring of pairs and removes the independence assumption between pairs. The network topology is built by scoring all triples (X, Y, Z) of genes and selecting the final graph by averaging how often a specific edge between two genes is inferred. The final graph is built from all edges which occur more often than a previously selected threshold (Markowetz et al., 2007). *Module networks* start with hierarchical clustering of the expressions profiles of intervention experiments. Effect profiles with a similar response are supposed to have a small distance within the network topology. These hierarchies are broken down into genes clusters of four genes at a time. Exhaustive search NEMs for the highest scoring models for these quadruples is performed and the modules are subsequently merged in a greedy hillclimbing fashion (Fröhlich et al., 2007a).

Further extensions to NEMs have been proposed lately: Niederberger et al. (2012) proposed a combination of Monte Carlo sampling and an Expectation-Maximization (EM) algorithm and Failmezger et al. (2013) introduced dynamic NEMs to analyze time laps cell images of RNAi knock downs.

2.2.2 Handling Prior Knowledge in Nested Effects Models

Werhli and Husmeier (2007) reason that network inference from sparse and noisy high-dimensional data leads to a poor reconstruction accuracy and suggest that the inclusion of complementary information might be indispensable. Two

ways of handling prior knowledge integration for the network topology Φ have been proposed for NEMs, following either a frequentist or a Bayesian formula (Fröhlich et al., 2007a, 2008b). Both approaches assume independent edge priors for all edges and model the likelihood of each specific edge using a Laplacian distribution with parameter λ.

The first approach uses the frequentist point of view (Fröhlich et al., 2009) and scales the belief into the prior as an regularization trade-off dependent on λ. Here, $\lambda = 0$ leads to a pure maximum likelihood estimate and $\lambda \to \infty$ leads to full belief into the prior edges. In order to select a balancing option between 0 and ∞ Fröhlich et al. (2007a) proposed to use the Akaike Information Criterion (AIC).

The second approach follows a Bayesian point of view and proposes the use of an inverse gamma distribution as prior on λ and marginalization to model the belief into prior knowledge edges (Fröhlich et al., 2008b).

2.3 Experimental Data

Previously generated experimental gene expression data of breast cancer cell lines is used within this thesis in order to demonstrate the integration of pathway data for network reconstruction purposes. The data consists of a number of gene perturbation experiments in the human estrogen receptor positive breast cancer cell line MCF7.

A total of 16 single knockout experiments was included for network reconstruction after leaving out double knockout experiments and treatments with stimuli or drugs. Table 2.1 lists the perturbed genes along with their specific identifiers. For every perturbed gene 2, 3 or 4 biological replicates, indicating 4, 6 or 8 microarrays per knockdown, were measured.

Within the scope of the thesis, this dataset is used to demonstrate the integration of prior knowledge from pathway databases into network reconstruction approaches. Nested Effects Models, introduced in Section 2.2.1, are applied to reconstruct the signaling cascade derived from the effects of the gene perturbation observed within this experimental data. In essence, this reconstructs a network with the 16 perturbed genes as nodes and their signaling flow as edges.

The data itself is available via the online repositories Gene Expression Omnibus [(2)(3)] (Edgar et al., 2002) and ArrayExpress [(4)(5)] (Brazma et al., 2003) and has been partially used in other network reconstruction publications (Fröhlich et al., 2007a, 2008b). Gene perturbation has been performed using siRNA knockdowns, the exact protocol is available along with the data via GEO and ArrayExpress.

The perturbation experiments have been measured using custom two-color microarrays following a dye-swap design. The mapping of microarray probes to mRNA identifiers is detailed in the GEO platform definition GPL3050[(6)]. The dye-swap design specifies that one biological replicate consists of two two-color chips, one chip using the green channel for treatment and the red channel as control and the other chip using the green channel for control and the red channel for treatment.

For each of the knockdown genes, a differential gene list is compiled representing the specific knockdown effects. The results of the statistical analysis of knockdown versus control can be found in Section 3.3.1 of Chapter 3 *Results*.

[(2)] GEO Accession GSE12291: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12291
[(3)] GEO Accession GSE7033: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7033
[(4)] ArrayExpress Accession E-GEOD-7033: http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-7033/
[(5)] ArrayExpress Accession E-GEOD-12291: http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-12291/
[(6)] GEO Platform definition GPL3050: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gpl3050

HUGO Gene Symbol	Entrez Gene ID	UniProt ID
AKT1	207	P31749
AKT2	208	P31751
BCL2	596	P10415
CCNG2	901	Q16589
ESR1	2099	P03372
FOXA1	3169	P55317
HSPB8	26353	Q9UJY1
MAPK1	5594	P28482
STAT5B	6777	P51692
STC2	8614	O76061
TMEM45B	120224	Q96B21
TP53	7157	P04637
XBP1	7494	P17861
DDR1	780	Q08345
GDF15	9518	Q99988
GPR30	2852	Q99527

TABLE 2.1 *Table of Perturbed Genes*

2.4 The R Project for Statistical Computing

The R scripting language and the R Project for Statistical Computing offer numerous ways for data processing, statistical testing, mathematical modeling and graphical plotting (Team, 2013). It is an implementation of the language S whose development begun at the Bell Laboratories in 1975. R is an open source software, part of the GNU project, and has been in development since 1997. One of its main advantages over other statistical computing environments like SAS, SPSS or Statistica is the portability and the extensibility with new software packages written in R, C or other languages. The quick succession of new discoveries in molecular biology, the open source approach and its extensibility have made R a very popular tool in many areas of bioinformatics. Several online repositories are available, the Comprehensive R Archive Network (CRAN, Hornik, 2012), Bioconductor (Gentleman et al., 2004) and the Omega Project for Statistical Computing (Lang, 2000) which currently contain 4705, 671 and 98 R packages, respectively.

Within this thesis R is used for several tasks: For the statistical analysis of gene expression of the experimental data, for the assessment and analysis of network reconstruction and for the implementation of a software package to integrate pathway data into R.

2.4.1 Packages for Statistical Bioinformatic Analyses

A great variety of packages is available for the user to pursue bioinformatic analysis in R, many of them available through the online repository Bioconductor (Gentleman et al., 2004). One example is the differential gene expression analysis, which essentially tests whether the expression levels of genes between two groups are deregulated. The results of a differential gene expression analysis of perturbation experiments can be used as input to compute Nested Effects Models, as described in Section 2.2.

Assessing differentially expressed genes is achieved by computing statistical tests to search for genes which show a significantly increased or decreased expression level under different experimental conditions. A differential gene

expression analysis is comprised by a number of sequential steps. An exemplified workflow from experimental data to differential gene list is depicted in Figure 2.5.

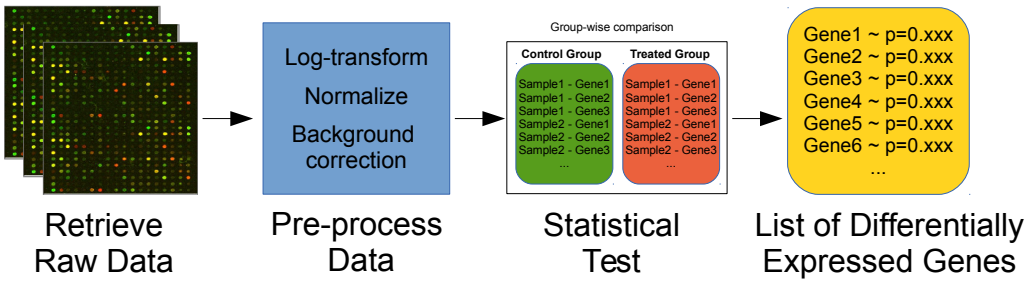

FIGURE 2.5 *This figure shows the testing for differentially expressed genes, a common analysis performed in statistical bioinformatics to find genes which differ significantly between two groups.*

Gene expression data must be read in from the file system or from an online data repository, e.g. the Gene Expression Omnibus (Edgar et al., 2002) or ArrayExpress (Brazma et al., 2003), which store and annotate data of microarray experiments. In order to retrieve data from these repositories, the R packages *ArrayExpress* (Brazma et al., 2003) or *GEOquery* can be used. Depending on the platform used to measure gene expression levels, the retrieved data files are often encoded using a proprietary format. However, packages exist to parse this data, for example the *affy* package (Gautier et al., 2004). After data has been retrieved and read in, it is usually log-transformed and a number of further pre-processing steps are available, i.e. background correction and normalization, depending on experimental design and microarray platform. Background correction can be useful for microarrays, where a scanner reads fluorescent light intensities from the arrays and always returns a minimum intensitity due to background noise. However, the necessity and use of background correction is disputed (Smyth et al., 2003). In order to account for different distributions of intensities, normalization can be applied depending on experimental design. Two kinds of normalizations are commonly applied, normalization for all microarrays of an experiment and/or for all microarrays of the specific groups within the experiment. Different types of normalizations can be used, common choices are quantile normalization and loess normalization (Dudoit et al., 2002; Bolstad et al., 2003). Nevertheless, the specific use cases and best practices are

disputed (Quackenbush, 2002; Qiu et al., 2013). Packages to parse proprietary microarray data formats often include background correction and normalization, for example the *affy* package (Gautier et al., 2004) for Affymetrix microarrays and the *lumi* package (Du et al., 2008) for the Illumina platform. Finally, a number of packages are available to test the pre-processed data for significantly regulated genes. The most commonly used package in this context is the *limma* package by Smyth and colleagues (Smyth, 2005). Its main procedure computes an empirical Bayes-moderated T-statistic of the gene expression levels between groups on a gene-by-gene basis. The resulting p-values are usually adjusted in order to account for the multiplicity problem, for example using Bonferroni correction to control the family-wise error rate. However, in high-throughput data analysis it is common to use the more liberal approaches, i.e. to control the false discovery rate (Benjamini and Hochberg, 1995).

2.4.2 Packages for the Integration of Pathway Data

The use of various pathway models, gene or protein identifiers and restrictions of the available R classes as well as slow execution times make the integration of pathway data into R a complex task. In order to integrate BioPAX-encoded pathway knowledge into R, a new software package was implemented, see Section 3.1 "rBiopaxParser" of the Results chapter. The *rBiopaxParser* parses BioPAX data, resembles its ontology within R and offers a general approach to work with pathway data in R (Kramer et al., 2013). The aim was to enable the user to integrate data from different sources and allow the merging of these different knowledge sources. BioPAX ontologies are encoded in the OWL format, which is based on XML/RDF encoding. The *XML* package (Lang, 2013), a wrapper for the Linux library libxml2, is used in order to read these XML/RDF files into R for further processing. The integrated data downloader for *rBiopaxParser* is based on the *RCurl* package (Lang, 2007), which is a wrapper for the libcurl library for data transfer using various network protocols. Mapping operations between different identifiers, e.g. UniProt ID and Entrez Gene ID, is performed using the *biomaRt* package (Durinck et al., 2009).

2.4.3 Nested Effects Models in R

Nested Effects Models, as described in section Section 2.2, are used for network reconstruction within the scope of this thesis. Nested Effects Models have been implemented in the R software package *nem*, which is available at the online repository Bioconductor and has been published (Fröhlich et al., 2008a).

A NEM is comprised of a network hypothesis and a bipartite graph attaching specific effected genes to perturbed genes. The various implementations of network inference differ regarding enumeration of search space for the network hypotheses and regarding the probabilistic model for attachment of the effected genes. Prior knowledge can be supplied in form of an adjacency matrix for the network hypothesis and as prior assumptions for the effected gene positions. Furthermore, a number of features are available, for example feature selection, which leaves out irrelevant effected genes, or statistical assessment of stability and robustness by applying bootstrap or jackknife methods. Finally, post-processing features enable the user to merge indistinguishable nodes and to compute the transitive reduction of the NEM graph.

The central function of the package is `nem`, which expects the data matrix, the inference model and a hyperparameter, containing all other relevant parameters, for example prior knowledge input. The `plot.nem` function offers a number of features and parameters to visualize the NEM.

Nested Effects Models expect prior network knowledge input in the form of an adjacency matrix of regulatory interactions between perturbed genes. Therefore, the common type of visualization for pathways within this thesis is the SBGN activity flow diagram, generated in R using Graphviz. Nodes within these graphs represent molecules, edges represent regulatory interactions with activating edges rendered green and inhibiting edges rendered red.

Chapter 3

Results

Following the workflow in Figure 1.1, as defined in Section 1.1 *Aim and Organization of the Thesis*, several steps are required to complete the task of integrating pathway data as prior knowledge into methods for network reconstruction. Figure 3.1 gives an overview of the results presented in this chapter.

Three central and novel results of this thesis are presented in this chapter:

First, the newly implemented software package `rBiopaxParser`, which enables the import of BioPAX-encoded pathway databases into R is described. Functionality and internal data model of the software are explained in this section.

Second, in Section 3.2 *Prior Knowledge Generation*, the features of the parser are used to exemplify the steps from pathway databases, to interactome and finally to a merged consensus network of prior knowledge. These steps are pursued with the aim to generate prior knowledge input from multiple data sources for network reconstruction purposes via NEMs. The prior knowledge is transformed into a graph depicting the directed signaling interactions between the genes perturbed in the experimental dataset introduced in Section 2.3.

Third, in Section 3.3 *Network Reconstruction*, the results of network reconstruction based on the generated prior knowledge and the experimental dataset are detailed. The results are reconstructed networks of 16 genes, based on the high-throughput data of the 16 knockdown experiments, with a network topology mirroring the signaling flow cascade of these genes.

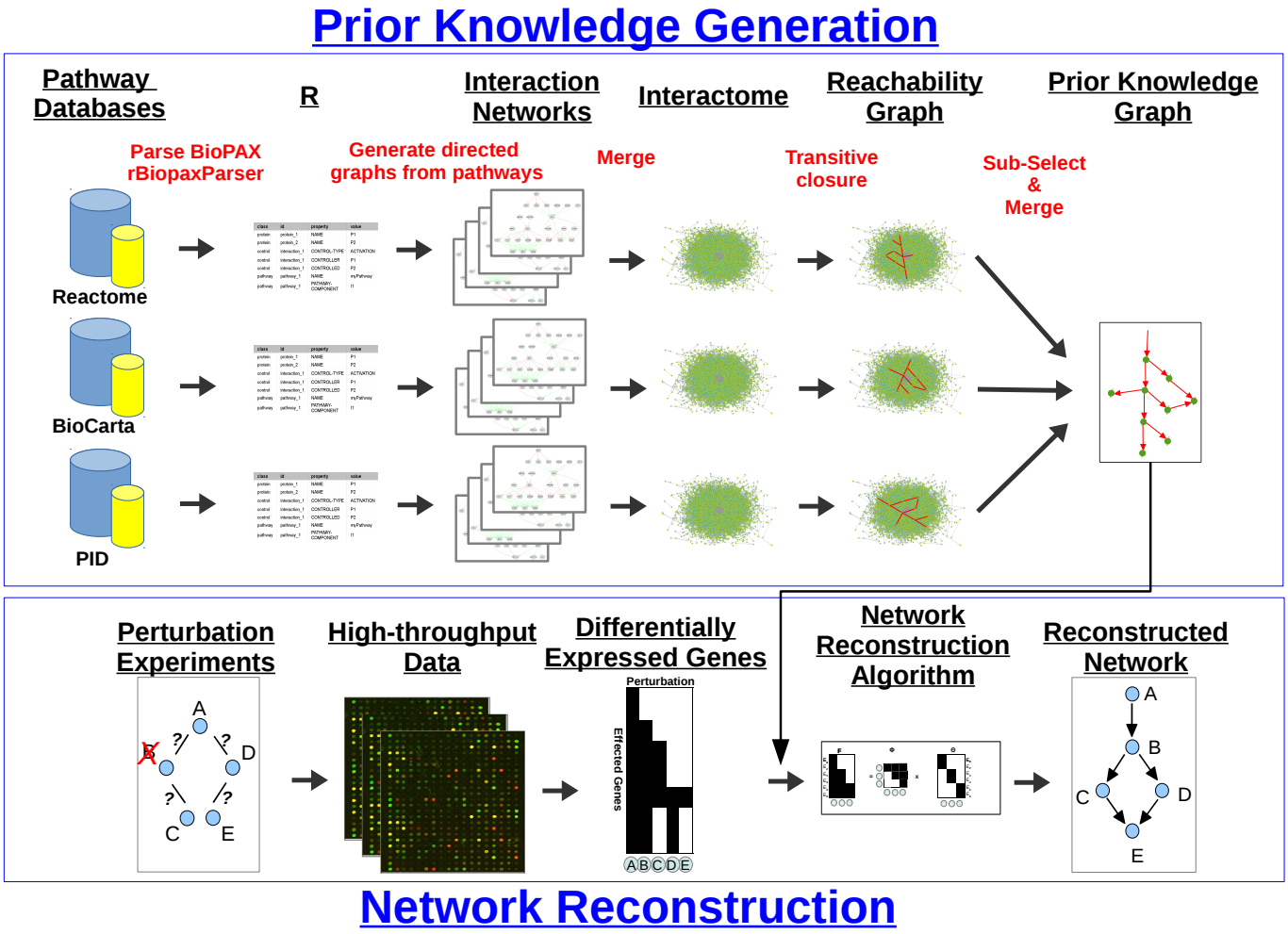

FIGURE 3.1 *Detailed workflow of the integration of prior knowledge into methods for network reconstruction within this thesis.*

3.1 rBiopaxParser

The software `rBiopaxParser` is an R package specifically implemented to make pathway data, which is encoded using the BioPAX ontology, available within the R Project for Statistical Computing. The software package has been published as open source on the online version control website GitHub[(1)] and has been released as part of Bioconductor 2.12[(2)]. An application note

[(1)] rBiopaxParser Repository on GitHub: https://github.com/frankkramer/rBiopaxParser
[(2)] rBiopaxParser Realease on Bioconductor:
http://www.bioconductor.org/packages/release/bioc/html/rBiopaxParser.html

describing the implementation and features of the package has been published in Bioinformatics (Kramer et al., 2013). As of December 2013, version 1.3 of the R package is available. The project includes about 7000 lines of source code in 66 functions, a reference manual (54 pages) documenting the available functions, and a vignette (17 pages) describing working examples. The documentation is not attached in an appendix, due to the extensive length, but can be readily downloaded from the GitHub repository as well as the Bioconductor website. This section describes the steps of parsing BioPAX encoded data, the internal data model used to represent the BioPAX ontology within R, an example on accessing, modifying and visualising a pathway, as well as a description how to create an interaction network from pathway data.

3.1.1 Retrieving Pathway Data

Several online pathway databases offer an export in BioPAX format. The *rBiopaxParser* package gives the user a shortcut to download BioPAX exports directly from database providers from the web. A list of links to commonly used databases is stored internally and the user can select from which source and which export to download. The data is stored in the working directory of the active R session. Currently, the website of the NCI[(3)], where exports of the Pathway Interaction Database (PID), BioCarta and Reactome are available, and the Reactome website[(4)] are linked. For example, the following command downloads the Pathway Interaction Database export from the NCI website.

```
> file = downloadBiopaxData("NCI","PID")
```

After the download is finished the on-screen output informs the user of success and name of the downloaded file. Subsequently, the downloaded database export can be parsed using the functionality described in Section 3.1.2. Another valid option to retrieve pathway data is to manually retrieve BioPAX encoded data from websites or via database providers.

(3) National Cancer Institute: http://pid.nci.nih.gov
(4) Reactome: http://www.reactome.org

3.1.2 Parsing of Pathway Data in BioPAX Format

The BioPAX ontology models biological pathway concepts and their relationships. Implemented in the Web Ontology Language OWL and encoded using an RDF/XML-based markup language, it allows the users to store and exchange pathway knowledge in a well-documented and standardized way. The *rBiopaxParser* can parse BioPAX encoded data into R from the local file system using the *XML* library. The readBiopax function reads in a BioPAX .owl file and generates the internal data format used within this package (see Section 3.1.3). As this function has to traverse the whole XML-tree of a database export, it is computationally intensive and may have a long run-time depending on the size of the BioPAX files. Large databases like the Pathway Interaction Database or Reactome contain millions of lines of XML-encoded data. Parsing this data using a system library wrapped into an R package and handling the data within R can result in parsing times between several minutes up to an hour.

The following command reads in the previously downloaded BioPAX file into variable biopax and prints its summary.

```
> biopax = readBiopax(file)
> print(biopax)
```

The latest released version of the ontology is BioPAX Level 3. This package currently supports BioPAX Level 2 and Level 3.

3.1.3 Internal Data Representation

The BioPAX-format definition and the data content are encoded as an ontology using the Ontology Web Language (OWL). The OWL-file is encoded using the Resource Description Framework (RDF) which in turn is encoded based on the Extensible Markup Language (XML).

The first element of the XML-file contains a tag specifying the used XML version and the character encoding of the file. This is followed by an RDF-element with various attributes specifying further element definitions and namespaces, i.e. the XMLSchema namespace, the RDF-schema namespace, and the OWL namespace. The root RDF-element further contains an Ontology-element pointing to the BioPAX definition hosted at biopax.org and the complete encoded

data encoded according to the specified BioPAX definition. These schemata and namespace definitions are saved in order to be able to reproduce a modified version of this parsed BioPAX file and write it out to the file system later on.

The BioPAX ontology models the domain of biological pathway knowledge. Classes like Protein, RNA, Interaction and Pathway are the defined entities in this domain. Their specific properties, for example, NAME, SEQUENCE, CONTROLLER and PATHWAY-COMPONENT, define the characteristics of and the links between the instances of these classes. An overview of the main classes in BioPAX Level 3 is shown in Section 2.1.2 of Chapter 2 *Material and Methods*. In simplified terms one can say that the main class, the pathway, is built up from a list of interactions. Interactions are linking controlling molecules to controlled reactions. Reactions are biochemical reactions which converse, transport or assemble educt molecules to product molecules. In order to illustrate the conversion from XML/RDF to R data.frame, a minimalistic example pathway in BioPAX representation with one enzyme regulating a biochemical reaction of two proteins modeled as instances of BioPAX classes is depicted in Figure 3.2. Of course, much more extensive and complex pathways can be constructed using the BioPAX ontology.

Table 3.1 shows the internal data of the minimal BioPAX example, as introduced in Figure 3.2. This internal data model uses an R data.frame to represent instances as a collection of their properties. The first column specifies the class and the second column specifies the id of the instance. The properties, for example "NAME", can either be of rdf:datatype, usually a string like "Small Pathway", or of type rdf:resource, which is a reference to another instance, like "#Reaction_1". For comprehensive databases, this data.frame can reach quite extensive sizes. The data.frame itself can be accessed either directly via the parsed object or by using one of the implemented functions to ease selection and modification of BioPAX instances.

The transformation needed in order to visualize the parsed pathway knowledge is illustrated in Section 3.1.5.

Mapping the XML/RDF representation of the BioPAX data from the OWL file to R is a time consuming task, especially considering the size of many complete exports of popular databases. The Pathway Interaction Database of the NCI currently includes more than 175000 BioPAX instances, for example.

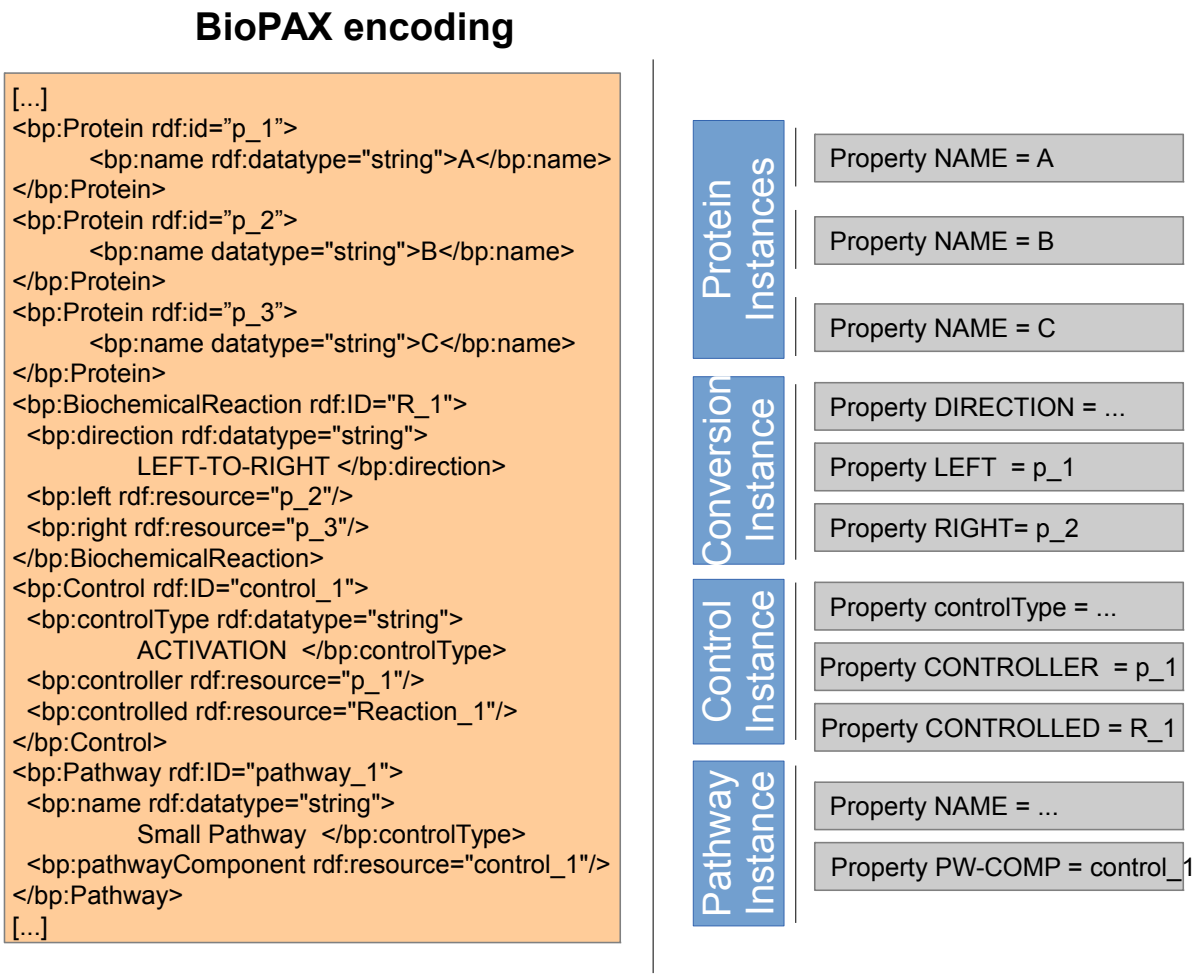

FIGURE 3.2 *A minimal example of a BioPAX pathway with a single pathway component. Protein A is acting as a catalysis for a biochemical conversion reaction of two other proteins.*

Mapping these instances to R objects and managing them within lists is not feasible due to the speed issues of searching inside of lists. List objects in R are generic and can contain any type of object. Subsequently, objects within a list can not be indexed for specific properties, for example a name variable, further obstructing quick search algorithms. Therefore, the classes and their respective properties are internally mapped to a single R matrix and then converted to a flat data.frame. This allows more efficient indexing and selecting of subsets from this data.frame when compared to lists.

class	id	property	property_attr	property_attr_value	property_value
pathway	pathway_1	NAME	rdf:datatype	string	Small Pathway
pathway	pathway_1	PATHWAY-COMPONENTS	rdf:resource	#control_1	
control	control_1	CONTROL-TYPE	rdf:datatype	string	ACTIVATION
control	control_1	CONTROLLER	rdf:resource	#p_1	
control	control_1	CONTROLLED	rdf:resource	#Reaction_1	
bioChem	Reaction_1	DIRECTION	rdf:datatype	string	LEFT-TO-RIGHT
bioChem	Reaction_1	LEFT	rdf:resource	#p_2	
bioChem	Reaction_1	RIGHT	rdf:resource	#p_3	
protein	p_1	NAME	rdf:datatype	string	A
protein	p_2	NAME	rdf:datatype	string	B
protein	p_3	NAME	rdf:datatype	string	C

TABLE 3.1 *Example of parsed BioPAX data encoding a small pathway.*

The conversion of BioPAX ontology data to the internal R data model is performed as revertible as possible, with one caveat, however. The XML structure of the original data would allow an infinite nesting of instance declarations. An example would be to instantiate an external publication reference within a protein instance, which could itself be instantiated in another instance. This is not desirable when attempting to map the data to a tabular format such as a data.frame. Identifying these nested instances is easy: the parser reaches an instance declaration within another instance. The trick here is to move these nested instances into the main XML tree and reference the specific instance with an rdf:resource attribute from within the parsed instance.

3.1.4 Accessing Pathway Data

A number of convenience functions are available, which aid the user in selecting specific parts or instances of the BioPAX model. Generally, these functions require the parsed BioPAX object as parameter and other parameters that differ from function to function.

The most basic function to select distinct instances is `selectInstances`. This function allows the user to specify conditions like class, ID or name to select a subset of the internal data.frame meeting these conditions. This function is vectorized to allow the user to select multiple instances. The user can extend the selection criteria by several parameters, for example to include inherited class types or all referenced instances.

The next type of functions return, in comparison to the internal data.frame, lists in a human-readable format: `listInstances`, `listPathways`,

`listPathwayComponents`, and `listComplexComponents`. These functions return a list of classes, IDs and names of instances.

The functions `getReferencedIDs` and `getReferencingIDs`, which can optionally be executed recursively, are passed the BioPAX object and an instance ID as parameters. The return value is a vector of IDs of all instances that are referenced by or are referencing the supplied instance. These functions can be used to traverse the database, retrieving molecules used within specific pathways or pathways including specific molecules.

While these functions cover the basic querying capabilities to the structured pathway data, more complex problems can be addressed by combining and extending these functions. Further operations, which can be used to modify, transform and merge pathway data are described in Section 3.2.

3.1.5 Visualizing Pathway Data

As described in Section 1.2.4 of Chapter 2 *Material and Methods*, different approaches to visualize biological pathways are possible offering differing granularity of details. Within this thesis the focus is on generating interaction graphs visualizing the different types of interactions within signaling pathways, similar to SBGN activity flow diagrams.

The function `pathway2RegulatoryGraph` transforms the BioPAX-encoded knowledge of a signaling pathway and compiles it into an interaction graph, which can be used as prior knowledge input for methods for network reconstruction. These graphs rely solely on the available BioPAX information about activations and inhibitions, by classes of or inheriting from, class *control*. Involved molecules, represented by nodes, are connected via edges depending on the encoded knowledge. This functions breaks up the inherent mechanistic representation of pathways in BioPAX, where an interaction connects a controlling molecule to a controlled biochemical reaction edge. Here, interaction graph is generated by connecting controller molecules and products of the controlled biochemical reaction via an interaction edge. The parameter "splitComplexMolecules" can be used to split all complexes into their most atomic members, with all members sharing the same in- and outgoing edges.

The generated graph objects can be layouted using `layoutRegulatoryGraph` and visualized using Rgraphviz or RCytoscape.

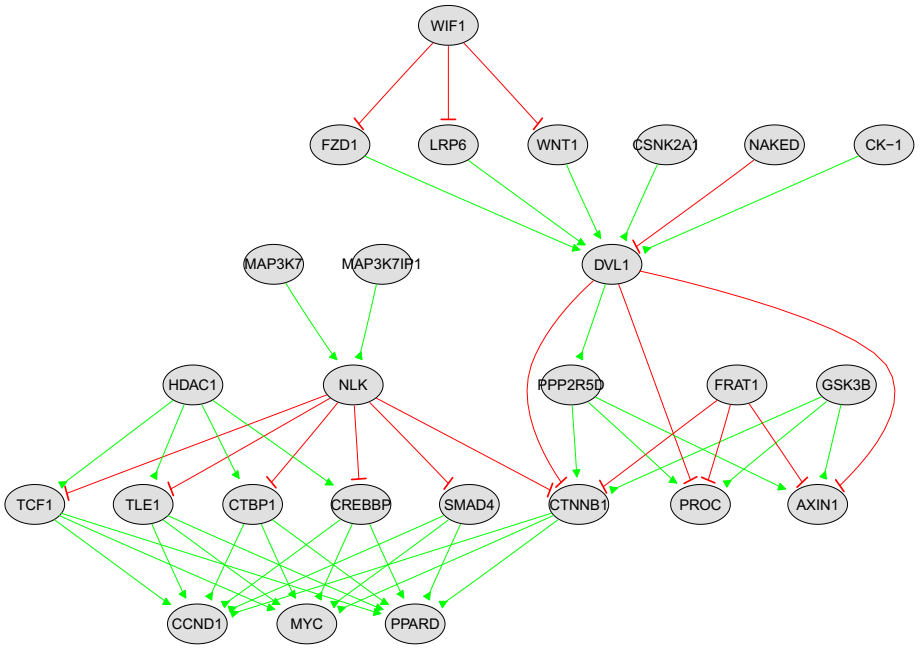

FIGURE 3.3 *Rgraphviz plot of the "wnt signaling pathway" parsed from BioCarta.*

Figure 3.3 shows the Rgraphviz plot of the "Wnt signaling network" pathway parsed from PID using the following commands:

```
> wnt_pwid = "pid_p_100002_wntpathway"
> wnt_pw_graph = pathway2RegulatoryGraph(biopax, wnt_pwid, splitComplexMolecules
> wnt_pw_graph_laidout = layoutRegulatoryGraph(wnt_pw_graph)
> plotRegulatoryGraph(wnt_pw_graph_laidout)
```

3.2 Generation of Prior Knowledge Networks from Pathway Databases

This section details the steps how a prior knowledge graph is built from interaction information by using the R software package *rBiopaxParser* to parse the pathway databases Reactome, Pathway Interaction Database and Biocarta.

Figure 3.4 illustrates the workflow for generating a prior knowledge network appropriate as input for network reconstruction from pathway databases.

The aim of this section is to transform the parsed prior knowledge into a graph depicting the directed signaling interactions between the genes perturbed in the experimental dataset introduced in Section 2.3. This yields a graph with 16 nodes (the knockdown genes) and their specific interactions extracted from the pathway databases.

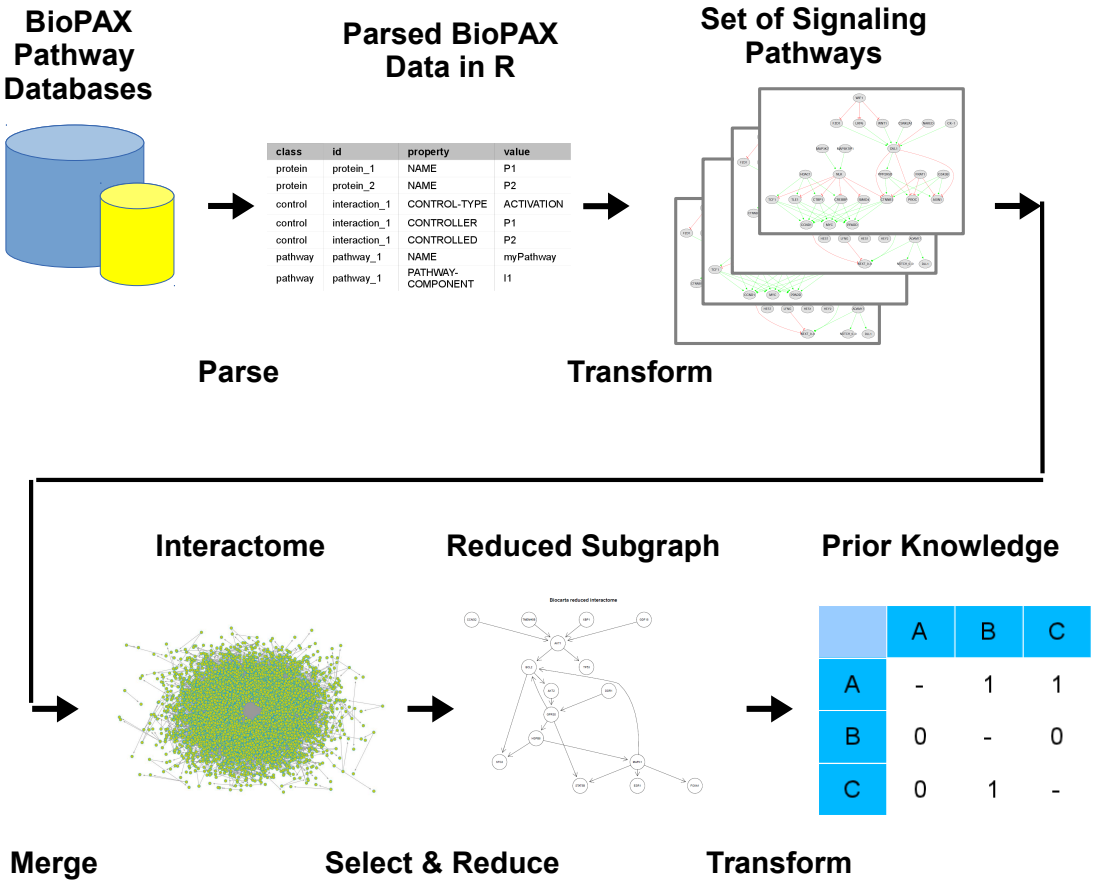

FIGURE 3.4 *This figure shows the steps to parse, transform and merge the prior knowledge retrieved from pathway databases to an input graph suitable for the given network reconstruction task.*

Within this section a detailed report of the extent and the compatibility of the parsed databases is given first. Second, the handling of identifiers for genes, overlaps of pathways and ambiguity of names and identifiers is reported. Finally, the steps of reducing the wealth of regulatory information to a network of the target genes are described. These include the steps of generating a comprehensive interactome for each database and subsequently reduction and merging of these interactomes.

3.2.1 Pathway Data from Reactome, Biocarta and PID

The pathway databases PID, Reactome and Biocarte serve as a basis of literature knowledge, which is compiled into a consensus network and used as prior knowledge input for network reconstruction. The three databases have been described in the corresponding sections of Chapter 2 *Material and Methods*. The exports of PID and Biocarta in BioPax Level 3 have been retrieved from the PID website[(5)] on March 8th, 2013. The BioPAX Level 3 data export of the Reactome database has been downloaded from the Reactome website[(6)] on March 3rd, 2013.

Database	PID	Reactome	BioCarta
Size (MB)	50.4	117.1	6.9
Parsed Size (Entries)	635331	965426	85588
Physical Entities	25241	24711	3988
Interactions	24067	9306	4727
Pathways	2047	1377	386

TABLE 3.2 *Overview of parsed BioPAX databases used for generating the consensus prior knowledge network.*

Table 3.2 lists the extent of the generated interactomes of the pathway databases Reactome, PID and BioCarta as nodes and edges. The pathway databases range from 7 to 117 MB file size resulting in parsed tables with 85000 to almost 1 million rows. Within these 4000 to 25000 entities are encoded taking part in 5000 to 25000 interactions organized into a total of 386 to 2047 pathways.

[(5)] National Cancer Institute: http://pid.nci.nih.gov
[(6)] Reactome: http://www.reactome.org

3.2.2 Identifier Handling

The three databases use different identifiers for proteins, molecules and pathways within their BioPAX models and each database has a different extent of data annotation. The mapping between identifiers is used for two operations within this work: First, the knockdown genes given by the experimental dataset introduced in Section 2.3, are identified within the pathway databases in order to select the revelant sub-networks to generate prior knowledge input. Second, in order to evaluate overlaps and discrepancies between the used databases, a generic comparison of interactions encoded in the pathway databases is performed.

3.2.2.1 Database-specific Identifiers

All BioPAX instances have a unique internal identifier. However, these are proprietary, non-functional, internal identifiers and might change between database exports.

Pathway names are not standardized and database curators use slightly different names and descriptions, e.g. the WNT signaling pathway is called *Signaling by Wnt* in Reactome, *wnt signaling pathway* in BioCarta and the PID contains this pathway split into *Wnt receptor activity*, *Signaling by Wnt* and *Wnt signaling network*. Interactions and reactions usually remain without a defined name. Furthermore, database curators are not bound by specific naming standards for physical entities. Instances can either be identified via their specific name property or using references to external annotations like UniProt or Entrez IDs.

In Reactome, protein instances have at least one or more gene names or symbols by HGNC nomenclature. Small molecules have one or more chemical names and might have the corresponding ChEBI ID as name. Protein instances in BioCarta are sometimes named using their HGNC gene name and sometimes using the gene symbol. Small molecules are named according to their chemical nomenclature. The Pathway Interaction Database uses HGNC gene symbols and sometimes additionally gene names and UniProt IDs for their protein names. Small molecules are named according to their chemical nomenclature and ChEBI ID. Within all databases, the names of complexes are compiled by

pasting complex component names, sometimes separated by white space and somtimes by character '\'. However, component names and order of complex components differ, which makes direct comparisons of complex entities between databases difficult.

The databases use different annotations to reference to external identifiers for encoded instances. In Reactome, the annotation of entities is performed by using Reactome proprietary database IDs for all entities. Additionally, proteins are annotated according to UniProt ID and small molecules with ChEBI IDs. BioCarta annotates proteins with Entrez Gene IDs and uses GO terms for some molecules. PID annotates proteins with Entrez Gene IDs, publications evidence for interactions with PubMed IDs and small molecules using CheBI IDs.

3.2.2.2 Identifier Mapping

Network reconstruction via Nested Effects Models requires a prior knowledge input containing the perturbed molecules as nodes and their interactions as edges. Therefore, the respective genes have to be identified in the pathway databases in order to compile fitting prior knowledge graphs. From the identifiers used in the experimental dataset, a mapping table of the 16 perturbed genes was compiled for HUGO gene symbols and names, Entrez Gene IDs and UniProt IDs. Table 2.1 in Section 2.3 *Experimental Data* lists the perturbed genes along with their specific identifiers. All perturbed genes are present in the parsed databases. In order to calculate the overlaps of interactions in pathway databases, all molecules in all databases were mapped to UniProt IDs. The mapping between the identifiers has been achieved by using the documented annotations within the pathway databases and the R package `biomaRt` (Durinck et al., 2009).

3.2.3 Generating the Interactome

In order to compile a comprehensive set of knowledge stemming from each database, a so-called interactome is generated for every database. These interactomes are the assembly of all known regulatory interactions of all pathways

into a single large graph. The process and the specific steps to generate an interactome can be formalized using pseudo-code as shown in Algorithm 1.

Algorithm 1: This algorithm describes the process to generate interactomes for the supplied pathway databases.

```
Data: Pathway Databases PID, BioCarta, Reactome
Result: Interactome of each Pathway Database
for every Pathway Database pwdb do
    Interactome ⟵ ∅;
    for all Pathways pw in pwdb do
        Interactome + = all Interactions of pw;
    end
    generate interaction graph from Interactome;
end
```

The output of these steps are three independent interactomes and the corresponding interaction graphs, one for every input database. These interaction graphs are very large (see Table 3.3), reflecting all regulatory interactions contained within the databases.

In Figure 3.5, the actual graph of the interactome of the PID database is depicted. It consists of 11364 nodes, representing molecular entities, and of 60921 edges, representing activating or inhibiting regulatory interactions. Table 3.3 lists the extent of the generated interactomes of the pathway databases Reactome, PID and BioCarta as nodes and edges. In order to break the graphs down to single gene level data, both controlling and controlled complexes were split into their complex components as described in Section 3.1 *rBiopaxParser*.

Interactome	PID	Reactome	BioCarta
Nodes	11364	6917	1980
Edges	60921	57492	6292

TABLE 3.3 *Table of Interactome Sizes*

Unfortunately, the extent and the hairball-like topology make a visual analysis of the graphs impossible and warrant further computational processing

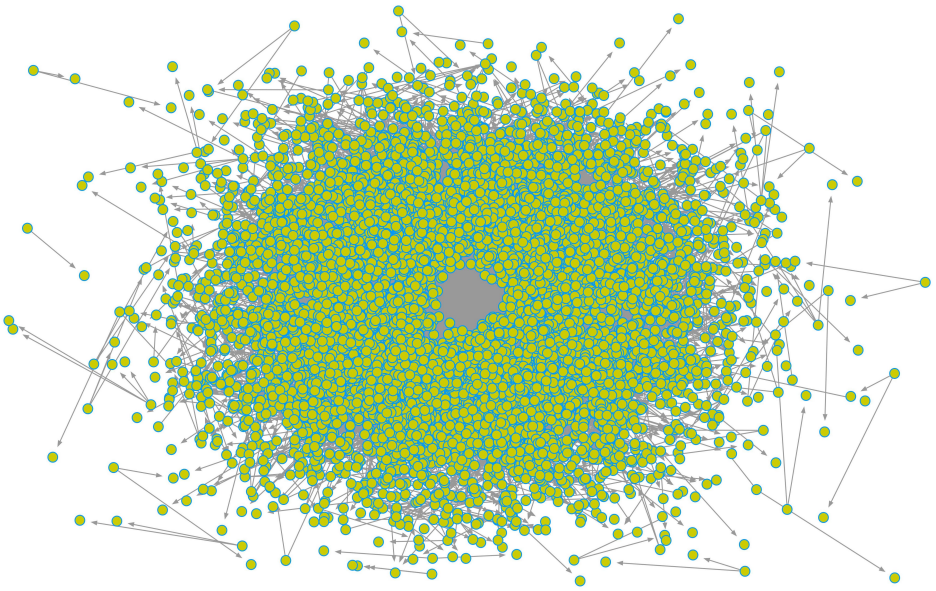

FIGURE 3.5 *The interaction graph of the interactome of the Pathway Interaction Database rendered using the RedeR software package(Castro et al., 2012).*

and reduction of the interactomes to retrieve a prior knowledge input graph suitable for network reconstruction via Nested Effects Models.

3.2.4 Graph Reduction

Depending on the required outcome many different approaches on how the interactomes can be handled are possible. Within this thesis, the prior knowledge input for network reconstruction using NEMs requires a directed graph with the intervention experiments as nodes and interactions as edges. The nodes are specified by the 16 genes knocked down in the experimental data introduced in Section 2.3.

In the setting of directed graphs the transitive closure means that whenever there are edges $X \to Y$ and $Y \to Z$ then there is also an edge $X \to Z$. For the purposes of the interactome the transitive closure answers the question of reachability. Therefore, applying transitive closure is a way to locate the paths between specific nodes within a graph. The subgraph containing only the perturbed genes is extracted from the transitively closed interactome. This subgraph contains only the perturbed genes as nodes and there is an edge from $G1 \to G2$ whenever there is any path of directed edges from G1 to G2 in the

interactome. This trait is used in order to compile the reachability for the perturbed genes of interest for all interactomes.

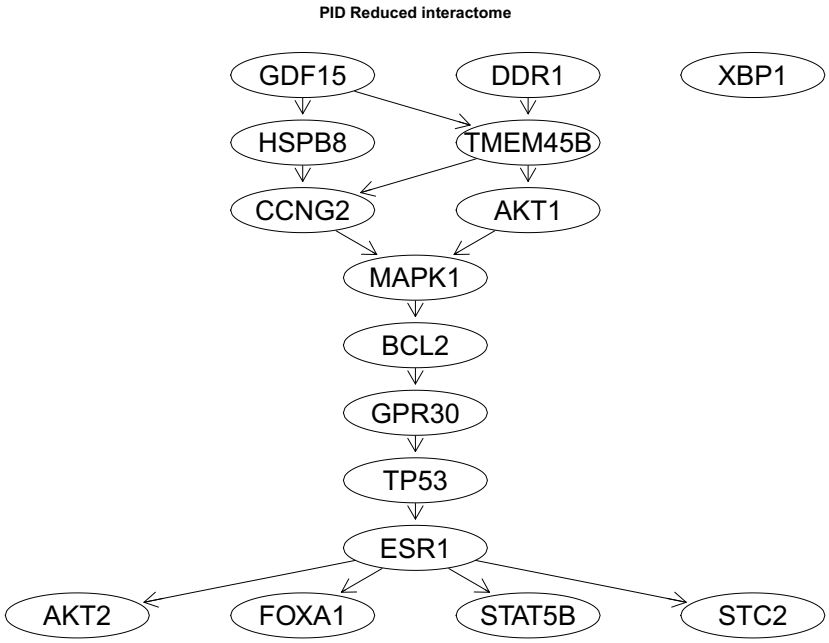

FIGURE 3.6 *The (transitively reduced) representation of the subgraph of 16 knockdown genes from the interactome of the Pathway Interaction Database.*

The first step is to calculate the transitive closure of the current interactomes. The transitive closure results in transitively closed interactomes with vastly increased number of edges: The transitively closed interactome of BioCarta, Reactome and PID have 719084, 3882911 and 18837458 edges, respectively. The subgraphs contain only the nodes corresponding to the genes knocked down in the experimental setting.

The output of these steps are three subgraphs of the transitive closure of the interactomes, one for every pathway database. Each of these graphs has the perturbed genes as nodes and their interactions as directed edges. The transitively closed subgraphs of Reactome, PID, and Biocarta have a total of 61, 94 and 63 edges, excluding self-loops, connecting the 16 knockdown gene nodes.

The results of the interactome subgraphs for PID, Reactome and BioCarta are illustrated in transitively reduced graphs in Figure 3.6, Figure 3.7 and Figure 3.8 in order to increase readability. However, for further calculations

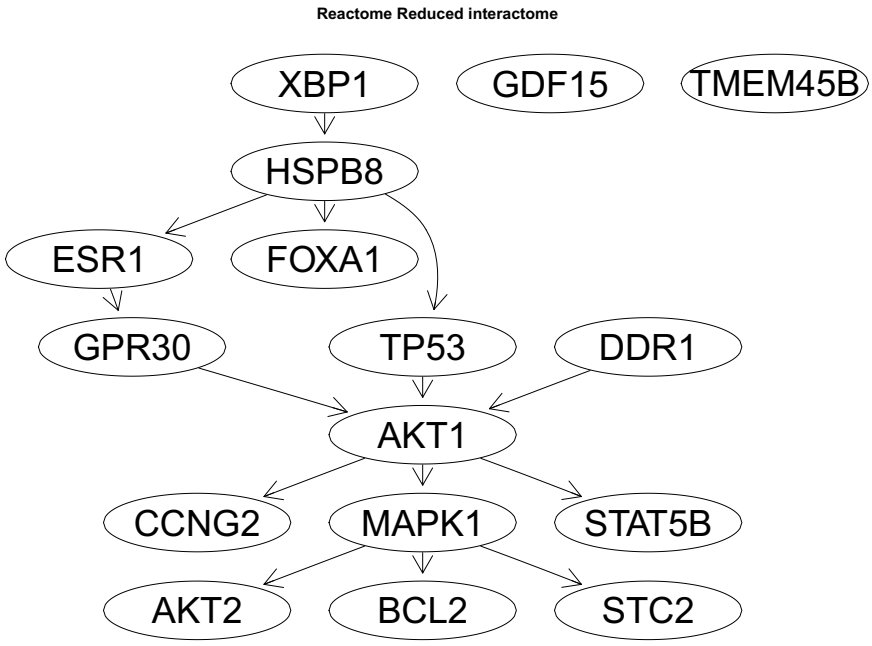

FIGURE 3.7 *The (transitively reduced) representation of the subgraph of 16 knockdown genes from the interactome of Reactome.*

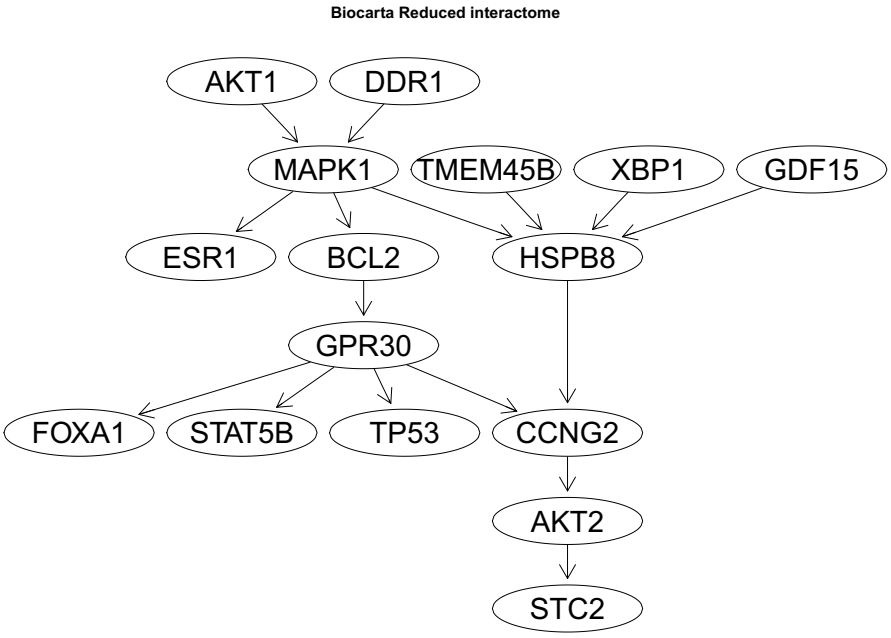

FIGURE 3.8 *The (transitively reduced) representation of the subgraph of 16 knockdown genes from the interactome of BioCarta.*

the transitively closed graphs are used. Different approaches than the one used here are possible to reduce the interactome according to the requirements.

	AKT1	AKT2	BCL2	CCNG2	DDR1	ESR1	FOXA1	GDF15	GPR30	HSPB8	MAPK1	STAT5B	STC2	TMEM45B	TP53	XBP1
AKT1		PBR	PBR	-BR	---	PB-	PB-	---	PB-	-B-	PBR	PBR	PBR	---	PB-	---
AKT2	---		---	---	---	---	---	---	---	---	---	---	-B-	---	---	---
BCL2	---	PB-		-B-	---	P--	PB-	---	PB-	---	---	PB-	PB-	---	PB-	---
CCNG2	---	PB-	P--		---	P--	P--	---	P--	---	P--	P--	PB-	---	P--	---
DDR1	P-R	PBR	PBR	PBR		PB-	PB-	---	PB-	-B-	PBR	PBR	PBR	P--	PB-	---
ESR1	--R	P-R	--R	--R	---		P--	---	--R	---	--R	P-R	P-R	---	---	---
FOXA1	---	---	---	---	---	---		---	---	---	---	---	---	---	---	---
GDF15	P--	PB-	P--	PB-	---	P--	P--		P--	PB-	P--	P--	PB-	P--	P--	---
GPR30	--R	PBR	--R	-BR	---	P--	PB-	---		---	--R	PBR	PBR	---	PB-	---
HSPB8	--R	PBR	P-R	PBR	---	P-R	P-R	---	P-R		P-R	P-R	PBR	---	P-R	---
MAPK1	---	PBR	PBR	-B-	---	PB-	PB-	---	PB-	-B-		PB-	PBR	---	PB-	---
STAT5B	---	---	---	---	---	---	---	---	---	---	---		---	---	---	---
STC2	---	---	---	---	---	---	---	---	---	---	---	---		---	---	---
TMEM45B	P--	PB-	P--	PB-	---	P--	P--	---	P--	-B-	P--	P--	PB-		P--	---
TP53	--R	P-R	--R	--R	---	P--	P--	---	---	---	--R	P-R	P-R	---		---
XBP1	--R	-BR	--R	-BR	---	--R	--R	---	--R	-BR	--R	--R	-BR	---	--R	

TABLE 3.4 *This matrix shows which edges were found in the pathway databases.* `P` *stands for PID,* `B` *for BioCarta, and* `R` *for Reactome.*

Subsequently, it is interesting to see the concordance of the parsed prior knowledge data. Table 3.4 shows in form of an adjacency matrix which of the 256 possible edges of the network of 16 perturbed genes are found in the pathway databases.

No Edge	128
Reactome	21
Biocarta	7
PID	30
Biocarta & Reactome	6
PID & Reactome	14
PID & Biocarta	30
PID & Biocarta & Reactome	20
Sum	256

TABLE 3.5 *This table shows the distribution of the 256 possible edges in the 16 nodes prior knowledge graphs.*

Out of 128 edges found in total, 58 are found in exactly one pathway database, 50 are found in exactly two pathway databases, and 20 edges are found in all three databases. A summary of overlaps is shows in Table 3.5.

3.2.5 Generated Prior Knowledge Regulatory Network

In a final step to assemble a consensus network of three pathway databases, the reduced interactomes are merged to a single graph. This step is a straightforward union of edge sets of the three graphs, no union of the node sets is required since these are already identical. The merged graph is shown in a transitively reduced fashion in Figure 3.9. It has to be noted that the merged graph is not necessarily transitively closed anymore.

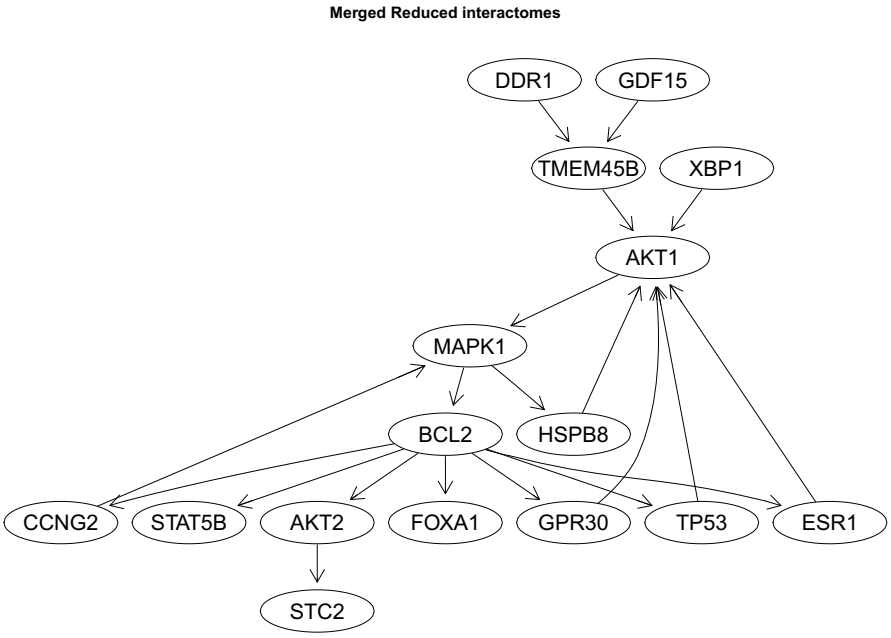

FIGURE 3.9 *The (transitively reduced) graph of the merged interactome subgraphs of PID, Reactome and BioCarta.*

The merged graph contains a total of 128 edges excluding self-loops of nodes. This merged consensus graph of the three databases is used as prior knowledge input for the network reconstruction described in the next section.

A thorough discussion of the used databases, their con- and discordance and the steps of prior knowledge generation is performed in Section 4.2 of Chapter *Discussion*.

3.3 Network Reconstruction

In this final section of the chapter, the main parts of this thesis are assembled and the results of reconstructing the regulatory structure of the pertubated genes using the generated prior knowledge network are described. First, the details of the statistical analysis for differential genes between the perturbation and the control groups for the experimental data are presented. Second, the merged consensus network generated in the previous section is integrated into the method for network reconstruction. Finally, the assessment of the nested effects model is described and the results of network reconstruction, with and without the integration of prior knowledge, are detailed.

3.3.1 Statistical Analysis of Experimental Data

In this thesis, network reconstruction is conducted based on observed experimental data of gene interventions with the aim to reconstruct the topology of a network of 16 genes. The experimental data has been introduced in Section 2.3. Two groups of breast cancer cell line samples have been grown independently for each of the knocked down genes: a knockdown group and a control group. Following a dyeswap design, one sample from each group has been measured using a two-color microarray chips with 26618 probes. For every gene 2, 3 or 4 microarray replicates were performed. A more detailed description of the experimental data is given in the corresponding section of Chapter 2 *Material and Methods*.

The log2 fold-changes of normalized gene expression values between the dyes were tested for differentially expressed genes by fitting linear models separately for each gene using the empirical Bayes method and the R package *limma* (Smyth, 2004). P-values were adjusted for multiple testing using the method by Benjamini-Hochberg (Benjamini and Hochberg, 1995). P-values < 0.05 were considered significant.

The result of this statistical analysis is a list of 26618 p-values and log-fold changes between knockdown and control samples for each of the 16 knocked out genes. These lists of differentially expressed genes represent the observed effects of each knockdown and is the basis for the network reconstruction.

In order to provide an overview of the results, Table 3.6 lists the number of differentially expressed genes for each knockdown experiment with $p < 0.05$ as well as $p < 0.05$ and the absolute value of log2 fold-change between knockdown and control bigger than 1.5.

Knockdown	$p < 0.05$	$p < 0.05$ & $abs(log2(FC)) > log2(1.5)$
AKT1	6489	50
AKT2	5461	1456
BCL2	1990	97
CCNG2	6016	618
ESR1	3012	521
FOXA1	4437	503
HSPB8	1728	31
MAPK1	1463	281
STAT5B	3978	43
STC2	4244	819
TMEM45B	4655	238
TP53	903	24
XBP1	4036	146
DDR1	7199	113
GDF15	5203	235
GPR30	4280	905

TABLE 3.6 *Table of Differentially Expressed Genes for all Perturbation Experiments.*

Only genes which showed a significant p-value ($p < 0.05$) and a high fold change ($abs(log2(FC)) > log2(1.5)$) in more than one knockdown experiment were used as effects for network reconstruction.

3.3.2 Prior Knowledge Network

The merged consensus network of the pathway databases PID, BioCarta and Reactome created in Section 3.2.5 is used as prior knowledge input for the network reconstruction algorithm.

3.3.3 Nested Effects Models

Network reconstruction is performed using Nested Effects Models described in Section 2.2.1 of Chapter 2 *Material and Methods*. Network reconstruction is performed on the introduced experimental data, once including and once

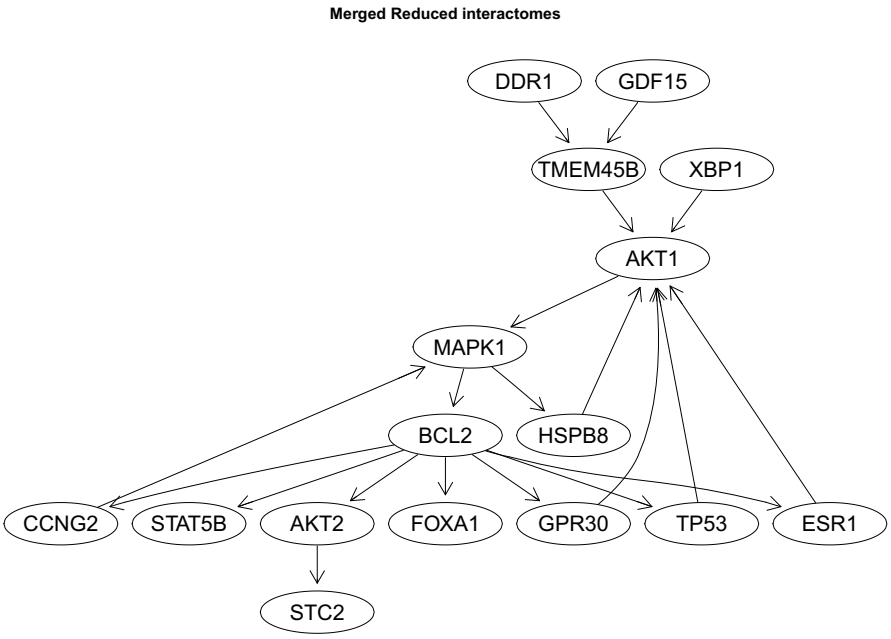

FIGURE 3.10 *The (transitively reduced) graph of the merged interactome subgraphs of PID, Reactome and BioCarta is used as prior knowledge for NEMs.*

excluding prior knowledge. Greedy hillclimbing is used as inference scheme for the network topology of all NEMs. The Bayesian inference scheme is used for the linking positions of effected genes to network nodes. Bootstrapping (100x) is performed on the linking positions of effected genes in order to assess the statistical stability of networks. A total of 1199 genes were identified as relevant based on the results of the statistical analyzes in Section 3.3.1, and are used as effected genes input. Prior knowledge is included as introduced in Section 2.2.2 by choosing the influence of prior knowledge via an inverse gamma distribution.

3.3.4 Reconstructed Network

Figure 3.11 shows the reconstructed network topologies with and without integrated prior knowledge in comparison. The numbers next to the edges indicate the percentage of times in which each edge was reconstructed over the total number of 100 bootstrap runs. Only edges reconstructed in at least 50% of the bootstrap runs are included in the figures.

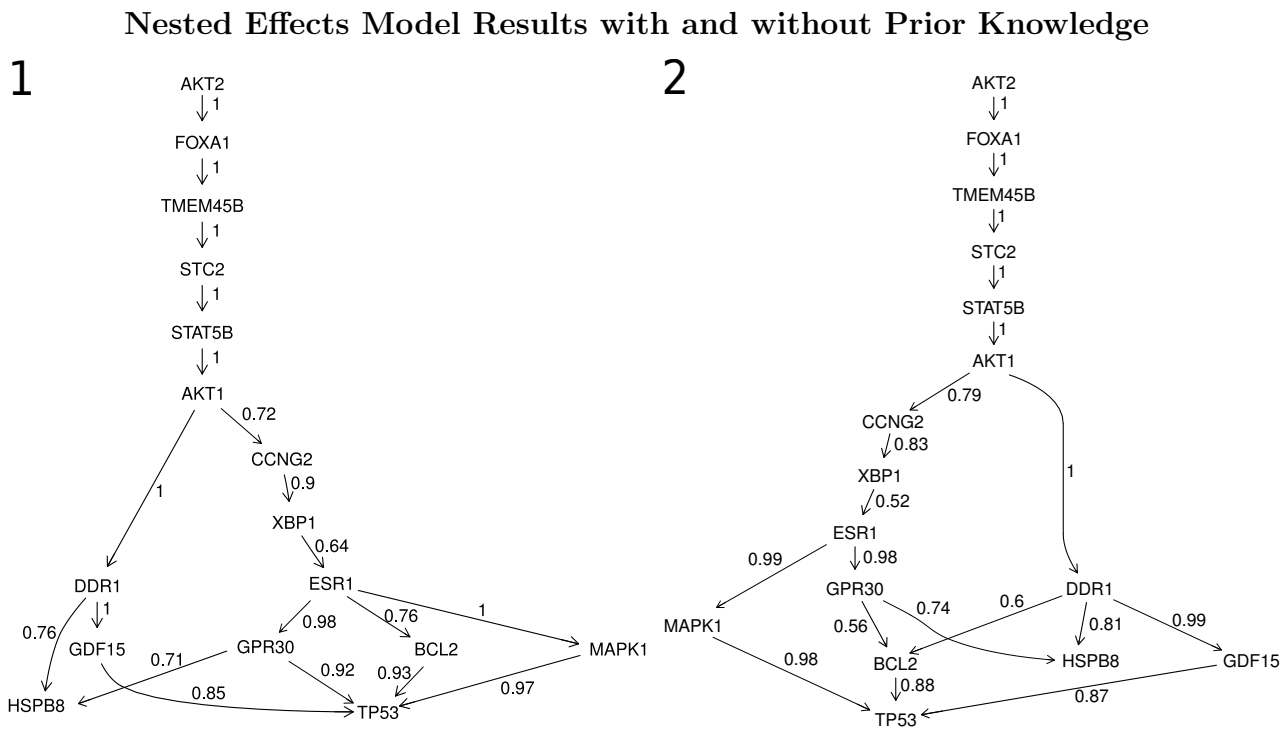

FIGURE 3.11 *The transitively reduced graphs of the computed NEMs:*
1) Bayesian inference scheme for effect positions, 100x bootstrap, greedy hillclimbing, without prior knowledge.
2) Bayesian inference scheme for effect positions, 100x bootstrap, greedy hillclimbing, with prior knowledge integrated.

3.3.4.1 Overlap of Literature Knowledge and Reconstructed Network

Table 3.7 shows how many reconstructed edges overlap with the ones found in the three pathway databases. The table shows the amount of edges present in literature knowledge in contrast to edges found in network reconstruction.

Literature Knowledge	Network Reconstruction			
	without PK		with PK	
	No Edge	Edge	No Edge	Edge
No Edge	65	63	65	63
Edge	90	38	88	40
Sum	155	101	153	103

TABLE 3.7 *Contingency table showing the overlaps and disagreements of the parsed literature knowledge and the network reconstruction results with and without integrated prior knowledge.*

For literature knowledge, the 256 possible edges of the 16 node network are divided into rows detailing which pathway database(s) an edge is present or "No Edge". The network reconstruction column contains the information for reconstruction with and without integrated prior knowledge. These edges are split corresponding where each edge is found in the specific literature knowledge.

Literature Knowledge		Network Reconstruction			
		without PK		with PK	
	Sum	No Edge	Edge	No Edge	Edge
No Edge	128	65	63	65	63
Reactome	21	13	8	12	9
Biocarta	7	3	4	3	4
PID	30	17	13	17	13
Biocarta & Reactome	6	4	2	4	2
PID & Reactome	14	14	0	14	0
PID & Biocarta	30	21	9	21	9
PID & Biocarta & Reactome	20	18	2	17	3
Sum	256	155	101	153	103

TABLE 3.8 *Detailed contingency table showing the overlaps and disagreements of the parsed literature knowledge and the network reconstruction results with and without integrated prior knowledge.*

Table 3.8 shows a more detailed contingency table differentiating between the literature knowledge extracted from the specific pathway databases. For the network reconstructed without integrated prior knowledge, 38 out of the 101 inferred edges are found in at least one of the pathway databases. 26 of these interactions are present in at least two databases and two reconstructed interactions are present in all databases. For the network reconstructed with integrated prior knowledge, 40 out of the 103 inferred edges are found in any of the pathway databases. 26 of these interactions are present in at least two databases and three edges are present in all pathway databases.

3.3.4.2 Influence of Prior Knowledge

The influence of prior knowledge on the reconstructed network can be assessed by comparing the results of the reconstructed networks with and without integrated prior knowledge when using a Bayesian prior.

It can be seen in Table 3.7 that the integration of prior knowledge into the network reconstruction approach led to two additionally inferred edges. As explained in Section 3.3.3, only edges which are inferred in at least 50% of the bootstrap runs are considered. In order to compare the results of NEMs with and without integrated prior knowledge, the differences in the frequencies of how often a certain edge was inferred can be analyzed.

	AKT1	AKT2	BCL2	CCNG2	DDR1	ESR1	FOXA1	GDF15	GPR30	HSPB8	MAPK1	STAT5B	STC2	TMEM45B	TP53	XBP1
AKT1	0.00	0.00	0.03	0.07	0.00	0.03	0.00	-0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	-0.08
AKT2	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
BCL2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.05	0.00
CCNG2	0.00	0.00	0.04	0.00	0.00	-0.12	0.00	0.00	0.01	0.01	-0.01	0.00	0.00	0.00	0.00	-0.07
DDR1	0.00	0.00	**0.60**	0.00	0.00	0.00	0.00	-0.01	0.00	0.05	0.00	0.00	0.00	0.00	-0.04	0.00
ESR1	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.03	-0.01	0.00	0.00	0.00	0.00	0.00
FOXA1	0.00	0.00	0.00	0.05	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.10
GDF15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
GPR30	0.00	0.00	**0.56**	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00
HSPB8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MAPK1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
STAT5B	0.00	0.00	0.03	0.07	0.00	0.03	0.00	-0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	-0.08
STC2	0.00	0.00	0.03	0.07	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	-0.08
TMEM45B	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.10
TP53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XBP1	0.00	0.00	0.08	0.00	0.00	-0.12	0.00	0.00	0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.00

TABLE 3.9 *Differences of network reconstruction with and without integrated prior knowledge.*

Table 3.9 presents the differences of reconstruction frequencies between the NEM bootstraps with integrated prior knowledge and the NEM bootstraps without integrated prior knowledge in matrix format. A value of 0 indicates that a certain edge is reconstructed as frequently with and without prior knowledge over 100 bootstrap runs. A negative value indicates that an edge was reconstructed in more runs without prior knowledge compared to runs with prior knowledge included. A positive value indicates that an edge was more often reconstructed in runs with prior knowledge included.

These network reconstruction results are further evaluated in the following Chapter 4 *Discussion* in Section 4.2.

Chapter 4

Discussion

This chapter contains the discussion of various points emerging from the methods used within this thesis, from the implemented software solution and from the generated results described in the previous chapters. First, the *rBiopaxParser* and its design is discussed in the light of current research and compared to similar approaches. Second, the generated prior knowledge is assessed concerning the integrated pathway databases and their overlaps and differences. Finally, the results of network reconstruction are analyzed with regard to the influence of integrated prior knowledge, the biological feasibility and the overlaps of the results of network reconstruction and literature knowledge.

4.1 rBiopaxParser

The use of various pathway models, gene or protein identifiers and restrictions of the available R classes, as well as slow execution times, make the integration of pathway data into R not a trivial task. In order to assess the *rBiopaxParser* in the context of current research and state-of-the-art software, it is discussed in two different directions. First, the design decision to use the BioPAX model for data encoding are discussed and compared to similar modeling approaches. Second, the implementation are compared to similar R packages which offer the integration of pathway data.

4.1.1 Data Model

Within this thesis the BioPAX ontology was chosen as a means to access pathway knowledge. The *rBiopaxParser* currently supports the parsing of data encoded in BioPAX Level 2 and Level 3 (Demir et al., 2010). However, a number of other approaches to standardized encoding for one or several types of pathways have been published, for example the KEGG Markup Language (KGML, Kanehisa et al., 2004), the Systems Biology Markup Language (SBML, Hucka et al., 2003) and the Human Proteome Organization's Proteomics Standards Initiative's Molecular Interaction format (PSI-MI, Hermjakob et al., 2004). Comparisons of pathway modeling approaches have been published by Strömbäck and Lambrix (2005) and by Cary et al. (2005).

KGML is a markup language for encoding pathways within the widespread KEGG database. It aims at encoding the pathway diagrams, including the layout of metabolic and signaling pathways. However, KEGG, as a proprietary database, restricted access to the bulk data download and introduced a paid subscription format in 2011, causing licensing worries[(1)]. Furthermore, the use of the KGML format is not explicitly supported for other databases and the development not openly documented (Strömbäck and Lambrix, 2005).

SBML aims at modeling metabolic pathways as reaction networks including mathematical relations for each reaction. SBML is mainly focused on quantitatively modeling the reactant levels and the cell state in system biology approaches. SBML does not support references to external databases and publications (Cary et al., 2005).

The main goal for PSI-MI is to standardize the encoding of protein-protein interaction knowledge. PSI-MI supports links to evidence, i.e. publications, and external databases. However, PSI-MI lacks the concept of pathways or networks and focuses entirely on protein binding and interaction information (Hermjakob et al., 2004).

The BioPAX ontology has the most complex model, enabling the encoding of metabolic and signaling pathways alike. Different biological concepts can be encoded either in detail or at a coarser granularity level. This allows for a generalization of concepts, for example using a control interaction instead

[(1)]KEGG: http://www.kegg.jp/kegg/docs/plea.html

of a catalysis (Demir et al., 2010). External databases and publications can be referenced. However, mathematical modeling of biochemical reactions is currently not possible (Cary et al., 2005).

Lately, the different standards have been extending from their niches to cover further concepts. For example, PSI-MI previously only supported protein interactions and has been extended to cover small molecules and RNA as well (Kerrien et al., 2007). Chaouiya et al. (2013) introduced SBML Qualitative Models as a means to model pathway knowledge in SBML, a qualitative instead of a quantitative way. Additionally, a number of tools attempting conversion between formats have been published. Wrzodek et al. (2013) generate SBML models from KEGG pathways. Büchel et al. (2012) automatically generate basic SBML models from BioPAX knowledge. Ruebenacker et al. (2009) published an intermediate model allowing conversion between SBML and BioPAX. Webb and Ma'ayan (2011) introduced a tool to integrate files from the Simple Interaction Format (SIF) into BioPAX models.

The choice of using the BioPAX model for the *rBiopaxParser* in order to import pathway data into R was based on several criteria: The model had to encode signaling pathways, be openly available and support linking to external databases for identifier matching and merging. Furthermore, BioPAX is well documented, actively developed and supported by a large number of databases (Bader et al., 2006). Additionally, using the strict aspects of ontologies to model pathway knowledge improves documentation, collaboration of research groups, and the use of external data sources. This eases the sharing and understanding of the underlying model as well as the modeled knowledge.

4.1.2 Comparison with other R Packages

The field of R software for pathway data integration is rapidly developing, even though identifier handling and incompatible encoding standards pose serious obstacles. A thorough review of R packages for the integration of pathway data has been published recently (Kramer et al., 2014).

Several packages are available that offer pathway data for R, each of them with different approaches to integration, storing and visualization of data. In general these packages can be divided into two categories: Packages which

directly supply parsed pathway data and packages which offer a generic parsing of encoded data.

Graphite (Sales et al., 2012) is an example from the former group of packages, as it supplies pathways of a number of different pathway databases as R graph objects. These objects are generated by parsing and conversion of the data export of the pathway databases. Subsequently the graphs are bundled into an R package and made available via Bioconductor. *Graphite* supplies the contents of the pathway databases PID, Biocarta, Reactome, KEGG, and Spike (Paz et al., 2011). In a similar fashion the package *CePa* (Gu and Wang, 2013) includes PID, BioCarta and Reactome.

The other group of packages are generic parsers for specific encoding standards. *KEGGgraph* (Zhang and Wiemann, 2009) enables the user to parse KGML files into layouted graphs in R. The package *rsbml* uses the linux library called libSBML (Bornstein et al., 2008) to parse and generate graphs for SBML data.

A completely different approach is taken by the *PSICQUIC* connector package. It enables the user to query PSI-MI-query compatible web services to retrieve PSI-MI interaction information (Aranda et al., 2011).

The *rBiopaxParser* is a generic parser for BioPAX-encoded data. The generic parsers have the advantage over pathway data supplied directly via packages that the user is able to independently load, archive and modify new versions of pathway data directly from the pathway database providers. Furthermore, the *rBiopaxParser* has the advantage over other parsers that it is not limited to supplying graph objects, but that its internal data model preserves the complete information, including annotations and links to external databases. Additionally, this internal data model allows the user to either modify the data on BioPAX-encoding level or to generate new graph objects from the modified data.

4.2 Prior Knowledge Generation

It is obvious that many possible ways exist to generate a consensus prior knowledge network from a number of different data sources. These possibilities are influenced by the target pathway type, the choice of pathway databases

and other factors. This also implies that the workflow within this thesis is by no means obligatory to generate and integrate prior knowledge. However, several noteworthy elements of the generation of prior knowledge within this thesis warrant a discussion. First, the overlaps of the interactomes of the used pathway databases are presented. Second, the steps for transforming and merging the pathway databases into the required format are discussed. The third part of this section analyzes the prior knowledge networks resulting from the chosen pathway databases concerning their concordance and discordance .

4.2.1 Pathway Databases

Several properties were considered for the choice of the pathway databases. A basic requirement for a database was to be available in BioPAX-encoding and to supply information on the interactions in signaling pathways. Furthermore, a certain level of good credibility or reputation, for example via publications or citations, is desireable. Another aspect is the curation and maintenance work, which must be continuously funded by a corporation or governmental institution. Moreover, databases which were too focused, for example the Rat Genome Database (RGD, Petri et al., 2011) and the Microbial Signal Transduction database (MiST, Ulrich and Zhulin, 2010), and meta-databases, merging already existing other databases, for example Pathway Commons (Cerami et al., 2011) and ConsensusPathDB (Kamburov et al., 2011), were not considered for prior knowledge generation.

Based on the properties mentioned above, the pathway databases PID (Schaefer et al., 2009), BioCarta (Nishimura, 2001), and Reactome (Croft et al., 2011) were chosen as suppliers of prior knowledge, see Section 2.1.3 of Chapter 2 *Material and Methods*. The databases differ in size and coverage, as shown in Table 3.2, the pathway databases range from 7 to 117 MB file size, resulting in parsed tables from 85000 up to almost 1 million rows. Within this data between 386 to over 2000 pathways are encoded, with 4000 to 25000 entities taking part in 5000 to 25000 interactions.

In order to put the accumulated data into context it is interesting to assess the overlaps of the pathway databases between each other and with a meta-database, i.e. Pathway Commons. Table 4.1 shows the overlaps of the generated interactomes from the pathway databases. The overlap of all interactions of

two molecules, where both molecules can be mapped to UniProt identifiers, was calculated.

	PID	BioCarta	Reactome	Pathway Commons
PID		5693/37380 (15%)	20362/37380 (54%)	27773/37380 (74%)
BioCarta	3757/4070 (92%)		447/4070 (11%)	3015/4070 (74%)
Reactome	21163/39830 (53%)	777/39830 (2%)		25166/39830 (63%)
Pathway Commons	53586/91939 (58%)	7435/91939 (8%)	50305/91939 (55%)	

TABLE 4.1 *This table shows for each database listed per row, the overlaps of all its interactions with the pathway databases listed per column.*

It can be seen that a main problem of comparing database contents are the differences in size: Almost 40000 interactions in Reactome and PID cannot possibly be contained in the 4000 interactions contained in BioCarta. Furthermore, direct interactions encoded within a smaller database might have been extended and only be contained as indirect interactions within another, more comprehensive, database.

Overall, the differences in the available literature knowledge are most probably also a result of different foci and curation processes. On the one hand, Reactome has the broadest focus of the three databases and is currently well-maintained, documented and growing quickly[(2)]. PID on the other hand set its focus on cancer research[(3)] and received its last updates in 2012. Lastly, Biocarta is a long-standing project mainly focused on pathway sketches and was manually curated into a database format in 2004 by the NCI[(4)].

4.2.2 Pathway Data Transformation

The process of generating prior knowledge is strongly dependant on the actual usage and purpose of the prior knowledge. Often the order in which certain actions are performed on the graphs might change the outcome, for example the merging of graphs from different databases or the removal of irrelevant nodes from a graph. This also means that the steps performed within this thesis cannot be seen as a strict manual for the generation of prior knowledge, but rather as one possible way to reach the desired outcome. The choice of transformations needed on the raw pathway data is influenced by the raw data

[(2)]Reactome: http://wiki.reactome.org/index.php/Past_Reactome_Calendar
[(3)]PID: http://pid.nci.nih.gov/userguide/introduction.shtml
[(4)]Biocarta:http://pid.nci.nih.gov/PID/userguide/database_content.shtml

itself, the decision whether metabolic, signaling or gene regulatory networks are of interest, and by the requirements of further algorithms or analyzes, for example concerning the cyclic-ness or directed-ness of graphs. For network reconstruction using NEMs, the prior knowledge input must be a directed graph, containing the perturbed targets as nodes and their interactions as edges.

The first design choice within this thesis was to include all pathways per pathway database. In order to achieve this the interactome for each database was generated, i.e. a single pathway consisting of all interactions found within all pathways of the pathway database, as described in Algorithm 1. The output of this transformation step is one interaction graph including all the pathways for each pathway database. Another possible solution would have been to restrict the integrated knowledge to specific pathways which are known to be associated with the perturbed genes. However, as this includes manual curation and might introduce a bias on already well-studied interactions, all pathways were considered. Furthermore, instead of using a specific interactome for every pathway database and moving on from there, it would have been possible to merge all interactomes into a consensus interactome of all pathway databases.

The second transformation applied to the data was to split the complexes which participate in interactions, into the corresponding genes and to treat these as controlling or controlled instances. This transformation is a direct consequence of the requirements of the network reconstruction algorithm. The perturbation experiments knocked down single genes, therefore the reconstructed network and the expected prior knowledge network consist of singular genes as nodes.

This also implies the third transformation: The removal of nodes not needed for the prior knowledge network. This was accomplished by applying transitive closure on the generated interactomes, answering the question of reachability in directed graphs. Selecting the subgraph of the 16 perturbed genes from this transitive closure inherently returns a graph with edges between nodes, which are connected via any path in the original interactomes. Another convenient aspect of using transitive closure for the graph reduction is that NEMs assume a nesting of effects and reconstruction results are transitively closed graphs. The output of this transformation step are three transitively closed subgraphs, one

per pathway database, with perturbation genes as nodes and edges between all nodes connected via a path within the interactome. These three subgraphs were shown in a transitively reduced manner in Figure 3.8, Figure 3.7, and Figure 3.6 in the previous chapter, *Results*. Naturally, endless possibilities exist at this point to select and transform the data extracted from the pathway databases. The transitive closure was chosen here due to simplicity and the elegance of being able to directly extract the reachability graphs for the knockdown genes. Another tested approach, not shown here, was the computation of shortest paths between the knockdowns and assembly of a graph based on this information.

The last transformation step is the straightforward union of the three transitively closed interactome subgraphs into a single prior knowledge consensus network. Figure 3.9 shows the result of this union.

4.3 Network Reconstruction

Several aspects of the network reconstruction and its results can be discussed. A first point to debate is the weighting of the prior knowledge network when performing NEMs in this thesis as well as network reconstruction approaches in general. The second aspect to be discussed is the overlap of the reconstructed networks with the integrated literature knowledge. Third, the impact of the integration of prior knowledge on the network reconstruction results is assessed. Finally, the sensibility of the additional edges reconstructed with integrated prior knowledge is discussed.

4.3.1 Weighting Prior Knowledge

In order to avoid a possibly biased outcome by choosing an arbitrary regularization parameter lambda, the NEMs within this thesis were computed with a Bayesian prior which was scaled using an inverse gamma distribution as proposed by Fröhlich et al. (2008b). This method was also found to work very well for the integration of prior knowledge (Fröhlich et al., 2009).

The methods presented for NEMs specify a topology prior, i.e. a prior whether specific edges should exist. The integration of prior knowledge in NEMs

is currently limited to priors on the network topology. However, further solutions for integrating prior knowledge into methods for network reconstruction are possible. One of these alternative approaches would be to specify structural priors, for example the preferral of sparse networks, i.e. a penalization for every additionally reconstructed edge (Husmeier, 2003; Werhli and Husmeier, 2007). Similarly, certain networks could be preferred or penalized depending on their structural properties like connectivity scores, for example the disconnectivity index.

4.3.2 Comparison of Network Reconstruction Results and Prior Knowledge

Another interesting aspect to be discussed is in how far the literature knowledge overlaps with the reconstructed network based on the experimental data. Table 3.7 in Chapter 3 *Results* illustrated these overlaps. It is shown that 38 out of 101 reconstructed edges of the NEM without integrated prior knowledge are found in the literature knowledge. The NEM with integrated prior knowledge reconstructed two additional edges (40 out of 103), both overlapping with the data retrieved from the pathway databases.

Neither an exact overlap of interactions between the pathway databases, discussed in Section 4.2.1, nor an exact overlap of reconstructed network and prior knowledge were to be expected. This is due to the fact that the prior knowledge also contains pathways which might only be active under specific conditions or in specific tissues. Examples of this are pathways which are only active in diseases like diabetes or cancerous cell, under stress conditions or in a specific phase of the cell cycle.

Unfortunately, meta-information concerning the relevant context for pathways, e.g. specific cell lines or diseases, cannot be stored in any of the current encoding standards. However, recent publications have moved the definition of context-specific pathways into the focus of research and might trigger further extensions to encoding standards (Mitra et al., 2013; Lan et al., 2013; Amar et al., 2013). This shift of focus might, in the long run, enable researchers to programmatically limit the integrated prior knowledge to specific pathways relevant to an experimental setting.

4.3.3 *Impact of the Integration of Prior Knowledge on Network Reconstruction Results*

As seen in Table 3.9, which shows the influence of integrated prior knowledge on the results, the overall small differences for almost all edges indicate a very robust network reconstruction result. The integration of prior knowledge leads to two additionally inferred edges, $DDR1 \rightarrow BCL2$ and $GPR30 \rightarrow BCL2$, when considering the threshold of only including edges which are inferred in at least 50% of the bootstrap runs.

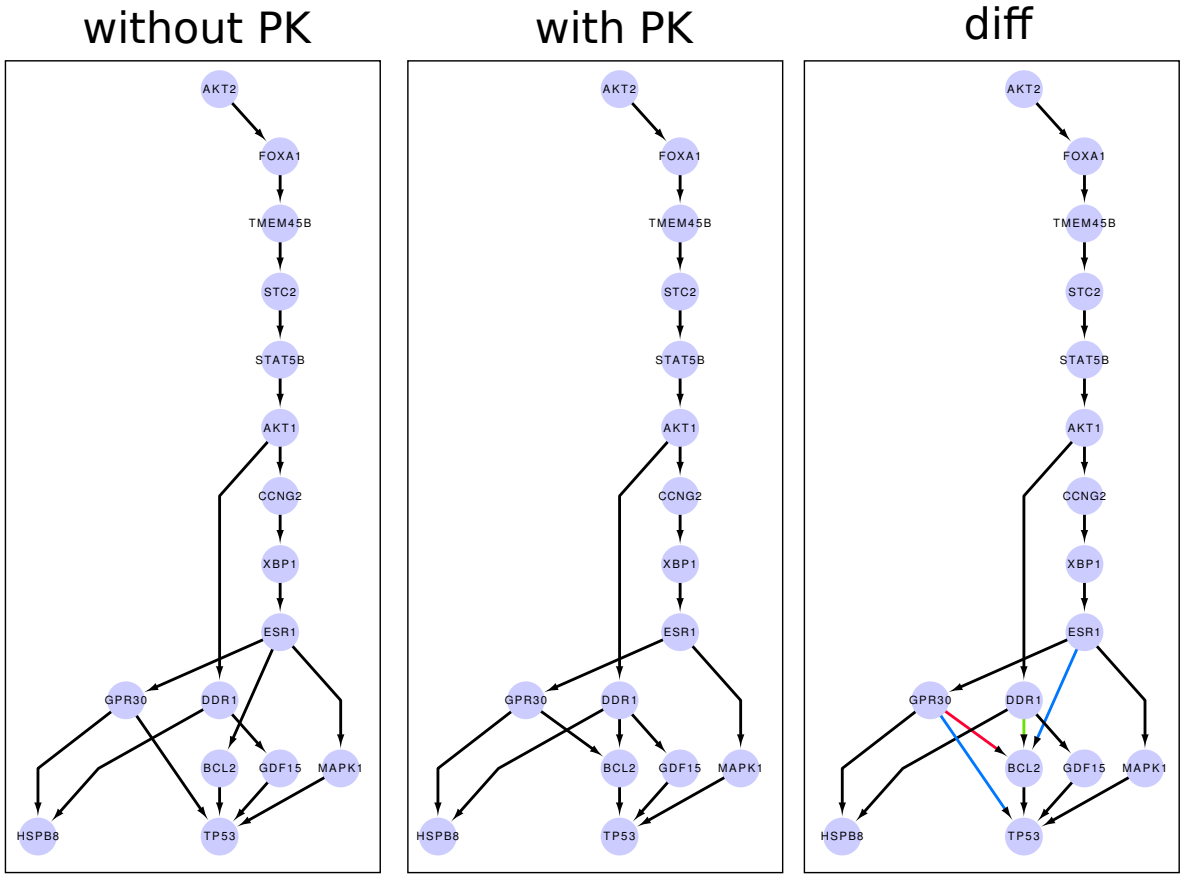

FIGURE 4.1 *Transitively reduced visualization of the overlaps and differences of reconstructed networks with and without integrated prior knowledge.*

The graphs in Figure 4.1 illustrate the transitively reduced results of network reconstruction. The first graph represents the network reconstructed without prior knowledge. The second graph shows the network reconstructed with prior knowledge. The third graph illustrates the differences between the first two

graphs: The green edge ($DDR1 \rightarrow BCL2$) and the red edge ($GPR30 \rightarrow BCL2$) are added. Due to the transitively reduced visualization, the blue edges are redundant with the red edge ($GPR30 \rightarrow BCL2$) and not visible in the second graph due to the transitive reduction.

In order to gain further insights, literature research is conducted using the pathway databases and the `rBiopaxParser` or by performing PubMed-based literature analyzes.

As demonstrated in Section 3.1 the `rBiopaxParser` can be used to retrieve numerous properties of pathways and molecules from pathway databases. For example in this case, interesting aspects about the inferred edges are the pathways these molecules participate in and their overlaps. Furthermore, it can be tested whether there is a direct edge between the individual molecules or if the path between these molecules spans several pathways.

A comparison with Table 3.8 in the previous section shows that the edge $DDR1 \rightarrow BCL2$ is present in all three databases, while the edge $GPR30 \rightarrow BCL2$ is only present in the Reactome database.

In PID the molecules $DDR1$ and $BCL2$ take part in 4 and 26 pathways respectively. The molecules have one pathway in common, the *il-2 receptor beta chain in t-cell activation* pathway. Although the molecules share a pathway, the shortest signaling path between these molecules is $DDR1 \rightarrow CDK1 \rightarrow PRKAR2A \rightarrow BCL2$, spanning across 3 pathways shown in Table 4.2.

Edge	Pathway
$DDR1 \rightarrow CDK1$	*estrogen responsive protein efp controls cell cycle and breast tumors growth*
$CDK1 \rightarrow PRKAR2A$	*stathmin and breast cancer resistance to antimicrotubule agents*
$PRKAR2A \rightarrow BCL2$	*regulation of bad phosphorylation*

TABLE 4.2 *Shortest path $DDR1 \rightarrow BCL2$ in PID.*

Similarly, in the BioCarta database $DDR1$ and $BCL2$ take part in 4 and 15 pathways respectively, sharing pathway *il-2 receptor beta chain in t-cell activation* as well. However, the shortest signaling path between these molecules differs, being $DDR1 \rightarrow CCNB1 \rightarrow BCL2$, spanning 2 pathways (see Table 4.3).

Within Reactome, $DDR1$ and $BCL2$ take part in 67 and 43 pathways respectively. Reactome has a different, very hierarchical organisation of pathways,

Edge	Pathway
$DDR1 \rightarrow CCNB1$	*cyclins and cell cycle regulation*
$CCNB1 \rightarrow BCL2$	*il-2 receptor beta chain in t-cell activation*

TABLE 4.3 *Shortest path $DDR1 \rightarrow BCL2$ in BioCarta.*

where both molecules are shared within the top-level *Disease* pathway. The shortest path is $DDR1 \rightarrow JNK1 \rightarrow BCL2$, connecting the *NRAGE signals death through JNK* and *Innate Immune System* pathways.

Edge	Pathway
$DDR1 \rightarrow JNK1$	*NRAGE signals death through JNK*
$JNK1 \rightarrow BCL2$	*Innate Immune System*

TABLE 4.4 *Shortest path $DDR1 \rightarrow BCL2$ in Reactome.*

Edge $GPR30 \rightarrow BCL2$ is only present in the Reactome database, with the molecules taking part in 224 and 43 pathways respectively. The two molecules share 19 pathways and have a shortest path within the *Activation of BAD and translocation to mitochondria* via $DDR1 \rightarrow PPP3CB \rightarrow BCL2$.

Edge	Pathway
$GPR30 \rightarrow PPP3CB$	*Activation of BAD and translocation to mitochondria*
$PPP3CB \rightarrow BCL2$	*Activation of BAD and translocation to mitochondria*

TABLE 4.5 *Shortest path $GPR30 \rightarrow BCL2$ in Reactome.*

Although it is only present in one database, the addition of edge $GPR30 \rightarrow BCL2$ is reasonable for the reconstructed network in so far as it merges the signaling strands $ESR1 \rightarrow GPR30 \rightarrow TP53$ and $ESR1 \rightarrow BCL2 \rightarrow TP53$ to $ESR1 \rightarrow GPR30 \rightarrow BCL2 \rightarrow TP53$. Furthermore, it coincides and overlaps with the other added edge $DDR1 \rightarrow BCL2(\rightarrow TP53)$, overlapping with the prior knowledge network.

Finally, PubMed analyzes reveal findings that link the gene expression levels of the genes of both edges, which have been observed in several peer-reviewed publications. Liu et al. (2011) found that "[...] the anti-apoptotic activity of GPR30 was dependent on the expression of Bcl-2 and pro-caspase-3." (Liu et al., 2011). Hsieh et al. (2007) report that they "[...] found that suppression of GPR30 but not ER-α prevented E2-BSA- or E2-induced PKA activation and Bcl-2 expression."(Hsieh et al., 2007). Berthier et al. (2005) reported a link

between $BCL2$ and $DDR1$ when studying the involvement of pro- and anti-apoptotic calcium-dependent transduction pathways. Additionally, Kanda and Watanabe (2003) published that "GPR30 anti-sense oligonucleotide did [...] suppress 17β-estradiol-induced cAMP signal, cAMP response element-binding protein phosphorylation, Bcl-2 expression, and apoptosis resistance."(Kanda and Watanabe, 2003).

These findings further strengthen the belief that the integration of prior knowledge into network reconstruction yields new insights into the inner workings of cells.

Chapter 5

Conclusion

With increasing amounts of literature knowledge available electronically and an information overflow in biology and medicine, searching and retrieving data poses a real problem for researchers nowadays. This has turned the focus on archiving complex knowledge in an organized and structured way by faciliating standardized encodings, for example using ontologies to model the knowledge domain. Extending the current knowledge on cellular processes and functions can help to develop new drugs and treatments to address currently lethal diseases and aim for new findings in the field of life sciences in general. The integration of prior knowledge into bioinformatic methods translates into using the accumulated knowledge of the last decades as building blocks for future discoveries. Ultimately, this has been the driving motivation for this thesis.

This thesis touches upon a number of important aspects in bioinformatics, for example the developing research fields of pathway knowledge modeling, pathway databases and the integration of this knowledge into bioinformatic methods. The thesis contains an introduction to methods and underlying concepts used to model pathway knowledge and network reconstruction approaches. Furthermore, a newly implemented open-source software package to work with BioPAX-encoded pathway data within R is presented. Additionally, a workflow to access, merge and transform literature knowledge from various sources into suitably-formated prior knowledge aims at showing possible approaches for the integration of prior knowledge.

Unfortunately, many hurdles in the usage of archived literature knowledge persist. Overall a trend to abiding by standards for encoding pathway knowledge

is noticeable and almost all popular pathway databases are available in one of the current encoding standards. However, the integration and sharing of structured data in medicine and biology remains an underdeveloped field given the current tools and documentation. Furthermore, in the short-term this situation is likely to persist due to rapid development of and changes to current standards. Nevertheless, fundamental steps have been made towards the archiving and reproducible use of structured data. Hopefully, these steps can be used as a leverage to enable new discoveries and findings in biology and medicine.

References

Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255. PMID: 16880171.

Akutsu, T., Kuhara, S., Maruyama, O., and Miyano, S. (1998). A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. *Genome Informatics Series*, (9):151–160.

Alberts, B. (2008). *Molecular biology of the cell*. Garland Science, New York.

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression. *PLoS Comput Biol*, 9(3):e1002955.

Anchang, B., Sadeh, M. J., Jacob, J., Tresch, A., Vlad, M. O., Oefner, P. J., and Spang, R. (2009). Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proceedings of the National Academy of Sciences*, 106(16):6447–6452. PMID: 19329492.

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., and Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, 12(4):357–368. PMID: 21712342.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E. W., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers,

M., Albrecht, M., and Hermjakob, H. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods*, 8(7):528–529.

Arnelh (2009). A diagram showing at which stages in the DNA-mRNA-protein pathway expression can be controlled. *Wikimedia Commons*, page http://commons.wikimedia.org/wiki/File:Gene_expression_control.png.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Augenlicht, L. H. and Kobrin, D. (1982). Cloning and Screening of Sequences Expressed in a Mouse Colon Tumor. *Cancer Research*, 42(3):1088–1093. PMID: 7059971.

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a Pathway Resource List. *Nucleic Acids Research*, 34(suppl 1):D504–D506.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(suppl 1):D154–D159. PMID: 15608167.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390.

Bauer-Mehren, A., Furlong, L. I., and Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology*, 5(1).

Beck, T., Free, R. C., Thorisson, G. A., and Brookes, A. J. (2012). Semantically enabling a genome-wide association study database. *Journal of Biomedical Semantics*, 3(1):9. PMID: 23244533.

Beckett, D. and McBride, B. (2004). RDF/XML syntax specification (revised). *W3C recommendation*, 10.

Beißbarth, T. (2006). Interpreting Experimental Results Using Gene Ontologies. In Kimmel, A. and Oliver, B., editors, *Methods in Enzymology*, volume 411, pages 340–352. Academic Press, Waltham.

Beißbarth, T. and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465. PMID: 14962934.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific American*, 284(5):28–37.

Berthier, A., Lemaire-Ewing, S., Prunet, C., Montange, T., Vejux, A., Pais de Barros, J. P., Monier, S., Gambert, P., Lizard, G., and Néel, D. (2005). 7-Ketocholesterol-induced apoptosis. *FEBS Journal*, 272(12):3093–3104.

Bickel, D. R. (2005). Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics*, 21(7):1121–1128. PMID: 15546939.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193. PMID: 12538238.

Bornstein, B. J., Keating, S. M., Jouraku, A., and Hucka, M. (2008). LibSBML: an API Library for SBML. *Bioinformatics*, 24(6):880–881. PMID: 18252737.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (1997). Extensible markup language (XML). *World Wide Web Journal*, 2(4):27–66.

Brazhnik, P., de la Fuente, A., and Mendes, P. (2002). Gene networks: how to put the function in genomics. *Trends in Biotechnology*, 20(11):467–472.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S. (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71. PMID: 12519949.

Büchel, F., Wrzodek, C., Mittag, F., Dräger, A., Eichner, J., Rodriguez, N., Novère, N. L., and Zell, A. (2012). Qualitative translation of relations from BioPAX to SBML qual. *Bioinformatics*, 28(20):2648–2653.

Burkhardt, H. and Smith, B. (1991). *Handbook of metaphysics and ontology*. Philosophia Verlag, Muenchen.

Cary, M. P., Bader, G. D., and Sander, C. (2005). Pathway information for systems biology. *FEBS Letters*, 579(8):1815–1820.

Castro, M. A., Wang, X., Fletcher, M. N., Meyer, K. B., and Markowetz, F. (2012). RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biology*, 13(4):R29. PMID: 22531049.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690. PMID: 21071392.

Chaouiya, C., Bérenguier, D., Keating, S. M., Naldi, A., Iersel, M. P. v., Rodriguez, N., Dräger, A., Büchel, F., Cokelaer, T., Kowal, B., Wicks, B., Gonçalves, E., Dorier, J., Page, M., Monteiro, P. T., Kamp, A. v., Xenarios, I., Jong, H. d., Hucka, M., Klamt, S., Thieffry, D., Novère, N. L., Saez-Rodriguez, J., and Helikar, T. (2013). SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Systems Biology*, 7(1):135. PMID: 24321545.

Coles, H. (1994). Nobel honours pursuit of G proteins. *Nature*, 371(6498):547. PMID: 7935774.

Consortium, T. G. O. (2008). The Gene Ontology project in 2008. *Nucleic Acids Research*, 36(suppl 1):D440–D444. PMID: 17984083.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl 1):D691–D697.

Degtyarenko, K., Matos, P. d., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl 1):D344–D350. PMID: 17932057.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Ruebenacker, O., Samwald, M., Iersel, M. v., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Novère, N. L., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13):1547–1548. PMID: 18467348.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, 12(1):111–140.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210. PMID: 11752295.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868. PMID: 9843981.

Ellson, J., Gansner, E., Koutsofios, L., North, S. C., and Woodhull, G. (2002). Graphviz— Open Source Graph Drawing Tools. In Mutzel, P., Jünger, M., and Leipert, S., editors, *Graph Drawing*, number 2265 in Lecture Notes in Computer Science, pages 483–484. Springer, Berlin Heidelberg.

Failmezger, H., Praveen, P., Tresch, A., and Fröhlich, H. (2013). Learning gene network structure from time laps cell imaging in RNAi Knock downs. *Bioinformatics*, 29(12):1534–1540. PMID: 23595660.

Fröhlich, H., Beißbarth, T., Tresch, A., Kostka, D., Jacob, J., Spang, R., and Markowetz, F. (2008a). Analyzing gene perturbation screens with nested effects models in R and bioconductor. *Bioinformatics*, 24(21):2549–2550. PMID: 18718939.

Fröhlich, H., Fellmann, M., Sueltmann, H., Poustka, A., and Beissbarth, T. (2007a). Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC Bioinformatics*, 8(1):386. PMID: 17937790.

Fröhlich, H., Fellmann, M., Sültmann, H., Poustka, A., and Beissbarth, T. (2008b). Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data. *Bioinformatics*, 24(22):2650–2656. PMID: 18227117.

Fröhlich, H., Praveen, P., and Tresch, A. (2011). Fast and efficient dynamic nested effects models. *Bioinformatics*, 27(2):238–244. PMID: 21068003.

Fröhlich, H., Speer, N., Poustka, A., and Beißbarth, T. (2007b). GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, 8(1):166. PMID: 17519018.

Fröhlich, H., Tresch, A., and Beißbarth, T. (2009). Nested effects models for learning signaling networks from perturbation data. *Biometrical Journal*, 51(2):304–323.

Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1(5):159–162.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315. PMID: 14960456.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80. PMID: 15461798.

Giardine, B., Borg, J., Higgs, D. R., Peterson, K. R., Philipsen, S., Maglott, D., Singleton, B. K., Anstee, D. J., Basak, A. N., Clark, B., Costa, F. C., Faustino, P., Fedosyuk, H., Felice, A. E., Francina, A., Galanello, R., Gallivan, M. V. E., Georgitsi, M., Gibbons, R. J., Giordano, P. C., Harteveld, C. L., Hoyer, J. D., Jarvis, M., Joly, P., Kanavakis, E., Kollia, P., Menzel, S., Miller, W., Moradkhani, K., Old, J., Papachatzopoulou, A., Papadakis, M. N., Papadopoulos, P., Pavlovic, S., Perseu, L., Radmilovic, M., Riemer, C., Satta, S., Schrijver, I., Stojiljkovic, M., Thein, S. L., Traeger-Synodinos, J., Tully, R., Wada, T., Waye, J. S., Wiemann, C., Zukic, B., Chui, D. H. K., Wajcman, H., Hardison, R. C., and Patrinos, G. P. (2011). Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics*, 43(4):295–301.

Groth, P., Gibson, A., and Velterop, J. (2010). The Anatomy of a Nanopublication. *Inf. Serv. Use*, 30(1-2):51–56.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5–6):907–928.

Gu, Z. and Wang, J. (2013). CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics*, 29(5):658–660. PMID: 23314125.

Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8):967–973. PMID: 16492685.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems*, 96(1):86–103.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2001). Dependency networks for inference, collaborative filtering, and data visualization. *The Journal of Machine Learning Research*, 1:49–75.

Hendricks, S. B. (1953). A discussion of photosynthesis. *Science*, 117(3041):370–373. PMID: 13048686.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004). The HUPO PSI's Molecular Interaction

format—a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183.

Hitzler, P., Krotzsch, M., and Rudolph, S. (2011). *Foundations of Semantic Web Technologies*. CRC Press.

Hornik, K. (2012). The Comprehensive R Archive Network. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4):394–398.

Hsieh, Y., Yu, H., Frink, M., Suzuki, T., Choudhry, M. A., Schwacha, M. G., and Chaudry, I. H. (2007). G Protein-Coupled Receptor 30-Dependent Protein Kinase A Pathway Is Critical in Nongenomic Effects of Estrogen in Attenuating Liver Injury after Trauma-Hemorrhage. *The American Journal of Pathology*, 170(4):1210–1218.

Hu, Z., Snitkin, E. S., and DeLisi, C. (2008). VisANT: an integrative framework for networks in systems biology. *Briefings in Bioinformatics*, 9(4):317–325. PMID: 18463131.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novère, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282. PMID: 14630656.

Ideker, T. and Lauffenburger, D. (2003). Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends in Biotechnology*, 21(6):255–262.

Ideker, T. E., Thorsson, V., and Karp, R. M. (2000). Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, pages 305–316. PMID: 10902179.

Iersel, M. P. v., Kelder, T., Pico, A. R., Hanspers, K., Coort, S., Conklin, B. R., and Evelo, C. (2008). Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, 9(1):399. PMID: 18817533.

Iersel, M. P. v., Villéger, A. C., Czauderna, T., Boyd, S. E., Bergmann, F. T., Luna, A., Demir, E., Sorokin, A., Dogrusoz, U., Matsuoka, Y., Funahashi, A., Aladjem, M. I., Mi, H., Moodie, S. L., Kitano, H., Novère, N. L., and Schreiber, F. (2012). Software support for SBGN maps: SBGN-ML and LibSBGN. *Bioinformatics*, 28(15):2016–2021. PMID: 22581176.

Jiang, R., Gan, M., and He, P. (2011). Constructing a gene semantic similarity network for the inference of disease genes. *BMC Systems Biology*, 5:S2. PMID: 22784573.

Kaderali, L. and Radde, N. (2008). Inferring Gene Regulatory Networks from Expression Data. In Kelemen, A., Abraham, A., and Chen, Y., editors, *Computational Intelligence in Bioinformatics*, number 94 in Studies in Computational Intelligence, pages 33–74. Springer, Berlin Heidelberg.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, 39(suppl 1):D712–D717. PMID: 21071422.

Kanda, N. and Watanabe, S. (2003). 17β-Estradiol Inhibits Oxidative Stress-Induced Apoptosis in Keratinocytes by Promoting Bcl-2 Expression. *Journal of Investigative Dermatology*, 121(6):1500–1509.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1):D277–D280.

Karp, G. (2010). *Cell and molecular biology: concepts and experiments*. John Wiley, Hoboken, NJ.

Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. (2002). The MetaCyc Database. *Nucleic Acids Research*, 30(1):59–61. PMID: 11752254.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2011). WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307.

Kerr, J. F., Wyllie, A. H., and Currie, A. R. (1972). Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer*, 26(4):239–257. PMID: 4561027.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., Bader, G. D., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J. J., Moore, S., Ceol, A., Chatr-aryamontri, A., Oesterheld, M., Stümpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M. E., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., and Hermjakob, H. (2007). Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, 5(1):44. PMID: 17925023.

Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005). Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, 23(8):961–966.

Klyne, G., Carroll, J. J., and McBride, B. (2004). Resource description framework (RDF): Concepts and abstract syntax. *W3C recommendation*, 10.

Kohn, K. W. (1999). Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems. *Molecular Biology of the Cell*, 10(8):2703–2734. PMID: 10436023.

Kostka, D. and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20(suppl 1):i194–i199. PMID: 15262799.

Kramer, F., Bayerlová, M., and Beißbarth, T. (2014). R-Based Software for the Integration of Pathway Data into Bioinformatic Algorithms. *Biology*, 3(1):85–100.

Kramer, F., Bayerlová, M., Klemm, F., Bleckmann, A., and Beißbarth, T. (2013). rBiopaxParser—an R package to parse, modify and visualize BioPAX data. *Bioinformatics*, 29(4):520–522. PMID: 23274212.

Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., and Wingender, E. (2006). TRANSPATH®: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Research*, 34(suppl 1):D546–D551. PMID: 16381929.

Lan, A., Ziv-Ukelson, M., and Yeger-Lotem, E. (2013). A context-sensitive framework for the analysis of human signalling pathways in molecular interaction networks. *Bioinformatics*, 29(13):i210–216. PMID: 23812986 PMCID: PMC3694656.

Lang, D. T. (2000). The Omegahat Environment: New Possibilities for Statistical Computing. *Journal of Computational and Graphical Statistics*, 9(3):423–451.

Lang, D. T. (2007). R as a Web Client–the RCurl package. *Journal of Statistical Software*.

Lang, D. T. (2013). *XML: Tools for parsing and generating XML within R and S-Plus*. R package version 3.95-0.2.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94(24):13057–13062. PMID: 9371799.

Liu, S., Han, J., Zhang, N., Tian, Z., Li, X., and Zhao, M. (2011). Neuroprotective effects of oestrogen against oxidative toxicity through activation of G-protein-coupled receptor 30 receptor. *Clinical and experimental pharmacology & physiology*, 38(9):577–585. PMID: 21645039.

Lotia, S., Montojo, J., Dong, Y., Bader, G. D., and Pico, A. R. (2013). Cytoscape app store. *Bioinformatics (Oxford, England)*, 29(10):1350–1351. PMID: 23595664.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(suppl 1):D54–D58. PMID: 15608257.

Marbach, D., Costello, J. C., Kuffner, R., Vega, N., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804. PMID: 22796662.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*. PMID: 20308593.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7. PMID: 16723010.

Markowetz, F. (2005). *Probabilistic Models for Gene Silencing Data*. PhD thesis, FU Berlin.

Markowetz, F. (2010). How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLoS Comput Biol*, 6(2):e1000655.

Markowetz, F., Bloch, J., and Spang, R. (2005). Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21(21):4026–4032. PMID: 16159925.

Markowetz, F., Kostka, D., Troyanskaya, O. G., and Spang, R. (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305–i312. PMID: 17646311.

Markowetz, F. and Spang, R. (2003). Evaluating the effect of perturbations in reconstructing network topologies. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks – a review. *BMC Bioinformatics*, 8(Suppl 6):S5. PMID: 17903286.

Maskos, U. and Southern, E. M. (1992). Oligonucleotide hybridisations on glass supports: a novel linker for oligonucleotide synthesis and hybridisation properties of oligonucleotides synthesised in situ. *Nucleic Acids Research*, 20(7):1679–1684. PMID: 1579459.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., Bono, B. d., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(suppl 1):D619–D622. PMID: 18981052.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378. PMID: 12520026.

McGuinness, D. L., Van Harmelen, F., et al. (2004). OWL web ontology language overview. *W3C recommendation*, 10(2004-03):10.

Meyerhof, O. (1927). RECENT INVESTIGATIONS ON THE AEROBIC AND AN-AEROBIC METABOLISM OF CARBOHYDRATES. *The Journal of general physiology*, 8(6):531–542. PMID: 19872214 PMCID: PMC2140809.

Mitra, K., Carvunis, A., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

Miura, Y. and Duncan, J. (1973). Wall charts and metabolic maps. *Biochemical Education*, 1(4):88–88.

Mons, B., van Haagen, H., Chichester, C., Hoen, P. t., den Dunnen, J. T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., and Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4):281–283.

Mukherjee, S. and Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105(38):14313–14318. PMID: 18799736.

NatGenEditorial (2008). Human Variome Microattribution Reviews. *Nature Genetics*, 40(1):1–1.

Niederberger, T., Etzold, S., Lidschreiber, M., Maier, K. C., Martin, D. E., Fröhlich, H., Cramer, P., and Tresch, A. (2012). MC EMiNEM Maps the Interaction Landscape of the Mediator. *PLoS Comput Biol*, 8(6):e1002568.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, 2(3):117–120.

Novère, N. L., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (2009). The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741.

Noy, N. F., McGuinness, D. L., et al. (2001). *Ontology development 101: A guide to creating your first ontology.* Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M., Chute, C. G., and Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Suppl 2):W170–W173. PMID: 19483092.

Ochs, R. S. and Conrow, K. (1991). A computerized metabolic map. *Journal of Chemical Information and Computer Sciences*, 31(1):132–137. PMID: 2026661.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34. PMID: 9847135.

Orchard, S., Albar, J. P., Deutsch, E. W., Eisenacher, M., Vizcaíno, J. A., and Hermjakob, H. (2011). Enabling BioSharing – a report on the Annual Spring Workshop of the HUPO-PSI April 11–13, 2011, EMBL-Heidelberg, Germany. *PROTEOMICS*, 11(22):4284–4290.

Pace, N. R. (2001). The universal nature of biochemistry. *Proceedings of the National Academy of Sciences*, 98(3):805–808. PMID: 11158550.

Patrinos, G. P., Cooper, D. N., van Mulligen, E., Gkantouna, V., Tzimas, G., Tatum, Z., Schultes, E., Roos, M., and Mons, B. (2012). Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Human Mutation*, 33(11):1503–1512.

Paz, A., Brownstein, Z., Ber, Y., Bialik, S., David, E., Sagir, D., Ulitsky, I., Elkon, R., Kimchi, A., Avraham, K. B., Shiloh, Y., and Shamir, R. (2011). SPIKE: a database of highly curated human signaling pathways. *Nucleic Acids Research*, 39(suppl 1):D793–D799. PMID: 21097778.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Pe'er, D., Regev, A., Elidan, G., and Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl 1):S215–S224. PMID: 11473012.

Penfold, C. A. and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870. PMID: 23226586.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4. PMID: 18460186.

Petri, V., Shimoyama, M., Hayman, G. T., Smith, J. R., Tutaj, M., de Pons, J., Dwinell, M. R., Munzenmaier, D. H., Twigger, S. N., Jacob, H. J., and

Team, R. (2011). The Rat Genome Database Pathway Portal. *Database*, 2011(0):bar010.

Plessis, L. d., Škunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology—a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735. PMID: 21330331.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. (2010). Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLoS ONE*, 5(2):e9202.

Qiu, X., Wu, H., and Hu, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, 14(1):124. PMID: 23578321.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, 32 Suppl:496–501. PMID: 12454644.

Rice, J. J., Tu, Y., and Stolovitzky, G. (2005). Reconstructing biological networks using conditional correlation analysis. *Bioinformatics*, 21(6):765–773. PMID: 15486043.

Rodbell, M., Birnbaumer, L., Pohl, S. L., and Sundby, F. (1971). The Reaction of Glucagon with Its Receptor: Evidence for Discrete Regions of Activity and Binding in the Glucagon Molecule. *Proceedings of the National Academy of Sciences*, 68(5):909–913. PMID: 5280527.

Rogers, S. and Girolami, M. (2005). A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, 21(14):3131–3137. PMID: 15879452.

Rubin, D. L., Shah, N. H., and Noy, N. F. (2008). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90. PMID: 18077472.

Ruebenacker, O., Moraru, I. I., Schaff, J. C., and Blinov, M. L. (2009). Integrating BioPAX pathway knowledge with SBML models. *IET systems biology*, 3(5):317–328. PMID: 21028923.

Sales, G., Calura, E., Cavalieri, D., and Romualdi, C. (2012). graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, 13(1):20. PMID: 22292714.

Sansone, S., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C. T., Forster, M. J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J. L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Sui, S. J. H., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C. E., Shang, C. A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., and Hide, W. (2012). Toward interoperable bioscience data. *Nature Genetics*, 44(2):121–126.

Schacherer, F., Choi, C., Götze, U., Krull, M., Pistor, S., and Wingender, E. (2001). The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, 17(11):1053–1057. PMID: 11724734.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(suppl 1):D674–D679.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764. PMID: 15479708.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:Article32. PMID: 16646851.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504. PMID: 14597658.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, page 397–420. Springer, New York.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003). Statistical Issues in cDNA Microarray Data Analysis. In Brownstein, M. J. and Khodursky, A. B., editors, *Functional Genomics*, number 224 in Methods in Molecular Biology, pages 111–136. Humana Press, Totowa.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297. PMID: 9843569.

Squidonius (2008). Steps involved in a DNA microarray experiment. *Wikimedia Commons*, page http://commons.wikimedia.org/wiki/File:DNA_microarray_experiment.svg.

Stanier, R. Y. (1947). Simultaneous Adaptation: A New Technique for the Study of Metabolic Pathways. *Journal of Bacteriology*, 54(3):339–348. PMID: 16561366.

Strömbäck, L. and Lambrix, P. (2005). Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24):4401–4407.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255. PMID: 12934013.

Suderman, M. and Hallett, M. (2007). Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659. PMID: 17720984.

Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. http://www.R-project.org.

Tresch, A. and Markowetz, F. (2008). Structure Learning in Nested Effects Models. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Ulrich, L. E. and Zhulin, I. B. (2010). The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Research*, 38(suppl 1):D401–D407. PMID: 19900966.

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., Bono, B. d., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., and Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, 8(3):R39. PMID: 17367534.

Virchow, R. (1855). Cellular-Pathologie. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, 8(1):3–39.

Webb, R. L. and Ma'ayan, A. (2011). Sig2BioPAX: Java tool for converting flat files to BioPAX Level 3 format. *Source Code for Biology and Medicine*, 6(1):5. PMID: 21418653.

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531. PMID: 16844710.

Werhli, A. V. and Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., Rohr, P. v., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology*, 5(11):R92. PMID: 15535868.

Wimburly, F. C., Heiman, T., Ramsey, J., and Glymour, C. (2003). Experiments on the accuracy of algorithms for inferring the structure of genetic regulatory networks from microarray expression levels. *Proc IJCAI 2003 Bioinformatics Workshop 2003*, 1(1).

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic acids research*, 35(Database issue):D521–526. PMID: 17202168.

Wrzodek, C., Büchel, F., Ruff, M., Dräger, A., and Zell, A. (2013). Precise generation of systems biology models from KEGG pathways. *BMC Systems Biology*, 7(1):15. PMID: 23433509.

Yamanishi, Y., Vert, J., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(suppl 1):i363–i370. PMID: 15262821.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 3(8):579–588.

Zak, D. E., Gonye, G. E., Schwaber, J. S., and Doyle, F. J. (2003). Importance of Input Perturbations and Stochastic Gene Expression in the Reverse Engineering of Genetic Regulatory Networks: Insights From an Identifiability Analysis of an In Silico Network. *Genome Research*, 13(11):2396–2405. PMID: 14597654.

Zhang, J. D. and Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, 25(11):1470–1471. PMID: 19307239.