Computational approaches for biological data integration

Nandal, U.

Publication date
2023

CHAPTER 1 INTRODUCTION

1.1 Background

Biological processes in a cell are highly dynamic and their regulation involves a multitude of molecular components such as DNA, genes, proteins, and metabolites. It is of critical importance to understand these entities not only as separate elements but also in terms of their interactions with one another. The elucidation of molecular interactions in a cell is essential for the identification of genes and pathways linked to disease. In specific, deeper knowledge and understanding of molecular interactions contribute to approaches towards disease prevention and control of disease progression, stratification in disease subtypes, and personalized treatment (Van Kampen and Moerland, 2016).

Advances in high-throughput technologies (HTTs) such as next generation sequencing (NGS) and quantitative mass spectrometry (MS) have allowed researchers to generate measurements for nearly all types of molecular entities from an individual. At a single level, one can measure the abundances (gene expression, protein expression, metabolite concentration) or states (DNA methylation, post-translational modifications such as histone modifications) or interactions (protein-protein complexes) of such molecular entities from different tissues, across conditions or more recently even in single cells. In some cases, interactions between levels can also be measured, for example using chromatic immunoprecipitation (ChIP) followed by sequencing for genome-wide profiling of DNA-protein interactions (Park, 2009). The collection of such measurements of various types of molecular entities generated using HTTs is referred to as 'omics' (Figure 1.1). Here, *genomics* involves the study of genes and their functions by determining whole-genome sequences with NGS. *Epigenomics* concerns the study of epigenetic events such as histone

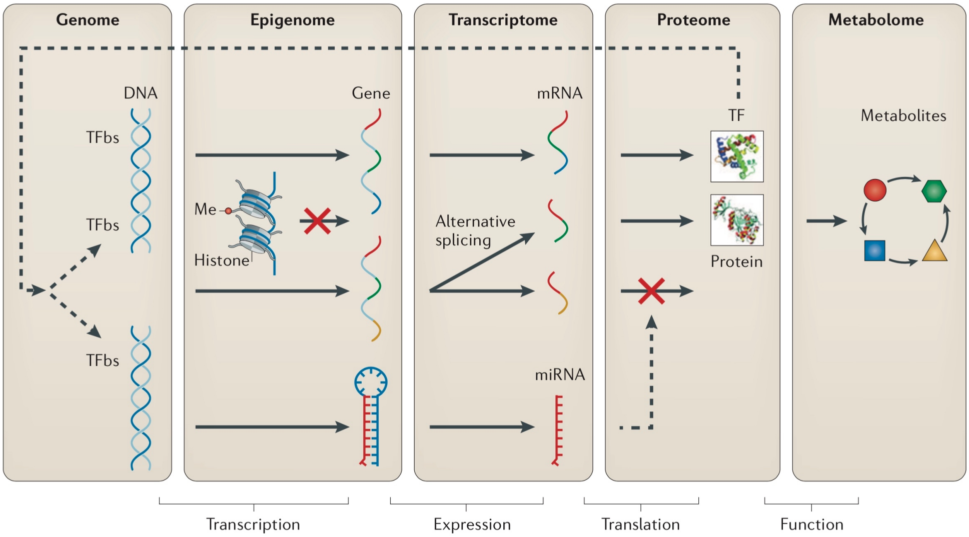

Figure 1.1: Different omics modalities in biology. Red crosses indicate repression. TFbs, transcription factor binding site; Me, methylation. Figure adapted from Ritchie *et al.* (2015a).

modifications (using ChiP-seq) and DNA methylation (using microarrays or NGS). *Transcriptomics* concerns the study of mRNA by the measurement of gene expression with microarrays or NGS. *Proteomics* involves the study of proteins. At the proteome level, one measures protein abundance, protein-protein interactions or protein complexes. Analytical techniques such as nuclear magnetic resonance, liquid or gas chromatography (LC/GC) mass spectrometry and two dimensional differential gel electrophoresis (2D-DIGE) have been commonly used to identify and/or quantify proteins. *Metabolomics* consists of the systematic quantification and identification of metabolites involved in cellular processes. For metabolomics, similar MS- and NMR-based techniques are used as for proteomics.

Technological developments have increasingly enabled multiple omics modalities to be measured in parallel. Integrative analyses that make use of information within and across these layers can provide a better understanding of the biological system of interest. However, the inherent heterogeneity of the data modalities and the rapid generation of vast volumes of data have made the analysis and interpretation of (multi-)omics data challenging. This thesis addresses three main challenges for integrative omics data analysis, (*i*) easy accessibility of high-throughput data so that it can be combined with in-house experimental data or used for reanalysis (Section 1.2), (*ii*) integration of information across different resources (Section 1.3.1) and (*iii*) integration within or across data modalities (Section 1.3.2 and 1.4).

1.2 Information management

The scientific community has generated a considerable amount of biological data over the past decades, which is distributed over hundreds to thousands of publicly available databases representing many different types of information (Rung and Brazma, 2013; Toomula *et al.*, 2012). The 2022 release of the NAR online Molecular Biology Database Collection lists 1,645 databases in the field of molecular biology (Rigden and Fernández, 2021)[1]. Even an entire journal (Database: The Journal of Biological Databases and Curation) dedicated to biological databases and their curation has been established. For functional genomics data such as gene expression, gene regulation and epigenomics data, more than 100,000 experiments are accessible via public repositories such as the Gene Expression Omnibus (Barrett *et al.*, 2013) and ArrayExpress (Rustici *et al.*, 2013). Although numerous experimental datasets in GEO and ArrayExpress are still based on microarrays, the contribution of NGS-based studies (RNA-seq and ChIP-seq, for example) is growing rapidly. In addition, centralized public repositories for proteomics data (Vizcaíno *et al.*, 2016) and metabolomics data (Haug *et al.*, 2013) have also been established. Most databases provide tools or platforms to search and navigate biological information stored in them. In addition, there are several other ways to retrieve information and datasets from such public repositories. First, via search queries directly on the corresponding website and then downloading the results by exporting them in pre-defined formats. Second, via programmatic access of web services to query databases such as representational state transfer application interfaces (REST-APIs)[2]. Third, by downloading the entire database as Oracle, SQL database or tab/csv delimited files[3].

Reuse of datasets from these repositories can be highly valuable and cost effective. However, retrieving, systematically storing, analyzing, and integrating datasets from public repositories to extract novel biological insights comes with many challenges. First, it is often challenging to easily access and mine public repositories to extract biological insights for a specific study of interest. Therefore, improved technological solutions are required for seamless access to information from public repositories via programmatic or API access such as

[1] `http://www.oxfordjournals.org/nar/database/c`

[2] See for example `https://www.ebi.ac.uk/arrayexpress/help/programmatic_access.html` for REST-style queries for ArrayExpress.

[3] See for example `https://chembl.gitbook.io/chembl-interface-documentation/downloads` for downloads of ChEMBL.

Google Dataset Search[4] or Zenodo[5], where researchers cannot only collaborate but also share their datasets. Second, most resources do not enable easy integration with the rich collection of software tools available for subsequent analyses, for example R/Bioconductor packages (Nandal *et al.*, 2016). Third, often data restructuring is needed, so that the data can be effectively analyzed. This also requires proper annotation of data with pre-defined vocabularies or ontologies. Many datasets lack complete meta-data or contain annotation errors and therefore require to manually extract (meta) data from the corresponding paper and possibly communication with the authors (Wang *et al.*, 2019). Moreover, given the vast heterogeneity of data within and across omics layers, transformation of data to common formats for integration can also be challenging. Finally, biases due to systematic differences between measurement platforms, laboratories and analysis methods complicate the integration and analysis of publicly available data.

Chapter 2 of this thesis describes our approach to build a compendium of functional genomics data retrieved from GEO. With the associated R package `compendiumdb` and the accompanying MySQL database, preprocessed GEO data from different studies and profiling platforms can be systematically retrieved and stored.

1.3 Integration of information sources and data

Integration can have different meanings depending on the scientific context. According to Gomez-Cabrero *et al.* (2014), data integration refers to an integrative study of multiple sources (for example, biological databases) and data modalities that can enhance knowledge discovery for a given system. We also view integration in this context. It is therefore not only limited to the combination of data using statistical methods, but also includes the integration with expert biological knowledge using various bioinformatics and computational tools (Hamid *et al.*, 2009). We, therefore, broadly categorize integration into two types, which are discussed in detail below. First, we describe the integration of *information sources*, *i.e.*, integration of biological information generated by the scientific community, which is distributed across different databases, for example, National Center for Biotechnology Information (NCBI) and Ensembl (Section 1.3.1). Next, we describe the integration of *data*, which we further categorize into (i) horizontal or homogeneous data integration, where the same type of data is integrated across multiple studies

[4] `https://datasetsearch.research.google.com/`

[5] `https://zenodo.org/`

(*e.g.*, experiments) and (*ii*) vertical or heterogeneous data integration, where multiple data modalities are integrated (Section 1.3.2).

1.3.1 Integration of information sources

Integration of information sources or databases is important for two main reasons. On the one hand, there is some degree of overlap between various databases that describe the same biological domain and database integration can therefore reduce data redundancy. However, this poses considerable challenges as illustrated in a systematic comparison of five frequently used human metabolic pathway databases (Stobbe *et al.*, 2011). Here, it was shown that the five databases agree on only a small subset of the metabolic network. In specific, only 3% of the metabolic reactions was included in all five databases. This implies that the contents of metabolic pathway databases are highly non-overlapping. However, discerning useful complementary information from disagreements and errors is difficult. On the other hand, many databases contain a different subset of biological knowledge that when integrated might enable answering more complex questions. For instance, for a gene its nucleotide sequence is stored in GenBank, pathways in which it is involved are catalogued in the KEGG Pathway Database, associated gene expression data can be retrieved from ArrayExpress, and its association with human diseases is annotated in MalaCards. Obtaining a unified view across these databases is therefore crucial to understand the role of a gene of interest.

Approaches to integrate information resources can be divided into three categories, (*i*) resource portals, (*ii*) warehouse integration and (*iii*) problem-driven integration. *Portals* of the major international bioinformatics institutes (NCBI, EBI[6], SIB[7]) offer access to a host of individual databases by a simple query mechanism. Portals are a useful first step to getting multiple views on a biological entity of interest. *Warehouse integration* involves merging of multiple databases into a data warehouse. To accommodate all the information that is contained in the source databases, the first step is to develop a unified data model. Next, one creates a series of software programs to retrieve data from the source databases, transforms the data to match the unified data model and then loads the data into the warehouse. The aggregated data can then be queried or analyzed using search or algorithmic tools that are integrated with the data warehouse. Examples of data warehousing frameworks that have been developed include BioMart, BioXRT, InterMine, and PathwayTools (Triplet and Butler, 2014). One of the biggest challenges is to keep a data warehouse

[6] The European Bioinformatics Institute

[7] Swiss Institute of Bioinformatics

up to date, since its source databases are continuously being updated with new information. *Problem-driven integration* refers to the *ad hoc* integration of heterogeneous data sources to solve a specific research question. This approach therefore does not require a structured data warehouse or regular updates of the underlying data sources. Most papers with a large bioinformatics component include examples of this type of integration.

Chapter 3 of this thesis describes a problem-driven integrative analysis approach across different data sources to rank candidate proteins for low-abundant spots in 2D-DIGE experiments.

1.3.2 Integration of data

Information on which multiple sources agree has a higher probability to be correct than information from a single source (Jansen *et al.*, 2002; Zhang *et al.*, 2013). Therefore, often greater statistical power and higher precision can be obtained by integrating data. Data integration is also an essential step to gain a coherent view of large (multi-)omics datasets. This section describes two key components of data integration: (*i*) integration types and (*ii*) integration stages.

Integration types

Data integration can be categorized as either horizontal or vertical depending on the type of data being integrated.

Horizontal data integration. Integration across homogeneous data from the same data modality or omics layer is referred to as horizontal integration (Figure 1.2). Horizontal integration includes integration of data from the same platform. One of the major obstacles for such integrative analyses are study-specific effects, for instance, differences in experimental conditions, sample processing and other batch effects between the original datasets. When performing horizontal data integration these effects have to be assessed and corrected for. For example, Sontrop *et al.* (2011) integrated ten breast cancer gene expression datasets measured on the same Affymetrix microarray platform using z-score normalization in order to compare subtype-specific breast cancer event predictors with predictors that do not take subtype information into account. Another example is Immuno-Navigator, a batch-corrected gene expression and coexpression database for immune cell types (Vandenbon *et al.*, 2016). Immuno-Navigator contains integrated expression datasets from 24 human and 19 mouse immune cell types, respectively. Datasets were measured on the same mouse/human Affymetrix microarray platform.

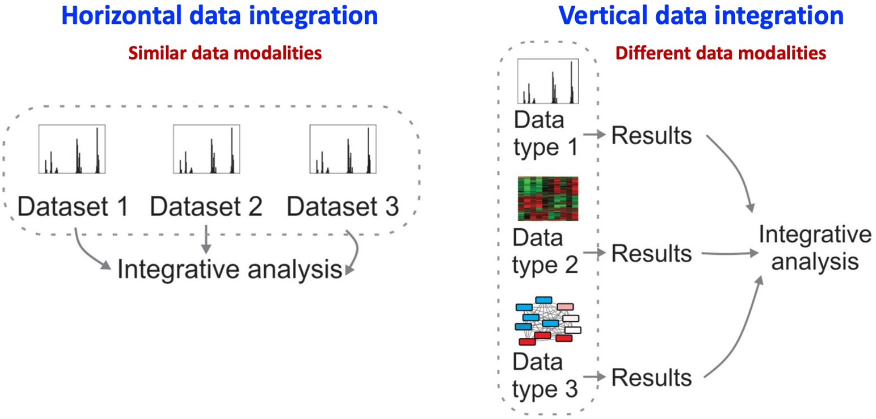

Figure 1.2: Horizontal and vertical data integration.

Study-specific differences are even more pronounced when integrating data across different platforms or measurement techniques for the same modality. A typical example is the integration of transcriptomics data captured using microarrays and RNA-seq. Separate analysis of microarray or RNA-seq data for experiments probing similar conditions reveals only part of the expression profile for a given gene. Therefore, by integrating data across microarray and RNA-seq technologies one can get a more complete picture of how genes are expressed. An example of such an approach is a recent study of gene expression in Down syndrome in which microarray and RNA-seq expression profiles of trisomy 21 and normal tissues were integrated (Pelleri *et al.*, 2018). This enabled not only the calculation of Down syndrome versus normal gene expression fold changes, but also the identification of differentially expressed genomic segments.

Vertical data integration. When datasets measured in different data modalities or omics layers are combined, this is referred to as vertical integration (Figure 1.2). A typical example is the integrated analysis of experiments in which multiple omics layers are probed in the same set of samples. For example, Roychowdhury *et al.* (2011) integrated data from genomics and transcriptomics layers in a clinical oncology setting. For each cancer patient, whole-genome sequencing of the tumour, targeted whole-exome sequencing of normal and tumour DNA, and RNA-seq of the tumour was performed to identify disease-related aberrations within several weeks. Integration over two

or more omics levels can further enhance understanding of cellular mechanisms and adaptive responses. Ebrahim *et al.* (2016), for example, gained insight into underlying biological mechanisms during protein translation in *Escherichia coli* by integrating multi-omics data.
Vertical integration is more challenging in comparison to horizontal integration. According to Ritchie *et al.* (2015a): "In particular, diversity in the size of data sets, patterns of missing data and noise across the different data types, and correspondence and correlation between measurements from different technologies can create substantial challenges".

Our R package `compendiumdb` and the accompanying MySQL database (Chapter 2) provide a flexible resource towards the horizontal integration of microarray-based functional genomics experiments across platforms and measurement modalities.

Integration stages

Data from different sources can be integrated at three different stages: (*i*) early, (*ii*) intermediate, or (*iii*) late (Figure 1.3). Which stage to choose depends, for example, on whether the original data are available and the biological question one tries to answer (Hamid *et al.*, 2009) .

Early integration consists of directly combining data either from different studies using the same data modality (horizontal) or from different modalities (vertical). The study of Sontrop *et al.* (2011) mentioned above, where ten breast cancer gene expression datasets measured on the same microarray platform were integrated, is an example of early horizontal integration. The main challenge with early integration is to identify the best approach for combining the data from different studies and platforms in a meaningful way, which requires cross-study normalization and cross-platform identifier mapping. Another challenge is the increase in number of features (vertical) after combining datasets. One way to cope with the high data dimensionality is by adding additional constraints, *e.g.*, using regularization. For example, iCluster takes an early vertical integration approach to integrative clustering of multiple genomic data modalities enforcing sparsity using a lasso penalty (Shen *et al.*, 2009).

Intermediate integration consists of first transforming the individual datasets and then combining them. For example, each individual dataset can be transformed into a similarity matrix such as a correlation matrix or a kernel matrix. Wang *et al.* (2014) vertically integrated mRNA expression, microRNA expression, and DNA methylation data for five cancer types by using similarity network fusion (SNF). SNF constructs a separate similarity network of samples for each data type and then combines these into one network. By using a network clustering algorithm cancer subtypes were identified from the fused similarity network. Advantages of intermediate integration are that data-type-specific properties are preserved and that the transformation is insensitive to different data measurement scales (Ritchie *et al.*, 2015a).

In *late integration*, a separate analysis is performed for each dataset and the final results are integrated. A typical example are meta-analytic techniques where one horizontally combines effect sizes or p-values across studies (Ramasamy *et al.*, 2008). For example, Ioannidis *et al.* (2007) integrated three type 2 diabetes genome-wide association studies. First, odds ratios (OR) were calculated for each single-nucleotide polymorphism (SNP) in each study and then a random effects model was used to combine the ORs. Another example of late integration is the ensemble-based method that Marbach *et al.* (2012) used for the inference of transcriptional gene regulatory networks. The authors integrated predictions from multiple network inference methods from a community of experts to construct consensus networks for *Escherichia coli* and *Staphylococcus aureus*. Advantages of late integration are that one can choose the most suitable method for analyzing each individual dataset and a lower impact of systematic biases between datasets.

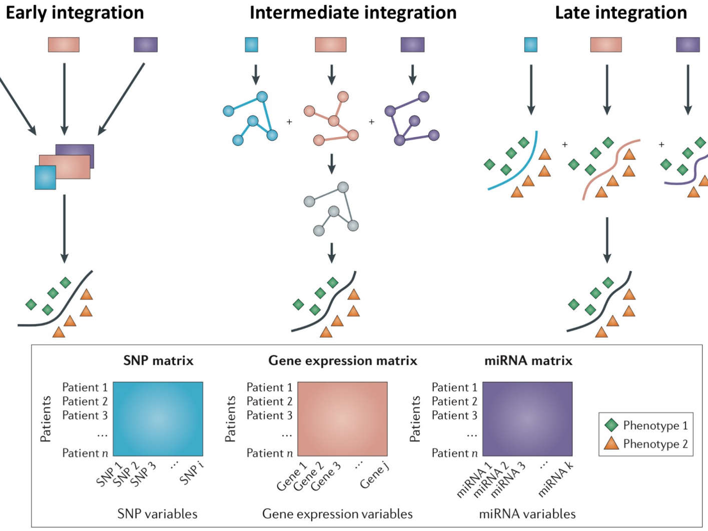

Figure 1.3: Stages of integration - vertical data integration. Data on SNPs, gene expression, and miRNA expression measured in the same patients are vertically integrated in order to classify each patient to one of two phenotypes. In the networks used for intermediate integration, a node corresponds to a patient. miRNA, microRNA; SNP, single-nucleotide polymorphism. Figure adapted from Ritchie *et al.* (2015a).

1.4 Network-based integration

Many intermediate integration approaches are network-based. Here, we give a brief overview on biological networks, mainly in the context of omics data, and then provide a detailed description of different *network-based integration* (NBI) approaches.

1.4.1 Biological networks

The coordinated actions and reactions of a set of molecules in a cell are referred to as a biological network. Different types of biological networks interact with each other, thereby defining the architecture of a cell. Each type of network reflects a different aspect of a cell's architecture. For example, *gene regulatory* and *signal transduction* networks describe how genes and proteins can be inhibited or activated, and therefore encode the steps leading to the expression of genes or proteins. *Protein-protein interaction* (PPI) networks represent interactions between proteins, whereas, *gene coexpression* networks encode the similarity of expression profiles across biological conditions between pairs of genes. *Metabolic* networks describe how metabolites are transformed via enzymatic reactions, for example to produce energy.

Biological networks are represented as a set of nodes connected to each other via edges. Here, nodes typically correspond to genes, proteins, or metabolites and edges correspond to a measured physical interaction or inferred functional link between nodes. For example, edges in a PPI network represent a physical contact between proteins in a cell measured using high-throughput techniques such as the yeast two-hybrid (Y2H) method or tandem affinity purification coupled with mass spectrometry (TAP-MS). On the other hand, edges in a gene coexpression network are inferred by calculating the pairwise similarity between gene expression profiles and do not represent physical interactions per se.

Most biological networks but also social networks, the world wide web and engineering systems are modular. In general, modules consist of physically or functionally related nodes that cooperate to achieve a specific function. Modularity in a network of interactions occurs when the network can be decomposed into components with many within-component connections and relatively few between-component connections. Biological networks display many examples of modules such as various stages of the cell cycle, nucleic acid synthesis, and DNA replication, that are governed by modularity of protein-protein interactions, protein-RNA complexes and temporally coregulated groups of genes (Barabási and Oltvai, 2004; Dong and Horvath, 2007; Hartwell *et al.*, 2002).

1.4.2 Network-based integration methods

Network modules have been successfully applied to get more insight in molecular mechanisms underlying physiology and disease phenotypes. Module-based integrative network analysis can be used to elucidate pathways and dynamic interactions underlying biological processes. NBI can be broadly grouped into four classes (Figure 1.4) related to the identification of: (*i*) active modules via the integration of a network and omics profiles, (*ii*) conserved modules across multiple species, (*iii*) differential modules between conditions and (*iv*) composite modules via the integration of different interaction types (Mitra *et al.*, 2013).

Active modules

The common principle of active module based approaches is to overlay a static biological network with omics profiles summarized by scores, for example the absolute fold changes between two conditions. Active modules are then defined as connected regions in the network which are enriched for high scoring nodes. Since the type of data used to determine the network is generally different

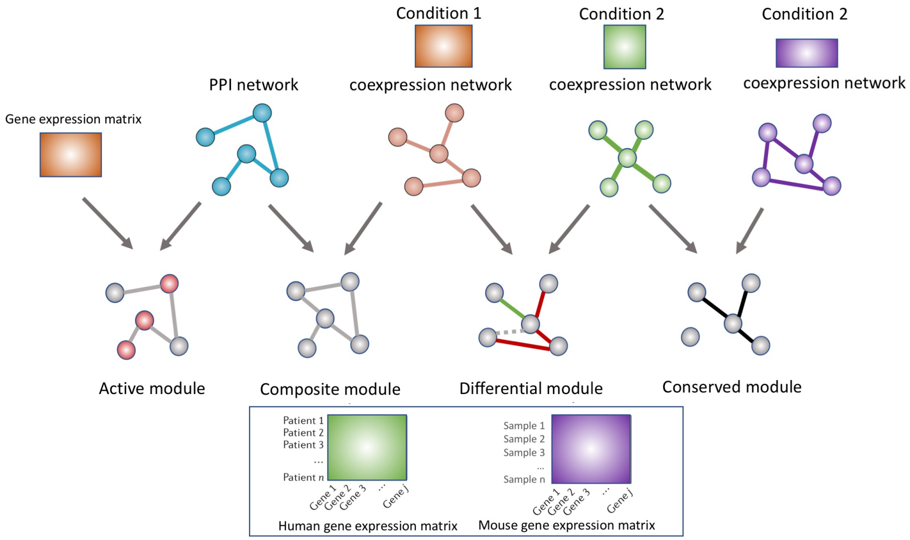

Figure 1.4: Overview of four module-based approaches for network-based integration. Human gene expression data are integrated with a human PPI network using gene-expression-based summary statistics (active module) or using a gene coexpression network (composite module). Human coexpression networks from two different conditions are integrated to detect differential modules. Human and mouse coexpression networks are integrated to detect conserved modules. PPI, protein-protein interaction.

from the omics data projected on the network, this is an example of vertical integration. The identification of active modules is relevant since it may reveal the dynamic nature of regulatory and signaling mechanisms associated with a given cellular response (Li *et al.*, 2017a). Active modules have also been referred to as responsive subnetworks or network hotspots (Begley *et al.*, 2004).

Several computational methods have been developed for the identification of active modules such as by defining this problem as an optimization problem, with the objective to identify subnetworks with high scores (Chen *et al.*, 2017; Li *et al.*, 2017b). Cowen *et al.* (2017) recently reviewed the use of network propagation methods to determine active modules. Here, a biological signal is magnified based on the observation that genes related to a similar phenotype have a tendency to interact with one another, often referred to as 'guilt by association'. Another group of approaches uses biclustering which not only clusters network interactions based on their topology, but also takes the conditions under which these interactions are active into account (Prelić *et al.*, 2006).

Conserved modules

With evolution, biological networks undergo substantial rewiring. However, basic cellular modules including the cell cycle, innate immunity, and gene regulatory interactions are often conserved across a large number of species. Therefore, by identifying conserved modules fundamental questions about core biological mechanisms can be addressed (Zinman *et al.*, 2011).

Over the past two decades, several approaches have been employed to identify conserved modules using horizontal integration. Stuart *et al.* (2003) and Bergmann *et al.* (2004), for example, used sequence similarity to associate genes from one species with orthologous genes in other species and then identified modules characterized by conserved coexpression, *i.e.*, pairs of genes whose expression profiles are similarly correlated across species. Another group of approaches uses local or global network alignment methods to identify conserved modules. These approaches in general employ a heuristic search algorithm that takes network topology into account (El-Kebir *et al.*, 2015a). Local network alignment aims to identify one or more shared subnetworks (Berg and Lässig, 2004; Flannick *et al.*, 2006; Kelley *et al.*, 2004; Liang *et al.*, 2006). By contrast, a global network alignment provides an overall comparison of complete networks by searching for an optimal one-to-one mapping between nodes, even though this may lead to suboptimal alignments in some local regions (Kollias *et al.*, 2013; Kuchaiev *et al.*, 2010; Mohammadi *et al.*, 2017).

Differential modules

Exploring how molecular networks change across conditions is of great importance for understanding the biological mechanisms underlying healthy and diseased states. Horizontal integration of networks across conditions to identify differences in their interactions and modules is referred to as differential network analysis (DiNA). DiNA complements differential expression analysis to identify changes in interactions between nodes (for example, genes) across different conditions.
The identification of differential networks often starts with the construction of condition-specific networks using correlation-based methods. Next, differential modules are determined by edge-wise subtraction of the correlations estimated for the condition-specific networks (Gambardella *et al.*, 2013; Ideker and Krogan, 2012; Mitra *et al.*, 2013). For example, Zhang *et al.* (2017) developed a node-based DiNA method where each condition-specific network is modelled as a precision matrix and the differential network as the difference between two precision matrices. Next, differential modules are determined that are driven by certain hub nodes. Also penalized statistical approaches for DiNA have been proposed. For example, the joint graphical lasso enables estimating condition-specific network models simultaneously and determining common and condition-specific interactions (Danaher *et al.*, 2014).

Composite modules

Most of the NBI approaches described above involve integration using a single type of interaction network. However, different aspects of cellular activity are represented by diverse biological interactions. Therefore, to generate a better understanding of biological interactions, vertical integration of different types of molecular interaction networks is required. Network modules that comprise more than one type of interaction are referred to as composite modules. Most methods for the detection of composite modules are extensions of module detection methods for single networks. For example, Bennett *et al.* (2015) proposed a mathematical programming model, SimMod, which takes multiple network types as input, optimises average modularity across all networks, clusters them and then returns composite modules. Detection of composite modules with SimMod was evaluated by integrating physical, genetic and gene coexpression networks in yeast.

Chapter 4 of this thesis describes a network alignment method to align a pair of gene coexpression networks generated from gene expression data measured across multiple conditions. The method is applied to gene expression data measured in human and mouse immune cell types to study conservation and divergence between the two species. In **Chapter 5** of this thesis a special case of this method is used to identify modules conserved between species for a single condition. The method is applied to gene expression data measured in human and mouse livers.

1.5 Thesis outline

Below we describe the main topics treated in the remaining chapters in this thesis.

In **Chapter 2** we address the problem of accessing and managing large volumes of transcriptomics data by creating a homogeneous framework for systematic retrieval and storage of functional genomics data that facilitates seamless integration of data from different profiling platforms. Public repositories such as GEO (Barrett *et al.*, 2013) and ArrayExpress (Rustici *et al.*, 2013) provide a large amount of functional genomics data from a wide range of studies performed in different organisms and on different platforms. However, retrieving and systematically storing these datasets to extract novel biological information is often challenging. We developed the `compendiumdb` R package to provide a homogeneous framework for constructing large compendia of functional genomics data by retrieving preprocessed GEO data from different studies and profiling platforms, and storing and maintaining these using a MySQL database.

Chapter 3 describes an integrated approach to rank candidate proteins for low-abundant spots in 2D-DIGE experiments. 2D-DIGE provides a powerful technique to separate proteins on their isoelectric point and apparent molecular mass and quantify changes in protein expression. Abundantly available proteins in spots can be identified using mass spectrometry-based approaches. However, identification is often not possible for low-abundant proteins. We present a prioritization method that generates ranked lists of candidate proteins for unidentified low-abundant (*i.e.*, only visible using fluorescence) spots from a 2D-DIGE experiment. Candidate proteins are proposed, based on the in-gel location of a spot, and resulting candidate lists are ranked, based on the strength of association of candidates with the MS-identified proteins using STRING functional association scores. The ranked list is further filtered based on gene expression data in expressed and unexpressed genes on a per sample basis. We assessed the performance of our approach on proteins differentially expressed at the peak of HIV-1 infection of T-cells. This is an example of problem-driven data integration, where information from ExPASy (Compute pI/Mw and TagIdent), STRING, the Gene Expression Barcode 3.0, and the

NIAID HIV database of (HIV-1)–human protein interactions was integrated with in-house experimental data.

In **Chapter 4** we present the results of applying a sparse global network alignment method for network-based integration across species. We use Natalie 2.0, which is a fast and robust aligner that can handle large networks with thousands of nodes and tens of thousands of edges (El-Kebir *et al.*, 2015a). Natalie 2.0 is based on an integer linear programming approach introduced by Klau (2009), which uses Lagrangian relaxation and provides an upper bound on how close a given solution to the network alignment problem is to the optimal solution. We apply Natalie 2.0 to align gene-gene coexpression networks of purified immune cell populations from various differentiation states in human (Novershtern *et al.*, 2011) and mouse (Heng *et al.*, 2008; Jojic *et al.*, 2013). Using the results obtained from the network alignment, we identify conservation and divergence at the transcriptional level in these immune cell types between the two species. We also show the advantages of using an objective function that explicitly takes conservation of expression, as introduced by Shay et al. (2013), into account to improve the network alignment and enable comparison of more than two conditions.

Chapter 5 focuses on the alignment of human and mouse liver coexpression networks. We apply Natalie 2.0 to align the networks as described in Chapter 4. Similar to the immune system in the previous chapter, we investigate conservation and divergence at the transcriptional level in human and mouse liver. In agreement with previous findings, we show that the aligned modules between human and mouse liver coexpression networks have overall similar biological functions. However, a considerable number of orthologous genes show poorly conserved coexpression suggesting functional divergence.

Chapter 6 offers a discussion of the various computational approaches for biological data integration presented in this thesis. Furthermore, we discuss related research directions and recent developments.