A systems biological perspective on complex human diseases: uncovering hidden relations via multi-scale network analyses

Mara Lena Hartsperger
Acknowledgements

This work would have not been possible without the help of various people who accompanied and strongly supported me during the last years.

First and foremost, I would like to thank Dr. Volker Stümpflen. Always open-minded for constructive discussion - also beyond pure science - he has been a mentor and leader with an ever open ear for any concern I had. I am very thankful for the trust and space I got to accomplish this thesis and his support in initiating contacts and independent research projects. Big thanks for encouraging me to see the bigger picture and promoting my personal development.

I would also like to cordially thank Prof. Hans-Werner Mewes, head of the Institute of Bioinformatics and Systems Biology (IBIS), for giving me the opportunity to do my doctorate at his institute. Besides for his professional support and ideas, I would like to thank him for enabling me numerous trips to workshops and conferences, for the trust and opportunities I got, and also for his time to chat on things not related to science.

Furthermore, I want to thank Prof. Fabian Theis for being part of my thesis committee and accompanying me on my long way towards finishing this thesis. His enthusiasm and support facilitated to put my ideas on the fuzzy clustering algorithm into practice. Without him, the Chapter on this would be missing from this thesis.

Dr. Florian Blöchl for our nice collaboration on the fuzzy clustering algorithm and its applications. Thanks also for pushing forward the laborious paper revisions and his help with any mathematical problem I had.

Jörn Leonhardt and Matthias Arnold for doing their diploma theses under my supervision. I learned a lot during that time.

Dr. Susanne Neschen and Melanie Kahle for providing experimental data, our fruitful discussions, and introducing me to the world of biologists.

Prof. Stephan Weidinger for initiating and pushing forward the GWAS disease network project. Thanks also for inviting me to workshops to present our work there. At this point, I would also like to thank Dr. Elke Rodríguez and Hansjörg Baurecht for their enormous support in manual data curation for this project.

Dr. Korbinian Grote for providing the necessary data for the gene regulatory networks. Thanks also for checking the thesis progress every once in a while.

My particular thank goes to the entire BIS group and the rest of IBIS for providing a excellent working atmosphere. Thank you all for the discussions, feedback, conference
trips, coffee breaks, and nice chats. Especially, Florian Büttner - my room mate and fellow-sufferer, Benedikt Wachinger, Dr. Michael Greeff, Daniel Ellwanger, Robert Strache, Dr. Filka Nenova, and Dr. Thorsten Barnickel.

A special thank goes also to Elisabeth Noheimer and Petra Fuhrmann for always being helpful in all respects. Thanks for the nice chats on real-life stories and the goodies in your office.

Finally, I want to thank my family and friends for being there when I needed them. Above all, I want to thank Robert for motivation when I saw no light, for his ongoing support, and his trust in me.

Thank you all.

Mara
Abstract

The identification of the underlying causes of complex human diseases represents one of the major challenges medical research is facing in the 21st century. Complex diseases, including obesity, mental health disorders, diabetes, and cancers, place a major health burden and the emerging field of personalized medicine presents a new paradigm in health care and the treatment of these diseases. In the last decade, systems biology has been established as a successful approach towards personalized medicine as it aims at finding mechanistic explanations of the relationship between a systems components, their interactions, and its physiological state. By considering complex systems from different perspectives, systems biology allows revealing novel insights into the molecular basis of diseases and therefore contributes to a more dynamic comprehension. For reaching this, not only algorithms capable to analyze complex systems and to cope with the ever increasing flood of information are necessary. Also methods are required that allow to find missing pieces – often the needle in the haystack – in the complicated puzzle of understanding complex disease.

The objective of this thesis was to address both by approaching complex systems via multi-scale network analyses and thereby allowing focusing on them from different perspectives in a systems-biological fashion. On a global scale, methods were developed that allow to structure networks and thus to better understand complex systems by decomposing them into interrelated components. Focusing on the interrelations within networks light was shed on disease relations in general. Concentrating on small scale interactions led to causative explanations and novel hypotheses, and accordingly, towards a better understanding of diseases on a systems-level. In a nutshell, starting from a general, static, and descriptive view of the global organization, switching to the local level allowed to head towards a more specific, dynamic, and predictive perspective of complex diseases. By resulting in experimentally testable hypotheses about potential causative explanations, both perspectives together represent pieces in the puzzle and strategies to solve it. Consequently, they allow to narrow down the gap between genotype and phenotype.
Zusammenfassung


Im Laufe der letzten Jahre etablierte sich die Systembiologie als ein sehr erfolgreicher Ansatz für die personalisierte Medizin. Durch die ganzheitliche Betrachtung biologischer Systeme – auch aus unterschiedlichen Perspektiven – ermöglicht es die Systembiologie neue Erkenntnisse über molekulare Ursachen von Erkrankungen zu gewinnen. Auf der theoretischen Seite sind hierzu Algorithmen erforderlich, die nicht nur in der Lage sind komplexe Systeme analysieren zu können, sondern die auch die zunehmende Datenflut bewältigen können. Zudem bedarf es geeigneter Methoden, welche Informationen liefern, die für ein besseres Verständnis komplexer Erkrankungen erforderlich sind.

Contents

1 Introduction ................................................. 1
  1.1 Motivation ............................................. 1
  1.2 Objective .............................................. 2
  1.3 Outline ................................................ 4

2 Preliminaries .................................................. 9
  2.1 Biological networks .................................... 9
    2.1.1 Basic definitions in graph theory ................. 9
    2.1.2 Network models ................................... 12
    2.1.3 Random vs. scale-free networks: robustness and small-world property .... 14
    2.1.4 Topology and modularity in scale-free networks .... 15
  2.2 Community detection .................................. 15
    2.2.1 Basic definitions .................................. 16
    2.2.2 Clustering methods ................................. 16
    2.2.3 Non-negative matrix factorization ................. 19
    2.2.4 Detecting overlapping communities ................ 20
  2.3 Constraint-based modeling of metabolomic systems ........ 21
    2.3.1 The constraint-based modeling approach .......... 23
    2.3.2 Genome-scale metabolic network reconstruction ... 24
    2.3.3 Mathematical definition of the genome-scale model and constraints .... 24
    2.3.4 Optimization theory and objective function ....... 26
  2.4 Genome-wide association studies and complex human diseases ....... 28
    2.4.1 Genetic mapping in human disease ................ 28
    2.4.2 The haplotypic structure of the genome .......... 29
    2.4.3 GWAS design: searching for association .......... 30
    2.4.4 Shortcomings of GWAS ............................ 32
  2.5 MicroRNAs and complex human diseases .................. 34
    2.5.1 Overview of microRNA biogenesis ................. 34
    2.5.2 Regulation of microRNA gene transcription ......... 37
    2.5.3 MicroRNAs in human disease ...................... 38
    2.5.4 The future of microRNA research ................... 38
## Contents

3  Insights into the structural organization of biological data via fuzzy clustering of $k$-partite graphs  
3.1  Modular decomposition by graph clustering  
3.2  A NMF-type community detection algorithm  
3.2.1  Graph approximation  
3.2.2  Derivation of the update rules  
3.2.3  Complexity analysis  
3.3  Algorithm evaluation  
3.3.1  Performance analysis  
3.3.2  Stability of clusters  
3.3.3  The cluster structure dependence on $m$  
3.4  Decomposition of a gene-disease-protein complex graph  
3.4.1  Parameter determination  
3.4.2  Large-scale clustering  
3.4.3  Small-scale clustering  
3.5  Summary and future perspectives  

4  The hierarchical organization of regulatory networks  
4.1  Extracting the hierarchical organization of regulatory networks  
4.1.1  Algorithm formulation  
4.1.2  Algorithm evaluation  
4.2  Hierarchical organization of mammalian TRNs  
4.2.1  Network construction  
4.2.2  General properties of TRNs  
4.2.3  Master-regulators are more likely to be essential  
4.2.4  Higher level TFs have more functions  
4.2.5  Master-regulators function as triggers in signaling cascades  
4.2.6  MicroRNAs act differently: from switches to fine-tuning  
4.2.7  Tissue-dependent TF expression correlates with their hierarchical position  
4.2.8  TFs and human diseases  
4.3  Summary and future perspectives  

5  Uncovering the role of pleiotropic genes via multi-scale disease network analyses  
5.1  Network medicine towards GWAS  
5.2  Data acquisition and preprocessing  
5.3  Analysis of GWAS-based disease networks  
5.3.1  The human GWAS disease network  
5.3.2  The human GWAS diseasome  
5.3.3  SNP- and allele-based sub-analysis  
5.4  Multi-scale analysis of widely studied traits  
5.4.1  The diseasome of widely studied traits  

ii
<table>
<thead>
<tr>
<th>5.4.2 Pathway-based view on diseases</th>
<th>117</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.3 Uncovering disease correlation from a SNP perspective</td>
<td>118</td>
</tr>
<tr>
<td>5.5 Summary and future perspectives</td>
<td>120</td>
</tr>
<tr>
<td>6 Analysis of the role of microRNAs in disease progression</td>
<td>123</td>
</tr>
<tr>
<td>6.1 Accessing the role of miRSNPs in complex human traits</td>
<td>126</td>
</tr>
<tr>
<td>6.1.1 GWAS-SNPs have no impact in microRNA biogenesis</td>
<td>126</td>
</tr>
<tr>
<td>6.1.2 GWAS-SNPs tend to be located in the 3'-UTR</td>
<td>127</td>
</tr>
<tr>
<td>6.1.3 In-depth investigation of 3'-UTR associated GWAS-SNPs</td>
<td>130</td>
</tr>
<tr>
<td>6.2 The impact of miRSNPs - a systems biological perspective on impaired microRNA-regulation</td>
<td>134</td>
</tr>
<tr>
<td>6.2.1 SMC4 - miRSNP links primary biliary cirrhosis and cancer</td>
<td>135</td>
</tr>
<tr>
<td>6.2.2 HMGCR - the role of microRNAs in LDL cholesterol related traits</td>
<td>138</td>
</tr>
<tr>
<td>6.3 Summary and future perspectives</td>
<td>142</td>
</tr>
<tr>
<td>7 Systems biology of metabotypes</td>
<td>145</td>
</tr>
<tr>
<td>7.1 Non-alcoholic fatty liver disease and its cross-links</td>
<td>146</td>
</tr>
<tr>
<td>7.2 Network-based prediction of metabolic behavior in NAFLD</td>
<td>147</td>
</tr>
<tr>
<td>7.2.1 Study design and experimental setup</td>
<td>147</td>
</tr>
<tr>
<td>7.2.2 In silico approach to model metabolic behavior</td>
<td>148</td>
</tr>
<tr>
<td>7.2.3 Metabolic activity in NAFLD</td>
<td>151</td>
</tr>
<tr>
<td>7.2.4 Interpretation of metabolic activity profiles</td>
<td>151</td>
</tr>
<tr>
<td>7.3 Large-scale analysis – global trends of disease specific metabolism</td>
<td>154</td>
</tr>
<tr>
<td>7.3.1 Metabolic activity decreases due to HFD</td>
<td>154</td>
</tr>
<tr>
<td>7.3.2 Reduced metabolic activity is accompanied by reduced lipid accumulation in hepatocytes</td>
<td>154</td>
</tr>
<tr>
<td>7.3.3 Relation between progression in NAFLD and metabolic flux predictions</td>
<td>157</td>
</tr>
<tr>
<td>7.3.4 Affected enzymes show strong bias to lipogenic processes and disorders in hepatic lipid homeostasis</td>
<td>158</td>
</tr>
<tr>
<td>7.4 Mid-scale analysis – changes in hepatic lipid metabolism</td>
<td>159</td>
</tr>
<tr>
<td>7.4.1 Metabolic trends in lipogenic processes</td>
<td>160</td>
</tr>
<tr>
<td>7.4.2 Metabolic trends in hepatic fatty acid oxidation</td>
<td>163</td>
</tr>
<tr>
<td>7.5 Small-scale analysis – from observations to systems biological models</td>
<td>166</td>
</tr>
<tr>
<td>7.5.1 Downregulation of Ppap2a improves hepatic steatosis and IR</td>
<td>166</td>
</tr>
<tr>
<td>7.5.2 Ppara and Fgf21 mediated regulation of hepatic fatty acid oxidation</td>
<td>171</td>
</tr>
<tr>
<td>7.5.3 MicroRNA mediated control of Foxo1 improves the effects of IR</td>
<td>175</td>
</tr>
<tr>
<td>7.6 Summary and future perspectives</td>
<td>179</td>
</tr>
<tr>
<td>8 Conclusion and outlook</td>
<td>183</td>
</tr>
</tbody>
</table>

Abbreviations
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>193</td>
</tr>
<tr>
<td>List of Tables</td>
<td>197</td>
</tr>
<tr>
<td>Bibliography</td>
<td>199</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

1.1 Motivation

“The behavior of most complex systems, from the cell to the Internet, emerges from the orchestrated activity of many components that interact with each other. At a highly abstract level, the components can be reduced to a series of nodes that are connected to each other by links, with each link representing the interaction between two components. The nodes and links together form a network, or, in more formal mathematical language, a graph.”[1] In other words, by using this representation network science deals with complexity by simplifying complex systems [2].

An important realization of the past decade has been the fact that networks[1] are not random, but follow a series of basic organizing principles in their structure and evolution that distinguish them from randomly linked networks [3]. Structure[2], topology[3], network usage, robustness and function are deeply interconnected [1]. Consequently, an in-depth understanding of this still partially puzzling interplay is essential for a thorough understanding of complex systems (see Figure 1.1).

The study of complex biological systems results in a tremendous amount of data that allows to view at the pathogenesis of human diseases in a different way as one is looking more on the interaction and dynamics of biological systems as at the individual components of biological processes [6]: trying to understanding diseases in the context of network principles allows to address some fundamental properties of the components (e.g. genes) that are involved in disease [3]. The central point in this network medicine approach is the think globally, act locally paradigm [3]: the heart of understanding, e.g., a disease on a systems-level is a detailed understanding of a systems global organization. This is best addressed by the intensely multidisciplinary approach known as systems biology. Here, diseases are considered not in a static way, but as dynamic processes. These operate on widely different scales – both in space and time – between biological states which are

---

1 Complex systems are simplified as networks. In the following, we use both terms as synonyms.
2 Structure is the architecture of a network, meaning its logical design [1].
3 Topology is the shape or layout of the network, meaning the physical elements required to enable the architectural design [4].
Figure 1.1: From the particular to the universal. The bottom of the pyramid shows the traditional representation of the cell’s functional organization: genome, transcriptome, proteome, and metabolome (level 1). There is remarkable integration of the various layers both at the regulatory and the structural level. Insights into the logic of cellular organization can be achieved when we view the cell as a complex network in which the components are connected by functional links. At the lowest level, these components form genetic-regulatory motifs or metabolic pathways (level 2), which in turn are the building blocks of functional modules (level 3). These modules are nested, generating a scale-free hierarchical architecture (level 4) [5] (see Section 2.1.2). The different levels that are addressed within this thesis correspond, on the one hand, to the vertical axis in the pyramid (Figure 1.1), where a system is analyzed on different resolution/granularity levels (see each Chapter). On the other hand, they correspond to the horizontal axis that reflects the elements of a cell’s functional organization and their potential perturbations: genes/TFs, proteins, miRNA, metabolites, and SNP (see according Chapters). Illustration adapted from Oltvai and Barabási [5].

1.2 Objective

The ultimate goal in systems biology is to find mechanistic explanations of the relationship between a systems components, their interactions, and its physiological state (phenotype).
In other words, the aim of systems biology is to bridge and finally close the gap between genotype\(^5\) and phenotype\(^6\). An important feature of systems biology is thereby a clear distinction between the so-called macro- and microscopic view. The macroscopic view treats the system as a whole, while the microscopic view focuses on small-scale mechanistic components of a system. Using the pyramidal illustration from before, this corresponds to the large-scale organization on the top and the single components in the lower levels, respectively (Figure 1.1). From this it becomes apparent that both views are tightly interwoven and thus highly interdependent as interaction and coordination are taking place within and between levels of organization. Not only interdependencies at one organizational level, but between different ones are of interest as the whole is to some degree constrained by the parts (upward causation) and at the same time the parts are to some degree constrained by the whole (downward causation).\(^7\) Consequently, as one level of explanation will hardly suffice, the challenge lies in reconciling both perspectives. From both layers much knowledge is inferable that often complements one another. Combining both perspectives in a multi-level approach represents a way of not only bridging between micro- and macro-level constructs, but also to significantly amplify rather than duplicate the knowledge. In keeping with the motto: one plus one equals more than two. Generic bioinformatics concepts on this, however, hardly exist and current multi-level approaches still struggle to cope with the flood of information.

In this thesis, a multi-level concept is presented that combines the macro- and microscopic view. This is achieved by developing appropriate analysis methods that enable to gradually and dynamically investigate the system on different levels.\(^8\) Starting from the macroscopic view, which concentrates on the topology and architecture of networks, and by developing network analysis methods, networks can be structured even if they are of vast complexity. The decomposition of networks reveals novel insights and reflects a static and descriptive perspective on their global organization. Thereby, it contributes to an overall understanding of the system. Focusing on disease relations and disease causations in more detail, the global organization is gradually broken down into smaller pieces. Subsequently, the system can be analyzed from a microscopic view, whereby - using different layers - one heads towards a hypothesis-generating perspective that allows for a more dynamic and predictive perspective and finally for a better understanding of human diseases. In the end, pieces in the puzzle are received that are of different size and consequently, this multi-level concept represents a small step towards bridging the gap.

\(^5\) Genotype is the genetic constitution (the genome) of a cell, an individual or an organism. It is distinct from its expressed features, or phenotype [10].

\(^6\) The phenotype results from the interaction between the genotype and the environment. It is the composite of the characteristics shown by the cell, individual or organism under a particular set of environmental conditions [10].


\(^8\) Different levels correspond, on the one hand, to the vertical axis in the pyramid (Figure 1.1) where a system is analyzed on different resolution granularity levels. On the other hand, they correspond to the horizontal axis that reflects the elements of a cell’s functional organization and their potential perturbations: genes [TF], proteins, mRNA, metabolites, and SNP.
1 Introduction

1.3 Outline

The goal of this thesis is to reveal novel insights into complex human diseases using a multi-level concept. Starting with the development of network analysis methods and data integration, disease relations can be uncovered from a macroscopic perspective. Going into more detail, the concentration lies on the microscopic perspective. This focus leads to causative explanations and novel hypotheses, and accordingly, towards a better understanding of diseases on a systems-level. In a nutshell, starting from a general, static, and descriptive view of the global organization, using experimental data I head towards a more specific, dynamic, predictive, and hypothesis-generating direction. For an abstract illustration on this see Figure 1.2.

At this point, it has to be mentioned that this is achieved by successively focusing on this goal. To simplify matters and to contribute to a better understanding, each Chapter of this thesis encapsulates one of the consecutive steps towards capturing the whole, and therefore is intended to be as self-contained as possible. In the following, an outline on the organization of this thesis is given.

In the Preliminaries the necessary theoretical and biological background is provided (Chapter 2). First, the concept of biological networks is presented by introducing basic definitions from graph theory and giving details on the properties of different network models and their inherent topology. Next, an overview of community detection is given that is used to identify and interpret the internal network organization. Furthermore, the conceptional basis of constraint-based modeling (CBM) and genome-wide association studies (GWAS) is introduced. Finally, an overview of microRNA (miRNA) biogenesis and function is given.

As mentioned before, a key question is how to analyze and interpret the internal organization of (biological) networks. A crucial step when studying the structure and dynamics of networks is to identify overlapping communities, but methods addressing the common biological case of $k$-partite graphs – biological networks are complex and highly diverse – are faintly available. In Chapter 3 we develop a novel fuzzy clustering algorithm based on non-negative matrix factorization (NMF). This algorithm clusters each node type of the graph separately and then connects clusters via a smaller, weighted $k$-partite graph in an alternating minimization procedure. After presenting the idea of graph approximation, we derive and validate the fuzzy clustering algorithm. Next, I decompose a tripartite disease-gene-protein complex graph on different resolution levels. First, I structure this complex graph into biologically meaningful clusters on a large-scale, which corresponds to the macroscopic view. Finally, focusing on the small-scale architecture, I exemplify how overlapping clusters allow for reclassification, annotation, or even detection of misclassified elements on a local level reflecting the microscopic view.

Although clustering analysis provides insights into the correlation among, e.g., genes and biological phenomena, it does not reveal the interdependencies of regulatory rela-

---

9 Chapter 3 led to two publications, namely, Hartsperger et al. [11] and Blöchi et al. [12].
In this thesis, a multi-level concept is presented that combines the macro- and microscopic view. This is achieved by developing appropriate analysis methods that enable to gradually and dynamically investigate the system on different levels. (a) Heterogeneous - experimental and computational - information is integrated, thereby allowing the construction of large-scale networks. The large-scale organization (b) is analyzed on a macroscopic level by developing network analysis methods (c), e.g., the decomposition of a network into clusters (indicated by the zoom). Focusing on the microscopic perspective (d), small clusters are detected that allow a further analysis from a microscopic perspective, or, systems biological hypotheses can be derived that have to be tested experimentally in the next round within the systems biological cycle (indicated by the zoom).

Figure 1.2: Illustration of the thesis outline. In this thesis, a multi-level concept is presented that combines the macro- and microscopic view. This is achieved by developing appropriate analysis methods that enable to gradually and dynamically investigate the system on different levels. (a) Heterogeneous - experimental and computational - information is integrated, thereby allowing the construction of large-scale networks. The large-scale organization (b) is analyzed on a macroscopic level by developing network analysis methods (c), e.g., the decomposition of a network into clusters (indicated by the zoom). Focusing on the microscopic perspective (d), small clusters are detected that allow a further analysis from a microscopic perspective, or, systems biological hypotheses can be derived that have to be tested experimentally in the next round within the systems biological cycle (indicated by the zoom).

In this thesis, a multi-level concept is presented that combines the macro- and microscopic view. This is achieved by developing appropriate analysis methods that enable to gradually and dynamically investigate the system on different levels. (a) Heterogeneous - experimental and computational - information is integrated, thereby allowing the construction of large-scale networks. The large-scale organization (b) is analyzed on a macroscopic level by developing network analysis methods (c), e.g., the decomposition of a network into clusters (indicated by the zoom). Focusing on the microscopic perspective (d), small clusters are detected that allow a further analysis from a microscopic perspective, or, systems biological hypotheses can be derived that have to be tested experimentally in the next round within the systems biological cycle (indicated by the zoom).

Figure 1.2: Illustration of the thesis outline. In this thesis, a multi-level concept is presented that combines the macro- and microscopic view. This is achieved by developing appropriate analysis methods that enable to gradually and dynamically investigate the system on different levels. (a) Heterogeneous - experimental and computational - information is integrated, thereby allowing the construction of large-scale networks. The large-scale organization (b) is analyzed on a macroscopic level by developing network analysis methods (c), e.g., the decomposition of a network into clusters (indicated by the zoom). Focusing on the microscopic perspective (d), small clusters are detected that allow a further analysis from a microscopic perspective, or, systems biological hypotheses can be derived that have to be tested experimentally in the next round within the systems biological cycle (indicated by the zoom).

In this thesis, a multi-level concept is presented that combines the macro- and microscopic view. This is achieved by developing appropriate analysis methods that enable to gradually and dynamically investigate the system on different levels. (a) Heterogeneous - experimental and computational - information is integrated, thereby allowing the construction of large-scale networks. The large-scale organization (b) is analyzed on a macroscopic level by developing network analysis methods (c), e.g., the decomposition of a network into clusters (indicated by the zoom). Focusing on the microscopic perspective (d), small clusters are detected that allow a further analysis from a microscopic perspective, or, systems biological hypotheses can be derived that have to be tested experimentally in the next round within the systems biological cycle (indicated by the zoom).
First, I introduce HiNO, differentiate it from existing methods and evaluate it using gene regulatory networks (GRNs) from *Saccharomyces cerevisiae* and *Escherichia coli*. Next, I apply HiNO to the GRNs available for *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. From a macroscopic perspective, I resolve the GRNs’ global organization by analyzing their general properties. By systematically investigating the properties of the individual layers, I retrieve novel insights into - among other things - disease associations from a microscopic perspective.

Most of the current systems-level approaches to study diseases focus on a single disease, relying on network-based methods to gain insights on the molecules and pathways relevant for the specific disease. A conceptually different approach has been proposed by Goh et al. [17] in order to study the pleiotropic relationships between different human diseases, leading, for example, to comorbidity, instead of focusing the attention on one disease [6]. To account for this, the focus lies on the single interactions within the underlying disease network in Chapter 3 after having analyzed its overall structure in Chapter 3 [12]. Besides a systematic analysis of the role of pleiotropic genes in disease networks, I present how diseases themselves are cross-connected. To elucidate disease relationships and key genetic loci and pathways, a systematic multi-scale analysis of common human diseases based on published genome-wide association data is performed. By yielding both biologically suspected as well as unexpected disease-disease relationships, I show that combining and correlating data from GWAS on multiple scales allows for a better insight into complex human diseases and their interrelations on single nucleotide polymorphism (SNP), gene, and even pathway level. Based on the results of this multi-level approach many candidate loci are provided that represent potentials to experimentally explore them and thus also the genomic architecture of complex diseases in more detail.

Despite efforts to understand the interrelations between different diseases based on cellular network characteristics, and in order to understand the genotype–phenotype relationships in diseases, the modeling of disease-causing mutation effects on molecular networks is required on a systems-level [6]. This is the focus of Chapter 6 where we investigate the impact of impaired miRNA-mediated regulation from the standpoint of population genetics. Starting with analyzing the global impact of miRSNPs (for a definition see Chapter 6), we look on disease causation from a macroscopic perspective. By contextualizing the results from a systems biological perspective, we show that miRSNPs are of high relevance and demonstrate that they have the potential to be used in disease prognosis and diagnosis. Finally, on the microscopic view, we end up with small-scale models that lead to hypotheses giving explanations about disease causations and representing suggestions about further experiments [13].

Besides genetic predisposition, environmental factors such as lifestyle and diet are supposed to play a major role in the onset of diseases, which are not accounted for by

---

12 Chapter 3 led to the following manuscript: Hartsperger et al. [18]
13 Basis for this Chapter was the diploma thesis of Matthias Arnold [19]. This Chapter led to the following manuscript: Arnold et al. [20].
GWAS. For developing appropriate therapies for their cure an in-depth understanding on disease development and its underlying causal mechanisms is essential. However, in many cases the detailed connections between genotype, phenotype, lifestyle, diet and (impaired) metabolism are not well understood yet. The metabolic state of a system represents the most sensitive level of organization. Consequently, the study of metabolic phenotypes (metabotypes) represents a promising field capable to detect and investigate a disease’s underlying causations. In Chapter 7 a systems biological perspective on metabotypes is presented by systematically analyzing the impact of a high fat diet (HFD) in the context of non-alcoholic fatty liver disease (NAFLD) induction and progression. Based on a CBM approach originally presented by Shlomi et al. [22], normal and abnormal metabolism in response to the diet can be studied in a time-resolved manner. Starting from a macroscopic perspective, significant changes of the overall metabolism in response to the HFD are identified. Zooming in, the focus lies on specific metabolic categories, which have turned out to undergo most significant changes, that are then analyzed on a more detailed level. In the end, we generate comprehensive systems biological models, which reflect our observations from a microscopic perspective. Again, they lead to novel hypotheses giving potential causative explanations from a microscopic perspective, representing promising candidates for experiments.

The last Chapter 8 concludes the thesis, summarizes its main contributions, and gives an outlook.

---

14 Basis for this Chapter was the diploma thesis of Jörn Leonhardt [21].
CHAPTER 2

Preliminaries

In this Chapter, the basic theoretical and biological concepts necessary for a full understanding of this thesis are introduced. First, the concept of biological networks (Section 2.1) is presented and an overview of community detection (Section 2.2) is given. Next, the conceptional basis of constraint-based modeling (CBM) (Section 2.3) and genome-wide association studies (GWAS) (Section 2.4) is introduced. Finally, a general review of microRNA (miRNA) biogenesis and function (Section 2.5) is given.

2.1 Biological networks

"The behavior of most complex systems, from the cell to the Internet, emerges from the orchestral activity of many components that interact with each other through pairwise interactions. At a highly abstract level, the components can be reduced to a series of nodes that are connected to each other by links, with each link representing the interactions between two components. The nodes and links together form a network, or, in more formal mathematical language, a graph." [1]

In the following, an introduction of the basic definitions in graph theory is given and different network models and their inherent properties are presented. For a more detailed overview on networks and graph theory see also the reviews of Barabási et al. [1] and Fortunato [23].

2.1.1 Basic definitions in graph theory

A graph $G$ is a pair of sets $(V,E)$, where $V$ is a set of vertices or nodes and $E$ is a subset of $V^2$, the set of unordered pairs of elements of $V$. The elements of $E$ are called edges or links. The number of vertices in a graph determines its size and is indicated with $n$.

Depending on the nature of the interactions, a graph can be either directed or undirected (Figure 2.1). In directed graphs, the interaction between any two nodes has thus a well-defined direction that represents, e.g., the direction of information flow from a transcription factor (TF) to the gene it regulates. In this case an ordered pair $(i,j)$ is an edge directed from vertex $i$ to vertex $j$ [23]. In undirected networks, the links do not have an assigned
2 Preliminaries

A matrix $A_{ij}$ with elements

\[
A_{ij} = \begin{cases} 
1 & \text{if there is an edge from } i \text{ to } j, \\
0 & \text{otherwise.}
\end{cases}
\]

is an $n \times n$ matrix that is symmetric for undirected graphs. If edges are weighted, one additionally defines the weight matrix $W$ whose element $W_{ij}$ expresses the weight of the edge between vertices $i$ and $j$ [23].

2.1.1.1 Degree and degree distribution

Two vertices are neighbors (or adjacent) if they are connected by an edge (Figure 2.1). The degree (or connectivity) $k_i$ of a vertex $i$ is the number of its neighbors and represents the most elementary characteristic of a node. On directed graphs, one distinguishes two types of degree on node $i$:

1. The in-degree $k_{in}$, which denotes the number of its predecessors, and
2. the out-degree $k_{out}$, which indicates the number of its successors.

The degree distribution $P(k)$ represents an important network measure. It gives the probability that a randomly selected node has exactly degree $k$:

\[
P(k) = \frac{n_k}{n},
\]

where $n_k$ is the number of nodes having degree $k$. 

---

**Figure 2.1: A sample graph with eight vertices and ten edges.** (a) is an undirected network with node $A$ (red node) having degree $k = 5$, whereas (b) is a directed network with $k_{in} = 4$ and $k_{out} = 1$ for node $A$. Only one pair of node $A$’s five neighbors in part (b) of the figure is linked together ($B$ and $C$, green edges). This gives $n_A = 1$ and thus a clustering coefficient of $C_A = 0.1$. Besides the graphical representation, also the corresponding adjacency matrices are shown [1]. Illustration adapted from Barabási et al. [1].

direction [1], e.g., the interaction between proteins, and thus $(i,j) = (j,i)$. An edge $(i,j)$ is called self-loop if $i = j$.

Besides the widely used graphical representation, the complete information about a graph’s topology can be represented by an adjacency matrix $A$ with elements

\[
A_{ij} = \begin{cases} 
1 & \text{if there is an edge from } i \text{ to } j, \\
0 & \text{otherwise.}
\end{cases}
\]

$A$ is an $n \times n$ matrix that is symmetric for undirected graphs. If edges are weighted, one additionally defines the weight matrix $W$ whose element $W_{ij}$ expresses the weight of the edge between vertices $i$ and $j$ [23].

---
2.1 Biological networks

2.1.1.2 Clustering coefficient

The *clustering coefficient* of a vertex in a graph quantifies to which extent its neighbors are inter-connected as well and is defined as

$$c_i = \frac{2n_i}{k_i(k_i - 1)},$$

(2.2)

where $n_i$ is the number of links connecting the $k_i$ neighbors of node $i$ to each other. In other words, it gives the ratio between the number of edges joining pairs of neighbors of $i$ and the total number of possible edges [24]. According to this definition, $c_i$ measures the probability that a pair of neighbors of $i$ is connected and thus how close they are to being a clique.\(^{15}\)

The clustering coefficient distribution $C(k)$ represents an important measure of the network's structure as it characterizes the overall tendency of nodes to form clusters or groups. The clustering coefficient $< C >$ for the whole network is given by Watts and Strogatz [24] as the average of the clustering coefficients of all vertices:

$$< C > = \frac{1}{n} \sum_{i \in V} c_i$$

(2.3)

2.1.1.3 Shortest path and connectivity

A *path* is an alternating sequence of vertices such that from each of its vertices there is an edge to the next in the sequence. The *length* $l$ of a path is the number of edges that it uses. Paths allow to define the concept of *connectivity* and *distance* in graphs. A graph is *connected* if, given any pair of vertices, there is at least one path going from one vertex to the other. In general, there may be multiple paths connecting two vertices with different lengths. A *shortest path* between two vertices of a graph is a path of minimal length. Such minimal length is the *distance* between the two vertices. The *diameter* of a connected graph is the maximal distance between two vertices.

If there is no path between two vertices, the graph is divided in at least two connected subgraphs. Each maximal connected subgraph of a graph is called *connected component*. A graph is *strongly connected* if every vertex is reachable from every other following the directions of the edges. On the contrary, a graph is *weakly connected* if its underlying undirected graph is connected. A weakly connected graph can be thought of as a graph in which every vertex is reachable from every other but not necessarily following the directions of the edges.

2.1.1.4 Graph partiteness

A graph $G$ is *bipartite* if the vertex set $V$ is separated in two disjoint subsets $V_1$ and $V_2$ and every edge joins a vertex of $V_1$ with a vertex of $V_2$. The definition can be extended to that

\(^{15}\) Clique is a subset of a graph’s vertices such that every two vertices in the subset are connected by an edge [29].
of $k$-partition\(^\text{16}\), where there are $k$ vertex classes (also called type or color) and no edge joins vertices within the same class (Figure 3.1). In this case one speaks of multi-partite graphs \[^{22}\]. In the following, the term partition comprises all vertices of one class. For more details see also Section 3.2.

Multi-partite graphs can be reduced to unipartite projections of each vertex class. For instance, from a bipartite network of TFs and their target genes one can extract a network of target genes only that are linked by co-regulation.

2.1.2 Network models

Any empirical investigation starts with the question whether observed phenomena have emerged by chance or whether those observations are due to random effects \[^{2}\].

In network science, the previously mentioned characteristics are used to distinguish whether the underlying network is either random or structured, meaning that the network follows certain characteristics. Average degree $< k >$, average path length $< l >$, and average clustering coefficient $< C >$ all depend on the number of nodes and links in the network. In contrast, the degree distribution $P(k)$ and cluster coefficient distribution $C(k)$ functions are independent of the network’s size. Consequently, capturing a network’s generic features, they can be used to classify a network \[^{1}\].

There exist three network types that are inherently different and had a direct impact on our understanding of biological networks: random networks, scale-free networks and hierarchical networks (Figure 2.2).

2.1.2.1 Random networks

The Erdös-Rényi (ER) model of a random network \[^{26}\] constructs networks by connecting nodes completely at random (Figure 2.2.1). Starting with a fixed number of $n$ nodes, each pair of nodes is connected with equal probability $p$. This creates a graph with approximately $pn(n - 1)/2$ randomly placed links. The node degrees follow a Poisson distribution, which indicates that most nodes have roughly the same degree (close to the average degree $< k >$). The tail of the degree distribution $P(k)$ decreases exponentially, showing that nodes that significantly deviate from the average (highly connected nodes, so-called hubs) are extremely rare. As the clustering coefficient is independent of a node’s degree, the cluster coefficient distribution $C(k)$ appears as a horizontal line if plotted as a function of $k$ \[^{1}\].

2.1.2.2 Scale-free networks

In contrast, many real networks are scale-free (Figure 2.2.2). Here, most nodes have only a few interactions that coexist with a few highly connected nodes, the so-called hubs that hold the whole network together. Scale-free networks are characterized by a power-law degree distribution

\[^{16}\) Here $k$ describes the number of disjoint sets, not the degree.\]
2.1 Biological networks

Network models are crucial for shaping our understanding of complex networks. There are three different models that allow us to distinguish whether a network is random or structured: random, scale-free and hierarchical network model. Besides examples how the different network models look like (1-3a), their corresponding degree distributions \( P(k) \) (1-3b) and clustering coefficient distributions \( C(k) \) (1-3c) are shown. Whereas in random networks the degree distribution follows a Poisson distribution, it follows a power law distribution in scale-free and hierarchical networks due the existence of hubs, which are indicated by blue nodes in (2a) [1]. Illustration taken from Barabási et al. [1].

\[
P(k) \propto k^{-\gamma},
\]

where \( \gamma \) is the degree exponent.\(^{17}\) In scale-free networks, the probability that a node is highly connected is statistically more significant than in a random graph, thereby the network’s properties are mainly determined by only a relatively small number of highly connected hub nodes [1]. Barabási and Albert [27] identified two central mechanisms that

\(^{17}\) The value of the degree exponent \( \gamma \) determines many properties of the system: a smaller the value of \( \gamma \) corresponds that hubs have a more important the role in the network [1].
jointly led to the appearance of the hub hierarchy exemplifying the scale-free structure:

- **Growth**, meaning that the number of nodes is not fixed, and
- **Preferential attachment**. A node with more links increases its connectivity faster than nodes with fewer links, since incoming nodes tend to connect to it with higher probability. This is also known as the *rich-get-richer* positive-feedback phenomenon [1, 27].

In the Barabási-Albert (BA) model of a scale-free network [27], at each time point a node with \( m \) links is added to the network, which connects to an already existing node \( i \) with probability

\[
\prod(i) = \frac{k_i}{\sum_{j=1}^{n} k_j},
\]

where \( k_i \) is the degree of node \( i \) and \( j \) is the index denoting the sum over network nodes.

Whereas the degree distribution \( P(k) \) follows a straight line on a log-log plot, the network that is created by the BA model does not have an inherent *modularity*, thus \( C(k) \) is independent of \( k \).

### 2.1.2.3 Hierarchical networks

Biological networks have an inherent hierarchical architecture (Figures 1.1 and 2.2.3): sparsely connected nodes are part of highly clustered areas and only a few hubs maintain the communication between the different highly clustered neighborhoods [1]. As the human body is composed by organs, organs are composed by tissues, tissues by cells, and so forth, it represents a perfect, paradigmatic example of hierarchical organization [23].

Hierarchical networks are characterized by the coexistence of a scale-free topology and *modularity*, an iterative combination of clusters – groups of physically or functionally linked molecules (nodes) that work together to achieve a (relatively) distinct function [1, 28, 29]. The hierarchical network model shows a power-law degree distribution and a large, system-size independent average clustering coefficient \( <C> \). It therefore seamlessly integrates a scale-free topology with an inherent modular structure. Consequently, the scaling of the clustering coefficient, which follows \( C(k) \propto k^{-1} \) a straight line of slope \(-1\) on a log-log plot represents the most important signature of hierarchical modularity [1].

### 2.1.3 Random vs. scale-free networks: robustness and small-world property

A common feature of all complex networks is that they have a relatively small *diameter*, even if the number of nodes is large. This fact is commonly referred to as the *small-world effect*, meaning that any two nodes can be connected with a relatively short path of a few links only. The best known example is *six degrees of separation* found by Milgram [30] who showed that an average number of six acquaintances links two arbitrary people from the United States. As a consequence, the perturbation of a given node can affect the activity of most nodes in their vicinity and also have a more or less drastic influence on the behavior of the network itself. To quantify this, the average path length \( <l> \) is used. In a random
2.2 Community detection

The network is $<l> \propto \log n$, indicating that it is characterized by the small-world property. In contrast, scale-free networks are even ultra-small with $<l> \propto \log \log n$.

Another important feature of scale-free networks is their robustness due to the existence of hubs. The robustness of a network is defined as the ability of its nodes to communicate being unaffected even by unrealistically high failure rates, e.g., disease, deregulation or mutation. It makes a system stable against random failures and thus enables the system to maintain its functionalities against external and internal perturbations. Logically, the chance that accidental failure affects one of the few hubs is relatively low. But the price of being tolerant to errors is quite high as these networks are extremely vulnerable (or sensitive) to directed attack, which is defined as the targeted selection and removal of a few nodes (hubs) that play a vital role in maintaining the network's connectivity.

In the end, attack on those might bring a system to collapse.

2.1.4 Topology and modularity in scale-free networks

As described before, hierarchical networks are characterized by a scale-free topology and an inherent modularity. These topological modules are often believed to carry specific cellular functions, hence leading to the concept of functional modules, an aggregation of nodes of similar or related function in the same network neighborhood. To address the modularity of networks, it is necessary (i) to develop appropriate tools and measures allowing to establish if a network is modular at all and if so (ii) to identify inherent modules and their relationships. For more details see also Chapters 3 and 4.

"High clustering indicates that networks are locally sprinkled with various subgraphs of highly interlinked groups of nodes, which is a condition for the emergence of isolated functional modules. Subgraphs capture specific patterns of interconnections that characterize a given network at the local level. Some subgraphs, which are known as motifs, are overrepresented when compared to a randomized version of the same network. Motifs constitute the basic building blocks of cellular networks and they are likely to be associated with some optimized biological function. The most important motifs are: feed-forward loops (FFLs), feedback loop (FBL) or bifans (see Figure 2.9). For more details on this see the publications of Shen-Orr et al. and Alon.

2.2 Community detection

From the previous Section it became clear that real-world networks exhibit an inhomogeneous distribution of edges among vertices, creating hot spots of heightened connectivity. These modules, which are densely intra- and weakly inter-connected, are relatively independent compartments of the network that account for its form and function as a system. Modularity in biological and other empirical networks can stem from two sources: (i) it can arise as a byproduct of other, more fundamental, topological properties such as the degree distribution or degree correlations (the dependence of a node’s degree on
its neighbors’ degrees), or (ii) it can be generated directly by some inherent property or mechanism within the system, e.g., the friends of my friends tend to become my friends in social networks \[36\]. The identification of these high order structures in networks unveils not only insights into their functional organization \[37\], but brings also significant advances to our understanding of complex systems \[23\]. A key question in network analysis is therefore how to detect and interpret this inherent internal organization.

Identifying modules and their boundaries allows for a classification of vertices according to their structural position in the modules. So, vertices with a central position in their clusters, i.e., sharing a large number of edges with the other group partners, may have an important function of control and stability within the group. Vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities\[23\]. A well-established approach on this is modular decomposition, which implies the coexistence of structural sub-units associated with more highly interconnected parts. The identification of these a priori unknown building blocks (e.g., the identification of functional modules in protein-protein interaction (PPI) networks), which are essential for understanding functional properties, is known as community detection or (graph) clustering. The goal of community detection in graphs is therefore to identify (i) the modules and also (ii) their hierarchical organization by only using the information encoded in the graph topology \[23\].

In this Section, basic definitions, different forms of community detection methods, and ways to evaluate their stability are introduced. For more details on this see also the corresponding original publications and the recent review of Fortunato \[23\].

2.2.1 Basic definitions

Communities, also called clusters or modules, are groups of vertices\[18\], which probably share common properties and/or play similar roles within the graph. Clustering is a form of unsupervised learning, meaning that no prior knowledge exists on any object classifications \[23\, 38\]. It allows to partition data into clusters such that objects with similar characteristics\[19\] are clustered together and dissimilar objects are in different clusters (Figure 2.3).

2.2.2 Clustering methods

There exists a range of different approaches towards community detection. Traditional community finding techniques are hierarchical and partitional clustering, where vertices are joined into groups according to their mutual similarity \[23\]. Whereas hierarchical algorithms successively identify clusters using previously established ones (Section 2.2.2.1), partitional clustering attempts to directly decompose the data set into a set of clusters (Section 2.2.2.2).

---

18 In the following we use the terms vertices, data points and objects as synonyms.
19 Similarity is defined as the distance between two objects. The distance measure or criterion might depend on the problem or the data used. For more details see Section 2.2.2.1.
2.2 Community detection

Figure 2.3: A simple graph with four communities. Clustering allows to partition data points into communities, such that objects with similar characteristics are clustered together and dissimilar objects are in different clusters. Similarity is usually defined as the distance between two data points. Here, the network is decomposed into four communities indicated by different colors: purple, blue, green, and yellow. In real-world networks communities often overlap, e.g., a protein may have different functions or in a social network a person may have distinct roles according to its environment. For details see Section 2.2.4. Here, the yellow community overlaps with the blue one in a single node, whereas it shares two nodes and a link with the green one. These overlapping regions are emphasized in red. Illustration taken from Palla et al. [39].

2.2.2.1 Hierarchical clustering

Hierarchical clustering algorithms aim at identifying groups of vertices with high similarity and group them over a variety of scales by creating a so-called cluster tree \(^{20}\) or dendrogram (see Figures 3.12, 4.15, or 5.11). The tree is not a single set of clusters but rather a multilevel hierarchy where clusters at one level are joined as clusters at the next level. Horizontal slices of the tree at a given level then indicate the communities that exist above and below a certain threshold. Hierarchical clustering techniques can be classified in two categories:

- **Agglomerative algorithms**: clusters are iteratively merged if their similarity is sufficiently high. In this bottom-up approach one starts from the vertices as separate clusters (singletons) and ends up with the graph as a unique cluster.

- **Divisive algorithms**: clusters are iteratively split by removing edges connecting vertices with low similarity. These methods are top-down as they follow the opposite direction.

The starting point of any hierarchical clustering method is the definition of a similarity measure between two data points \(x_1, x_2 \in \mathbb{R}^n\). Commonly used (dis-)similarity measures are the Euclidean distance, where

---

\(^{20}\) A tree is an undirected graph in which any two vertices are connected by exactly one simple path. In other words, any connected graph without cycles is a tree [23].
the maximum distance or the Manhattan distance. Furthermore, distances based on the Pearson's (Equation 2.7) or Spearman’s correlation coefficients are also often used to quantify similarity.

\[ d(x_1, x_2) = 1 - \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} , \tag{2.7} \]

where \( \bar{x}_y = \frac{1}{n} \sum_{i=1}^{n} x_{yi} \) with \( y \in [1, 2] \) is the mean.

The linkage criterion then specifies the dissimilarity of sets using the previously defined distance measure in different ways:

- **Single linkage** where the distance between two clusters is the distance between their two closest members.
- **Complete linkage** where the distance between two clusters is defined as the distance between their two farthest members.
- **Average linkage** where the distance between two clusters is calculated based on the average values using all elements of each cluster.

Hierarchical clustering has the advantage that it does not require a preliminary knowledge on the number and size of the clusters.

### 2.2.2.2 Partitional clustering

In contrast, in partitional clustering methods the number of clusters is preassigned. This means that they have to be specified in advance and are not determined by the algorithm itself. The goal is to separate the data points \((x_1, x_2, \ldots, x_n)\) in \(m\) disjoint clusters \(C = \{C_1, \ldots, C_m\}\) such that the given cost function is minimized. Here, this is the total intra-cluster distance, or squared error function

\[ f(C) = \sum_{j=1}^{m} \sum_{x_i \in C_j} \|x_i - \mathbf{v}_j\|^2 , \tag{2.8} \]

where \(C_j\) indicates the subset of points of the \(j\)-th cluster and \(\mathbf{v}_j\) its centroid that represents the cluster’s center

\[ \mathbf{v}_j = \frac{\sum_{x_i \in C_j} x_i}{\sum_{x_i \in C_j}} , \]

with \(j = 1, \ldots, m\). The minimization is achieved by optimizing the given cost function in an iterative manner, starting with randomly chosen centroids such that they are as far as possible from each other. Then, each data point is assigned to the nearest centroid, and new cluster centers can be estimated. This new set of centroids allows for a new
2.2 Community detection

classification of the observations, and so on. The positions of the centroids are stable after a small number of iterations and the algorithm has converged.

In real-world data, however, there is often no sharp boundary between clusters as data points may belong to different groups simultaneously. Therefore, fuzzy clustering techniques are better suited as they allow a point to belong to two or more clusters at the same time by introducing continuous degrees of membership between zero and one, instead of crisp assignment where every object is assigned to exactly one cluster. These can be aggregated in a membership matrix \( C \), where \( C_{ij} \) quantifies the assignment of point \( i \) to cluster \( j \). The matrix \( C \) is normalized to be right-stochastic, i.e., the degrees of membership of each data point sum up to one.

A popular approach, similar to \( k \)-means clustering, is fuzzy \( k \)-means clustering \([40, 41]\). It is based on the minimization of the associated cost function

\[
f(C) = \sum_{j=1}^{m} \sum_{i=1}^{n} (C_{ij})^{f} \|x_i - \upsilon_j\|^2 ,
\]

where additionally a fuzzification factor \( f \geq 1 \) is included. This is commonly done to extend cost functions in data clustering to include fuzzy clusters \([40, 41]\). When \( f \) is close to 1, the cluster center closest to the observation is given much more weight than the others and the algorithm is similar to \( k \)-means.

Besides their dependency on random initialization, for both methods the number of clusters to be identified has to be specified before. For more details see Fortunato \([23]\) and Chapter 3.

2.2.3 Non-negative matrix factorization

High-throughput "-omics" technologies result in large amounts of biological data requiring analysis and interpretation. Dimensionality reduction and visualization are key aspects in effectively analyzing and interpreting these high-dimensional data \([42]\). Non-negative matrix factorization (NMF), first introduced by Lee and Seung \([43, 44]\), is an unsupervised, parts-based learning paradigm in which a non-negative matrix \( V \) is decomposed into two non-negative matrices \( W \) and \( H \) such that

\[
V \propto WH
\]

by a multiplicative updates algorithm \([32, 45]\). By not allowing negative entries, NMF enables a purely additive combination of parts that together reconstruct the original data. The most prominent example is the parts-based decomposition of face images from Lee and Seung \([43]\), or their usage for deriving the semantic features of texts to cluster related documents.

To find an approximative factorization for the matrix \( V \), cost functions that quantify the quality of the approximation need to be defined. Such a cost function can be constructed using some measure of the divergence between \( V \) and \( WH \). The decomposition is reached
by minimizing the cost function defined by the choice of the distance measure. In case of using the squared error (or \textit{Frobenius norm}) as divergence function this means that given a matrix $V$ find nonnegative matrices $W$ and $H$ that minimize the function

$$f(W,H) = \|V - WH\|^2_F,$$  

where $\| \cdot \|^2_F$ denotes the squared \textit{Frobenius norm}. Starting with random initial values for $W$ and $H$, the algorithm simultaneously updates these matrices via multiplicative rules until convergence to a local minimum is attained \cite{Lee2001}. For more details see Lee and Seung \cite{Lee2001} and Chapter 3 where the key ideas of NMF are adapted to develop a fuzzy graph-partitioning algorithm \cite{Lee2001}.

\subsection*{2.2.4 Detecting overlapping communities}

In real networks, vertices may have different functions/roles and thus belong to more than one group at the same time. Therefore, approaches are required that allow for the detection of overlapping communities. Besides methods such as fuzzy k-means clustering, where the overlap is represented as a degree of membership, approaches have been developed that assign a vertex to different clusters simultaneously with equal weight.

The most popular technique to find overlapping communities is the \textit{clique percolation method (CPM)} developed by Palla et al. \cite{Palla2005}. It is based on the concept that the internal edges of a community are likely to form $k$-cliques due to their high density and the fact it is unlikely that inter-community edges form cliques \cite{Palla2005, Fortunato2006}. For details see Palla et al. \cite{Palla2005} and Fortunato \cite{Fortunato2006}.

\subsection*{2.2.5 Comparing clusters: measures}

Checking the performance of an algorithm involves the definition of a criterion that establishes how similar or robust the clusters delivered by the algorithm are compared to the (optimal) clusters one wishes to recover. Especially in cases of non-deterministic\footnote{Determinism is the concept that events within a given paradigm are bound by causality in such a way that any state is completely, or at least to some large degree, determined by prior states. Mathematical models that are not deterministic because they involve randomness are called stochastic.} algorithms – such as $k$-means or NMF type-based methods – it is crucial to understand the stability of the cluster assignments towards the random initialization. To compare the clustering results of different runs - characterized by different random initializations - and to quantify their stability, an appropriate similarity measure (or stability score) has to be used. Several measures for the similarity of clusters exist.

Garge et al. \cite{Garge2014} proposed to use Cramer’s $v^2$ as a measure for the degree of replicability. Cramer’s $v^2$ makes use of $\chi^2$ statistics and quantifies the degree of association in contingency tables larger than $2 \times 2$ reflecting the frequency distribution of data classification (or cluster assignment) by two systems simultaneously. Cramer’s $v^2$ is the squared canonical correlation between two sets of nominal variables that define the rows and columns of
the contingency table. It indicates the proportion of variance in one classification scheme (or reference clustering result) that can be predicted by the other classification scheme (or clustering result obtained by a different initialization).

\[
Cramer's \, \nu^2 = \frac{\chi^2}{n(m-1)},
\]

where \(\chi^2\) is the ordinary \(\chi^2\) test statistic for independency in contingency tables, \(n\) the number of items cross classified (i.e., the total number of elements to be clustered), and \(m\) the number of clusters extracted. The index ranges from 0 to 1, with 0 indicating no relationship and 1 indicating a perfect reproducibility.

To use Cramer’s \(\nu^2\) in the case of fuzzy clustering, crisp assignment of data points to clusters – every data point is assigned to the cluster showing maximum degree of membership – is required. As the process of crisp assignment may affect the stability score and an important feature of several methods is the detection of overlapping clusters, a different similarity measure, the so-called fuzzy rand index (FRI) can be used [48].

The FRI allows to compare fuzzy clusters and therefore to determine how close the fuzzy memberships obtained in different runs are. For details see Equations 2.13 and 2.14. Let \(\mathbf{C}\) be the matrix of the degrees of membership of a fuzzy clustering of a data set. A fuzzy equivalence relation on \(\mathbf{C}\) is defined in terms of a similarity measure on the associated degrees of membership vectors. Generally, this relation can be defined via any distance measure on \([0,1]^m\) that yields values in \([0,1]\). The maximum norm is employed and the distance \(d_{\mathbf{C}}(x_r, x_s)\) between the degrees of memberships of two data points \(x_r, x_s\) is defined as

\[
d_{\mathbf{C}}(x_r, x_s) := \max_t |C_{st} - C_{rt}|.
\]

Now, given the fuzzy cluster assignments \(\mathbf{C}\) and \(\mathbf{C}'\) resulting from two different random initializations the FRI can be calculated. The FRI is defined as the degree of concordance:

\[
FRI(\mathbf{C}, \mathbf{C}') = 1 - \frac{\sum_{r<s} |d_{\mathbf{C}}(x_r, x_s) - d_{\mathbf{C}'}(x_r, x_s)|}{n(n-1)/2},
\]

where \(n\) is the number of data points. It ranges from 0 to 1, with 0 indicating no relationship and 1 indicating a perfect reproducibility [46]. For more detailed information on this measure see the original publication by Hüllermeier and Rifqi [48].

2.3 Constraint-based modeling of metabolomic systems

With the continuous development of ever better high-throughput technologies, various “-omics” datasets are becoming available in ever increasing sizes. In the two previous sections, first the idea of biological networks a concept that is widely used to represent complex systems was presented (Section 2.1). Furthermore, a short introduction in community detection, which is used to systematically analyze them in a qualitative manner,
was given (Section 2.2). Now, an alternative approach oriented more towards quantitative network analysis is presented.

Metabolomics is considered to provide a direct functional readout of the physiological state of an organism [49]. Changes in the concentrations of enzymes have been shown to have only small effects on metabolic pathways, but significant impact on the concentrations of metabolic intermediates [50–52]. The metabolic state of a system thus represents a more sensitive level of organization than information from genomics, transcriptomics, or proteomics (Figure 1.1). Modern techniques in metabolomics make it possible to precisely quantify the metabolites present in biological systems [53, 54]. Then, metabolic networks allow to systematically study the (dynamic) behavior of a complex system in response to diverse stimuli, and thus represent a powerful tool for studying and modeling metabolism. The global metabolic network comprising all metabolic pathways is illustrated in Figure 2.4.

![Global map of metabolic pathways](http://www.genome.jp/kegg/pathway/map/map01100.html)

**Figure 2.4: Global map of metabolic pathways.** All metabolic pathways (reference pathways) from KEGG are shown [55]. The colors indicate different pathway categories, e.g., blue corresponds to carbohydrate metabolism and red to nucleotide metabolism. For more details and the original illustration see [http://www.genome.jp/kegg/pathway/map/map01100.html](http://www.genome.jp/kegg/pathway/map/map01100.html).

Many mathematical modeling approaches are currently being used to model cellular processes and the metabolic behavior, respectively. In addition to stochastic [56, 57] and

---

22 A metabolic network is the complete set of metabolic and physical processes – driven by chemistry and mass action – that determine the physiological and biochemical properties of a cell. These networks comprise the chemical reactions of metabolism as well as their underlying regulatory interactions.

23 From a biochemical point of view, metabolic pathways are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by a series of chemical reactions and numerous distinct pathways co-exist within a cell. This collection of pathways is called the metabolic network. Although pathways - which are important to maintenance homeostasis within an organism - themselves are not existing in this way in the cell, they rather represent a virtual concept that simplifies complexity by pooling subsequent, functionally linked chemical reactions.
cybernetic [58, 59], especially kinetic models [60, 61] are well established in the metabolic context. The latter describe the metabolic state of a cell and its variation over time by metabolite concentrations and reaction rates that can be viewed as the endpoints of metabolic operation. Each enzymatic reaction is characterized as a complex function of metabolite concentrations and enzymatic activity that – as a result of transcriptional and metabolic regulation – vary over time [449].

Although these methods provide useful results, it is currently difficult to use them to model networks on a genome-scale as a large number of parameters is required resulting in high computational complexity. Genome-scale models, however, are essential for a systematic analysis of metabolic behavior and a full understanding. To date, genome-scale models of metabolism have only been analyzed with constraint-based modeling approaches [62].

2.3.1 The constraint-based modeling approach

The classical starting point of CBM is flux balance analysis (FBA) of metabolic networks at steady state (see also Section 2.3.3). The flux is a the mathematical representation of a given metabolic reaction and represents the amount of substrate that is converted to a product within a unit of time [1]. By allowing the flux for each reaction to be calculated, metabolic FBA have recently significantly improved the ability to make quantifiable predictions on the relative importance of various reactions which in the end represents the basis to experimentally testable hypotheses [63, 64].

The constraint-based model (CBM) approach facilitates integration of experimental data of various types and from disparate sources while increasing the accuracy in its quantitative prediction of biological hypotheses formulated in terms of objective functions (Figure 2.5). Constraint-based modeling requires no a priori knowledge (or assumptions) regarding all of the mechanisms and parameters for a given system. But when a priori knowledge exists, such as data on enzyme kinetics or measurement of in situ concentrations and fluxes, this knowledge can be introduced in the form of constraints, equatable with molecular genetic observations on the topology and information flow in a biochemical network [65].

There exist four different types of constraints limiting cellular functions: fundamental physico-chemical, spatial or topological constraints, condition-dependent environmental constraints, and regulatory or self-imposed constraints [66].

- **Physico-chemical constraints** are inviolable and provide ‘hard’ constraints on cell function. Mass, energy and momentum must be conserved.
- **Topobiological constraints** (or three-dimensional) affect both the form and function of biological systems.

---

24 Genome-scale is used to describe these models because all of the metabolic reactions that could be determined to take place in an organism based on genome annotation and biochemical literature are included in the model [62].
o **Environmental constraints** on cells are time and condition dependent (e.g. nutrition, pH, temperature, ...).

o **Regulatory constraints** are self-imposed and subject to evolutionary change. They allow the cell to confine itself to behaviors of increased fitness and implemented in various ways, including the amount of gene products made (transcriptional and translational regulation) and their activity (enzyme regulation) [66].

In contrast to, e.g., PPI networks, the main constraints in metabolic networks are chemistry and mass action. In the following, the focus lies therefore only on physico-chemical constraints.

### 2.3.2 Genome-scale metabolic network reconstruction

The essential basis for any CBM approach is the underlying metabolic network. The reconstruction of metabolic pathways [25] has been a strong focus of biochemistry and due to modern genome-sequencing techniques, pathway reconstructions have been increasingly integrated into genome-scale metabolic models [67]. During the past decade, genome-scale metabolic network reconstructions (GENREs) have been built for a variety of single- and multi-cellular organisms, from bacteria to human [67, 68]. Several of these networks are available online, e.g., at the BiGG [26] repository [69].

The reconstruction involves a four-step process [67] and breaks down metabolic pathways into their respective reactions and enzymes, and analyzes them within the perspective of the entire network. In simplified terms, a reconstruction starts with collecting all of the relevant metabolic information of an organism and then compiles it in a way appropriate for various types of analyses to be performed. For more details see the recent review of Oberhardt et al. [67].

### 2.3.3 Mathematical definition of the genome-scale model and constraints

After a GENRE has been defined, all metabolic reactions have to be represented in a mathematical way fulfilling the central assumption and essential constraint of all CBM approaches:

1. All biochemical reactions follow *mass-balance*, and
2. the system is in *steady-state*, which means that metabolite concentrations in the system stay constant over time.

Metabolic reactions are [therefore] represented as a stoichiometric [27] matrix $S$ of size $m \times n$, where $m$ is the number of compounds/metabolites and $n$ is the number

---

25 For a definition see Footnote 23.
26 BiGG: Biochemical Genetic and Genomic knowledge base of large-scale metabolic reconstructions
27 Stoichiometry rests upon the law of conservation of mass, the law of definite proportions (i.e., the law of constant composition) and the law of multiple proportions. In general, chemical reactions combine in definite ratios of chemicals. Due to the nature of chemical reactions, the amount of each element must be the same throughout the overall reaction.
of reactions in the system.] Every row of this matrix represents one unique compound (...), and every column represents one reaction (...). The entries in each column are the stoichiometric coefficients of the metabolites participating in a reaction. Negative coefficients correspond to metabolites that are consumed, positive coefficients to metabolites that are produced. For every metabolite that does not participate in a particular reaction, a stoichiometric coefficient of zero is used. As most biochemical reactions involve only a few different metabolites, S is a sparse matrix. The flux through all of the reactions in a network is represented by the vector v of length n. The concentrations of all metabolites are represented by the vector x with length m.\footnote{After the recognition and definition of constraints - imposed by the stoichiometries, they need to be described mathematically. Once in a mathematical form, they can be used to perform an \textit{in silico} analysis, e.g., to predict optimal states such as growth and ATP production.}

Constraints can generally be classified as either \textit{balances} or \textit{bounds}. Balances are constraints that are associated with conserved quantities, such as energy, mass, redox potential, and momentum, as well as with phenomena such as solvent capacity, electroneutrality, and osmotic pressure. Bounds are constraints that limit numerical ranges of individual variables and parameters such as concentrations, which must always be non-negative fluxes, or kinetic constants. (...) At steady-state, there is no accumulation or depletion of metabolites in a metabolic network, so the rate of production of each metabolite in the network must equal its rate of consumption. This balance of fluxes can be represented mathematically as equality constraint:

$$\frac{dx}{dt} = Sv = 0.$$ \hfill (2.15)

In addition to that, thermodynamic and enzyme capacity constraints are imposed on the CBM, regarding the directionality of and maximal metabolic flux through single reactions. This is achieved by the definition of finite lower and upper bounds that base on the knowledge of cellular thermodynamics or experimental measurements.\footnote{The null space of the $m \times n$ matrix S is the collection of those vectors in $\mathbb{R}^n$ that S maps to the zero vector in $\mathbb{R}^m$. More precisely, $N(S) = \{x \in \mathbb{R}^n | Sv = 0\}$.} These can be stated as follows:

$$v_{\text{min}} \leq v_i \leq v_{\text{max}},$$ \hfill (2.16)

where $v_i$ is the actual flux through the reaction, $v_{\text{min}}$ denotes the reaction's lower bound (0 for irreversible reactions), and $v_{\text{max}}$ is the corresponding upper bound. Both, the upper and lower bound, are based on enzyme capacity measurements. Any $v$ that satisfies these equations is said to be in the \textit{null space} of S (Figure 2.5).
2.3.4 Optimization theory and objective function

Taken together, the allowable functional states of reconstructed networks are limited by both, bound and balance constraints. With the imposition of appropriate constraints on the reactions in the GENRE, including their exact stoichiometry and reversibility, a so-called genome-scale model (GEM) is formulated. A GEM reflects allowable network states (or in a less abstract way the phenotypes of a cell) by defining a range of permissible solutions that are consistent with its mathematical representation [62]. In mathematical terms, the range of allowable network states is described by a solution space (or null space) that represents the phenotypic potential of an organism [62, 73]. All allowable network states are contained in this solution space.

In any realistic large-scale metabolic model, one finds more reactions than there are compounds \((n > m)\). This means that there are more unknown variables than equations and consequently there is no unique solution to this system of equations.

Although a range of solutions is defined by the constraints, it is still possible to identify and analyze the desired optimal solution [70]. "Therefore, in CBM the objective functions play a crucial role. [Although all of the metabolic fluxes in the solution space are generally feasible by the model, not all of them are biologically meaningful.] A given objective function can thus be thought of as a mathematical formulation of a working hypothesis for the function of particular cell or cellular system." [65] "Flux balance analysis seeks to maximize or minimize the given objective function \(f(Z) = c^T v\) \text{ (2.17)}

which can be any linear combination of fluxes, where \(c\) is a vector of weights indicating how much each reaction contributes to the objective function. Optimization of such a system is accomplished by linear programming \([\text{LP}]\) \([74, 75]\). Flux balance analysis can thus be defined as the use of linear programming to solve the equation \(Sv = 0\), given a set of upper and lower bounds on \(v\) and a linear combination of fluxes as an objective function. \([70]\) The maximization of the chosen objective function - that is a single network state (in the form of a flux distribution) - represents the result of the linear optimization \([70]\). "Interestingly, for genome-scale networks in particular, there can be multiple network states or flux distributions with the same optimal value of the objective function. Therefore, the need for enumerating alternate optima arises. The number of such alternate optima varies depending on the size of the metabolic network, the chosen objective function, and the environmental conditions \([76, 77, 65]\). For exhaustively enumerating all alternate optima, a recursive mixed-integer LP (MILP) algorithm has been developed \([77]\)."
2.3 Constraint-based modeling of metabolomic systems

Figure 2.5: The conceptional basis of constraint-based modeling. (a) A metabolic network reconstruction consists of a list of stoichiometrically balanced biochemical reactions. (b) This reconstruction is converted into a mathematical model by forming a matrix $S$, in which each row represents a metabolite and each column represents a reaction. (c) With no constraints, the flux distribution of a biological network may lie at any point in a solution space. (d) When mass balance constraints imposed by (i) the stoichiometric matrix $S$ and (ii) capacity constraints imposed by the lower and upper bounds ($v_{\text{min}}$ and $v_{\text{max}}$) are applied to a network, it defines an allowable solution space. The network may acquire any flux distribution within this space, but points outside this space are denied by the constraints. (e) Through optimization of an objective function $Z = c^T v$, FBA can identify a single optimal flux distribution that lies on the edge of the allowable solution space. Illustration adapted from Orth et al.
2 Preliminaries

2.4 Genome-wide association studies and complex human diseases

Network-based approaches have made it possible to link multiple genes to a single disease as well as to connect multiple diseases to one another by way of the genes are associated with them. Genome-wide association studies (GWAS) play a large role in unraveling these complex relationships [78] and thereby allow to identify the relevant information required for the construction of these networks of diseases. After the theoretical introduction in network science (Section 2.1) and concepts of qualitative and quantitative approaches to analyze them (Section 2.2 and 2.3), in this Section the basics of GWAS are introduced. For more details there are many excellent reviews available [79–82].

2.4.1 Genetic mapping in human disease

In contrast to monogenic disorders (also Mendelian disorders\footnote{Mendelian disorder is a genetic disease showing a mendelian pattern of inheritance, caused by a single mutation in the structure of DNA, which causes a single basic defect with pathologic consequences [83].}), complex traits are caused by many genetic and environmental factors that work together, whereat each of those has only a relatively small effect\footnote{Effect size is defined as the magnitude of the inferred effect of one variable on another. The effect size of a SNP is the difference in phenotype between genotypes with and without one of the nucleotides [81].}, and only a few if any are absolutely required for the occurrence of a disease [82, 85].

To understand complex diseases and to develop appropriate therapies, not only their inheritability\footnote{Heritability is the proportion of the observed phenotypic variation that is attributable to genetic variation [81].}, but rather the determination of genetic predisposition to disease development is of great concern. "Genome-wide association studies – in which hundreds of thousands of common variants\footnote{Classically defined as having a minor allele frequency (MAF) of > 1% [80]. [MAF] refers to the frequency at which the less common allele occurs in a given population.} are tested for association with a disease in hundreds or thousands of persons [(Figure 2.7)] – have revolutionized the search for genetic influences on complex traits:"\footnote{See also illustration in McCarthy et al. [85].} Instead of mapping disease genes by tracing transmission in families – as it is done in linkage mapping\footnote{Genes are mapped by typing genetic markers in families to identify regions that are associated with disease or trait values within pedigrees more often than are expected by chance [79].} – one might localize them through association studies\footnote{A genetic variant is genotyped in a population for which phenotypic information is available (such as disease occurrence, or a range of different trait values). If a correlation is observed between genotype and phenotype, there is said to be an association between the variant and the disease or trait [79].} where frequencies of genetic variants among affected and unaffected individuals are compared [80].

The basis of GWAS was created by the definition of the common disease - common variant (CD/CV) hypothesis. Here, genetic influences on susceptibility to common diseases are attributable to a limited number of common variants present in more than 1% to 5%
of the population that individually have little effect, but in concert confer a high risk for a certain disease [86, 87].

Each disease-causing mutation arises on a particular copy of the human genome and bears a specific set of common alleles in cis at nearby loci termed haplotype.36 Because recombination rate is low (...), disease alleles in the population typically show association with nearby marker alleles for many generations, a phenomenon termed linkage disequilibrium37 (LD).38 For details see Figure 2.6.

The GWAS approach relies on the foundation of data produced by the International haplotype map (HapMap) Project38 [88, 89] and the fact that genetic variance at one locus can predict with high probability genetic variance at an adjacent locus. „[The] (...) haplotypic structure of the human genome [...] the vast majority of common SNP’s are strongly correlated to one or more nearby proxies (Section 2.4.2) [...] means that it is possible to survey the genome for common variability associated with the risk of disease simply by genotyping approximately 500,000 [well-chosen] (...) markers in the genome of several thousand case subjects and control subject. Consequently, it is now routine to identify common, low-risk variants (i.e., those that are present in more than 5% of the population) that confer a small risk of disease.“39 In these studies, two groups of participants are compared: people with the disease (cases) and similar people without (controls). If genetic variations are more frequent (measured by odds ratio39) in cases, the variations are said to be associated with the disease. The associated genetic variations are then considered as pointers to the region of the human genome where the disease-causing problem is likely to be located.

When it comes to the search for disease-associated mutations one distinguishes between hypothesis-driven and non-hypothesis driven methods. Hypothesis-driven methods start with the hypothesis that a particular gene may be associated with a particular disease and the goal is to find the association. Non-hypothesis-driven studies scan the entire genome in order to see which of those genes demonstrate an association. GWAS are generally non-hypothesis-driven [91, 92].

2.4.2 The haplotypic structure of the genome

The HapMap project revealed not only a great amount of common variants, but yielded the awareness that those variants are not independently distributed over the genome but rather fall into blocks of variation (Figure 2.6) [89]. The SNP’s located within such blocks were found to occur in more or less perfect correlation to each other, i.e., the genotype of one SNP predicts those of correlated neighboring SNP’s. The goodness of correlation is

---

36 Haplotype is a set of DNA variations, or polymorphisms, that tend to be inherited together [79, 82].
37 Linkage disequilibrium (LD) is the correlation between nearby variants such that the alleles at neighboring markers (observed on the same chromosome) are associated within a population more often than if they were unlinked [79].
38 HapMap is a catalog of common genetic variation in humans compiled by an international partnership of scientists and funding agencies. Its goal was to characterize SNP frequencies and local linkage disequilibrium (LD) pattern across the genome in different human populations [80].
39 A measure of relative risk that is usually estimated from case-control studies [79].
tested by calculating the pairwise LD of SNPs: the higher the LD is, the more two SNPs correlate [93, 94].

When considering two biallelic loci, one of the simplest measures of the degree of non-random association is to calculate the difference between the observed and expected frequencies of a haplotype. It is defined as

$$D_{AB} = p_{AB} - p_A p_B ,$$

where $p_{AB}$ is the frequency of $AB$ gametes and $p_X$ is the frequency of allele $X \in \{A,B,a,b\}$ at locus $X$. If the LD parameter $D$ significantly differs from zero, then it may be said that the LD exists. Therefore, it seems to be more intuitive to use $D$ as a measure of independent association between markers in the two loci. The value 0 is corresponding to linkage equilibrium and 1 is considered as perfect LD. $R^2$ is known as the measure of correlation of alleles at two loci and defined as

$$R^2 = \frac{D^2}{p_A p_B p_a p_b} ,$$

where the nominator is the square of the LD parameter $D$ and the denominator is the product of the four allele frequencies. The range of $R^2$ is between 0 and 1, with $R^2 = 1$ being known as perfect LD and occurring if exactly two of the four possible haplotypes existed [95, 96].

2.4.3 GWAS design: searching for association

Population-based association analyses, resulting in the suggestion of a genome-wide LD mapping [97, 98], facilitate the detection of those variants with greater statistical power leading to robust associations between specific chromosomal loci and complex human diseases. In the following, the typical GWAS design is described (Figure 2.7).

First, a suitable disease/trait has to be selected, which has to be sensitively and specifically diagnosable or measurable in order to yield reliable results (see also Chapter 5). The fact that GWAS are generally aimed at the identification of SNPs featuring relatively small effects – odds ratios (OR) range approximately from 1.1 to 1.5 [99] – thousands of samples from case and control subjects are required to detect such variants [100]. As a great number of hypotheses are tested in GWAS, large sample sizes are essential to reach robust findings. Stringent thresholds for statistical significance are needed to avoid false positives due to this multiple hypothesis testing. However, sample sizes and genotyping numbers rapidly exceed affordability when it comes to the identification of variants with more moderate effects with adequate statistical power. The consequence of this are more relaxed p-values resulting in a far less extent of sample sizes for a study. Simulations indicated that a dense genome-wide scan of common variants involves the equivalent of

---

40 Each SNP tested in a GWAS is hypothesized to be associated with the considered trait, resulting in 500,000 hypotheses when using genotyping assays with sufficient coverage of the genome [79].
2.4 Genome-wide association studies and complex human diseases

Figure 2.6: DNA sequence variation in the human genome. (a) „Common and rare genetic variation in 10 individuals, carrying 20 distinct copies of the human genome. The 12 common variations include 10 SNPs, an insertiondeletion polymorphism (indel), and a tetranucleotide repeat polymorphism. The six common polymorphisms on the left side are strongly correlated. Although these six polymorphisms could theoretically occur in \(2^6\) possible patterns, only three patterns are observed (indicated by pink, orange, and green). These patterns are called haplotypes. Similarly, the six common polymorphisms on the right side are strongly correlated and reside on only two haplotypes (indicated by blue and purple). The haplotypes occur because there has not been much genetic recombination between the sites. By contrast, there is little correlation between the two groups of polymorphisms, because a hotspot of genetic recombination lies between them. The pairwise correlation between the common sites is shown by the red and white boxes below, with red indicating strong correlation and white indicating weak correlation. In addition to the common polymorphisms, lower-frequency polymorphisms also occur in the human genome. Five rare SNPs are shown, with the variant nucleotide marked in red and the reference nucleotide not shown. (...) (b) Small regions such as in a are often embedded in genomic regions with much greater extents of LD. The diagram shows actual data from the International HapMap Project, showing 420 genetic variants in a region of 500 kb on human chromosome 5q31. Positions of the variants and the pairwise correlations are shown below. Blocks of strong correlation are indicated by the black outlines.“[80] Illustration taken from Altshuler et al. [80].

1 million independent hypotheses.“[80] A significance level of \(p = 5 \times 10^{-08}\) – equivalent to a p-value of 0.05 after correction for multiple testing for 1 million independent tests – thus results in wrong associations for 5% of all genotyped SNPs. When using a standard genotyping assay containing 500,000 SNPs, this means that 25,000 false-positive associations contain only a few correctly associated causal alleles.[79]

The consequence of a liberal p-value threshold are therefore follow-up studies (replications) to iteratively exclude false-positives [101]. Another approach to reduce false-positive finding is a so-called meta-analysis, that seeks to pool information from multiple GWAS to increase the chances of finding true positives among the false positives and provides a way
to combine associations across GWAS even when the original data are unavailable.\[102\]

In the end, after significant candidate SNP\s have been identified, those which seem most promising are then closely analyzed along with fine-mapping\[41\] of the associated region. Finally, the discovered associations are set into the biological context.

### 2.4.4 Shortcomings of GWAS

"Genome-wide association studies have uncovered many genetic loci that associate with human diseases, but two fundamental limitations have hampered [the] (...) ability to translate these results into clinically useful predictors of disease and drug targets.

First, the genetic loci associated with disease generally explain very little of the disease risk. The odds of having a risk genotype at a particular disease locus given that one has the disease, divided by the odds of having a risk genotype given that one does not have the disease, are typically less than 1.5 \[99\].\[103\]

Second, as shown in Chapter 6, SNPs reported in GWAS are far from being unbiased. In most cases, the SNPs or loci selected for replication experiments are chosen based on their biologic plausibility. SNPs in coding regions are implicitly accepted as "the answer" when an association with a missense variant is detected, often without the functional scrutiny that is required for a SNP in a non-coding region, and often despite the presence of many nearby variants that might be equally or more strongly associated with disease.\[79\]

"Non-coding mutations with roles in disease susceptibility will (...) open new doors to understanding genome biology and gene regulation. Regulatory variation also suggests different therapeutic strategies: modulating levels of gene expression may prove more tractable than replacing a fully defective protein or turning off a gain-of-function allele.\[80\]

Understanding the biological context in which a given causal gene for disease operates is therefore the necessary step in understanding the underlying mechanisms and in identifying the best drug targets \[103\].

---

41 An experimental approach to narrowing a genome-wide association (GWA) signal by typing all known SNPs in the haplotype block containing the tag SNP. If successful, this approach results in the identification of a subsegment of the block that has a stronger association than the surrounding areas \[82\].
2.4 Genome-wide association studies and complex human diseases

Figure 2.7: The genome-wide association study. The GWAS is typically based on a case-control design in which SNPs across the human genome are genotyped. (a) depicts a very small fragment of the genome on chromosome 9, (b) illustrates the case-control populations having homo- and heterozygotic alleles. (c), the strength of association between each SNP and disease is calculated on the basis of the prevalence of each SNP in cases and controls, with each chromosome shown in a different color. In this example, SNP 1 and 2 on chromosome 9 are associated with disease, with p-values of $e^{-0.12}$ and $e^{-0.08}$, respectively. Next, these candidate SNPs are chosen out for iterative replication experiments. (d) The SNPs or loci which seem most promising are then closely analyzed along with fine-mapping of the associated region. The study is then completed by setting the discovered associations into biological context. Figure adapted from Manolio and Hardy and Singleton.
2.5 MicroRNAs and complex human diseases

In the previous Section, the concept of GWAS was introduced (Section 2.4). In the end, it was emphasized that often SNPs in coding regions are regarded as "the answer" when an association with a missense variant is detected. However, meanwhile it has been shown for many cases that a regulatory polymorphism in a non-coding region — which is in strong LD with a SNP in the coding region — is more strongly associated with disease and is therefore more likely to be causal [79]. One class of these functional polymorphisms are so-called miRSNPs that affect the miRNA regulatory pathway inter alia in non-coding regions such as the 3'-untranslated region (UTR) [104]. For more details see also Chapter 6.

MicroRNAs (miRNAs) comprise a large family of \(\sim 21\)-nucleotide-long RNAs that have emerged as key post-transcriptional regulators of gene expression in a wide range of biological processes, including animal and plant development, cell proliferation and differentiation, apoptosis, and metabolism [105–109]. In mammals, miRNAs are estimated to control the activity of \(\sim 30 - 50\%\) of all protein-coding genes, and changes in miRNA expression have been associated with the pathogenesis of many human diseases such as cancer and metabolic disorders [107, 108]. For a detailed analysis of the role of miRNAs in diseases see Chapters 6 and 7.

During the past decade, much about the basic mechanisms of miRNA biogenesis and function has been identified [105, 107, 110–112]. More recently, it has become apparent that miRNAs themselves are subject to sophisticated control, which takes place at the levels of both, miRNA metabolism and function. The numbers of individual miRNAs expressed in different organisms (e.g., \(\sim 1,424\) in humans) are comparable to those of TFs or RNA-binding proteins (RBPs). Besides, they are also expressed in a tissue- or developmental stage-specific and thus time- and spatial-dependent manner, thereby greatly contributing to cell-type specific profiles of protein expression [117]. The nature of miRNA interactions with their messenger RNA (mRNA) targets, which involve short sequence signatures, makes them well suited for combinatorial effects with other miRNAs or RBPs that associate with the same mRNA. With the potential to target dozens or even hundreds of different mRNAs, individual miRNAs can coordinate or fine-tune the expression of proteins in a cell [108] or act as global switches in, e.g., developmental processes [118].

2.5.1 Overview of microRNA biogenesis

In the following, a detailed overview of miRNA biogenesis is given, which is also illustrated in Figure 2.8. For more details on this see the recent reviews of Krol et al. [108] and Winter et al. [119].

\[\text{miRBase version 17.0 [113-116]}\]
2.5 MicroRNAs and complex human diseases

Figure 2.8: microRNA biogenesis. MiRNAs are processed from RNA polymerase II specific transcripts of independent genes or from introns of protein-coding genes. In the canonical pathway, primary precursor (pri-miRNA) processing occurs in two steps, catalyzed by two members of the RNase III family of enzymes, Drosha and Dicer, operating in complexes with dsRBPs, e.g., DGCR8 and TRBP in mammals. In the first nuclear step, the Drosha-DGCR8 complex processes pri-miRNA into a ∼ 70-nucleotide precursor hairpin (pre-miRNA), which is exported to the cytoplasm. Some pre-miRNAs are produced from very short introns (mirtrons) as a result of splicing and debranching, thereby bypassing the Drosha-DGCR8 step. In either case, cleavage by Dicer, assisted by TRBP, in the cytoplasm yields a ∼ 20-bp miRNA/miRNA* duplex. In mammals, AGO2, which has robust RNaseH-like endonuclease activity, can support Dicer processing by cleaving the 3’ arm of some pre-miRNA, thus forming an additional processing intermediate called AGO2-cleaved precursor miRNA (ac-pre-miRNA). Following processing, one strand of the miRNA/miRNA* duplex (the guide strand) is preferentially incorporated into a miRISC, whereas the other strand (passenger or miRNA*) is released and degraded. Generally, the retained strand is the one that has the less stably base-paired 5’ end in the miRNA/miRNA* duplex. MiRNA* strands are not always by-products of miRNA biogenesis and can also be loaded into miRISC to function as miRNA. Illustration taken from Krol et al. [108].
MicroRNAs are processed from precursor molecules (pri-miRNAs), which are either transcribed by RNA polymerase II from independent genes or represent introns of protein-coding genes. The pri-miRNAs fold into hairpins, which act as substrates for two members of the RNase III family of enzymes, Drosha and Dicer. The product of Drosha cleavage, a 70-nucleotide pre-miRNA, is exported from the nucleus [by Exportin-5-Ran-GTP] to the cytoplasm where the RNase Dicer in complex with the double-stranded RNA-binding protein TRBP (...) processes it to a 20-bp miRNA/miRNA* duplex. [43] One strand [the functional one –] of this duplex, representing the mature miRNA, is then incorporated into the miRNA-induced silencing complex (miRISC). Argonaute (AGO) proteins, which directly interact with miRNAs, and GW182[44] (...) proteins, which act as downstream effectors in the repression, are key factors in the assembly and function of miRISC.[108] As part of miRISC, miRNAs basepair to target mRNA, and induce their translational repression or deadenylation and degradation.[108, 119].

Most animal miRNAs imperfectly [wobble] base-pair with sequences in the 3'-UTR of target mRNA, and inhibit protein synthesis by either repressing translation or promoting mRNA deadenylation and decay. Efficient mRNA targeting requires continuous [perfect Watson-Crick] base-pairing of miRNA nucleotides 2 – 8[108], which represent the so-called seed sequence[45][105, 108]. Processing of some precursors by Drosha and Dicer is not uniform and generates miRNA isoforms that present different termini [107, 120]. In particular, heterogeneity at the 5'-end has important functional consequences, as it affects the seed region of the miRNA and consequently, changes the identity of targeted mRNAs[121]. Modifications of the 3'-end, on the other hand, were shown to influence the stability and abundance of the miRNA.[108].

Very recently, approaches towards large-scale experimental verification of exact AGO binding sites have been developed[122][123]. These so-called CLIP[46] methodologies allow to exactly determine the footprints of the miRISC and therefore allow to infer miRNA binding sites from that. A footprint is a defined region of miRNA complexed with AGO determined by AGO/miRNA clusters, where AGO bound with 62 nucleotides (nts) of cluster peaks ≥ 95% of the time[124]. The experimental setup of PAR-CLIP was performed in HEK293 cells[47] and delivers targets of 221 human miRNAs. This corresponds to only around 22% of all human miRNAs.[48] This study, however, provides the most comprehensive data set of the highest quality and accuracy on microRNA binding sites available up to now.

---

[43] MiRNA* is the so-called passenger microRNA - the strand of RNA in the duplex which is not preferentially incorporated in the silencing complex after cleavage by the Dicer complex.

[44] GW182: glycine-tryptophan protein of 182 kDa

[45] Seed sequence: Nucleotide positions 2 – 8 from the 5'-end of the miRNA, which generally perfectly base-pair with target mRNA and are important for defining the target repertoire of a miRNA.[108].


[47] HEK293: human embryonic kidney

[48] miRBase release 16 (Sept 2010)[116] contains information of more than 1,000 human miRNAs.
2.5.2 Regulation of microRNA gene transcription

"The promoter regions of autonomously expressed microRNA genes are highly similar to those of protein-coding genes [125, 126]. The presence of CpG islands, TATA box sequences, initiation elements, and certain histone modifications indicate that the promoters of microRNA genes are controlled by TFs, enhancers, silencing elements, and chromatin modifications, which is similar to protein-coding genes. Many TFs regulate microRNA expression positively or negatively in a tissue-specific or developmental-specific manner, leading to a complex regulatory network (Figure 2.9). Control of gene expression by auto-regulatory feedback loops is a common regulatory mechanism that is particularly important during cell fate determination and development. MiRNAs are uniquely suited to participate in feedback circuits owing to their potential to directly base-pair with and repress mRNAs that encode factors involved in the biogenesis or function of the same microRNA. Indeed, many examples have been described of microRNAs regulating their own transcription through single-negative or double-negative (or positive) feedback loops with specific transcription factors. By fine-tuning microRNA expression and adjusting it to physiologically optimal levels, the circuits described have a strong impact on the precise spatio-temporal expression of microRNA targets." [108]

**Figure 2.9: MicroRNA regulation.** (a) "Transcription of microRNA genes is regulated in a similar manner to that of protein-coding genes, and is a major level of control responsible for tissue-specific or development-specific expression of microRNA. Many TFs regulate microRNA expression positively or negatively in a tissue-specific or developmental-specific manner (transcriptional activators or repressors are shown in green and red, respectively). For instance (...) MYC and MYCN both stimulate expression of the miR-17-92 oncogenic cluster in lymphoma cells [127] and miR-9 in neuroblastoma cells [128], but inhibit expression of several tumor suppressor microRNAs (e.g. miR-15a), which promote MYC-mediated tumorigenesis [129]. MiRNAs frequently act in regulatory networks with TFs, which can drive or repress the expression of the microRNAs." [108]

In (b) a few examples of auto-regulatory feedback loops are shown. Illustration adapted from Krol et al. [108].
2.5.3 MicroRNAs in human disease

Being aware of the complex interplay between TFs, miRNAs, their corresponding targets, and the heterogeneous function of miRNAs, it becomes obvious that perturbation of miRNA regulation can have great and diversified impact on complex systems. Indeed, most of the miRNA-disease associations rely on miRNA expression measurements in diseased tissues \[130\], e.g., deregulated miRNA expression has been shown to be associated with various types of diseases ranging from cancer to metabolic and neurodegenerative disorders \[131–137\]. In the last decade, a multitude of studies uncovering the role of miRNA in diseases have been published. Besides various online resources such as TransmiR \[138\], PhenomiR \[139\], miR2Disease \[140\], or DIANA-microT \[141\] have been created that systematically collect such information.

2.5.4 The future of microRNA research

“Considering the fundamental role of miRNAs in organismal development, cellular differentiation, metabolism, and oncogenesis, we can anticipate many more sophisticated mechanisms for the regulation of their biogenesis, function, and catabolism to emerge in coming years. A few examples of post-translational modifications of miRISC proteins and factors controlling miRNA biogenesis are already known and it will be important to determine how these, and other modifications that might be discovered, affect miRISC function and what signalling pathways are responsible for them. Such information will be particularly important for the understanding of the basis of miRNA dysregulation known to occur in human pathologies.” \[108\]

Another important area of research will be understanding control at the level of the 3’-UTRs of mRNAs, a point that is addressed in Chapter 6. Functionally, the mRNA’s 3’-UTR can be considered as a post-transcriptional equivalent of the gene promoter at which most transcription-related decisions are made. With hundreds of RBPs being expressed in eukaryotic cells, it is clear that the interplay between miRISC and RBPs, which both bind along kilobases of the 3’-UTR will be of great importance for fine-tuning protein synthesis. Especially new technologies such as CLIP which allows miRISC and RBP binding site mapping on a global scale, promise rapid progress in this area \[122, 123\].
CHAPTER 3

Insights into the structural organization of biological data via fuzzy clustering of \( k \)-partite graphs

High-throughput biological methods, namely methods that perform thousands of simultaneous measurements of biological molecules, have rapidly transformed the landscape of biomedical research during the past decade. These \"-omics\" technologies such as next generation sequencing, proteomics, or metabolic profiling generate a large volume of information, which is even increased by the enormous amount of textual data accessible in the biomedical literature and difficult to manage. Methods able to structure these heterogeneous data and to extract new knowledge gain therefore more and more importance.

Learning approaches often focus on the analysis of homogeneous data sets that can be represented as graphs having vertices of a single type only. Biological networks, however, are complex and highly diverse, and therefore often involve objects of multiple types, forming \( k \)-partite graphs that consist of different kinds of vertices (Section 2.1.1). We use this representation as it provides a more comprehensive picture of the underlying structure compared to the widely used graph transformations. These so-called projections – e.g. of a bipartite network into a unipartite version – discard important information [142]. For instance, Montañez et al. [143] showed that in the case of metabolism the use of projections leads to wrong interpretations of some of the most relevant graph attributes, whereas the bipartite view offers a cleaner interpretation of its topological features.

The human disease network presented by Goh et al. [17] is an example for a bipartite graph having two disjoint sets of vertices. Here, structural questions need to be addressed outside of the unipartite graph setting. One set of nodes represents all known genetic disorders, the vertices of the other partition correspond to all known disease genes in the human genome. A disorder and a gene are connected if mutations in that gene are implicated in that disorder. Other examples of bipartite networks are protein complex or gene-localization, gene-function or \( \text{miRNA} \)-target networks. The integration of such network data then leads to complex \( k \)-partite graphs.
3 Insights into the structural organization of biological data via fuzzy clustering of \( k \)-partite graphs

### 3.1 Modular decomposition by graph clustering

A key question is how to interpret the internal organization of these networks (Section 2.1.4). A possible answer may be a modular decomposition, which implies the coexistence of structural subunits associated with more highly interconnected parts. We regard the identification of these \textit{a priori} unknown building blocks — such as for instance functional modules in PPI networks — as clustering methods. The clusters and their interconnections are essential for understanding the underlying functional properties. They structure biological data by compressing their information into a condensed form (Section 2.2).

Most available clustering methods do not treat the single node types (partitions) separately and therefore do not represent the global cluster structure of \( k \)-partite networks correctly. While this has been addressed in terms of algorithms for some time now [144–146], not many applications were successfully implemented in bioinformatics yet, although the field commonly deals with such networks [142]. A particular issue that may hamper application to biological data is that most existing algorithms identify separated, disjoint clusters by assigning each point to exactly one cluster [147, 148]. This is unrealistic for biological systems as, e.g., genes or proteins commonly participate in multiple processes or pathways [149]. So far, only a few approaches exist that allow the detection of overlapping clusters. These either assign each data point to several hard clusters [39] or determine a degree of membership to each cluster [41, 150] (Section 2.2.4). Such methods are known as fuzzy clustering, but have not been applied to the common biological case of \( k \)-partite graphs.

To overcome these difficulties, we developed a novel fuzzy clustering algorithm based on a non-negative matrix factorization (NMF) model [43] (Section 2.2.3). Our algorithm extends a hard clustering algorithm recently put forward by Long et al. [151]. This algorithm clusters each node type of the graph separately and then connects clusters via a smaller, weighted \( k \)-partite graph in an alternating minimization procedure. Thereby, the cluster assignment in the first step is made in a binary fashion. This disjoint clustering is a feature that is often achieved by soft clustering algorithms when not forcing explicit cluster overlap [11]. But it can be easily seen that the cost function proposed is not fully minimized. Our computationally efficient and scalable algorithm avoids this problem. It is similar in structure to multiplicative algorithms for NMF, with the difference that we address a three-matrix factorization problem (see e.g. Dhillon and Sra [152]), and have to deal with a multi-summand cost function. As our cost function is monotonous with respect to the number of clusters, our algorithm allows the detection of clusters on different scales (Section 2.1.2.3). Hence, we are able to decompose the network on different resolution levels.

This Chapter is organized as follows. In the next Section, we develop the fuzzy clustering algorithm and validate it on a toy example and graphs with known modular structure. 

\[ \text{Chapter 3 led to two publications, namely, Hartsperger et al. [11] and Blöchi et al. [12].} \]
tute. Then, we apply it to a tripartite disease-gene-protein complex graph representing an expanded view of the human disease network from Goh et al. [17] extended by protein complexes [153]. By integrating functional annotation, we demonstrate that we are able to structure this complex graph into biologically meaningful clusters on a large-scale, which corresponds to the macroscopic view. Finally, focusing on the small-scale architecture, we identify overlapping clusters that give a more comprehensive picture about gene-disease connections rather than looking at disjoint clusters alone. We exemplify how this clustering allows for reclassification, annotation or even detection of misclassified elements on a local level, which reflects the microscopic view.

3.2 A NMF-type community detection algorithm

A $k$-partite graph is a graph $G = (V,E)$ of edges $E$ between a set of vertices $V$ together with a partition of the vertices into $k$ disjoint subsets $V_i$ such that no two vertices in the same subset are adjacent. For $k = 1$ this reduces to the standard graph, where we do not take into account different node types. Graphs with two partitions are called bipartite (Section 2.1.1.4). Let $n_i := |V_i|$ be the number of vertices in the partition $i$, $i = 1\ldots k$. We represent the graph as a set of $n_i \times n_j$-dimensional adjacency matrices $A^{(ij)}$ for all $i,j$ with $1 \leq i < j \leq k$. Typically, each matrix element is either 0 or 1, but we only restrict the matrices to have non-negative coefficients, thereby allowing weighted graphs as well.

3.2.1 Graph approximation

We want to approximate $G$ by a smaller $k$-partite cluster network $H$, which we call backbone network. It is defined on the fuzzy clusters of each $G$-partition $V_i$. We fix the number of clusters in partition $i$ to $m_i$. We denote a non-negative $n_i \times m_i$-dimensional matrix $C^{(i)}$ to be a fuzzy clustering of $V_i$, if it is right-stochastic, i.e. $\sum_{l=1}^{m_i} C^{(i)}_{kl} = 1$ for all $k$. Its $(k,l)$-th element $C^{(i)}_{kl}$ gives the degree of membership of the original node $k$ to the backbone node $l$.

Then we search for a $k$-partite graph $H$ with $m_i \times m_j$ adjacency matrices $B^{(ij)}$ and a fuzzy clustering $C := (C^{(i)})_{i=1,\ldots,k}$ such that the connectivity explained by $H$ is as close as possible to $G$ after clustering according to $C$. Figure 3.1 shows an example graph and its approximation by a backbone network.

From the approximation, we can easily reconstruct an edge $A^{(ij)}_{uv}$ between two nodes $u$ and $v$ from partitions $i$ and $j$ in the original graph. To this end, we have to sum up all edge weights $B^{(ij)}$ in the backbone graph that connect the communities $u$ and $v$ are assigned to. Of course, in a fuzzy environment these contributions have to be weighted by the nodes’ degrees of membership $C^{(i)}$ and $C^{(j)}$, respectively. Taken together, the entry of the adjacency matrix can be reconstructed as the double sum

$$A^{(ij)}_{uv} \approx \sum_{x=1}^{m_i} \sum_{y=1}^{m_j} C^{(i)}_{ux} B^{(ij)}_{xy} C^{(j)}_{vy}.$$
Figure 3.1: Illustration of the fuzzy clustering approach. We want to approximate the tripartite example graph $G$ in (a) by a smaller tripartite cluster network $H$, the so-called backbone graph (b). The decomposition into fuzzy clusters connected by this backbone must explain the original connectivity as good as possible. The edges of $G$ are collected in adjacency matrices $A^{(ij)}$ connecting the elements of the partitions $i$ and $j$. The approximation of $G$ by the backbone graph is encoded in the adjacency matrices $B^{(ij)}$ connecting the fuzzy node clusters $C^{(i)}$. These matrices $C^{(i)}$ collect the degrees of membership of each node of partition $V_i$ to each cluster of this type. Its $(k,l)$-th element $C_{kl}^{(i)}$ specifies how much node $k$ belongs to the backbone node $l$.

Writing this in matrix notation, we see that the requirement of explaining maximum possible connectivity means that the adjacency matrices $A^{(ij)}$ are best possible approximated by factorizations of the form

$$A^{(ij)} \approx C^{(i)} B^{(ij)} (C^{(j)})^\top.$$

We can measure the difference between the two graphs $H$ and $G$ in a variety of ways. In Long et al. [151], this choice has been circumvented by focusing on arbitrary Bregman divergences, which still allow efficient reformulation of gradient-type algorithms without knowing the specific formula (see, e.g., Banerjee et al. [154]). This is also possible in our case of multiplicative update rules, as has been shown for NMF by Dhillon and Sra [152]. Here, we choose the minimum square distance as the cost function. This implies minimization of

$$f(H,C) := \sum_{i<j} \left\| A^{(ij)} - C^{(i)} B^{(ij)} (C^{(j)})^\top \right\|_F^2,$$  \hspace{1cm} (3.1)

where $\| \cdot \|_F^2$ denotes the squared Frobenius norm, i.e., the square sum of the matrix elements. This cost function is obviously monotonous with respect to the number of clusters in each partition.
3.2 A NMF-type community detection algorithm

3.2.2 Derivation of the update rules

We want to minimize $f(H, C)$ in Equation 3.1 using a local algorithm extending gradient descent. Let $D^{(ij)} := A^{(ij)} - C^{(i)}B^{(ij)}(C^{(j)})^\top$ denote the residuals, then

$$ f = \sum_{i<j,k,l} (d_{kl}^{(ij)})^2. $$

Hence

$$ \frac{\partial f}{\partial b^{(ij)}_{rs}} = -2 \sum_{kl} d_{kl}^{(ij)} c_{kr}^{(i)} c_{ls}^{(j)} \rho_{s}^{(i)} \nabla_{r}^{(i)} C^{(i)}(C^{(j)})^\top. $$

$$ \frac{\partial f}{\partial c^{(i)}_{rs}} = -2 \sum_{j \neq i} d_{ji}^{(ij)} b_{sk}^{(ij)} c_{lk}^{(j)} c_{ls}^{(j)} \rho_{s}^{(i)} \nabla_{r}^{(i)} C^{(i)}(B^{(ij)})^\top. $$

We assume an undirected $k$-partite graph, so $A^{(ij)}$ is undefined for $i > j$. For simplicity of notation, we now set $A^{(ij)} := (A^{(ij)})^\top$ for $i > j$ (and similarly for the $k$-partite graph $H$). Then $D^{(ij)} = (D^{(ji)})^\top$, and the differential simplifies to

$$ \frac{\partial f}{\partial c^{(i)}_{rs}} = -2 \sum_{j \neq i} \rho_{s}^{(i)} \nabla_{r}^{(i)} C^{(i)}(A^{(ij)})^\top C^{(j)}(B^{(ij)})^\top. $$

Altogether, by replacing the residuals, we have shown

$$ \frac{\partial f}{\partial b^{(ij)}_{rs}} = -2 \rho_{s}^{(i)} \nabla_{r}^{(i)} C^{(i)}(A^{(ij)})^\top C^{(j)}(B^{(ij)})^\top. $$

If we are to minimize $f$ by alternating gradient descent, we start from an initial guess of $B^{(ij)}$, $C^{(i)}$. Then, we alternate between updates of the $B^{(ij)}$ and the $C^{(i)}$ with learning rates $\eta_{rs}^{(ij)}$ and $\eta_{rs}^{(i)}$, respectively:

$$ b^{(ij)}_{rs} \leftarrow b^{(ij)}_{rs} - \eta_{rs}^{(ij)} \frac{\partial f}{\partial b^{(ij)}_{rs}}, \quad \forall i, j : i < j $$

$$ c^{(i)}_{rs} \leftarrow c^{(i)}_{rs} - \eta_{rs}^{(i)} \frac{\partial f}{\partial c^{(i)}_{rs}}, \quad \forall i $$

These update rules have two disadvantages:
1. The choice of update rate $\eta$ (possibly different for $B$, $C$, and $i,j$) is unclear. In particular, for too small $\eta$, convergence may take too long or may not be achieved at all, whereas for too large $\eta$, we may easily overshoot the minimum.

2. The resulting matrices may become negative. To avoid the choice of update rates and to ensure positivity of both the backbone network and the degrees of membership of all nodes, we employ multiplicative update rules. This strategy is widely used in algorithms for NMF [44]. Hence, we follow Lee and Seung’s idea for NMF [43] and rewrite this into multiplicative update rules. We therefore choose update rates

$$
\eta_{rs}^{(ij)} := \frac{b_{rs}^{(ij)}}{2 \left( \sum_{j \neq i} C(i)^\top B(ij)^\top C(j) \right)_{rs}}
$$

and

$$
\eta_{rs}^{(i)} := \frac{c_{rs}^{(i)}}{2 \left( \sum_{j \neq i} A(ij)^\top C(j)^\top C(j) \right)_{rs}}
$$

Plugging this into the gradient descent equations, we finally get:

$$
b_{rs}^{(ij)} \leftarrow b_{rs}^{(ij)} \frac{\left( C(i)^\top A(ij)^\top C(j) \right)_{rs}}{\left( C(i)^\top C(i)^\top B(ij) \right)_{rs}}
$$

and

$$
c_{rs}^{(i)} \leftarrow c_{rs}^{(i)} \frac{\left( \sum_{j \neq i} A(ij)^\top C(j)^\top B(ij)^\top C(j) \right)_{rs}}{\left( \sum_{j \neq i} A(ij)^\top C(i)^\top C(j) \right)_{rs}}
$$

Commonly, to extend cost functions in (unipartite) data clustering to include fuzzy clusters, a so-called fuzzification factor is introduced [40, 41] (Section 2.2.2.2). Instead of squared norm minimization of the residuals $D^{(ij)}$, a higher residual power is minimized, which results in overlapping non-trivial cluster assignments. Nevertheless, we see that in our examples, already the standard case is sufficient. This is because we are interested in co-clustering, which is different from standard data clustering where only a unipartite graph and hence $C(i) = C(1)$ is assumed.

We note that these update rules do not increase the cost function (Equation 3.1). This can be shown via auxiliary functions similar to NMF [44] and multi-factor NMF [152], which has been applied in a related model for co-clustering of microarray data [155]. This theoretical result implies convergences of the update rules. But, in contrast to early statements in NMF [44], it does not necessarily imply convergence to stationary points of the Euclidean norm (zero of the differential from Equation (3.1)), since the update steps may be too small to reach those points. Another possible drawback of such multiplicative updates is the fact that once a matrix entry has been set to zero (which may happen due to zeros in $A^{(ij)}$ or to numerics), the coefficient will never then be able to become positive again during learning. This is one of the reasons, why sometimes alternating least-squares
3.2 A NMF-type community detection algorithm

algorithms are chosen [156].

We have not yet taken into account the constraint that the fuzzy clusterings $C^{(i)}$ are required to be right-stochastic (Section 3.2.1). We force this constraint by regularly projecting each row of $C^{(i)}$ onto the sphere of the 1-norm. The final fuzzy $k$-partite clustering algorithm is summarized in Figure 3.2.

### Algorithm 1 fuzzy $k$-partite clustering

**Input:** $k$-partite graph $G$ with possibly non-negatively weighted edge matrices $A^{(ij)}$, $i < j$, number of clusters $m_1, \ldots, m_k$

**Output:** fuzzy clustering $C^{(i)}$ and $k$-partite cluster graph $H$ given by matrices $B^{(ij)}$

1. Initialize $C^{(i)}$, $B^{(ij)}$ to random non-negative matrices.
2. Normalize $c_{rs}^{(i)} \leftarrow c_{rs}^{(i)}/(\sum_t c_{rt}^{(i)})$ for all $i, r, s$

repeat

3. **update fuzzy clusters**
   
   for $i \leftarrow 1, \ldots, k$ do
   
   $C^{(i)} \leftarrow C^{(i)} \odot (\sum_{j \neq i} A^{(ij)} C^{(j)} B^{(ij)\top}) \odot (\sum_{j \neq i} C^{(ij)} B^{(ij)} C^{(j)} B^{(ij)\top})$

   Normalize $c_{rs}^{(i)} \leftarrow c_{rs}^{(i)}/(\sum_t c_{rt}^{(i)})$ for all $r, s$

   end

4. **update $k$-partite cluster graph $H$**
   
   for $i \leftarrow 1, \ldots, k - 1$ do
   
   for $j \leftarrow i + 1, \ldots, k$ do
   
   $B^{(ij)} \leftarrow B^{(ij)} \odot (C^{(i)} \top A^{(ij)} C^{(j)}) \odot (C^{(i)} \top C^{(ij)} B^{(ij)} C^{(j)})$

   end

   end

until convergence;

Note: $\otimes$ and $\oslash$ symbolize element-wise multiplication and division, respectively.

#### Figure 3.2: Fuzzy clustering algorithm

Pseudo code summarization of the final fuzzy $k$-partite clustering algorithm.

3.2.3 Complexity analysis

The massive amount of data on real networks currently available makes the issue of the efficiency of clustering algorithms essential. The computational complexity of an algorithm is the estimate of the amount of resources required by the algorithm to perform a task. This involves both, the number of computation steps needed and the number of memory units that need to be simultaneously allocated to run the computation [23].

Our algorithm has two nested loops over the number of partitions $k$, hence its runtime depends quadratically on this number. The update rules for $C^{(i)}$ and $B^{(ij)}$, however, are fully vectorized and contain only matrix operations with non-square matrices. Their time complexity is dominated by the occurring matrix products: multiplying two matrices of sizes $s_1 \times s_2$ and $s_2 \times s_3$ is of complexity $\mathcal{O}(s_1 s_2 s_3)$. Assuming that the cluster numbers $m_i$ are smaller than the largest two partition sizes, the total time complexity of the fuzzy clustering algorithm can then be estimated as
Here, \( n_{\text{max}1} \) and \( n_{\text{max}2} \) denote the sizes of the largest and the second-largest partition, \( m_{\text{max}} \) is the maximum number of clusters within any partition. Hence, the runtime grows only quadratically in the total number of nodes in the case of graphs with similarly large partitions. In general, the runtime is linear in each partition’s size \( n_i \) and cluster number \( m_i \).

### 3.3 Algorithm evaluation

For illustration, we applied our algorithm to a bipartite graph having several vertices connected with all vertices of the other partition (e.g., nodes 1 and 10). Figure 3.3 shows that these vertices are assigned to two clusters with distinct degree of membership, whereas vertices partially connected are element of a single cluster only (e.g., node 5). This demonstrates the idea and importance of using a fuzzy clustering approach that allows for overlapping clusters.

**Figure 3.3: Illustration of the cluster decomposition of a bipartite toy example.** (a) We demonstrate the graph decomposition with our algorithm on a small bipartite graph with overlapping cluster structure. The original graph consists of partitions \( V_1 = \{1 \ldots 4\} \) (red filled nodes) and \( V_2 = \{5 \ldots 10\} \) (blue filled nodes) connected by edges \( A^{(12)} \) colored in black. We decomposed it into two clusters for partition \( V_1 \) and three clusters for partition \( V_2 \). The resulting fuzzy clustering is illustrated as a weighted graph connecting original nodes to cluster nodes (framed red and blue). The cluster assignments \( C^{(1)} \) and \( C^{(2)} \) are indicated by dashed lines, where the coloring corresponds to the degree of cluster membership. The interconnections of the clusters form the backbone graph, encoded in the adjacency matrix \( B^{(12)} \) which we denote by continuous lines where color indicates the edge weight. Another way of illustrating the graph decomposition is shown in (b). It is clearer especially for larger graphs. First, we plot hierarchical clusterings of the nodes’ degrees of membership in partitions \( V_1 \) and \( V_2 \) (encoded by \( C^{(1)} \) and \( C^{(2)} \)). This facilitates the identification of overlapping clusters (e.g., nodes 1 and 10 are assigned to more than one cluster) or hard cluster assignments (e.g., node 5). The backbone graph \( B^{(12)} \) is shown bottom right. This backbone graph is densely connected in our example.
3.3 Algorithm evaluation

3.3.1 Performance analysis

Before applying our algorithm to real-world data, we tested its behavior on simulated data with controlled cluster structure. In particular, we compared it to the hard clustering algorithm from Long et al. [151] using exactly the same stopping criteria for both algorithms.

We generated a random, modularly structured $k$-partite network as follows: First, we fix the number of clusters $m_i$ of nodes with color $i$, $i = 1, \ldots, k$. The backbone graph is initialized by $m_i \times m_j$-matrices $B^{(ij)}$ filled with zeros. Then, we added uniformly random ones in each column according to a set percentage $\alpha$ (here on average $\alpha \geq 1$ ones in each column) such that each row has at least a single non-zero entry. To construct the actual network $A$, we split up $A^{(ij)}$ into $m_i \cdot m_j$ blocks of a fixed chosen clustersize (here 10). We fixed a cluster connectivity $\beta$ and a random connectivity $\gamma < \beta$. Now, for each non-zero entry in $B^{(ij)}$, we set the corresponding block of $A^{(ij)}$ to an Erdös-Rényi (ER) graph [157] with density $\beta$ (Section 2.1.2.1). Finally, the clusters are connected by replacing each zero block of $A^{(ij)}$ with an ER graph of the lower connectivity $\gamma$.

To compare algorithm performance, we determined algorithm runtime, final cost function value, and quality of cluster estimation. Cluster estimation quality was measured by the summed up Frobenius norms of the difference between the true $C^{(i)}$ and the estimated $\hat{C}^{(i)}$, where clusters have been permuted such as to give minimal difference (permutation indeterminacy). We analyzed 1,000 realizations of four network prototypes with increasing complexity. The corresponding parameters are given in Table 3.1. We restricted ourselves to bipartite and layered tripartite graphs with two different noise settings because Long et al. [151] provided code for analyzing these special cases only.

<table>
<thead>
<tr>
<th>model</th>
<th>$k$</th>
<th>$m$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>(3,3)</td>
<td>1</td>
<td>0.7</td>
<td>0.2</td>
<td>equal-sized, no overlap</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(3,4)</td>
<td>1</td>
<td>0.7</td>
<td>0.2</td>
<td>no cluster overlap</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>(3,4,5)</td>
<td>1.2</td>
<td>0.6</td>
<td>0.1</td>
<td>3-partite, low-noise</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>(3,4,5)</td>
<td>1.2</td>
<td>0.8</td>
<td>0.2</td>
<td>3-partite, noisy</td>
</tr>
</tbody>
</table>

Table 3.1: Parameters for the simulated data models. $k$ denotes the number of partitions of the network, $m$ is a vector with the number of clusters in each partition, $\alpha$ the backbone connectivity, $\beta$ the cluster and $\gamma$ the noise connectivity.

We found that while the method of Long et al. [151] performed around two times faster, our algorithm produced an around 10% lower cost function and was able to estimate the cluster structure better (Figure 3.4). This difference in algorithm runtime originates from the much more fine-tuning of the continuous degrees of membership compared to hard cluster assignments. These require less update steps until convergence.
Figure 3.4: Performance on toy models. We validated our algorithm on graphs with predefined cluster structure. To this end, we compared it with the hard clustering method by Long et al. [151] on four different random toy models, see Table 3.1. The plot shows the mean relative deviation between the two algorithms relative to the results of the hard clustering. Error bars denote standard deviations over 1,000 runs. We see that the fuzzy cluster assignments of our method require much more runtime, but both cost function and data estimation error are significantly smaller. The large standard deviations show the dependency of the decomposition on the random initial conditions. Therefore, by default we perform multiple restarts with different initializations.

3.3.2 Stability of clusters

In contrast to deterministic methods such as for instance singular value decomposition (SVD), NMF-based methods have problems concerning robust computation. Even for standard unipartite NMF there is no unique global minimum of the cost function [158]. Our algorithm aims to minimize the cost function using a local optimization strategy extending gradient descent. This implies that the algorithm only converges to a local minimum. The algorithm is indeterministic, it does not converge to the same solution on each run due to the stochastic nature of initial conditions. However, the validity of specific solutions represents an important criterium to determine the performance of clustering algorithms. The reproducibility and therefore stability of clustering results has to be analyzed since different outcomes may have significant impacts on the interpretation of the results. It is widely known that NMF-based methods are sensitive to initializations and therefore may converge to different solutions due to different (random) initializations. To identify stable results multiple restarts are generally used [42, 158]. Thus, we compared the local minima from several different starting points (multiple restarts), using the results of the best local minimum found.

To determine how stable the clustering results are (or in other words, to analyze the stability and reproducibility of the fuzzy clustering algorithm), we applied our algorithm to a bipartite toy network with well defined cluster structure (Figure 3.5). The graph
with $n_1 = 900$ nodes in partition 1 and $n_2 = 800$ nodes in partition 2 consists of three unconnected components: we chose to have three hard clusters ($k = 3$) in each partition, each of them connected to only one cluster of the other partition. Then, we employed a stability score to compare the clustering results and to quantify their stability (or reproducibility). Garge et al. [47] proposed to use Cramer’s $\nu^2$ (Equation 2.12) as a measure for the degree of replicability. But since crisp assignment of data points to clusters is required, we used the FRI instead (Equation 2.14). The FRI allows to compare fuzzy partitions and therefore to determine how close the fuzzy memberships obtained in different runs are. See Section 2.2.5 for more details.

Figure 3.5: Adjacency matrix for cluster stability analysis. The adjacency matrix (with 900 nodes in partition 1 and 800 nodes in partition 2) used for the cluster stability analysis is illustrated as heat-map, where white denotes zeros and black denotes ones.

We calculated the pairwise of all clustering results to the reference fuzzy memberships for both partitions separately. The resulting FRI distribution is shown in Figure 3.6. In more than 70% we reach a FRI of 1, indicating that our algorithm produces stable results with close fuzzy memberships. But although the input graph has a well defined cluster structure, our algorithm does not always converge to a meaningful decomposition. Sometimes, only one cluster could be determined correctly, which shows that we can not guarantee that the local optimization finds a global minimum of the cost function. This illustrates the critical need for multiple restarts.
3 Insights into the structural organization of biological data via fuzzy clustering of \(k\)-partite graphs

3.3.3 The cluster structure dependence on \(m\)

Using a bipartite graph with six well defined, yet not hard clusters (Figure 3.7a), we studied the dependency of the clusters on the cluster number \(m\) (Figure 3.8). In this example, we see that extracting less clusters than really present, our algorithm identifies clusters composed by the union of true ones. If we want to extract more clusters than the correct number, it splits up true clusters into strongly overlapping sub-clusters. Hence, even when not extracting the exactly correct number of clusters, the decomposition obtained is meaningful and allows for a deeper interpretation of the extracted clusters.

The histograms of the obtained degrees of membership show a characteristic behavior. In graphs containing a well defined cluster structure, they have an \(U\)-like shape. We found a large number of high degrees of membership almost up to one (Figure 3.8). In graphs without cluster structure, compare Figure 3.9, these large values are completely missing. Hence, such histograms can be employed to give an indication whether a graph is modularly organized or not.

Figure 3.6: Cluster stability analysis (a) Good decomposition of partition 1 into the true cluster structure, illustrated as heat-map, where white denotes zeros and black denotes ones. We compared the clustering results of 100 runs with the extracted cluster structure to evaluate the reproducibility of our algorithm’s outcome. The FRI was used to quantify the stability of the clustering results. (b) Distribution of the fuzzy rand index (FRI) over 100 runs. We reach a FRI of 1 in more than 70\%, indicating a perfect reproducibility. (c) Partition 2 is well decomposed into its predefined cluster structure with 3 clusters. Again, in more than 70\% a perfect reproducibility score with FRI of 1 is reached. The FRI distribution of the 100 comparisons of partition 2 is shown in (d).
To analyze the profile of the cost function, we generated two bipartite example graphs. Their adjacency matrices are illustrated in (a) and (c) as heat-maps (black codes for ones, white for zeros). The first graph has six obvious (yet not hard) clusters in each partition, each of them connected to only one cluster of the other partition. It contains 75 nodes per cluster and two nodes of different color stemming from linked clusters are connected with a probability of 0.4. Additionally, we introduced random connections between the other nodes with a probability of 0.05. The second graph also contains 450 nodes per partition, but has no cluster structure (all pairs of nodes connected with probability 0.12). (b) and (d) show the profile of the cost function after algorithm convergence (average and minimum value over 25 runs) when extracting between two and 15 clusters. While in the second, cluster-free example there is no structure in the profile of the cost function, the well defined cluster structure of the first example has a sharp break. For $m = 1 \ldots 6$, i.e., until all present six clusters detected, we observe steep drops in the cost function, followed by a flat plateau with little refinement for $m > 6$. 
Figure 3.8: Cluster structure - membership behavior. (a) to (d) give hierarchical clusterings of the degree of membership matrices obtained for the true $m = 6$ and also for $m = 5$ and $m = 7, 8$. In each case, we show the best of 25 runs. The algorithm shows a nice behavior: If we extract only five clusters, it detects four of the six clusters, the fifth cluster being the union of the last two true clusters. If we extract more than six clusters, it splits up true clusters into two strongly overlapping sub-clusters. In (e) to (h) we give histograms of the degrees of membership obtained in these situations (for better recognizability, we counted only entries $\geq 0.01$). These histograms show a typical U-like shape, with a peak at small entries and a second peak at large degrees of membership around one indicating the well defined cluster structure of the studied graph.
3.4 Decomposition of a gene-disease-protein complex graph

To exemplify the analysis of biological networks, we applied our algorithm to a layered tripartite disease-gene-protein complex network, see Figure 3.11a for an abstract illustration. In this graph, a disorder and a gene are connected if mutations in that gene are implicated in that disorder. A complex and a gene are linked if the gene is coding for a protein part of the complex. We constructed this graph by integrating the human gene–disease network from Goh et al. [17] and protein complexes from the CORUM database. Integrating both manually curated data sets resulted in a graph of 5,672 nodes and 7,795 edges with all genetic disorders, all known disease genes, and human protein complexes. We extracted the largest connected component (Section 2.1.1.3) resulting in a network with $|V| = 3,737$ and $|E| = 6,219$. It consists of 854 complexes ($V_c$), 590 diseases ($V_d$) and 2,293 genes ($V_g$), an illustration is given in Figure 3.10.

3.4.1 Parameter determination

An important feature of many biological networks is their hierarchical organization, where higher-level structure is composed of multiple instances of a lower-level structure of different types [13] (Section 2.1.2.3 and Chapter 4). A good example is shown in the illustrations to modularity, where systems are repeatedly connected to form other systems [1]. This implies that small groups of nodes organize in a hierarchical manner to increasingly larger groups on many different scales [1, 29]. To account for this topological characteristic we have to be able to extract relevant information on an appropriate, pre-defined resolution level.

We used the CORUM core set as of July 2009 [153].
Figure 3.10: Illustration of the gene-disease-protein complex graph. Here, the largest connected component of the layered, tripartite graph gene-disease-protein complex network is shown. It consists of 2,293 gene (green), 590 disease (red) and 854 complex (blue) nodes connected by 6,219 edges.

We addressed this issue by analyzing the very global structure and a detailed local level of the disease-gene-protein complex network. In the following, we first present the results of a decomposition into large clusters, which demonstrates that our method is generally applicable to biological data. Then, we discuss smaller clusters that allowed for a precise interpretation of single elements.

As discussed before, due to its random initialization our algorithm is inherently indeterministic. Different clustering results have of course a significant impact on the interpretation of the biological meaning of the results. We already showed that our algorithm is quite stable on graphs with well defined cluster structure (Section 3.3.2). Hence, we verify that the disease-gene-protein complex network has indeed a defined cluster structure. To avoid analyzing a local minimum, we discuss performance over 10 runs.
3.4 Decomposition of a gene-disease-protein complex graph

Dealing with a theoretically monotonous cost function, it is hard to determine the optimal numbers of clusters for each node type in which the graph has to be partitioned. Even in the case of unipartite $k$-means there is no direct and computationally simple answer (Section 2.2.2.2). Appropriate values are not apparent from prior knowledge about our data set. We therefore chose desired approximate resolutions $m_g$ for the gene partition. For large-scale clustering, we approximated the number of clusters to be found for each node type by limiting the maximal number of gene clusters $m_g$ for $V_g$ to

$$m_g = \lfloor \sqrt{|V_g|}/2 \rfloor$$

as suggested by Mardia et al. [159]. The number of complex clusters $m_c$ for $V_c$ and disease clusters $m_d$ for $V_d$ were then scaled according to $m_g$ by

$$m_i = \lceil m_g \sqrt{|V_i|/|V_g|} \rceil,$$

where $i \in \{c, d\}$. We use this heuristic, since a brute-force\(^{51}\) sampling of the three-dimensional parameter space is computationally out of reach. Then, we looked for plateaus and steep drops in the cost function within a certain range around this value $m_g$, and chose a local optimum of the algorithmically found decompositions. The simulations – illustrated in Figure 3.7 – show that the profile of the cost function may indeed indicate for a proper number of clusters in graphs with known cluster structure.

3.4.2 Large-scale clustering

First, we focused on the identification of large clusters. Figure 3.11b shows the distribution of the cost function values after algorithm convergence for each parameter setting. In the following discussion, we used $(m_g,m_c,m_d) = (10,5,6)$ as it showed the first steep drop in the cost function. Moreover, here we observed a significant local minimum of the cost function values of the algorithmically determined decompositions.

From the illustration of the decomposition in Figure 3.12, we see that the resulting clusters vary strongly in size. For all partitions, the majority of elements was assigned to a single cluster with degree of membership $\mu > = 0.9$. This demonstrates that the analyzed graph has a well defined cluster structure at the desired resolution level. The corresponding histograms are given in Figure 3.13.

As discussed before, such large degrees of membership are rarely found in graphs lacking any cluster structure. This gives evidence for the presence of a well defined cluster structure on the desired resolution level. But there exists also a considerable amount of elements assigned to several clusters simultaneously, e.g., in complex clusters 3 and 5, gene clusters 1 and 3, or disease clusters 3 and 4. This shows the need for a fuzzy approach.

\(^{51}\) Brute-force is a trivial, but very general problem-solving technique that consists of systematically enumerating all possible candidates for the solution [4].
3 Insights into the structural organization of biological data via fuzzy clustering of $k$-partite graphs

Figure 3.11: Decomposition of a gene-disease-protein complex network. We integrated the gene-disease network from [17] with human protein complexes from the CORUM database [153]. This resulted in a layered tripartite graph, which is schematically drawn in (a). We performed a 10-fold approximation of this graph to estimate appropriate numbers of clusters. The boxplot curve (b) shows how the cost function $f(H,C)$ from Equation 3.1 depends on the number of gene clusters $m_g$. The true minima of the cost function are decreasing with $m_g$, and this is also visible in the approximated minima using our proposed algorithm. Therefore, we are able to identify structures on various resolution levels. The details represent the cost function course for large-scale clustering (i) and a decomposition on small scale (ii), respectively. For our detailed analyses, we used the decompositions showing steep drops in the cost function marked by the red and green boxes.

3.4.2.1 Cluster evaluation

To determine whether the resulting clusters are biologically reasonable, we applied GO enrichment analysis to the clusters of the gene partition. To this end, the genes used in the analysis (degree of membership $\mu > 0.2$) were tagged with their respective GO categories and analyzed within each cluster for overrepresentation of certain categories versus the "background" level of the population (in this case, all genes in the tripartite graph).

We used Ontologizer [160] with the setting "Parent-Child-Intersection" restricting the analysis to the biological process category. To correct for multiple testing we employed Bonferroni correction. To assign GO terms to gene sets, a $p$-value cutoff of 0.05 was used.

We found, for instance, that for the genes in the two overlapping clusters 1 and 3 significantly enriched GO-terms are cell cycle and cellular response to stimulus/stress. Genes in cluster 4 can be related to, e.g., death, cell proliferation and developmental processes, whereas cluster 6 represents translation. Gene expression-associated GO-terms such as RNA processing and splicing were detected in cluster 7. This shows that our method is able to identify biologically meaningful functionally enriched clusters.
Figure 3.12: Illustration of large-scale cluster structures in the gene-disease-protein complex network. The large-scale decomposition of the gene-disease-protein complex network is illustrated. The hierarchical clustering of the nodes’ degrees of membership of the (a) complex, (c) gene and the (d) disease partition show that the majority of elements was assigned to single clusters. However, a considerable amount of cluster overlaps exists, e.g., for the disease clusters 3 and 4. The backbones for gene-complex (b) and for gene-disease (e) are sparsely connected, but show that locally overlapping clusters tend to interconnect with the same clusters of the other partition; e.g., disease cluster 3 and 4 are both connected to gene cluster 9 with large weights.
3 Insights into the structural organization of biological data via fuzzy clustering of \( k \)-partite graphs

Figure 3.13: The degrees of membership from the large-scale clustering of the gene-disease-protein complex graph: (a) genes, (b) protein complexes, (c) diseases. The U-like shape gives evidence for the existence of a well defined cluster structure. Again, for better recognizability, we counted only entries \( \geq 0.01 \).

3.4.2.2 Backbone evaluation

The interconnectivity of the in total 21 clusters is sparse (Figure 3.12). The skeleton for the global cluster structure for both underlying bipartite graphs (gene-complex and gene-disease) demonstrates that locally overlapping clusters also tend to interconnect with the same clusters of the other partition, e.g., disease clusters 3 and 4 are both connected with gene cluster 9. To evaluate the extracted backbone graph, we tested the hypothesis that interconnected clusters of different partitions are also functionally correlated. If the resulting interconnected gene and complex or disease clusters are functionally related, one expects to see a similar profile for functional annotation and backbone interconnectivity of each cluster.

**Evaluation strategy** For backbone (cluster interconnectivity) evaluation, we employed FunCat\(^{52}\) classifications for all genes and protein complexes. We used FunCat\(^{52}\) because gene ontology (GO) associations for genes could be mapped to their according FunCat categories, but not vice versa. A subset of 13 main categories\(^{53}\) was used, subcategory annotations were mapped to corresponding main category terms. Disorder classifications\(^{54}\) for genes and diseases were taken from Goh et al.\(^ {17} \), where classification classes grey and multiple were combined for pleiotropic\(^ {55} \) genes. We calculated Pearson’s correlation coefficients between cluster FunCat/disorder annotations by weighting a cluster element’s classification by its degree of membership to the particular cluster. The difference score between normalized backbone interconnectivity and annotation correlation was determ-

---

52 http://mips.gsf.de/proj/funcatDB/
53 F02 Energy; F10 Cell Cycle and DNA Processing; F11 Transcription; F12 Protein Synthesis; F18 Regulation of Metabolism and Protein Function; F20 Cellular Transport, Transport Facilities and Transport Routes; F30 Cellular Communication/Signal Transduction Mechanism; F32 Cell Rescue, Defense and Virulence; F41 Development systemic; F42 Biogenesis of Cellular Components; F43 Cell Type Differentiation; F45 Tissue Differentiation; F47 Organ Differentiation
54 Bone; Cancer; Cardiovascular; Connective tissue; Dermatological; Development; Ear,Nose,Throat; Endocrine; Gastrointestinal; Hematological; Immunological; Metabolic; Multiple; Muscular; Neurological; Nutritional; Ophthalmological; Psychiatric; Renal; Respiratory; Skeletal; Unclassified; Grey
55 A gene is defined to be pleiotropic if it is associated with more than one disease independent of causal relationships or its function (see also Chapter\(^{5}\)).
in using the Frobenius norm of their difference.

The null models for the evaluation of the backbone graph and the quantification of the results’ statistical significance were generated by applying a weighted bipartite randomization procedure to each partition-cluster subgraph $C^{(i)}$. To this end, we generalized the degree preserving rewiring of complex networks first introduced by Maslov and Sneppen [162]. In the weighted case, one has to decide between preserving either the number of neighbors of all nodes, or the total weight of their adjacent edges. We chose to maintain the first quantity: In every randomization step we randomly picked two edges and exchange their endpoints of the partition type, thereby keeping the weights attached to the edges. With this we also conserved the weighted degree of the partition nodes which reflects the right-stochasticity of the fuzzy clusterings. The degree of randomization can be monitored by a loss of degree-correlations between first and second neighbors. In practice, correlations vanish after about one randomization step per edge. So, for our analyses we used five times this number as suggested in [163]. The $p$-values were calculated over 10,000 runs.

**Evaluation results**

**Gene–complex interconnections:** The evaluation results are shown in Figure 3.14. Here, we find, for instance, that complex cluster 3 and the interconnected gene clusters 1 and 3 show a high binary FunCat correlation. The difference score between backbone interconnectivity and annotation correlation is 2.48, resulting in a $p$-value $< e^{-0.05}$. To compare the results of the fuzzy clustering approach with the results for the disjoint clustering method from Long et al. [151] we applied the algorithm with the same parameter settings and identical annotation and randomization procedure to the obtained clusters. For hard clustering, we achieved a larger difference score of 2.99 which corresponds to a significant $p$-value of 0.0015.

**Gene–disease interconnections:** To ascertain that our method is able to detect biological feasible clusters in all partitions, we also determined for each gene and disease cluster disorder class profiles. Again, we observed a high similarity between backbone interconnectivity and disorder correlation having a difference score of 1.09 ($p$-value $< e^{-0.05}$). For instance, gene cluster 1 and 10 and the interconnected disease clusters 1 and 5 show a high disorder correlation (Figure 3.14). The results of the hard clustering approach show a larger difference score of 2.24 and a still significant $p$-value of $2 \cdot e^{-0.04}$. We see that the annotation correlations between connected clusters of the backbone graph is higher when applying the fuzzy approach.

**3.4.3 Small-scale clustering**

We showed that our method is able to detect and interconnect biologically meaningful clusters. However, due to their size of about 279 genes on average the single clusters are
3 Insights into the structural organization of biological data via fuzzy clustering of \( k \)-partite graphs

Figure 3.14: Evaluation of the backbone of the gene-disease-protein complex network. To evaluate the large-scale clustering we additionally included functional annotations. (a) and (b) compare the gene-complex backbone graph with the functional correlations of the extracted clusters according to FunCat annotation. Similarly, (d) and (e) show the gene-disease backbone and the clusters’ disorder class correlations. We see that interconnected clusters also seem to correlate in their annotations. To test this hypothesis rigorously, we calculated difference scores to quantify the correlation of the backbones and their annotations, respectively. Vertical lines in (c) and (f) correspond to these difference scores for the fuzzy (black) and the hard (red) clustering. Comparing these values to the difference scores for \( e^{-0.05} \) randomized cluster assignments we obtain significant \( p \)-values, both \( < e^{-0.05} \). The correlations between annotations of connected clusters of the backbone is higher when applying the fuzzy approach.

hard-to-interpret. The detection of smaller clusters representing biological units enables a more precise biological interpretation. To detect smaller clusters, we set the maximum number of gene clusters to \( m_g \) for \( V_g \) according to

\[
m_g = \lceil \frac{|V_g|}{10} \rceil.
\]

Parameters \( m_c \) and \( m_d \) for \( V_c \) and \( V_d \) were scaled as previously. This resulted in a minimum average cluster size of 10 genes. This number seems to be biological meaningful as most functional complexes contain 10 to 30 protein components [164,165]. Furthermore, genes (e.g. TP53) are involved in up to 10 disorders [17]. In the following, we describe results for \( (m_g,m_c,m_d) = (222,135,112) \), where we found the lowest value of the cost function (Figure 3.11b). This setting accounts for an average cluster size of 10 genes in the gene partition.

To make use of the cluster overlaps, we looked for genes assigned to more than one cluster with a degree of membership of \( \mu > 0.2 \). We considered this threshold as significant as it is 50-fold higher than assigning each gene uniformly to all 222 gene clusters with equal
degree of membership of $\mu = 0.0045$.

As a showcase, we chose MECP2\textsuperscript{56}, a protein that functions as a key factor in epigenetic transcriptional regulation. It is known to be involved in neurodevelopmental and psychiatric disorders such as autism, mental retardation, and Angelman syndrome\textsuperscript{17, 166, 167}, and was assigned to three distinct gene clusters: 25 ($\mu = 0.42$), 32 ($\mu = 0.31$), and 200 ($\mu = 0.24$). These clusters mainly cover neurological (23%), psychiatric (81%), and pleiotropic (7%) genes having a degree of membership of $\mu > 0.2$. This is illustrated in Figure 3.15 where we visualized the backbone interconnectivity and the fuzzy clustering of the nodes in the neighborhood of MECP2.

We then analyzed the nine disease clusters interconnected with the three gene clusters in the backbone network. In total, 45 disorders representing mainly psychiatric (66%) and neurological (20%) disorders were assigned to eight disease clusters with a degree of membership of $\mu > 0.2$. Six out of nine psychiatric disorders available in the network are present in three disease clusters.

Another large fraction of these disease clusters are disorders classified as multiple. Most of them, e.g., Shprintzen-Goldberg syndrome or Aarskog Scott syndrome, show also neurological phenotypes such as mental retardation\textsuperscript{168, 169}. We also identified the ophthalmological disorder blepharospasm, an adult-onset focal dystonia that causes involuntary blinking and eyelid spasms\textsuperscript{170} for that a known polymorphism in the DRD5\textsuperscript{57} is associated with\textsuperscript{171}. Blepharospasm is a subform of dystonia and classified as a neurological disorder (ICD-10 G24.5)\textsuperscript{58} by the WHO\textsuperscript{59, 172}.

Furthermore, we found anorexia nervosa to be present in the analyzed clusters. It is annotated as a nutritional disorder by Goh et al.\textsuperscript{17}, however, it represents a life-threatening complex psychiatric disorder\textsuperscript{173}. Another so far unclassified disease alcohol dependence was assigned to the interconnected cluster. It is classified as a mental and behavioral disorder (ICD-10 F10.2), but in a broader sense, it can be considered also as psychiatric disorder. For alcoholism a high comorbidity rate with psychiatric disorders has been described\textsuperscript{174, 175}.

In contrast, applying the hard clustering algorithm, MECP2\textsuperscript{56} was assigned only to a single gene cluster, which is connected to two disease clusters. Although all associated disorders were identified correctly, no further information could be obtained from the clusters. Samaco et al.\textsuperscript{166}, however, reported an epigenetic overlap in autism-spectrum neurodevelopmental disorders as MECP2 affects the regulation of UBE3A\textsuperscript{60} expression. These relations became immediately apparent in the cluster result of our fuzzy approach: both genes were mutually assigned to gene cluster 25 that identifies the phenotypic and genotypic overlaps, whereas direct links to known connected genes are missing in the hard clustering (Figure 3.16).
3. Insights into the structural organization of biological data via fuzzy clustering of k-partite graphs

Figure 3.15: The small-scale clustering in the neighborhood of MECP2. We draw the results – the backbone network and the nodes’ degrees of membership to clusters, thresholded by $\mu > 0.2$ – of the small-scale clustering in the neighborhood of MECP2 using the fuzzy and the hard clustering (Figure 3.16). Nodes are colored according to their disorder class annotations. Blue edges indicate backbone interconnectivity, grey edges cluster assignment. Edge thickness indicates the degree of membership. MECP2 is connected to three gene clusters mainly covering neurological (red) and psychiatric (purple) genes. The seven interconnected disease clusters also represent mainly psychiatric and neurological disorders. Also unclassified disorders are present such as, e.g., Alcohol dependence (white), which is classified as a mental and behavioral disorder. In a broader sense, however, it can be considered as psychiatric disorder. For a color legend see Figure 3.16.
3.4 Decomposition of a gene-disease-protein complex graph

Figure 3.16: The small-scale clustering in the neighborhood of MECP2. Applying the hard clustering, MECP2 is assigned to gene cluster 209, which is connected to two disease clusters only. Although all associated disorders are identified correctly, no further information can be obtained from the decomposition in contrast to the fuzzy clustering.
3.5 Summary and future perspectives

The widespread application of high-throughput methods such as microarrays or next generation sequencing has considerably increased the amount of experimental data and the information available in biomedical literature, which is accessible to text-mining approaches \[176\]. These data can usually be represented in terms of networks. Over the last years, networks have emerged as an invaluable tool for describing and analyzing complex systems. However, we need to take into account that network information is commonly available for various types of nodes. Especially integrative biological networks are \(k\)-partite \[17, 177\].

Another important feature of biological networks is their hierarchical organization (Section 2.1.2.3 and Chapter 4), implying that small groups of nodes organize in a hierarchical manner to increasingly larger groups on many different scales \[1, 13, 29\]. This necessitates the analysis of these objects on various resolution levels. Furthermore, many proteins or genes are pleiotropic, and often associated with many functions (Chapter 5). Hence, clustering algorithms that assign elements into several functional modules are essential \[37, 39, 150\].

In this Chapter, we presented a novel, computationally efficient and scalable graph clustering algorithm that is capable to deal with all these described issues. Further, it does not require any \textit{a priori} knowledge about the data set to be analyzed. Results on a tripartite network, constructed by integrating the human disease network with protein complexes, demonstrated that we could identify and interconnect biologically meaningful clusters on different scales. Overlapping modules gave a more comprehensive picture of, e.g., gene-disease connections than looking at disjoint clusters alone. Summarizing, the proposed fuzzy clustering algorithm is suitable to compress and approximate the underlying topology of heterogeneous biological networks, which facilitates the understanding of such networks on multiple scales.

In the future, this algorithm can be generalized to allow links within a partition enabling the application to uni-partite and not solely \(k\)-partite graphs. This will allow to include information on PPI or regulatory interactions between TF's into the studied biological networks. Another possible extension is to allow a mixture of directed and undirected links.

To explore whether the algorithm may also help to detect synonyms in large networks extracted by text mining strategies \[176\], we have to address another important issue: the inclusion of a procedure to overcome the parameter determination problem in order to scale down the calculation time and to improve the robustness of our method. Besides parallelization, extension to Bayesian NMF models \[35\], or the utilization of Ensemble methods \[178\], would be the first steps towards this direction.
CHAPTER 4

The hierarchical organization of regulatory networks

A basic way to understand the architecture of biological networks is to consider their topology. Topological analysis is able to capture functional properties and important architectural features of the network (Section 2.1.4) [179], and brings significant advances to our understanding of complex systems. As introduced in Section 2.2 modular decomposition is a well-established approach on this [23].

Previously in Chapter 3, we presented a novel method for decomposing $k$-partite networks into overlapping communities, thereby allowing to structure heterogeneous information even on a large-scale. Hierarchy, which captures inherent dependencies, is a central organizing principle of complex networks, capable of offering insights into many network phenomena (Figure 1.1). Hierarchical structure, however, goes beyond simple clustering by explicitly including organization at all scales in a network simultaneously [13]. Topological analyses have suggested the existence of hierarchical modularity in TRNs of many organisms, and previous works have proposed methodologies from which this organization could be inferred suggesting the existence of a pyramidal top-down hierarchy [14].

In this Chapter, I present HiNO (Hierarchical Network Organization) - a method that directly reveals the hierarchical structure from regulatory networks by considering the occurrence of network motifs (Section 2.1.4), and thus overcomes the limitations from existing methods.61 In Section 4.1 I introduce the method and evaluate it using GRNs from *Saccharomyces cerevisiae* and *Escherichia coli*, respectively. Subsequently in Section 4.2 I apply HiNO to GRNs from *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. Focusing on the global organization, I show that the deduced hierarchical structures have a pyramidal shape and are not random. By systematically analyzing the single hierarchical layers and thus focusing on the microscopic view, I find that their inherent properties/characteristics are quite different. The hierarchical structure correlates with existing biological processes and thereby allows for novel classifications of TFs and their associated elements. Taken together, this represents a step forward towards a better understanding of complex systems in response to, e.g., network perturbations.

---

61 Chapter 4 led to the publication Hartsperger et al. [15].
4 The hierarchical organization of regulatory networks

4.1 Extracting the hierarchical organization of regulatory networks

One of the fundamental problems in systems biology is the understanding of regulatory networks and the interdependencies between their components. Of particular importance is how TFs mutually coordinate the expression of thousands of genes and non-coding RNAs (ncRNAs) such as miRNAs or long ncRNAs in response to various stimuli. The qualitative interaction between TFs, miRNAs, and their targets can be modeled in terms of directed regulatory networks (Section 2.1).

The determination of hierarchical interdependencies within GRNs is important for the understanding of the potential impact of perturbations on underlying cellular processes and their correlated diseases [6, 117, 181, 182]. Yu and Gerstein [16] pointed out that in contrast to well investigated undirected biological networks the superior feature of GRNs is their directedness and hence their orientation towards control rather than communication. The direct implication is a hierarchical organization, where a higher-level structure results from multiple instances of lower-level structure of different types [13]. As a consequence, small groups of nodes organize in a hierarchical manner to increasingly large groups on many different scales [1, 29, 183] (Section 2.1.2.3).

It is evident that the correct determination of the overall topology of the network as well as the unambiguous prediction of the position of individual regulatory elements are essential for the comprehensive understanding of regulatory effects on a systems-level (Figure 4.1). For instance, master-regulators such as HOXA10 (homeobox A10) display significant downstream effects in postnatal hematopoietic development depending on different concentrations of the key regulator [184], see also Section 4.2.4.2. This can lead to complete blocking of erythroid and megakaryocyte development for high concentrations of HOXA10 whereas intermediate concentrations result in increased stem cell proliferation. In contrast to master-regulators, the so-called mid-level regulators can not only be mediators of different incoming regulatory signals, but also influence multiple downstream components or pathways. In the case of TP53 (tumor protein p53) – an important factor in cancer development – it has long been known to be influenced by multiple upstream signals such as stress leading to different downstream effects such as apoptosis or development of cancer [185, 186].

The detailed understanding of the hierarchical topology is essential to understanding the dynamic behavior and the possible malfunction of regulatory networks. This can be achieved only by the accurate determination of the hierarchical organization.

Several methods exist that identify the hierarchical structure within directed regulatory networks such as the leaf removal algorithm [187, 188], or the breadth first search (BFS) method [16]. Although these methods are capable of extracting hierarchical structures from GRNs, they have several shortcomings as discussed by Jothi et al. [189]: they either extract the hierarchical structure incorrectly or they are not scalable. Another important drawback is that they are only partly applicable to networks containing loops. Loops, however, are an important feature of regulatory networks [33] (Section 2.1.4). Network
4.1 Extracting the hierarchical organization of regulatory networks

**Figure 4.1: Hierarchical structure clarifies regulatory interdependencies.** An abstract example of a directed regulatory network is shown (a) unstructured and (b) hierarchically structured. We schematically show in (c) and (d) that for master regulators such as \( \text{HOXA10} \) significant downstream effects in postnatal hematopoietic development depending on different concentrations can be observed. In contrast to master regulators so called mid-level regulators can not only be mediators of different incoming regulatory signals, but also influence multiple downstream components or pathways. We schematically display that \( \text{TP53} \) is influenced by various upstream signals such as (e) DNA damage or (f) oncogenes leading to different downstream effects such as apoptosis or development of cancer. The detailed understanding of the hierarchical topology is necessary to comprehend the dynamic behavior and the possible malfunction of regulatory networks. For a similar illustration see also del Sol et al. [6].
motifs represent the building blocks of complex networks and their properties determine the local and global organization [5, 34]. Common networks motifs are FFLs or FBLs. The leaf-removal algorithm that was used for network decomposition (Section 2.2) and to infer hierarchical structures in biological networks [187, 188] is therefore not applicable to GRNs containing loops. The BFS approach presented by Yu and Gerstein [16] can be applied to cyclic networks, but the inferred hierarchical structure may contain conflicts in the level assignment. Jothi et al. [189] presented a method capable of dealing with these shortcomings, however, it does not determine a clear hierarchical position for each vertex in the network, rather vertices are assigned to an interval of possible positions. The algorithm does not count or enumerate all feasible topological orderings of nodes in a network because this is a non-deterministic polynomial-time hard (NP-hard) problem. It only outputs a linear ordering of nodes containing all feasible solutions rather than reporting just a single solution [189]. Subsequently the final transformation into a graph has to be done manually, a quite cumbersome procedure for large regulatory networks.

Here, I present HiNO an algorithm for inferring the hierarchical organization from regulatory networks. This algorithm represents a significant improvement of the BFS approach and directly reveals the hierarchical structure from GRNs by considering the occurrence of network motifs. This idea is implemented by expanding the BFS method with two correction steps, a downward and an upgrade procedure as shown in the next Section.

Therefore, the accurate hierarchy of a GRN that we define as its hierarchical structure without having any inconsistencies in the level assignments can be determined.

4.1.1 Algorithm formulation

HiNO – an algorithm for inferring hierarchical organization from regulatory networks

A directed network is a graph \( G = (V; E) \) of directed edges \( E \) between a set of vertices \( V \) (Section 2.1.1). Given a directed network, our algorithm first identifies all vertices at the bottom representing the lowest level in the hierarchy. A vertex is assigned to the bottom level if and only if it does not regulate any other vertices or if it is only regulating itself (i.e., auto-regulation). In the next step, we traverse the network into a shortest-path-tree [62, 25]. Here, we define the level in the hierarchy for any non-bottom vertex as its shortest distance to a bottom vertex. Now, each vertex is assigned to a certain level within the hierarchy. However, in the case of occurring loops it can be observed that regulators have a lower level annotation than their targets [16]. To account for the occurrence of loops in the network (such as feed-forward loops) and to extract the underlying hierarchical structure accurately, we added a recursive correction procedure consisting of two steps:

1. Vertices are assigned to the lowest level that this and its regulating vertices are assigned to. This downward step resolves conflicts in level assignments where regulators have a lower level annotation than their targets.

---

62 A shortest path tree, in graph theory, is a subgraph of a given (possibly weighted) graph constructed so that the distance between a selected root node and all other nodes is minimal [25] (Section 2.1.3).
2. Vertices with no predecessors are placed to the next higher level if some of their successors are located on the same level. In this upgrade step, vertices that are regulators only get the next higher level annotation than their targets.

Finally, the hierarchical structure of the directed regulatory network is determined without conflicts in the vertices’ level assignments. The hierarchy extraction procedure is illustrated in Figure 4.2, a pseudo code representation of HiNO is shown in Figure 4.3.

Figure 4.2: Illustration of the hierarchy extraction approach. We illustrate the approach of hierarchy extraction from directed networks. In the first step, the original network (grey) is traversed into a shortest-path-tree (step1) with an initial hierarchical organization consisting of three layers. As this alone does not capture all feasible solutions we introduce two correction steps to resolve conflicts in the level assignments (e.g. node 1 is located in level 2, but it is regulating node 4 that is located in level 3). First, in the downgrade step each vertex is placed on the lowest level that this vertex and its regulating vertices are assigned to (step2). Thereby it is assured that each regulator has at least the same level assignment as its targets. Here, node 4 and node 5 are downgraded from level 3 to level 2 as its regulators - node 1 and node 3 - are assigned to level 2. Second, vertices are upgraded if they have successors with the same level assignment and have no regulators themselves. Here, node 1 has no predecessor and targets on the same level (level 2). Consequently it is upgraded from level 2 to level 3 (step3). In the end, HiNO determined a three layered hierarchical structure of the directed network without conflicts in the level assignment.
Algorithm 1 HiNO - Hierarchical network organization of directed networks

Input: Graph \( G = (V, E) \)
Output: Hierarchical structure of \( G \)

//determine set of bottom vertices
1 \( B \leftarrow \emptyset \);
   foreach \( v \in V(G) \) do
      if \( \text{outdegree}(v) = 0 \lor (\text{outdegree}(v) = 1 \land \text{selfloop}(v)) \) then
         \( B \leftarrow B \cup \{v\} \)
      \( \text{level}(v) = 0 \)
   end
   //calculate shortest path to bottom vertices
2 foreach \( v \in V(G) \land v \notin B \) do
   \( l = \text{shortestpath}(v, B) \)
   \( \text{level}(v) = l \)
end
//downgrade step
3 \( D \leftarrow V(G) \);
   while \( D \neq \emptyset \) do
      foreach \( v \in D \) do
         if \( \text{level}(v) > \min(\text{level}(\text{predecessor}(v))) \) then
            \( \text{level}(v) = \min(\text{level}(\text{predecessor}(v))) \)
         else
            \( D \leftarrow D \setminus \{v\} \)
         end
      end
   end
//upgrade step
4 \( U \leftarrow V(G) \);
   while \( U \neq \emptyset \) do
      foreach \( v \in U \) do
         if \( \text{numpredecessors}(v) = 0 \land \text{level}(v) = \max(\text{level}(\text{successors}(v))) \) then
            \( \text{level}(v) = \max(\text{level}(\text{successors}(v))) + 1 \)
         else
            \( U \leftarrow U \setminus \{v\} \)
         end
      end
   end

Figure 4.3: Pseudo code of the algorithm. Pseudo code for extracting the hierarchical organization of directed networks.
4.1 Extracting the hierarchical organization of regulatory networks

4.1.2 Algorithm evaluation

To evaluate HiNO the GRNs for *E. coli* and *S. cerevisiae* provided by Yu and Gerstein were reconstructed. The E. coli gene regulatory network (EGRN) consists of 2,044 regulatory interactions between 143 TFs and 1,051 targets and the yeast gene regulatory network (YGRN) contains 8,371 interactions involving 286 TFs and 3,369 targets.

Subsequently, sub-graphs were extracted containing TFs and regulatory interactions between TFs only. The hierarchical topology is deduced only from the regulatory interdependencies of transcription factors. The extracted TF-subgraph of *E. coli* (or ETRN) consists of 200 interactions between 143 TFs with a total of 77 self-loops (auto-regulatory edges). The extracted TF-subgraph of *S. cerevisiae* (or YTRN) consists of 604 interactions involving 286 TFs with a total of 30 self-loops. In Figure 4.4 the different analysis steps are shown, precise numbers are given in Table 4.1.

To evaluate our method, we compared the hierarchical structures extracted by HiNO with the results when applying the BFS method of Yu and Gerstein. For detailed numbers see Table 4.2.

<table>
<thead>
<tr>
<th>Organism</th>
<th>GRN nodes</th>
<th>GRN edges</th>
<th>GRN TFs</th>
<th>GRN genes</th>
<th>GRN targets</th>
<th>GRN TFs</th>
<th>GRN edges</th>
<th>GRN selfloops</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>1,095</td>
<td>2,044</td>
<td>143</td>
<td>952</td>
<td>1,052</td>
<td>143</td>
<td>200</td>
<td>77</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>3,458</td>
<td>8,371</td>
<td>286</td>
<td>3,172</td>
<td>3,369</td>
<td>286</td>
<td>604</td>
<td>30</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>3,307</td>
<td>4,039</td>
<td>1,804</td>
<td>1,503</td>
<td>1,050</td>
<td>1,804</td>
<td>1,119</td>
<td>–</td>
</tr>
<tr>
<td>M. musculus</td>
<td>2,653</td>
<td>2,391</td>
<td>1,587</td>
<td>1,066</td>
<td>716</td>
<td>1,587</td>
<td>695</td>
<td>–</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>2,092</td>
<td>1,689</td>
<td>1,439</td>
<td>653</td>
<td>523</td>
<td>1,439</td>
<td>456</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 4.1: Number of nodes and edges for GRNs and TRNs. For the five organisms detailed numbers are shown. For the GRNs the number of nodes and edges are shown. Furthermore, how nodes split into TFs and genes and how many of the nodes are targets meaning they get regulated independently if they are TFs or not. For the TRNs the number of edges between TFs only is given. Besides, for *E. coli* and *S. cerevisiae* the networks contain self-loops as in those organisms auto-regulation exists.

---

63 Yeast
64 Downloaded from [http://www.gersteinlab.org/proj/nethierarchy/](http://www.gersteinlab.org/proj/nethierarchy/)
65 ETRN E. coli transcriptional regulatory network
66 YTRN yeast transcriptional regulatory network
E. coli & S. cerevisiae

<table>
<thead>
<tr>
<th>Level</th>
<th>Elements(BFS)</th>
<th>Elements(HiNO)</th>
<th>Elements(BFS)</th>
<th>Elements(HiNO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>10</td>
<td>30</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>42</td>
<td>104</td>
<td>115</td>
</tr>
<tr>
<td>1</td>
<td>89</td>
<td>89</td>
<td>144</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 4.2: Level distribution of the TRNs. Comparison of the level distribution of the TRN in E. coli and S. cerevisiae retrieved by applying the BFS method and HiNO.

Figure 4.4: From the GRN to the hierarchical structure. Illustration of (a) the GRN of E. coli with 1,095 nodes and 2,044 edges and (d) the GRN of S. cerevisiae with 3,458 nodes and 8,371 edges. Red nodes indicate transcription factors, green nodes genes. For extracting the hierarchical structure we deduced transcriptional regulation only. The TRNs are shown in (b) for E. coli and (e) S. cerevisiae only. A snapshot of the hierarchical structure of the TRN is shown for E. coli in (c) and for S. cerevisiae in (f). The TRN of E. coli has a four-layered pyramidal-shaped hierarchical structure, whereas the TRN of S. cerevisiae consists only of three layers. The different colors represent the distinct hierarchical levels: level 4 – yellow; level 3 – red; level 2 – green; level 1 – blue.
4.1 Extracting the hierarchical organization of regulatory networks

4.1.2.1 Hierarchical organization in E. coli

Applying HiNO to the ETRN reveals a pyramidal-shaped hierarchical structure consisting of four levels with most TFs on the bottom level (here level 1) and only a few regulators on the top (level 4), details on the level distribution are shown in Table 4.2.

The global findings are consistent with the results when applying the method of Yu and Gerstein [16] where also four levels have been identified. The BFS method, however, incorrectly assigns four TFs to the top level, whereas our method only assigns two TFs to this layer. The misassigned two elements, gntR and yhiW are placed by HiNO on the next lower level due to the correction step procedure. Both TFs regulate yhiE assigned to level 3 by Yu and Gerstein [16]. This TF is regulated by several TFs assigned to level 2 and therefore the hierarchical structure is incorrect as a target has a higher level annotation than its regulator. In contrast, HiNO directly assigns yhiE to level 2 and subsequently downgrades the level annotation for its regulators gntR and yhiW to level 3.

In the upgrade step TFs are assigned to the next higher level if they are not regulated at all and if they have targets on the same hierarchical layer. Exemplarily, this is the case for yicE which is upgraded from level 2 to level 3. For an illustration see Figure 4.5.

4.1.2.2 Hierarchical organization in S. cerevisiae

Applying HiNO to the YTRN we extracted a pyramidal-shaped hierarchical structure consisting of three levels, for details see Table 4.2.

In comparison to this the method of Yu and Gerstein [16] reveals a four layered topology. The BFS method placed ADA2 on the top level (level 4), although it is regulated by RAP1 that is assigned to the second lowest level (level 2). This hierarchical assignment is inconsistent to the fact that a regulator has to be at least on the same hierarchical layer as its target. Similar conflicts can also be observed for SPT23, NGG1, GAT1 or MOT3 having all predecessors with lower level annotations (Figure 4.6). In contrast to this HiNO assigns the correct level in the hierarchy to these TFs. In the downgrade step, ADA2 is placed on the same level as its regulators RAP1 and FHL1 (level 2). Interdependencies are resolved subsequently by downgrading other factors such as RTG3 or MSN4. Afterwards, elements are upgraded to the next higher level if they have no upstream elements and if they are positioned at the same layer as one of their targets. This step is illustrated in Figure 4.6 for HAL9, which is upgraded from level 2 to level 3.
Figure 4.5: Comparison BFS-method vs. HiNO using the TRN of E. coli. A snapshot of the results (the comparison of the BFS-method and HiNO) using the ETRN is shown. Node colors indicate the level assignment in HiNO: blue - level 1; green - level 2; red - level 3; yellow - level 4. Dashed edges represent additional regulatory interdependencies (further up- and/or down-stream factors) that are not shown. The results of the BFS method are shown in (a). The node coloring indicates conflicts in level assignments. In (b) the result of the \textit{downgrade} step of HiNO is shown. In (c) the result of the \textit{upgrade} step is shown. This is also the final hierarchical assignment of the elements by HiNO.
Figure 4.6: Comparison BFS-method vs. HiNO using the TRN of S. cerevisiae. A snapshot of the results (the comparison of the BFS-approach and HiNO) using the YTRN is shown. Node colors indicate the level assignment in HiNO: blue - level 1; green - level 2; red - level 3; yellow - level 4. Dashed edges represent additional regulatory interdependencies (further up- and/or down-stream factors) that are not shown. The results of the BFS method are shown in (a). The node coloring indicates conflicts in level assignment. In (b) the result of the downgrade step of HiNO is shown. In (c) the result of the upgrade step is shown. This is also the final hierarchical assignment of the elements by HiNO.
4.2 Hierarchical organization of mammalian TRNs

In the previous Section, HiNO an approach that accurately deduces the hierarchical structure from GRNs without having any inconsistencies in the level assignments is presented (Section 4.1.1). Significant improvements of the BFS approach are achieved by expanding the BFS method with two correction steps, a downgrade and an upgrade procedure. Evaluating HiNO against two publicly available GRNs from *S. cerevisiae* and *E. coli* showed that the resolution of conflicting assignments clearly improves the BFS method (Section 4.1).

While the BFS method returns the same global hierarchical organization of the network, inconsistencies on the individual levels in local neighborhoods can be resolved by HiNO. Small local motifs and modules define the basic dynamic properties of regulatory processes (Section 2.1.4). Their correct identification is necessary for the deduction of experimentally testable hypotheses inferred from large data sets. Deducing the general topology of mammalian regulatory networks and analyzing systematically the properties of their hierarchical layers facilitates to structure and characterize regulatory networks. Especially the aim to understand complex systems and diseases – their pathobiology – requires an indepth-understanding of the underlying regulatory structure and its (inter-)dependencies.

A variety of GRNs is publicly available that are known to be hierarchical in their overall structure [182, 190]. At present the *sea urchin embryo endomesoderm* GRN is the only most nearly complete, predictively useful, and validated large-scale developmental GRN [182]. But we want to go beyond solely developmental GRNs as we are interested in a holistic, rather abstract picture of the underlying hierarchical organization. Here, we look from a more global perspective on averaged functional processes that are not resolved over time and space. Consequently, we neither consider any dynamics nor differentiate between tissues and cell types and states.

In the following, I analyze the hierarchical organization of the GRNs available for human, mouse, and rat. The deduced general properties and the distribution of essential and disease phenotypes over the hierarchy give first static and descriptive impressions towards an understanding of complex diseases.

4.2.1 Network construction

The GRNs for human, mouse and rat were reconstructed from the MatBase library. Transcription factors and genes represent the nodes and two nodes are connected by an...

---

67 Pathology, with emphasis on the pathophysiological and pathogenetic mechanisms underlying human disease [10].
68 MatBase library version 8.2 as of January 2010 (Genomatix Software GmbH). MatBase is a database containing manually curated information on regulatory interactions between transcription factors and their targets [191].
edge if a regulatory\textsuperscript{69} interaction between them has been described in the literature. The resulting GRNs\textsuperscript{70} (Table 4.1) are as follows:

- **HGRN**: 3,307 regulatory interactions between 1,804 TFs and 1,050 targets.
- **MGRN**: 2,653 regulatory interactions between 1,587 TFs and 716 targets.
- **RGRN**: 2,092 regulatory interactions between 1,689 TFs and 523 targets.

As before, we extracted TF-subnetworks consisting of regulatory interactions between TFs only. The resulting TRNs\textsuperscript{71} are:

- **HTRN**: 1,119 regulatory interactions between 1,804 TFs.
- **MTRN**: 695 regulatory interactions between 1,587 TFs.
- **RTRN**: 456 regulatory interactions between 1,689 TFs.

### 4.2.2 General properties of TRNs

After constructing the networks, we applied HiNO to the HTRN, the MTRN and the RTRN to deduce their hierarchical structure. All three TRNs show a pyramidal-shaped hierarchical structure where the number of TFs on each level is smaller than that of the previous level (Table 4.3). This is in line with the topology found in the ETRN and the YTRN (Section 4.1.2). Both, the HTRN and the MTRN have 5 layers, whereas the RTRN has a hierarchical structure consisting of 4 layers only. This is very likely due to the smaller data set available for the rat.

<table>
<thead>
<tr>
<th>Level</th>
<th>H. sapiens</th>
<th>M. musculus</th>
<th>R. norvegicus</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>116</td>
<td>88</td>
</tr>
<tr>
<td>2</td>
<td>216</td>
<td>144</td>
<td>111</td>
</tr>
<tr>
<td>1</td>
<td>1,432</td>
<td>1,314</td>
<td>1,234</td>
</tr>
</tbody>
</table>

Table 4.3: Level distribution of the TRNs. We applied HiNO to the TRNs of human, mouse and rat. For the HTRN and the MTRN 5 regulatory layers could be derived, whereas the smaller RTRN consists of 4 levels only. For each TRN the number of TFs per level (with 1 representing the bottom and 5 the top layer) are shown.

#### 4.2.2.1 Regulatory hierarchies are well organized

To investigate whether the observed hierarchical organization could be achieved also by chance, we compared our findings with the topology of random networks (results shown

---

\textsuperscript{69} We do not differentiate between the type of regulation and therefore merge \textit{activation}, \textit{inhibition} and \textit{regulation}.

\textsuperscript{70} HGRN: human gene regulatory network; MGRN: mouse gene regulatory network; RGRN: rat gene regulatory network

\textsuperscript{71} HTRN: human transcriptional regulatory network; MTRN: mouse transcriptional regulatory network; RTRN: rat transcriptional regulatory network
for human only). We therefore generated 1,000 random networks by randomly connecting TFs with target genes, while keeping the total number of TFs, target genes\(^72\) and edges constant. Important to note here is that this randomization procedure does not preserve the original degree distribution as in this case we expect to observe a similar topology, see also Liu et al.\(^{192}\).

We applied HiNO to these random networks and averaged the results. The illustration in Figure 4.7a demonstrates that in random networks the pyramidal-shaped hierarchical structure does not exist \(p = 0; \chi^2\)-Test). In random networks, layers are organized in an onion-shaped structure with most TFs in the second layer (Table 4.4) being consistent with the findings reported by Yu and Gerstein\(^{16}\). Furthermore, in more than 2/3 of the random networks more layers (up to 9) were found. This observation indicates that the identified pyramidal organization is not random and therefore suggests that the existing inherent topology of regulatory networks is of real biological importance.

![Figure 4.7: Level and target distribution of the HTRN.](image)

(a) For each level the number of elements (indicated by #) assigned to this layer are shown. The HTRN (blue) shows a pyramidal-shaped hierarchical organization, whereas the randomized networks (red) show an onion-shaped topology with up to 9 layers. (b) The average number of regulated elements (indicated by #) meaning how many targets a TF has on average in each level. In the HTRN (blue) the second layer has significantly more targets than the other layers, whereas random networks (red) show on average a more or less equal target distribution over the single hierarchical layers. For random networks averaged results and standard deviation (indicated as error bars) are shown for 1,000 runs each.

---

\(^{72}\) A target gene can be either a TF or a non-TF gene. Required is only that it is regulated by a TF.
4.2 Hierarchical organization of mammalian TRNs

Level H. sapiens Random Std

<table>
<thead>
<tr>
<th>Level</th>
<th>H. sapiens</th>
<th>Random</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>–</td>
<td>0.001</td>
<td>0.1</td>
</tr>
<tr>
<td>8</td>
<td>–</td>
<td>0.01</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>–</td>
<td>0.12</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>10.1</td>
<td>3.6</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>69.9</td>
<td>9.7</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>318.8</td>
<td>13.6</td>
</tr>
<tr>
<td>2</td>
<td>216</td>
<td>872.1</td>
<td>21.3</td>
</tr>
<tr>
<td>1</td>
<td>1,432</td>
<td>531.7</td>
<td>16.7</td>
</tr>
</tbody>
</table>

Table 4.4: Level distribution of the HTRN and comparison with random networks. The level distribution of 1,000 random networks compared to the original level distribution of the HTRN. Results were averaged and standard deviation (Std) was calculated. The numbers indicate that the level distribution found in the original data differs significantly from that of random networks.

4.2.2.2 Middle-layer regulators are more influential

We then examined the average number of targets for TFs at the different levels of the regulatory hierarchy to analyze the downstream impact of each layer to the network (results shown for human only). Again, we confirmed the findings of Yu and Gerstein as we found a similar relationship to the typical organization of social networks. For a detailed description on this we refer to the original publication [16]. Transcription factors located in the second level have on average the most targets (downstream regulatory interactions), whereas bottom layer TFs and those from the top layers have all fewer targets by and large. This suggests that TFs in the middle layer have a broader influence on the system and consequently, we expect them to be associated with more functions (Section 4.2.4). In contrast, TFs in the upper levels are expected to be more likely to be essential as they regulate only fewer targets, but as one can imagine they are more likely to be more important for the overall topology (Figure 4.1 and Section 4.2.3). Bottom layer TFs might be more specific as they represent the triggers for specific processes and are the last link in the downward causation process.

Again, these findings are not random. In rewired networks the target distribution is significantly different \(p < 8.2e^{-007}; \chi^2\text{-Test}\). Here, the average out-degree is almost constant in the different hierarchical levels (Figure 4.7b).

Another interesting aspect that has to be mentioned here is the controllability of complex systems. According to control theory, a dynamical system is controllable if, with a suitable choice of inputs, it can be driven from any initial state to any desired state within finite time [192]. Controllability is therefore a measure of what states can be achieved from a given set of initial states [193].
fully control the system and those so-called *driver nodes*\textsuperscript{74} tend to avoid the high-degree nodes \textsuperscript{192}. Interestingly, we found that hubs within the HTRN are mainly located in the second layer. This is plausible as the TFs in the higher levels have less regulatory interactions (Figure 4.7b).

4.2.3 Master-regulators are more likely to be essential

A gene that is absolutely required for survival, or a gene that strongly contributes to fitness and robust competitive growth is considered to be an *essential* gene \textsuperscript{194}. To analyze the correlation of essentiality with the hierarchical topology, we obtained essential genes for human and mouse from the Mouse Genome Database (MGD) \textsuperscript{193, 196}. Human and mouse orthologous genes (‘orthologs’) were identified through the Mammalian Orthology section of Mouse Genome Informatics (MGI). In the following, we use the phenotypes *embryonic lethal* (EL)\textsuperscript{75} and *postnatal lethal* (PL)\textsuperscript{76} to describe essentiality.

We mapped EL and PL phenotypes to all TFs in the HTRN and the MTRN and found that in total 670 (37\%) human and 664 (42\%) mouse TFs are annotated to be essential. More precisely, 31\%/35\% of the TFs have an EL phenotype and 17\%/19\% a PL phenotype in human and mouse, respectively. In general, it becomes obvious that for both organisms the overall fraction of TFs showing an EL phenotype is almost twice as high as for a PL phenotype.

Next, we compared these findings to expected values, meaning how many essential genes one would expect in each layer by chance. We observed for both organisms that in the bottom layer less and in the upper layers more essential TFs are located, for EL and PL phenotypes, respectively (Figure 4.8). This can be explained by the strong negative selection of EL genes in early developmental processes. They are very likely to cause the death of the organism, i.e., spontaneous abortions, which may account for as much as 20\% of recognized pregnancies \textsuperscript{197}. Transcription factors in the second layer show a high tendency to be essential as on average more than 50\% of the TFs have an EL or PL phenotype. This is in line with the previous findings that TFs in the second level have most targets underpinning their importance for the network organization. Besides, from this layer more specific processes might be regulated where no alternative routes are possible to overcome, e.g., misregulation or defects (see also Section 4.2.4). These regulators have no backup genes within the same level suggesting that in the event of deletion or mutation of these nodes there are no upper-level regulators that can compensate for their loss \textsuperscript{198}. This tendency of essentiality reflects also the role of master regulators that are intuitively more influential as they have a multitude of indirect targets \textsuperscript{198}. The consequence is a so-called *cascade effect* and in case of causative mutations\textsuperscript{77} all downstream genes are

\textsuperscript{74} Driver nodes can direct the network to a given behavior \textsuperscript{193}.
\textsuperscript{75} Embryonic lethal: death of an animal within the embryonic period prior to organogenesis (mouse: prior to E14) \textsuperscript{193}.
\textsuperscript{76} Postnatal lethal: premature death anytime between the neonatal period and weaning age (mouse: p1 to approximately 3 weeks of age) \textsuperscript{196}.
\textsuperscript{77} Causative means that somehow the function is influenced.
affected as well (see also Figure 4.1).

Figure 4.8: Distribution of lethal phenotypes. Distribution of the embryonic lethal (EL) and postnatal lethal (PL) phenotypes in the HTRN and the MTRN respectively. (a)-(d) show the absolute numbers (indicated by #) of essential TFs in each of the hierarchical layers, whereas (e)-(h) indicate the relative fraction (indicated by %) of TFs with an essential phenotype in each layer. The upper panel (blue) corresponds to the HTRN, the lower panel (green) to the MTRN. From this distribution it becomes evident that a considerable difference exists between level 1 and all other hierarchical layers. This indicates that any of the higher levels is more relevant than the bottom layer.
Although most essential TFs are located at the bottom level (in absolute numbers), we observed that the relative fraction of essential TFs in each layer is higher in the upper levels. In general, we observed a negative correlation between essentiality (the relative fraction of TFs in a layer that are essential) and the size of the hierarchical layers for both, human and mouse, with $R^2 = -0.73/R^2 = -0.84$ and PL: $R^2 = -0.40/R^2 = -0.59$). The obvious difference is between level 1 and all other hierarchical layers indicating that any of the higher levels is more relevant than the bottom layer. This is also in line with the observations by Jothi et al. [189] who found that the higher a TF is located in the hierarchy the more likely it is to be essential.

### 4.2.4 Higher level TFs have more functions

Intuitively and based on our previous results, we expect that TFs in higher levels are associated with more functions than bottom layer TFs. To test this hypothesis we mapped functional annotation in the form of GO terms to all TFs in the HTRN in order to analyze if a different pattern in functional activity and/or diversity can be identified. Mapping information was obtained using the BioMart [199] Web service at Ensembl [200]. For 95% of the TFs (1722/1804) functional annotation was available, whereat all unannotated TFs are located in the bottom layer. As mentioned before, we neither consider any dynamics nor differentiate between tissues and cell types and states here. Consequently, we look at a rather global perspective and therefore on averaged functional processes that are not resolved over time and space.

As expected, we observed that TFs in the second layer – that also showed the most targets (Section 4.2.2) – are associated with on average 18 associated GO terms and consequently with most functions. The high amount of hubs localized in this layer might account for that (Section 4.2.2.2). Compared to the bottom layer, where TFs are annotated with on average five GO terms, higher level TFs are associated with more functional terms (Figure 4.9a). This indicates that TFs in the upper levels tend to be functionally more versatile. They are more important and essential for the survival and adaption of an organism to its environment (see Section 4.2.3). Besides, the so-called master regulators affect much more downstream targets (both, directly and indirectly) than TFs on the bottom layer and therefore fulfill in total more and probably also more general functions, whereas bottom layer TFs are more likely to be rather specific.

#### 4.2.4.1 Hierarchy reflects general-to-specific pattern

Higher level TFs are expected to be involved in processes in early developmental and at large more general ones, whereas TFs in the bottom layer are supposed to be more specific and involved in final processes representing, e.g., the end-points in regulatory or signaling cascades.

---

78 Pearson correlation
4.2 Hierarchical organization of mammalian TRNs

Figure 4.9: Distribution of HTRN associated functions and pathways. (a) Distribution of the average number (indicated by #) of functions associated with the TFs in the single layers in the HTRN is shown. We used GO category biological processes. (b) Distribution of the average number (indicated by #) of KEGG pathways that are associated with the TFs in the single layers in the HTRN. From both distributions it becomes apparent that TFs in the second layer are associated with most functions and pathways, respectively.

To test this hypothesis, we extracted GO terms that are exclusively associated with bottom layer TFs and higher level TFs, respectively. For higher level TFs, we identified many functional terms describing developmental processes, regulation of differentiation/morphogenesis as well as response to different internal/external stimuli, but also terms involved in the reproduction process. Terms associated with the bottom layer are more specific, e.g., we identified processes such as fertilization, glucose transport, histone demethylation or zinc ion transport.

To analyze the TFs’ functional composition and to identify significantly overrepresented functions in each layer, we calculated an enrichment of GO functional categories of all layers relative to those in the bottom layer using Fisher’s Exact Test. Layer 5 with its master-regulator GZF1 (GDNF-inducible zinc finger protein 1), which is required for renal branching morphogenesis, is significantly enriched in branching involved in ureteric bud morphogenesis reflecting its important role in early developmental processes [201, 202]. In level 4, we found processes such as embryonic limb morphogenesis and the regulation of several differentiation processes. Similar processes are enriched in level 3, where we also identified processes involved in hematopoiesis or osteoblast development to be overrepresented. The second level is significantly associated with even more specific processes such as osteoblast differentiation, regulation of MAPKK, regulation of fatty acid oxidation or gluconeogenesis as well as different signaling pathways. These findings fit the hierarchical organization where information is processed from top to bottom.

Taken together, these observations support the hypothesis that higher level TFs fulfill

---

79 We used Benjamini & Hochberg procedure for multiple testing correction and as significance level a $p$-value $< 0.05$.
80 MAPKK: mitogen-activated protein kinase kinase cascade
more central and general functions (e.g., *early development*, *regulation of differentiation processes*), whereas bottom layered TFs are involved in more specific processes such as triggering *signaling pathways* or *cell proliferation*. Developmental genes seem to be overrepresented in the upper part of the hierarchical pyramid. We expect that their exclusion or their consideration from a dynamic point of view (in contrast to the static view described here) would have an influence on the shape of the hierarchical pyramid. This would lead to a more shallow hierarchical structure as developmental genes are often only used at specific developmental stages. More and especially time- and space-resolved GRNs are required to investigate this in more detail. Unfortunately, they are not available yet for the organisms analyzed here.

### 4.2.4.2 Hematopoiesis - information propagation correlates with the hierarchical structure

A nice example that and how information propagation in TRNs is correlated with their hierarchical organization is *hematopoiesis* – the formation of blood cellular components that are derived from hematopoietic stem cells (HSCs). Hematopoiesis comprises the differentiation of hematopoietic cells from at least 11 lineages from a small pool of self-renewing stem cells and the production of each cell type is highly regulated and responsive to environmental stimuli [203]. Hematopoietic stem cells are derived from ventral mesoderm and the definitive hematopoiesis involves the colonization of the fetal liver, thymus, spleen, and ultimately the bone marrow. In definitive hematopoiesis, long term hematopoietic stem cells (LT-HSC) give rise to short term hematopoietic stem cells (ST-HSC) that produce common myeloid progenitors (CMP) and common lymphocyte progenitors (CLP). Common lymphocyte progenitors are the source of committed precursors of B and T lymphocytes, whereas CMP give rise to megakaryocyte/erythroid progenitors (MEP) and granulocyte/macrophage progenitors (GMP). Granulocyte/macrophage progenitors develop to the committed precursors of mast cells, eosinophils, neutrophils, and macrophages [204]. An abstract illustration of these maturation stages is depicted in Figure 4.10, for more details see [203–205].

Until very recently, the dominant model of hematopoiesis posited that it is controlled by a hierarchy of a relatively small number of critical TFs that are sequentially expressed, largely restricted to a specific lineage, and can interact directly to mediate and reinforce cell fate decisions [203, 206]. The key TFs essential for the maturation stages have been recently summarized by Orkin and Zon [204]. We used this snapshot representation to overlay the overall hierarchical topology of the HTRN with the defined hierarchy in hematopoietic stem cell differentiation (Figure 4.10). At this point we want to mention once more that – due to lack of completeness of regulatory interactions and the fact that we do not analyze these processes on a time- and space-dependent scale — only a general, static picture can be depicted.

As described before, TFs starting from layer 3 – e.g. *RUNX1* and *EMX2* – and

---

81 *RUNX1* runt-related transcription factor 1
82 *EMX2* empty spiracles homeobox 2
4.2 Hierarchical organization of mammalian TRNs

Figure 4.10: Abstract overlay of the HTRN with the hierarchy in hematopoiesis. The HTRN is overlayed with the hierarchy that defines hematopoiesis. The single layers in the HTRN correspond to the developmental and maturating steps of hematopoiesis. TFs in the upper layers (e.g., CBFB) are associated with GO terms describing stem cell differentiation and hematopoiesis, whereas TFs in the second layer such as GATA3 already are associated with more specific processes such as regulation of interleukin production, B and T cell differentiation. Bottom layer TFs such as e.g. ZBTB32, SMAD3 or RCL3 are even more specific as they are mostly associated with processes describing the immune response itself, such as activated T cell and natural killer cell proliferation. In general the HTRN hierarchical organization fits nicely the hierarchy in hematopoiesis. HSC differentiation was adapted from Orkin and Zon [204].

downwards are significantly enriched in processes related to hematopoiesis. But also upper layer TFs can be directly and/or indirectly linked to hematopoiesis. GZF1 – the master regulator of the HTRN and important in the early development of kidney [201, 202] – regulates HOXA10 which plays a role in morphogenesis of the ureteric bud in the developing kidney [207]. Kidneys develop from intermediate mesoderm that is derived from ventral mesoderm precursors [208] that in turn represents the origin of HSCs. HOXA10 regulates PAX6 [209], which then controls PAX6 in layer 2 that is associated with immune response function. Furthermore, CBFB in level 4 – associated with lymphocyte and myeloid cell differentiation – regulates RUNX1 [211], which is important for the self-renewal of HSCs and the process from LT-HSC to ST-HSC [203]. RUNX1 regulates GATA3, which is important for the development of T lymphocytes representing the final stage of hematopoiesis, the mature cell production.

In summary, TFs in layer 4 and 3 are associated with GO terms related to stem cell differentiation and hematopoiesis, whereas TFs in the second layer such as GATA3 are already annotated with more specific processes such as regulation of interleukin production.

---

83 PAX6: paired box gene 6
84 CBFB: core binding factor beta
85 GATA3: GATA binding protein 3
The hierarchical organization of regulatory networks

**B and T cell differentiation.** Bottom layer TFs – e.g., BCL3, SMAD3, or ZBTB32\(^{86}\) – are even more specific. They are mostly associated with processes describing the immune response itself such as activated T cell and natural killer cell proliferation.

### 4.2.5 Master-regulators function as triggers in signaling cascades

Analyzing the functional annotation of TFs revealed that higher level TFs fulfill more central and general functions, whereas bottom layered TFs are involved in more specific processes such as triggering signaling pathways or cell proliferation. To examine if this pattern is also reflected on pathway level, we analyzed the correlation between the hierarchical organization and pathway association of the TFs in the single layers. To this end, we mapped KEGG pathway information \[^{55, 213}\] to all TFs in the human transcriptional regulatory network (HTRN). In total, only for 285 (15\%) of all TFs pathway annotations were available.

As previously in the case of functional association, we found that TFs in level 2 are associated with most pathways (on average around four) reflecting their important role in signal propagation and differentiation in more specific functionality (Figure 4.9b). To determine the relevance of our findings, we extracted pathways significantly enriched relative to those in the bottom layer using Fisher’s Exact Test.\(^{87}\) Interestingly, we found no enriched pathways for level 5 and 4, probably due to the small amount of annotated TFs in these layers. For TFs in level 3 we found three significantly enriched pathways: Pathways in cancer, MAPK signaling pathway and JAK-STAT signaling pathway. Compared to that TFs in level 2 are enriched in more specific pathways such as cancer pathways\(^{88}\) and several signaling pathways. We found that many end-of-cascade pathways are exclusively associated with bottom layer TFs, e.g., DNA replication, RNA degradation, or Mismatch repair.

Furthermore, 90\% (31/35) of the basal transcription factors are associated with the bottom layer. This is not contradictory to our findings that the bottom layer is associated with more specific functions. Basal (or general) transcription factors are intimately involved in the process of gene regulation as they are required either for transcriptional initiation or elongation. Though they have a general, but rather fundamental role, time- and tissue-specific combination with other factors makes them more specific again.

Concluding, the observations on pathway level indicate that in TRNs information is processed from top to bottom: transcription factors in the top trigger (regulatory/signaling) cascades that lead in the end to specific functions and pathways.

---

\(^{86}\) BCL3: B-cell leukemia/lymphoma 3; SMAD3: SMAD homolog 3 (Drosophila); ZBTB32: zinc finger and BTB domain containing 32

\(^{87}\) We used Benjamini & Hochberg procedure for multiple hypothesis testing and a significance level of \(p - value = 0.05\).

\(^{88}\) Pathways in cancer represents the overview, whereas here cancer pathways stands for the individual cancer pathways such as Colorectal cancer, Prostate Cancer, and so on.
4.2.6 MicroRNAs act differently: from switches to fine-tuning

The relation between miRNAs and their targets in higher eukaryotes is part of the highly complex GRN (Section 2.5.2), whereby here we concentrate on the TRN only. Information about miRNA-mediated post-transcriptional regulation was therefore included to account for this additional regulatory layer. So far, only a limited number of experimentally verified miRNA-target interactions exists. To get insights on how miRNAs have an influence on the regulatory organization we used the most comprehensive data set on microRNA binding sites available up to now. The mapping is based on the PAR-CLIP data set [122], which resolves miRNA-target interactions on the target site level by the precise determination of argonaute (AGO)-mRNA footprints. The experimental setup of PAR-CLIP was performed in HEK293 cells and delivers targets of 221 human miRNAs (Section 2.5.1). Although this study provides information of a very small set only (∼22% of the miRNAs and a single cell line), it represents the most comprehensive and reliable data set available today. Consequently, from this set general conclusions and statements can be drawn.

4.2.6.1 Higher level TFs are under stronger microRNA control

We found for all 221 miRNAs present in the PAR-CLIP data set experimentally validated miRNA-target interactions with in total 28% (508/1804) of the TFs in the HTRN. This is in line with the current opinion that miRNAs regulate more than one-third of all mRNA transcripts [214–217]. Very recently, Krol et al. [108] reported that even 50% of all genes are supposed to be under miRNA control. The observed gap can be explained by tissue-specific expression as not all TFs are ubiquitously expressed [89] (Section 4.2.7), or the fact that the current data set is far from being complete. Nevertheless, it is appropriate to get first informative impressions on the impact of miRNAs on the TRN and to draw general conclusions from that. The same is true for the single hierarchical layers where between 26% and 37% of the TFs in each layer are under miRNA control (Figure 4.11a). An exception is the top layer: for GZF1 no miRNA-target interaction has been measured so far. This is of special interest as GZF1 is known to be expressed in kidney [90] and represents an important factor in kidney development (Section 4.2.4). We conclude that GZF1 is either not controlled by miRNAs at all, or rather no regulatory interaction has been detected so far.

AGO-mRNA footprint perspective To get more insights into miRNA-mediated regulation, we analyzed the distribution of AGO-mRNA footprints. The number of footprints (or binding sites) per TF shows a U-shaped distribution: transcription factors in the top and the bottom layer have on average more footprints than TFs located in the middle of the hierarchy (Figure 4.11b). In general, higher level TFs tend to have more binding sites (Figure 4.11b). Multiple binding sites are known to have a stronger regulatory effect and

[89] The experimental setup was performed in human embryonic kidney (HEK293) cells only [122].
[90] http://biogps.gnf.org
lead to fast state switches necessary, e.g., in differentiation processes [218–220]. Furthermore, they enable co-operative binding of different miRNAs resulting in synergistic gene repression [218]. This is in line with our observations on the functional layer (Section 4.2.4), where we observed that higher level TFs are associated with important differentiation processes and bottom level TFs are required to be under more specific miRNA control. Fewer footprints in the middle layers indicate lower miRNA control of those TFs. A reason for that might be that these TFs are triggers of signaling cascades that are under weaker miRNA control (Section 4.2.5). Cui et al. [221] showed that miRNAs target downstream signaling components more often than upstream components. Figuratively, this is similar to the so-called snowball effect: high effort has to be put on triggering the initial reaction, e.g., an avalanche, and especially to finally control and/or stop this cascade precisely. In-between, it works almost like magic with only little interference required.

MicroRNA perspective One AGO footprint can correspond to up to 34 miRNAs. Concentrating on the miRNA level, we see a similar U-shaped distribution as before as on average a footprint corresponds to 11 miRNAs. Transcription factors that are regulated by
many miRNAs are located in the bottom layer and in the top layer (Figure 4.11). Higher level TFs – associated with around 27 distinct miRNAs – require regulation by different miRNAs to trigger different downstream pathways. The synergistic gene repression by co-operating miRNAs even increases their complex mode of functioning [218].

4.2.6.2 Hierarchical organization reveals different classes of microRNAs

Next, we investigated the impact of miRNAs on the HTRN to identify groups of miRNAs showing a similar regulation pattern. All 221 miRNAs have targets in the bottom layer, however, concerning the other higher layers differences can be observed. To measure these differences, we calculated Pearson’s correlation coefficients between miRNAs regulation profiles over the hierarchical layers. Results on this are shown in the heat-map [91] in Figure 4.12 where well-defined miRNA clusters become apparent. The cluster in the lower left corresponds to the 77 miRNAs that have been identified to regulate TFs in all 4 layers. The remaining 144 miRNAs do not control any of the TFs in layer 4. Another well-defined cluster exists consisting of five miRNAs (shown in the upper right) that have targets in the two lower levels (1 and 2) only: hsa-miR-208a, hsa-miR-411, hsa-miR-126, hsa-miR-487b, and hsa-miR-499-5p. Interestingly, three of these miRNAs have been reported with diseases [140]. For instance, hsa-miR-126 has been described with several diseases ranging from different types of cancer to metabolic diseases such as type 2 diabetes [222, 223]. Meanwhile, we know that miRNAs have been found to be causally associated with a broad range of diseases [140] (see also Section 2.5.3). To investigate this in more detail, we will focus on miRNAs and their impact on disease development and progression in more detail later on in Chapters 6 and 7.

We found four miRNAs that do not control TFs in the second layer, but have targets in the bottom layer and in layer 3: hsa-miR-517a, hsa-miR-875-5p, hsa-miR-1180, and hsa-miR-1279. The 77 targets of these miRNAs are significantly enriched in signaling pathways, cell cycle and cancer. [92] Besides miRNAs with similar regulatory pattern there exist also some quite specific miRNAs. For instance, more than 90% of the targets of hsa-miR-218 are located in the bottom layer. We conclude that by knowing the hierarchical organization either the elements in the single hierarchical layers or in this case even their regulators can be annotated or classified.

4.2.7 Tissue-dependent TF expression correlates with their hierarchical position

Lately, Bhardwaj et al. [226] reported that essential genes are more likely to be abundantly and ubiquitously expressed in cells and tissues. Goh et al. [17] found that essential genes that are not associated with disease are expressed in multiple tissues. In order to examine tissue-dependent expression patterns of TFs in the HTRN we used the Genome Novartis Foundation SymAtlas dataset [227], which contains measurement of transcript levels for

---

91 We used hierarchical clustering with average linkage and Pearson’s correlation as distance measure.
92 For enrichment analysis we used GeneCodis [224, 225], http://genecodis.dacya.ucm.es/
Figure 4.12: Clustergram of miRNA regulation profiles. We calculated the correlation of miRNA regulatory patterns over the hierarchical structure. We used hierarchical clustering with Euclidean distance. In the lower left (light blue box) a cluster of 77 miRNAs is shown that have targets over the complete hierarchical structure. The cluster in the upper right (light blue box) corresponds to five miRNAs that control layers 1 and 2 only. The color indicates the degree of correlation, ranging from white (negative correlation) to dark red (perfect correlation of 1).

79 human tissues, tumor samples and cell lines.

Transcription factor activity highly depends on post-translational events, and gene expression does not necessarily indicate regulatory activity. However, it is still useful to assess the extent of TF expression as it provides the first line of evidence for the locations in which they may function. In the following, we focus on 32 healthy and major tissues and organs according to Vaquerizas et al. [228].

4.2.7.1 Levels of TF expression

In Figure 4.13 we illustrate the average expression of probe sets mapping to 82% (1,475) of the TFs and 13,021 of the non-TF genes represented on Affymetrix GeneChips across the 32 human samples examined. The plot confirms observations from previous molecular studies that TFs tend to be expressed at lower levels than non-TF genes ($p < 2.5e^{-9}$; T-test) [229]. Mechanistically this makes sense: the effect of a single TF molecule is amplified by transcribing many copies of mRNA from a target gene. Moreover, it is easy to trigger a regulatory event by altering TF concentrations or activity if their

---

93 Obtained using the Affymetrix GeneChip HG-U133A
expression levels are kept low. Finally, cells need to ensure that TFs recognize the correct target sites in the genome. Maintaining lower expression levels would allow TFs to bind the highest affinity sites, and keep lower affinity sites free for activation under special conditions, or nonfunctional sites free from undesired binding [228, 230].

Figure 4.13: Distribution of gene expression levels. Gene expression levels for TFs (red) and non TFs (blue) are shown for all 32 different tissue types, shown as a box plot.

By comparing the expression profiles of the single hierarchical layers a similar picture appears. The expression levels for lower level TFs tend to be higher than those of the upper layer TFs (Figure 4.14). This is in agreement with our previous observations as higher level TFs tend to function as switches and triggers of functional cascades that are more easily to influence or change by altering TF concentrations or activity if their expression levels are kept low. As described in Section 4.2.5, most general TFs are associated with the bottom layer that overall shows higher expression compared to the other layers. This fits perfectly as general TFs are classified as housekeeping genes [231] that have been reported to show a high level and breadth of expression [232].
Figure 4.14: Distribution of gene expression levels in single hierarchies. Distribution of gene expression levels for TFs of the different hierarchical layers, shown as a box plot. (a) shows level 1, (b) level 2, (c) level 3, and (d) level 4. In all samples, higher level TFs have lower average expression than lower level TFs.
4.2 Hierarchical organization of mammalian TRNs

![Graph showing expression levels of various tissues](image-url)
4.2.7.2 TF expression patterns

The heat-map\(^94\) in Figure 4.15 displays the pattern of TFs expression across the 32 major tissues examined. As proposed by Vaquerizas et al. \(^228\), we calculated *propensity* values as a measure for tissue-specific expression for each TF. The propensity score normalizes a probeset’s expression relative to all other probesets. For comparison, we also calculated propensities for a randomized gene expression dataset, in which expression values are shuffled among all tissues. Tissue-specific TFs were then identified as those in which any of their probesets had propensity values higher than the top 5%. For more details on this we refer to Vaquerizas et al. \(^228\). Based on the propensity values, the TFs can be grouped into two categories:

1. *Ubiquitous* TFs that are present in all or most tissues with similar expression levels, and
2. *specific* TFs that are selectively expressed in a few tissues.

\(^{94}\) We used hierarchical clustering with average linkage and Euclidean distance.

Figure 4.15: Heat-map representation of TF expression in 32 human organs and tissues. Heat-map of TF expression (rows) in 32 organs and tissues (columns). Intersecting cells are shaded according to expression level (light yellow for low expression and dark red for high expression). Ubiquitous and specific TFs are grouped according to their expression profiles using hierarchical clustering. Ubiquitous regulators are expressed at similar levels across most tissues, whereas specific regulators are expressed at significantly different levels in certain tissues.
4.2 Hierarchical organization of mammalian TRNs

The overall fraction of ubiquitous TFs is quite high in all hierarchical layers (Figure 4.16). The two-tier system of general and specific TFs suggests different potential regulatory scenarios (Figure 4.15). Ubiquitous TFs alone - in isolation or in combination with each other - might control the general cellular machinery, and combination of specific TFs might regulate tissue-specific genes. Alternatively, ubiquitous TFs might serve as a platform to regulate a broad set of genes, which are then fine-tuned by specific regulators [228]. Both scenarios go along with our previous findings.

![Bar chart showing the average number of tissues associated with TFs at each hierarchical level.](image)

**Figure 4.16: Tissue association of the HTRN.** The average number (indicated by #) of tissues a TF is associated with is shown for each hierarchical layer.

4.2.8 TFs and human diseases

Numerous diseases arise from a breakdown in the regulatory system: transcription factors are overrepresented among oncogenes [233], and a third of human developmental disorders has been attributed to dysfunctional TFs [234]. To evaluate the overall impact of TFs in human diseases, and especially to analyze if certain hierarchical layers emerge to be significantly associated with disease development or progression, we examined the proportion of TF genes included within the OMIM database, which contains information of Mendelian-inherited monogenic diseases [228, 235].

We identified 269 TFs (14%; \( p = 0.035 \) for association of TFs with diseases using a Fisher’s Exact Test) that are directly responsible for 457 diseases or syndromes. Among these, a significant proportion is related to developmental defects, highlighting the importance of TFs during the early stages of development [234]. Considering the distribution of essential TFs in Section 4.2.3, we expect that disease associated TFs tend to be located in the lower levels. Indeed, as shown in Figure 4.17, the relative fraction of OMIM TFs per level is highest in the second layer where around 40% of the TFs are associated with a monogenic disorder, whereas only 10% of the bottom layer TFs. Higher level TFs seem to be associated with more diseases than bottom layer TFs. This is not necessarily contradictory to our hypothesis as human disease genes tend to display also characteristics of

---

95 OMIM: Online Mendelian Inheritance in Man
essential genes [17, 197].

Figure 4.17: Distribution of disease associations of the HTRN. (a) illustrates the relative fraction of TFs in each layer that are associated with at least one OMIM disorder. (b) shows the distribution of OMIM TFs over the single layers, in (c) the average number of OMIM diseases of a TF is associated is shown. (d) illustrates the relative fraction of TFs in each layer that are associated with at least one disorder that has been focus of a GWAS. (e) shows the distribution of GWAS associated TFs over the single layers, in (f) the average number of in GWAS reported diseases of a TF is associated is shown. Numbers are indicated by #, fractions by %.

Perturbation of TF genes themselves has also important implications for more complex, or multigenic, diseases such as cancer or Parkinson’s disease [PD] [236]. To examine the impact of TFs on such disorders we analyzed the distribution of TFs that have been identified to be associated with them with genome-wide significance (Section 2.4). To this end, we used associations of TFs with 98 traits, see also Chapter 5 for more details.

We identified 191 TFs (11%; $p = 1.8e^{-05}$ for association of TFs with diseases using a Fisher’s Exact Test) that are associated with 43 diseases. From Figure 4.17 we see that the relative fraction of TFs with disease association is highest in the middle layers. The average number of diseases a TF is associated with, however, is highest for bottom layer TFs and decreases with higher levels. Interestingly, in the bottom layer many TFs are associated with several types of cancer (e.g., nine with lymphoid leukemia), whereas many TFs in the second layer are associated with autoimmune disorders (e.g., 20 TFs are linked
to T2D).

Altogether, we see that TFs identified in GWAS tend to be located in the lower layers as these common variants do not bring the complete regulatory system to collapse. In contrast to that, rare variants (not identified in GWAS) and mutations responsible in monogenetic disorders tend to be more essential as they might lead to more severe diseases. Drawing final conclusions from these findings remains difficult and thus results have to be interpreted carefully. A more detailed analysis on this is presented in the following Chapters 5 and 6.

### 4.3 Summary and future perspectives

Hierarchy is a central organizing principle of complex networks (Section 2.1.4), capable of offering novel insights into many network phenomena and via this even contributing to an enhanced understanding of complex diseases [6, 14]. At present, the organization of system modules - the building blocks of any cellular network - is typically limited to either a multilevel hierarchy that describes the vertical relationships between modules at different scales, or a single-scale graph that represents the horizontal relationships among modules. Both types of organizations fail to provide a broader and deeper view of the complex systems that arise from an integration of vertical and horizontal relationships [237].

In this Chapter, I presented HiNO a method that allows to extract the hierarchical organization from regulatory networks under consideration of network motifs and thereby overcomes the limitations of existing methods. This accounts also for vertical (intra-layer) and horizontal (inter-layer) relationships. Consequently, analyses at various granularity levels are possible [237], which is important as discussed before in Chapter 3. First, I showed in Section 4.1 that - via implementing a correction procedure consisting of a downgrade and an upgrade step - HiNO significantly improves the BFS originally presented by Yu and Gerstein [16]. Second, in Section 4.2 I applied HiNO to the GRNs of human, mouse, and rat, and systematically investigated the properties of the revealed network topology. I could confirm that these - though not complete - GRNs have an inherent pyramidal-shaped hierarchical organization. Simulations revealed that this shape is not random at all indicating that the underlying topology in regulatory networks is of real biological importance. Besides, I observed that TFs in the single hierarchical layers show in fact distinct characteristics. In a nutshell, we found that in principle higher level TFs tend to be more essential, functionally more versatile and ubiquitously expressed, whereas lower level TFs seem to be involved in more specific processes and are more likely to be associated with common diseases. Furthermore, using miRNAs as an example, I successfully showed that the hierarchical structure of GRNs is suitable to be used for the classification and characterization of both, TFs and their associated elements, e.g., based on the hierarchical structure we identified several groups of miRNAs regulating on and over different hierarchical layers. These general findings can provide (i) clear hypotheses for detailed analyses of relevant subnetworks, and (ii) arguments for the interpretation of
high-throughput results in terms of rules, or as supporting information.

Based on these results, further applications are conceivable where the hierarchical organization represents the basis for network comparison and characterization, which in the end might greatly improve our current knowledge. Populations are aging\(^{96}\), and the burden of neurodegenerative diseases - especially the two most common disorders Alzheimer’s disease (AD) and PD - is increasing. Effective disease-modifying treatments remain elusive as the current understanding of these complex disorders is still an unsolved miracle. Interestingly, AD and PD are both characterized by insoluble protein deposits: β-amyloid plaques and tau-containing neurofibrillary lesions in AD and α-synuclein-containing Lewy bodies in PD. As a significant percentage of patients have clinical and pathological features of both diseases, the patho-cascades of the two diseases might overlap \(^{239, 240}\) (Chapter 5).

Research communities work separately, however, one drags behind the other finding more and more commonalities. A comparison of the underlying regulatory networks of both disorders, and especially a systematic analysis of their topology might therefore be beneficial to dissect the pathogenic pathways and to differentiate between additive and synergistic effects \(^{239}\).

Besides the TRN also the relation between miRNA and their targets is part of the highly complex GRN (Section 2.5.2, Chapter 6 and 7). Accordingly, it stands to reason to include this additional regulatory layer, especially as the combinatorial and cooperative properties of TFs and miRNAs become more and more obvious \(^{241–243}\).

Furthermore, one has to think of extending HiNO towards resolving the hierarchical organization under the consideration of combinatorial regulation where regulators partner with each other. Thereby, they control common targets and this in turn allows a small number of regulators to govern many targets \(^{226}\). This is a challenging task as the transcriptional modules are not completely known and their time- and space-dependent existence and (in-)activity even increases the degree of complexity that has to be considered. Considering the direction of the regulatory event, or in other words to distinguish between activating and inhibitory links (at the moment this is merged to regulation only), would also be a benefit. Both together open a perspective towards bringing dynamics into the (at the moment) static picture of GRNs. Dynamic aspects such as time- and tissue-dependent regulation strongly influence and especially alter the density of the overall underlying regulatory networks. After all, only their incorporation allows for an enhanced understanding how network perturbations lead to diseases \(^{6, 244}\).

\(^{96}\) E.g. the aging of Europe, also known as the graying of Europe, is a demographic phenomenon in Europe characterized by a decrease in fertility, a decrease in mortality rate, and a higher life expectancy among Europeans \(^{4, 238}\). This has to be distinguished from senescence or biological aging which is the change in the biology of an organism as it ages after its maturation \(^{4}\).
In the previous Chapter, I showed how directed regulatory networks are organized and at which regulatory layers risk factors for common diseases have been identified (Section 4.2). In this Chapter, I shift the focus from the network structure to the individual interactions within the underlying network and present a systematic analysis of the role of pleiotropic genes in disease networks, and how diseases themselves are cross-connected. Most of the current systems-level approaches to study diseases focus on a single disease, relying on network-based methods to gain insights on the molecules and pathways relevant for the specific disease. A conceptually different approach - the so-called diseasome - has been proposed by Goh et al. [17] in order to study the pleiotropic relationships between different human diseases, leading, e.g., to comorbidity, instead of focusing the attention on one disease [6].

Genome-wide association studies greatly aided in the identification of common risk variants conferring small to moderate risks for specific diseases. There is increasing evidence that many of these loci are of relevance across diseases, and that there are common pathways involved in the development of disease spectra rather than isolated disorders. Obviously, genetic links exist not only between subtypes of the same disorder, but also between apparently disparate conditions [245]. In most cases it remains however unclear if (i) these links, and their mechanisms, are indeed pleiotropic in terms that a single gene has effect on multiple phenotypes, or if (ii) rather comorbidity accounts for the supposed “pleiotropy”. In case of frequent comorbidities such as obesity, diabetes, and hypertension, respective study cohorts will contain a substantial part of comorbid cases. Consequently, GWA signals for “one” disease might light up for the other diseases as well, and thereby falsify the results.

Being aware of this fact, I performed a systematic multi-scale analysis of common human diseases based on published GWA data to elucidate disease relationships and key

---

97 Chapter 5 led to the following manuscript: Hartsperger et al. [18]
98 The human diseasome is the synthesis of all human genetic disorders (the disease phenome) and all human disease genes (the disease genome) [17, 245].
99 Comorbidity is the co-occurrence of two or more disorders or diseases in an individual [245].
5 Uncovering the role of pleiotropic genes via multi-scale disease network analyses

We identified multiple genes and gene variants shared between common diseases, including diseases that are not obviously related, and identified common pathways of relevance across several traits. Existing GWA datasets thus contain highly valuable information beyond initially published results, and their systematic exploration and interpretation helps to enhance our understanding of disease interconnections and their underlying molecular mechanisms. However, as pleiotropy is not clearly defined and difficult to measure, the underlying molecular mechanisms explaining the observed phenomena have to be elucidated from different perspectives.

In the following, the results from a hypothesis-free and holistic network-based analysis of published GWA data are reported. We show that combining and correlating data from GWAS on multiple scales allows for a more detailed insight into complex human diseases and their interrelations. It yields both, biologically suspected as well as unexpected disease-disease relationships. Besides common susceptibility genes, we found that traits also highly correlate on the pathway-level. The analysis of shared pathways instead of single genes allows for identifying mechanisms that are involved in disease progression. Their application has proven to provide complementary information to conventional single-marker analysis in GWAS [138]. On SNP-level we were able to identify mutations common to several diseases and to study potential antagonistic effects. Based on the results of this multi-level approach many candidate loci are provided that represent potentials to experimentally explore them and thus the genomic architecture of complex diseases in more detail.

5.1 Network medicine towards GWAS

Genome-wide association studies improved our understanding of multi-factorial diseases over the last decade. Based on population studies and the application of a well established statistical framework, GWAS led to the identification of hundreds of common variants with small to moderate effects on common diseases such as immune and autoimmune disorders [246] (Section 2.4 and Figure 2.7).

In the past years, the number of published GWAS has grown rapidly [247] [248], but so far these GWAS investigated diseases one at a time on a single SNP basis only [249]. More recently, meta-analyses (Section 2.4.3) on specific traits are being broadly used as powerful approach to detect additional associated common variants with small effects. Despite much success, a large proportion of the total heritability remains however unexplained for most traits [85] [250]. It also has become evident that existing GWAS datasets have not been fully explored yet, e.g., for epistasis [100] and across phenotypes.

The massive repository of GWAS data compiled so far now allows for systematic comparative analyses across traditionally defined traits. Network-based approaches to human dis-
ease, recently coined as network medicine \[3\], facilitate the analysis of the physiology and pathophysiology of biological systems. They can be applied to the large-scale hypothesis-free genotyping data from different GWAS to unravel so far hidden relationships and links between phenotypes. Although yet incomprehensive, network medicine allows for a structured view on any type of interaction in complex systems \[3\] (Chapters \[3\] and \[4\]). Their application has inspired the concept of the human diseaseome - the synthesis of all human phenotype-genome associations \[17\]. This approach and follow-up studies \[251–253\] aid in uncovering interrelationships between diseases and disease mechanisms, not only between subtypes of the same disorder, but also between apparently disparate pathophenotypes \[3\]. Consequently, we hypothesized that when applied to GWAS data, network approaches based on the human diseaseome concept \[17\] would identify so far unknown interrelationships between diseases and disease mechanisms.

Disease phenotypes share risk genes \[89, 254\] and their multi-functional behavior has been denoted as pleiotropy, the capacity of a gene to affect a number of phenotypic traits \[255, 256\]. The traditional definition of pleiotropy distinguishes between

- Type I pleiotropy, where a gene product has multiple molecular functions, and
- Type II pleiotropy, where a single molecular function has multiple morphological and physiological consequences \[256\].

But our understanding of molecular disease mechanisms is far from being complete \[257\], and the action of shared risk factors cannot be explained as a common mechanism of gene pleiotropy. As pleiotropy is not clearly defined and its general patterns are poorly, the interconnections between diseases have to be elucidated from a different perspective.

To gain insights into disease interrelations and key disease mechanisms shared across common diseases, we thus explored publicly available GWAS data beyond conventional single-marker analysis and carried out a hypothesis-free and holistic network-based analysis of all published GWAS data that had to be preprocessed in the beginning.

5.2 Data acquisition and preprocessing

We downloaded A Catalog of Published Genome-Wide Association Studies\[101\] hosted by the National Human Genome Research Institute (NHGRI) \[258\]. This catalog, further referred to as GWAS catalog, includes information about 3,455 SNPs with reported trait associations having a genome wide significance level ($p$-value $< 1 \times 10^{-5}$) (Section \[2.4.3\]).

We preprocessed these data consisting of 700 publications and automatically mapped all traits to official disease terms according to the MeSH (Medical Subject Heading) ontology.\[102\] Where an automated mapping was not possible or the database-listed trait was found to be imprecise (e.g. waist circumference), respective publications were reviewed to determine the precise trait studied (e.g. if the waist circumference study rather describes

---

obesity). We excluded all quantitative and non-disorder traits (e.g. height or hair color), and thus restricted our analysis to disease phenotypes only. Synonymous trait terms were merged (e.g. adiposity and obesity). We further excluded copy number variation (CNV) studies and studies for which identified SNPs were either not reported (NR) or pending. For consistency reasons, all reported SNPs were mapped to official identifiers according to dbSNP build 131.103 For each SNP, we defined the risk locus as all genes that are associated with the corresponding block ($r^2 \geq 0.8$):104 using the SNAP tool.105 Where no genes could be assigned to a LD block, we defined the corresponding locus to be intergenic and used region information instead (Section 2.4.2).

To extend the data set - the GWAS catalog might report only limited number of SNP-trait associations for several papers [258] - we used information obtained from HuGE Navigator [260], a continuously updated database in human genome epidemiology, and associations extracted via Text Mining [176]. We manually evaluated all additional information from these resources to ensure that only SNP-trait associations fulfilling the predefined criteria [258] and requirements were included. For traits, SNPs, and risk loci we applied the same mapping procedure as before to ensure comparability and consistency.

In total, after completing data preprocessing, the final, manually validated data set is based on 311 studies and consists of reported associations between 98 traits and 1,384 SNPs corresponding to 942 loci and/or 797 genes. The complete data acquisition and preprocessing procedure is illustrated in Figure 5.1.

5.3 Analysis of GWAS-based disease networks

We then analyzed this curated data set on multiple scales to identify trait correlations and to investigate the role of pleiotropic genes. For this, we define the correlation or overlap of two traits on different levels: gene, pathway and SNP. Two traits overlap if they are associated with the same gene(s), pathway(s) or SNP(s). The more common associations exist the higher is the overlap of two traits.

5.3.1 The human GWAS disease network

First, we constructed a bipartite network, further referred to as the human GWAS disease network, consisting of two disjoint sets of nodes (Section 2.1.1.4). One set corresponds to all traits, whereas the other set corresponds to all loci that have been reported to be significantly associated with those traits. A trait and a locus are therefore connected by a link if this association has been identified in a GWA study.

The resulting human GWAS disease network consists of 1,040 nodes (98 traits and 942 loci) and 1,235 edges. From the illustration in Figure 5.2 it can be seen immediately that.

104 If two SNPs are linked via a LD of $r^2 \geq 0.8$ this means the minor alleles of both SNPs occurred in combination in 80% of the examined individuals.
105 SNAP: SNP Annotation and Proxy Search
5.3 Analysis of GWAS-based disease networks

Figure 5.1: Data acquisition and processing procedure. We extended the GWAS catalog [261] by additional SNP-trait associations obtained from HuGE Navigator [260] and associations extracted via Text Mining [176] where we applied the predefined filtering criteria and mapping procedure. Traits were mapped to official disease terms using MeSH ontology. Quantitative and non-disorder traits were excluded. Single nucleotide polymorphisms were mapped to official identifiers according to dbSNP build 131. Risk loci were defined as all genes associated with the corresponding LD block ($r^2 \geq 0.8$) of a SNP using the SNAP tool [259]. The final, manually validated data set is based on 311 studies and describes associations between 98 traits and 1,384 SNPs corresponding to 942 loci and/or 797 genes, respectively.

the majority of the nodes (95%) lie within a single weakly connected component (Section 2.1.1.3). Only 47 nodes are not connected to this component and thus represent isolated traits and their associated genes. This already indicates that traits are highly connected and obviously overlap among the genetic variants that affect different human diseases. A fact that was to be expected due to the nature of biological networks and their small-world property (Section 2.1.3).

5.3.1.1 The neighborhood of GWAS traits and risk loci

Next, we analyzed in detail how loci and traits are connected in general. We found that more than 80% of the loci have associations with a single trait only, whereas a smaller group of loci represent hubs (Section 2.1.2.2) that are connected with a multitude of traits. In particular, the HLA locus [106] on chromosome 6p21.32 is associated with several inflammatory and autoimmune diseases. This confirms the crucial role of the HLA locus across several diseases, yet the underlying mechanisms remain unclear. Many genes belong to this family, and their functional implication to disease pathogenesis is yet to be fully explained [254]. In other words, it is not known if the observed associations are indeed causal for an autoimmune reaction, or if they are rather the triggers for it and thus are a

106 The human leukocyte antigen (HLA) is the name of the major histocompatibility complex (MHC) in humans. Many genes belong to this family and a strong LD is operating in that region [254].
Figure 5.2: Illustration of the bipartite disease-locus association network. Nodes represent diseases and genetic risk loci. A link is placed between a disease and a locus if a SNP located in this region was identified to be associated with that trait with genome-wide significance. The network consists of 1,040 nodes (98 diseases and 942 loci) and 1,235 edges. The node colors correspond to disease classes according to the MeSH ontology. In cases where a trait is categorized with multiple disease classes, the corresponding node is multi-colored. Loci are colored according to the disease classes of their associated traits. The node size reflects the number of associated traits and loci respectively, whereas edge thickness indicates the number of studies an association has been reported and/or confirmed in.

secondary effect only. Apart from the complex HLA region, where different alleles are associated with different diseases, a number of loci shared between two or more diseases was observed,\textsuperscript{107} e.g., nine and six loci are shared by rheumatoid arthritis (RA) and type 1 diabetes (T1D), or systemic lupus erythematosus (SLE), respectively. In addition, various genes appear to play a central role in the predisposition towards several diseases, e.g., ZMIZ1\textsuperscript{108} and PTPN22\textsuperscript{109}, for which associations with six and five traits were reported, respectively. Whereas ZMIZ1 is a known risk factor for a quite heterogeneous spectrum of traits (from cancer to immune and mental disorders), PTPN22 appears to be specifically important for autoimmune disorders and also influences susceptibility towards bacterial infections \textsuperscript{262, 263}.

For the vast majority of traits more than one associated locus is reported and for one third of the traits more than 10 loci were identified. Traits with 50 or even more risk loci are, e.g., attention-deficit/hyperactivity disorder (ADHD), T1D T2D or multiple

\textsuperscript{107} Based on the assumption that these associations may be causal.
\textsuperscript{108} ZMIZ1: Zinc finger MIZ domain-containing protein 1
\textsuperscript{109} PTPN22: protein tyrosine phosphatase, non-receptor type 22 (lymphoid)
5.3 Analysis of GWAS-based disease networks

Sclerosis (MS). Details on this are shown in Figure 5.3.

![Figure 5.3: Trait and loci association distribution.](image)

Figure 5.3: Trait and loci association distribution. The absolute number of loci and with how many traits they are associated is given in (a), whereas (c) depicts the relative fraction of loci associations. For traits the absolute numbers and with how many loci they are associated are given in (b). In (d) the relative fraction of trait associations is shown. It can be seen that more than 80% of the loci are associated with a single trait only and that for the majority of the traits only a few risk factors have been identified so far. However, some pleiotropic loci become apparent that are associated with several traits. Furthermore, for several traits many more risk factors have been identified with genome wide significance, e.g., for ADHD more than 100 loci were found.

5.3.1.2 Focus of GWAS is biased towards certain traits

Analyzing how many studies were published for each of the traits, we found that the number of published GWAS largely differs for individual traits (Figure 5.4). For around 45% of the traits only a single study has been reported so far, whereas other traits, e.g., T2D or AD were subject of more than 15 published GWAS. An explanation for the obvious focus on neurodegenerative disorders and diseases related to the metabolic syndrome\(^\text{110}\) is probably both, the great medical and the enormous economical interest as the number of affected people is increasing dramatically (see also Chapter \(^\text{7}\)). As side effect, due to their high prevalence, required cohort sizes for case-control studies can be achieved more easily compared to less prevalent diseases (Section 2.4.1). More than 70% of the loci were reported in a single GWAS only, while around 15% of the loci were confirmed in at least

\(^{110}\) Metabolic syndrome is a combination of medical disorders that, in concert, increase the risk of developing cardiovascular disease and diabetes [264].

105
three GWAS. Several loci such as the HLA locus or TCF7L2\textsuperscript{111} were reported even in more than ten independent studies. Together, these findings indicate a high degree of pleiotropy.

![Figure 5.4: Study distribution.](image)

Of note, loci found to be associated with different traits are not necessarily pleiotropic in terms of being context dependent multi-functional and they may not even exert a significant risk for two related diseases. In case of frequent comorbidities such as obesity and diabetes, respective disease cohorts will contain a substantial part of comorbid cases (Section 2.4.4). Depending on the rate of comorbidity and the penetrance\textsuperscript{112} of a GWA signal for one disease, the signal might also light up in the other (see also Page 99). Likewise, the traditional phenotypic classifications may render a clear diagnostic separation between phenotypically and etiologically complex and heterogeneous diseases difficult. For example, neuropsychiatric phenotypes such as schizophrenia and bipolar disorder, which are characterized by high rates of comorbidity, show a large phenotypic variability with overlapping symptomatology and the absence of objective diagnostic biological markers.

\textsuperscript{111} TCF7L2: transcription factor 7-like 2 (T-cell specific, HMG-box)

\textsuperscript{112} The proportion of individuals with a specific genotype who manifest the genotype at the phenotypic level. For example, if all individuals with a specific disease genotype show the disease phenotype, then the genotype is said to be completely penetrant.
Thus, misclassification bias between two diseases may generate overlapping signals in GWAS – even if the diseases have at least in part distinct etiologies. Therefore, multivariate association approaches to study co-variation of related disease phenotypes are necessary. More recently, meta-analyses on specific traits are being broadly used as powerful approach to detect additional associated common variants with small effects. It also has to be considered that a common allele, which associates with multiple traits, may be a marker for several causal rare variants independently underlying the detected associations. In addition, as we see in the next Section, complex downstream or indirect effects might be affected by, e.g., a regulatory element disrupting DNA variant, and thus, the etiology of distinct diseases cannot be explained by solely defining the associated gene to be pleiotropic in the traditional sense.

So far, we have seen that a certain bias towards widespread diseases exist. To analyze if there is also a prevalence to certain disease classes, we assigned disease classes to all traits (Table 5.1). In cases where traits are associated with multiple different disease categories (e.g. T1D is classified as nutritional disorder, metabolic disorder, and as endocrine system disease) we use multiple disease class assignment. The majority of the traits are annotated with disease classes such as neoplasms, digestive system diseases, cardiovascular diseases, or nervous system diseases. Around 50% of the traits have assignments to multiple disease classes reflecting the current focus of GWA studies on complex diseases (Figure 5.5).

Figure 5.5: Disease class distribution. All 98 diseases where annotated with their corresponding disease classes. Disease classes are defined as the main categories (C01-F04) from the MeSH ontology. As traits can be associated with multiple different disease categories, we included the additional disease class multiple. The relative fraction (indicated by %) of traits associated with a certain disease class is shown. Disease classes are colored according to the legend in Figure 5.2.

5.3.2 The human GWAS diseasome

We discussed earlier in Chapter 3 that so-called network projections (Section 2.1.1.4) – e.g. of a bipartite network into its corresponding unipartite version – discard important

\[113\] C01-F04 main categories from the MeSH Ontology
5 Uncovering the role of pleiotropic genes via multi-scale disease network analyses

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01</td>
<td>Bacterial Infections and Mycoses</td>
</tr>
<tr>
<td>C02</td>
<td>Virus Diseases</td>
</tr>
<tr>
<td>C03</td>
<td>Parasitic Diseases</td>
</tr>
<tr>
<td>C04</td>
<td>Neoplasms</td>
</tr>
<tr>
<td>C05</td>
<td>Musculoskeletal Diseases</td>
</tr>
<tr>
<td>C06</td>
<td>Digestive System Diseases</td>
</tr>
<tr>
<td>C07</td>
<td>Stomatognathic Diseases</td>
</tr>
<tr>
<td>C08</td>
<td>Respiratory Tract Diseases</td>
</tr>
<tr>
<td>C09</td>
<td>Otorhinolaryngologic Diseases</td>
</tr>
<tr>
<td>C10</td>
<td>Nervous System Diseases</td>
</tr>
<tr>
<td>C11</td>
<td>Eye</td>
</tr>
<tr>
<td>C12</td>
<td>Male Urogenital Diseases</td>
</tr>
<tr>
<td>C13</td>
<td>Female Urogenital Diseases and Pregnancy Complications</td>
</tr>
<tr>
<td>C14</td>
<td>Cardiovascular Diseases</td>
</tr>
<tr>
<td>C15</td>
<td>Hemic and Lymphatic Diseases</td>
</tr>
<tr>
<td>C16</td>
<td>Congenital, Hereditary, and Neonatal Diseases and Abnormalities</td>
</tr>
<tr>
<td>C17</td>
<td>Skin and Connective Tissue Diseases</td>
</tr>
<tr>
<td>C18</td>
<td>Nutritional and Metabolic Diseases</td>
</tr>
<tr>
<td>C19</td>
<td>Endocrine System Diseases</td>
</tr>
<tr>
<td>C20</td>
<td>Immune System Diseases</td>
</tr>
<tr>
<td>C21</td>
<td>Disorders of Environmental Origin</td>
</tr>
<tr>
<td>C22</td>
<td>Animal Diseases</td>
</tr>
<tr>
<td>C23</td>
<td>Pathological Conditions, Signs and Symptoms</td>
</tr>
<tr>
<td>F01</td>
<td>Behavior and Behavior Mechanisms</td>
</tr>
<tr>
<td>F02</td>
<td>Psychological Phenomena and Processes</td>
</tr>
<tr>
<td>F03</td>
<td>Mental Disorders</td>
</tr>
<tr>
<td>F04</td>
<td>Behavioral Disciplines and Activities</td>
</tr>
</tbody>
</table>

Table 5.1: MeSH disease classes. The 27 disease classes correspond to the main categories from the MeSH ontology.

But we wont apply any network analysis method to the projected network. Here, we rather use the network projections to (visually) point out to what extent traits are based on general genetic predisposition.

5.3.2.1 The GWAS diseasome from a locus perspective

First, we generated a disease-centric projection of the human GWAS disease network where nodes represent traits and two traits are connected if they share at least one locus that is associated with both traits. The resulting network - the human GWAS diseasome - is sparsely connected and consists of 98 nodes and 402 edges. Most of the traits lie within a single weakly connected component (Section 2.1.1.3), whereas the remaining 16 traits are completely unconnected as for those no joint susceptibility loci have been identified so far (Figure 5.6).
5.3 Analysis of GWAS-based disease networks

Figure 5.6: The human GWAS diseasome. In the disease-centric projection of the human GWAS disease network - the human GWAS diseasome - nodes represent traits and two traits are connected if they have one or several candidate genes in common. In total 402 associations between the 98 traits were found. Most of the traits lie within a single weakly connected component, whereas the remaining 16 traits, e.g. hearing loss or osteoporosis, are completely unconnected. The node colors correspond to disease classes (according to the MeSH ontology) and in cases where a trait is assigned to several disease categories, the node is multi-colored. The node size corresponds to the number of traits with those common genes have been found, edge thickness indicates the number of joint risk factors. For instance for the autoimmune disorders such as T1D and RA or CD and UC the most overlap (9 and 8 common genes respectively), has been identified. For color legend see Figure 5.2.

Autoimmune disorders are highly connected. For almost 80% of the pairwise trait associations we identified single connecting genes – e.g., IL13 for asthma and psoriasis – whereas for the remaining 20% more links have been found. For instance, RA and T1D or celiac disease (CeD) and T1D respectively, are associated with more than five joint risk genes. In general, autoimmune disorders such as RA, CeD, ulcerative colitis (UC), MS, T1D, Crohn’s disease (CD), and SLE group together representing traits that show a high amount of common risk genes. Closely linked are also several autoimmune and inflammatory disorders such as ankylosing spondylitis (AS) and psoriasis that are both associated with the 6q21.33 locus and IL23R. Furthermore, we found that UC is also associated with IL23R and that for AS and RA an association with the 2q11.2 locus has been reported.

114 IL13: interleukin 13
115 IL23R: interleukin 23 receptor
Via overlap towards pleiotropy: what is behind?  Another locus implicated in a large number of traits is the 9p21 gene desert\textsuperscript{116}, where several SNPs have been reported to be associated with coronary artery disease (CAD) and its complications such as myocardial infarction and ischemic stroke \textsuperscript{269}. We identified a total of nine associated traits, including CAD and several types of cancers. The identified region itself harbors no obvious candidate gene, but is flanked by the CDKN2A/CDKN2B\textsuperscript{117} genes, which are key regulators of the cell cycle and known tumor suppressor genes. Furthermore, the 9p21 interval is one of the most dense enhancer-containing region in the human genome with the CAD associated SNPs (rs10811656 and 10757278) located in a newly identified enhancer. These variants disrupt the binding of the STAT1\textsuperscript{118} transcription factor, which causes changes in chromatin architecture and altered gene expression of a set of genes \textsuperscript{269}, thereby offering one potential molecular basis for observed pleiotropic effects.

The case is similar for the chromosomal region 8q24.21 where multiple markers within a large gene desert are associated with several types of cancer. The top scoring hits are positioned near or within loci of genes potentially implicated in cancer predisposition, namely MYC, PSCA, and TP63\textsuperscript{119}, but the precise source(s) of the signals remain to be elucidated. Current data suggest that the region might contain multiple independent variants rather than a unique one that predispose to different tumor specificities \textsuperscript{270, 271}.

Another gene with potential pleiotropic effect is RELN (reelin), which associates AD, schizophrenia, MS, and otosclerosis. Reelin, a serin protease encoded by the RELN gene, is part of the apoE pathway and involved in the phosphorylation of tau, the major component of neurofibrillary tangles, either directly or through \(\beta\)-amyloid pathways that influence tau phosphorylation \textsuperscript{272}. The female-specific association between RELN and schizophrenia is one of the few examples of a replicated sex-specific genetic association in any disease. Reelin is critical for brain development and synaptic plasticity. Post-mortem studies have shown lower reelin protein levels in the brains of patients with schizophrenia and bipolar disorder compared with controls \textsuperscript{273}.

\textbf{5.3.3 SNP- and allele-based sub-analysis}

Besides the overlap of traits indicated by joint candidate genes\textsuperscript{120}, we were interested to what extent traits correlate when applying a more stringent criterion of LD rather than physical proximity. Reported variants within one gene are not necessarily highly correlated in terms of \(r^2\) - within identical LD blocks, i.e., be part of an LD bin (Section 2.4.2). This is especially true for quite large genes, e.g., PTPN22 located at 1p13.2.

We therefore defined two traits to share a LD based locus if associated SNPs were within high correlation \((r^2 > 0.8)\) in the same LD block (Section 2.4.2). Again, we constructed a

\textsuperscript{116} Large intergenic regions \textsuperscript{52}.
\textsuperscript{117} CDKN2A: cyclin-dependent kinase inhibitor 2A; CDKN2B: cyclin-dependent kinase inhibitor 2B
\textsuperscript{118} STAT1: signal transducer and activator of transcription 1
\textsuperscript{119} MYC: v-myc myelocytomatosis viral oncogene homolog (avian); PSCA: prostate stem cell antigen; TP63: tumor protein p63
\textsuperscript{120} A gene for which there is evidence of its possible role in the trait or disease that is under study \textsuperscript{79}.
5.3 Analysis of GWAS-based disease networks

disease-centric network, thereby identifying 35 traits connected to 89 SNPs in 46 LD bins (Figure 5.7).

Figure 5.7: Illustration of the disease-centric projection of the human GWAS disease network. Two traits are linked if their associated SNPs belong to the same LD block ($r^2 \geq 0.8$). In total 46 associations were found between 35 traits. The node colors reflect the disease classes according to the MeSH ontology. In cases where a disease is not precisely annotated, the node is multi-colored. The node size corresponds to the number of traits with those correlating risk variants have been found, edge thickness indicates the number of joint LD-block SNPs. High similarity for the two primary forms of inflammatory bowel disease [CD] and [UC] can be seen. For color legend see Figure 5.2.
Our analysis confirms earlier observations from candidate gene approaches that there is a substantial genetic overlap among phenotypically closely related diseases, such as CD and UC. Both disorders share several clinical characteristics and the distinction can be difficult since there is no differentiating single marker, but they are considered distinct and treated differently [274, 275]. The high overlap of CD and UC on the genetic level that has been demonstrated in a very recent meta-analysis [267] is reflected by great analogy of 14 highly correlating SNPs located in six LD blocks. Such joint susceptibility loci mirror shared pathomechanisms, but may also indicate limitations of clinical disease classifications. Another high overlap on this layer are the two highly correlating SNPs rs2542151 and rs1893217 (LD $r^2 = 1.0$) in the PTPN22 gene that link CD, CeD, and T1D (see Table 5.2).

5.4 Multi-scale analysis of widely studied traits

To account for the varying number of GWAS carried out for different traits, we reduced the GWAS dis easeome to traits that have been studied in at least three independent GWA studies (Figure 5.8). This covers around 40% of the traits and results in a smaller bipartite disease-locus association network consisting of 816 nodes (38 traits and 778 loci) and 966 edges. Interestingly, all traits and their associated loci are part of a single weakly connected component where every trait is only a few links from any other trait (Figure 5.9).

Figure 5.8: Illustration of the 38 top diseases for that the most GWAS have been published. The size the disease term corresponds to the number of existing GWAS e.g. for osteoarthritis three GWAS have been reported, whereas T2D or AD were focus of more than 15 studies. The image was created using Wordle (http://www.wordle.net) by C. Gunter, HudsonAlpha Institute for Biotechnology, Alabama, USA.
### Table 5.2: SNPs that link multiple phenotypes are also highly correlated.

Besides the overlap of traits indicated by joint candidate genes, we were interested to what extent traits correlate when applying a more stringent criterion of LD rather than physical proximity. Reported variants within one gene are not necessarily highly correlated in terms of $r^2$ - within identical LD blocks, i.e., be part of an LD bin. This is especially true for quite large genes, e.g., PTPN22 located at 1p13.2. We therefore defined two traits to share an LD based locus if associated SNPs were within high correlation ($r^2 > 0.8$) in the same LD block. We found six SNPs pairs (out of 39 pairs) that highly correlate ($r^2 \geq 0.8$) and thereby link more phenotypes than considering them alone.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Trait</th>
<th>SNP-Risk Allele</th>
<th>Region</th>
<th>Locus</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10889677</td>
<td>UC</td>
<td>rs10889677-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>17 marker haplotype-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2201841</td>
<td>UC</td>
<td>rs2201841-C</td>
<td>1p31.3</td>
<td>IL23R</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>17 marker haplotype-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2476601</td>
<td>RA</td>
<td>rs2476601-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>rs2476601-G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs2476601-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SLE</td>
<td>rs2476601-A</td>
<td>1p13.2</td>
<td>PTPN22</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Vitiligo</td>
<td>rs2476601-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6679677</td>
<td>RA</td>
<td>rs6679677-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs6679677-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2542151</td>
<td>CD</td>
<td>rs2542151-G</td>
<td>18p11.21</td>
<td>PTPN22</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs2542151-G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1893217</td>
<td>CeD</td>
<td>rs1893217-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs1893217-C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2872507</td>
<td>RA</td>
<td>rs2872507-A</td>
<td>GSDML/IKZF3/...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>rs2872507-A</td>
<td>...ORMDL3/ZPBP2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs2872507-A</td>
<td>17q12</td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>rs2305480</td>
<td>Asthma</td>
<td>rs2305480-C</td>
<td>GSDMB/GSDML/...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UC</td>
<td>rs2305480-T</td>
<td>...IKZF3/ORMDL3/ZPBP2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs653178</td>
<td>CeD</td>
<td>rs653178-G</td>
<td>ATXN2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hypertension</td>
<td>rs653178-G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3184504</td>
<td>RA</td>
<td>rs3184504-T</td>
<td>LOC100287812/...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CeD</td>
<td>rs3184504-T</td>
<td>.../SH2B3</td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs3184504-T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9858542</td>
<td>UC</td>
<td>rs9858542-A</td>
<td>APEH/BSN/...</td>
<td></td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>rs9858542-A</td>
<td>3p21.31</td>
<td>.../DAG1/MST1/</td>
<td></td>
</tr>
<tr>
<td>rs3197999</td>
<td>UC</td>
<td>rs3197999-T</td>
<td>.../NCC1/TCTA/</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>rs3197999-T</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.4 Multi-scale analysis of widely studied traits
5 Uncovering the role of pleiotropic genes via multi-scale disease network analyses

Figure 5.9: Illustration of the human GWAS disease-locus association network of the 38 top traits. Nodes represent diseases and genetic risk loci. Only diseases that have been reported in at least three GWAS are shown. A link is placed between a disease and a risk locus if a SNP located in this region was identified with genome-wide significance. The network consists of 816 nodes (38 diseases and 778 Loci) and 966 edges. The node colors reflect the disease class (according to the MeSH ontology) of the diseases that correspond to that node. In cases where a disease is not precisely annotated, the node is multi-colored. Loci are colored according to the disease classes of their associated disorders. The node size corresponds to the number of associated diseases/loci, respectively. Edge thickness indicates the number of publications supporting this association. For color legend see Figure 5.2.

5.4.1 The diseasome of widely studied traits

The disease-centric projection of the GWAS disease network is illustrated in Figure 5.10 and shows, as before, that autoimmune disorders in general share several risk genes. For instance, CD and UC, the two primary forms of inflammatory bowel disease (IBD), have eight factors in common.

To identify groups of highly overlapping traits, we applied hierarchical clustering\textsuperscript{121} using the pairwise overlap of the traits in associated genes as similarity measure (Section 2.2.2.1). The heat-map in Figure 5.11a confirms the previous findings as CD and UC cluster together. Furthermore, a cluster of the traits obesity, CAD, hypertension, hypertension,

\textsuperscript{121} We used complete linkage analysis with Euclidean distance.
Figure 5.10: Illustration of the disease-centric projection of the human GWAS disease network. Two traits are linked if they are associated with the same locus. Nodes represent traits that have been studied in at least 3 GWA projects. In total 183 associations were found between 38 traits and all traits build a single weakly connected component indicating a high correlation in genetic predisposition. The node colors reflect the disease classes according to the MeSH ontology. In cases where a disease is not precisely annotated, the node is multi-colored. The node size corresponds to the number of traits with those common genes have been found, edge thickness indicates the number of joint risk factors. For instance, for the autoimmune disorders such as T1D and RA or CD and UC, the most overlap (9 and 8 common genes respectively), has been identified. For color legend see Figure 5.2.

...and depression emerges. Although they do not share any genetic risk factors, they are all genetically connected to T2D. Additionally, we found that this cluster is also linked to a group of different forms of cancer.
5 Uncovering the role of pleiotropic genes via multi-scale disease network analyses

Figure 5.11: Heat-map of trait associations. We used complete-linkage hierarchical clustering in order to determine groups of high similar traits. (a) Similarity of two traits is defined as the number of common candidate genes both diseases are associated with. Autoimmune disorders (lower left) represent a group of traits that show a high overlap in genetic risk factors. The color scheme indicates the number of joint elements: white represents no, whereas dark red indicates the highest overlap (e.g. RA and T1D have 9 common candidate genes). (b) Similarity of two traits is defined as the number of common pathways both diseases are associated with. Traits in the lower left (mainly autoimmune disorders) represent a group of traits that show a high overlap in pathways. The color scheme indicates the number of shared pathways: white represents no, whereas dark red indicates the highest overlap (e.g. MS and T1D are associated with 31 common pathways).
5.4.2 Pathway-based view on diseases

From the disease-centric analysis it became apparent that many traits share several genetic risk loci. However, we also observed that traits that cluster together do not necessarily overlap genetically. This begs the fundamental question: are they indirectly linked via another trait such as T2D as before? Or are common pathways involved?

Shared pathways between diseases are not obviously connected by joint risk loci, rather many different genes belonging to the same pathway have to be considered. We retrieved KEGG pathway information to determine if and how traits overlap in associated pathways, and consequently to identify ways how they correlate and play together on a molecular level [276]. In total, we found 104 different pathways to be affected by the 778 risk variants. Multiple sclerosis and T1D represent the hubs with more than 30 associated pathways, followed by hypertension and SLE that are linked to 20 and 18 pathways, respectively (Figure 5.12a). Pathways associated with most diseases represent mostly pathways related to the immune system (Figure 5.12b).

Figure 5.12: Pathway association of 38 top traits. Pathway information was linked to all 38 traits that have been reported in at least 3 GWAS. A pathway is associated with a disease if a reported risk factor is annotated with that pathway. (a) Pathway information was available 35 diseases only. The number of associated pathways per trait is shown. (b) Only the top pathways are illustrated.
The heat-map in Figure 5.11b illustrates the similarity of traits based on common pathways. Besides an obvious high overlap in pathways of genetically overlapping traits (e.g. 17 pathways for MS and T1D), we also found common pathways of traits without any direct genetic overlap. For instance, psoriasis and MS are associated with 13 common pathways. Obesity and hypertension are linked to three common pathways, and obesity and amyotrophic lateral sclerosis (ALS) have even 11 pathways in common. Interestingly, although we did not observe any joint loci between ALS and obesity, both disorders appear to share 11 common pathways such as the GnRH signaling pathway, Calcium signaling pathway or the Phosphatidylinositol signaling system. Both traits have been previously linked via the neuroprotective protein CNTF (ciliary neurotrophic factor), which has been originally targeted and developed for the treatment of ALS and in clinical trials produced marked weight loss in obese adults. CNTF mediates appetite suppression and weight loss via a similar signaling mechanism to leptin, making it a candidate for weight control in leptin resistant subjects. Additionally, we found no shared genetic risk loci for obesity, CAD, hypertension, and depression, but all these traits appear to display a high degree of comorbidity with T1D and also have several pathways in common. This is in line with epidemiological observations, e.g., that CAD is the major cause of morbidity and mortality in patients with T2D and that obesity is one of the most important risk factors for the development of CAD or hypertension. Parkinson’s disease shows a high overlap in pathways with many other disorders - mainly autoimmune diseases - than on the genetic level. Strikingly, PD and CeD are both associated with only one common gene (WNT3), but share four common pathways.

5.4.3 Uncovering disease correlation from a SNP perspective

Our primary analysis focused on genes rather than alleles. Joint genetic loci might reflect different LD bins within the same gene indicating different causal variants. Furthermore, an associated variant or LD-block might exert similar or antagonistic effects, i.e., an allele increases the risk for one disease, but is protective against another. We therefore carried out an allele-based sub-analysis focusing on traits with at least three published GWAS.

5.4.3.1 SNPs are correlated with multiple diseases

We determined all SNPs that have been reported for at least two traits to determine how traits correlate based on common variants. In total, we identified 39 SNPs to be associated with at least two different - mainly autoimmune - diseases; eight of them are associated with CD and UC. Three traits are associated with rs2201841 (CD, UC and psoriasis) and rs3184504 (CD, RA and T1D), whereas rs2476601 is associated with even five different traits (RA, CD, T1D, SLE and vitiligo).

The majority of the SNPs show homogeneous effects. For instance, rs7574865 (G/T), a
SNP in \textbf{STAT4}\textsuperscript{123}, or rs2230926 (G/T), a SNP in \textbf{TNFAIP3}\textsuperscript{124} have been shown to increase the risk for both RA and SLE\textsuperscript{266, 284–286}. Other polymorphisms near \textbf{TNFAIP3} not in linkage disequilibrium with rs2230926 have been reported to be associated with psoriasis\textsuperscript{287}. Furthermore, we identified for several SNPs a proxy SNP ($r^2 > 0.8$) that even increased the number of linked traits (Table 5.2). For instance, the highly correlating SNPs rs653178 and rs3184504 ($r^2 = 1.0$) are associated in total with four different traits (RA, CeD, T2D and hypertension), instead of two and three each.

5.4.3.2 SNPs show antagonistic effects

Interestingly, we also found six SNPs where opposite risk alleles were reported for different traits (Table 5.3). Specific variants in the \textbf{HNF1B}\textsuperscript{125} gene are associated both, with an increased risk of prostate cancer but a decreased risk of T2D\textsuperscript{288}. Furthermore, the SNP-allele (rs2476601) in the \textbf{PTPN22} gene – an important factor in negative control of T lymphocyte activation – increases the risk for both T1D and RA\textsuperscript{289, 290}, but has opposite effects on CD risk\textsuperscript{291}.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Trait</th>
<th>SNP-Risk Allele</th>
<th>OR (+/-)</th>
<th>Region</th>
<th>Locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2305480</td>
<td>Asthma</td>
<td>rs2305480-C</td>
<td>+</td>
<td>17q12</td>
<td>GSDMB/.../ZPBP2-*</td>
</tr>
<tr>
<td></td>
<td>UC</td>
<td>rs2305480-T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2476601</td>
<td>RA</td>
<td>rs2476601-A</td>
<td></td>
<td>1p13.2</td>
<td>PTPN22</td>
</tr>
<tr>
<td></td>
<td>T1D</td>
<td>rs2476601-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SLE</td>
<td>rs2476601-A</td>
<td>+</td>
<td>1p13.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vitiligo</td>
<td>rs2476601-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CD</td>
<td>rs2476601-G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2736100</td>
<td>Lung neoplasms</td>
<td>rs2736100-G</td>
<td>+</td>
<td>5p15.33</td>
<td>TERT</td>
</tr>
<tr>
<td></td>
<td>Testicular neoplasms</td>
<td>rs2736100-T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs401681</td>
<td>Lung neoplasms</td>
<td>rs401681-C</td>
<td>+</td>
<td>5p15.33</td>
<td>CLPTM1L</td>
</tr>
<tr>
<td></td>
<td>Pancreatic neoplasms</td>
<td>rs401681-T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs4430796</td>
<td>T2D</td>
<td>rs4430796-G</td>
<td>+</td>
<td>17q12</td>
<td>HNF1B</td>
</tr>
<tr>
<td></td>
<td>Prostatic neoplasms</td>
<td>rs4430796-A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs744166</td>
<td>CD</td>
<td>rs744166-T</td>
<td>+</td>
<td>17q21.2</td>
<td>STAT3</td>
</tr>
<tr>
<td></td>
<td>MS</td>
<td>rs744166-C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* GSDMB/GSDML/IKZF3/ORMDL4/ZPBP2

Table 5.3: SNPs with antagonistic effects in distinct traits We found six SNPs that have been reported with at least two traits where the risk allele showed antagonistic effects. For each SNP the associated trait, risk allele, sign (+/-) of the odds-ratio (OR) and the corresponding region, locus is shown.

\begin{itemize}
  \item \textbf{STAT4} signal transducer and activator of transcription 4
  \item \textbf{TNFAIP3} tumor necrosis factor, alpha-induced protein 3
  \item \textbf{HNF1B} HNF1 homeobox B
\end{itemize}
Results from this analysis confirmed that the genetic overlap between diseases reflects both, positive and negative relationships. Thus, the same allele can be associated with multiple phenotypes, the manifestation of which depends on environmental and other disease-specific genes, and at the same exert antagonistic effects on other traits. Such loci might represent important switch points in pathways that affect susceptibility for groups of diseases.

5.5 Summary and future perspectives

Genome-wide association studies have provided invaluable information on disease related individual risk. In case of common multi-factorial diseases, the etiology is without doubt complex interpretation of the genetic impact risks unjustified oversimplification. Thus, the statistical significance of any association found asks for a detailed causative mapping on the sequence level first and the understanding of its functional role on the molecular level. Genome-wide association studies thus opens the door to start to ask the right questions.

Our integrative and multi-scale analysis of GWAS data revealed multiple genes/loci, gene variants, and pathways shared between common diseases. While many of these disease correlations are well established, others have not been described so far. We identified key common pathways that represent interesting therapeutic targets applicable to more than one disease. Furthermore, our data supports the idea that the genetic architecture of common disorders comprises a mosaic of pleiotropic alleles and loci unique for one or a few diseases.

Of note, loci found to be associated with different traits are not necessarily pleiotropic in terms of being context-dependent multifunctional and they may not even exert a significant risk for two related diseases. In case of frequent comorbidities such as obesity and diabetes, respective disease cohorts will contain a substantial part of comorbid cases. Depending on the rate of comorbidity and the penetrance of a GWA signal for one disease, the signal might also light up in the other. Likewise, the traditional phenotypic classifications may render a clear diagnostic separation between phenotypically and etiologically complex and heterogeneous diseases difficult. For example, neuropsychiatric phenotypes such as schizophrenia and bipolar disorder, that are characterized by high rates of comorbidity, show a large phenotypic variability with overlapping symptomatology and the absence of objective diagnostic biological markers. Thus, misclassification bias between two diseases may generate overlapping signals in GWAS - even if the diseases have at least in part distinct etiologies. Therefore multivariate association approaches to study co-variation of related disease phenotypes are necessary. More recently, meta-analyses on specific traits are being broadly used as powerful approach to detect additional associated common variants with small effects. It also has to be considered that a common allele, which associates with multiple traits, may be a marker for several causal rare variants independently underlying the detected associations. In addition, complex up-/downstream or indirect effects might be affected by, e.g., a regulatory element dis-
rupting DNA variant [269], and thus, the etiology of distinct diseases cannot be explained by solely defining the associated gene to be pleiotropic in the traditional sense.

Rather, to answer the central question of functional genetic overlaps in diseases, one has to determine their functional role in diseases. Since there is a broad variety of possible causes for finding a gene significantly enriched in a certain disease, each genetic connection needs to be characterized in detail. This is a rather challenging task as most complex diseases are not well understood in terms of molecular function [292]. Therefore, within the scope of GWAS the term pleiotropy has to be used with care. In general, causality can almost never be resolved by large-scale association studies [293], necessitating a systems-biological framework dealing with that problem as we show in Chapter 6.

Furthermore, we anticipate that fine-mapping and re-sequencing will enable a higher resolution view of pleiotropy, and further refine whether the same or distinct alleles in target regions contribute to multiple traits [294]. The value of this approach would be greatly improved by a broader sharing of complete GWAS results among members of the scientific community with appropriate safeguards [295], and the improvements in resources with more dense maps of SNP provided by the 1,000 Genomes and other sequencing projects that are underway. This new approach, together with efforts that pull together GWAS data into publicly accessible centralized repositories, is well suited to propel the study of the genetic architecture of complex diseases beyond individual investigators or individual datasets [258, 296, 297].
CHAPTER 6

Analysis of the role of microRNAs in disease progression

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits. Thereby, they provided valuable insights into their genetic architecture. Most variants that have been identified so far confer, however, relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, *missing heritability*\(^\text{126}\) can be explained [250]. A striking example is human height, for which 180 stringently validated loci collectively explain only about 10% of the genetic variation [81, 299].

The details of genetic causation are turning out to be complex. The definition of the molecular mechanisms which determine the pathogenic interrelations, the disease risk and the crucial factors for phenotype development stays challenging. Furthermore, most of the causative genetic variation need to be uncovered [250]. "[Consequently, n]ew approaches are required for the identification of candidate genes, variants, and affected pathways in order to provide more target-oriented approaches for the experimental setup and especially the interpretation of GWAS [97]." [19]

As demonstrated previously in Chapter 5, GWAS data are not fully explored yet. By systematically analyzing publicly available GWAS data, we are able to show that traits considerably correlate on different scales yielding both, biologically suspected as well as unexpected disease-disease relationships. But despite efforts to understand the interrelations between different diseases based on cellular network characteristics, and to understand the genotype–phenotype relationships in diseases, the modeling of disease-causing mutation effects on molecular networks is required [6]. We point out many regions, which get uncovered by multiple GWAS [297]. The underlying effects, however, are not always obvious. This is especially the case for the large fraction of risk variants that have been identified within intergenic and/or non protein-coding regions [82]. Recently, Harismendy et al. [269] found that variants in the gene desert 9p21 disrupt enhancer binding sites and similar findings were reported by Ahmadiyeh et al. [300] for the 8q24.2 locus (Section 5.3.2.1). Keeping in mind that SNPs are likely to affect (unknown) regulatory elements, we have to face the

\(^{126}\) GWA studies revealed well-supported associations, but these generally explain only a small proportion of the phenotypic variance. Most variants are tiny because the variants are rare or their effect sizes are small [298] (Section 2.4.1).
most striking challenge: the development of frameworks which enable the large-scale biological interpretation of GWAS results, and therefore allow for a better understanding of complex diseases [19].

Recently, Cowper-Sal Lari and colleagues proposed an approach based on a genome-wide regulatory network (GWRN) which is composed of experimental data with the purpose to provide a framework that facilitates the ability to automatically predict the putative-causality of the variants detected by GWAS [294]. However, the authors strongly demand the incorporation of other (putatively) regulatory elements with experimental data available in similar target-oriented frameworks (Section 2.5.2) [19]. A first attempt towards this direction is, e.g., the condition-specific mRNA-miRNA network integrator which has been recently published by Huang et al. [297].

MiRNAs are key regulators of most known cellular processes by post-transcriptionally regulating their target genes (Section 2.5). Deregulation of miRNAs has often been reported to be involved in disease development. This increased attention is mirrored in the rising number of platforms (e.g., PhenomiR [139] or miR2Disease [140]), which systematically collect information about miRNA-disease interrelations. Polymorphisms in the miRNA regulatory pathway (so-called miR-polymorphisms or SNPs that interfere with mRNA function – miRSNPs) are a novel class of functional polymorphisms present in the human genome, making them a prime candidate for correlation analysis in the context of disease-driving mutations. In the last years, several groups have acknowledged the role of miRSNPs, and suggested a strong association of miRSNPs with disease progression and drug response [302–304].

The ability of miRNAs to locate and bind mRNA is critical for regulating mRNA level and protein expression. Polymorphisms that may potentially affect miRNA-mediated regulation of the cell can be present not only in the 3'-UTR of a miRNA target gene, but also in the genes involved in miRNA biogenesis and in pri-, pre- and mature miRNA sequences [305–308]. A polymorphism in processed miRNA may affect expression of several genes and have serious consequences, whereas a polymorphism in a miRNA target site, in the 3'-UTR of the target miRNA, may be more target and/or pathway specific. MiRSNPs can cause a gain or loss of miRNA function. Functional miRSNPs can create or destroy a miRNA binding site within a target mRNA and affect gene expression by interfering with the function of a miRNA [302]. The potential effects of SNPs on miRNA processing and function greatly diverge and are summarized in Figure 6.1. For more details on this we refer to the review of Mishra and Bertino [302].

Polymorphisms in the miRNA pathway are emerging as powerful tools to study the biology of a disease and they have the potential to be used in disease prognosis and diagnosis. Detection of miRSNPs holds promise in the field of pharmacogenomics, molecular epidemiology, and for individualized medicine. To validate functional

---

127 Putative-causality describes the ability of a genetic variation to significantly change gene expression. [19]

128 Beyond TFs and their target genes [301].
miRSNPs, a combination of computational predictions and functional studies is needed [302]. Meanwhile, high-throughput technologies such as AGO PAR-CLIP (Section 2.5.1) afford a transcriptome-wide study of miRNA:mRNA duplicates [124].

In this Chapter, we investigate the impact of impaired miRNA-mediated regulation from the standpoint of population genetics. Starting with analyzing the global impact of miRSNPs, we look on disease causation from a macroscopic perspective. For this, we used - as before in Chapter 5 - information of common variants reported in GWAS and combine them with experimentally validated miRNA binding regions. This allows us to determine the significance of miRSNPs much more precisely than using computational predictions. By contextualizing the results from a systems biological perspective, we show that common SNPs tend to be miRSNPs and are thus of high relevance (Section 6.1). Furthermore, we are able to demonstrate that miRSNPs have indeed the potential to be used in disease prognosis and diagnosis (Section 6.2). Finally on the microscopic layer, we end up with small-scale models that lead to hypotheses, which give potential explanations about disease causations and represent suggestions for further experiments. Consequently, they initiate a novel round in the systems biological cycle.

Figure 6.1: Potential effects of SNPs on miRNA processing and function. SNPs (indicated as red nucleotides) can have either effects on the transcriptional or post-transcriptional layer. (a) „The SNP changes the secondary structure of the pre-microRNA, thus the Dicer-complex is not able to cleave at the designated sites. [The] same can apply for pri-miRNA and the Drosha-complex. Furthermore, the SNP can disrupt the promoter of the microRNA gene, which completely prohibits expression.“ [19] (b) „The SNP is located in the seed region of the mature microRNA. Thus, the control over the usual targets is lost. Moreover, other mRNAs could be targeted due to the SNP, giving rise to even greater perturbation of mRNA levels.“ [19] (c) „The SNP lies within or close to a target site in the 3’-UTR which results either in disrupting the target site or in changing the folding of the mRNA such that the RISC complex can not bind to the target site. Conversely, mutation of a 3’-UTR can create a new target site in a transcript, which beforehand is not targeted by this microRNA.“ [19] (d) „The mutation lies within or close to a RNA element in the 3’-UTR regulating (indicated by green arrows) miRNA-targeting, which thus loses its function. Here, also the reverse effect can apply and a new regulatory element can be created.“ [19] See also Mishra and Bertino [302] for more details. Illustration adapted from Arnold [19].

129 Basis for this Chapter was the diploma thesis of Matthias Arnold [19]. This Chapter led to the following manuscript: Arnold et al. [20].
6.1 Accessing the role of miRSNPs in complex human traits

To access the impact of miRSNPs in the development of complex traits, we systematically investigated the role of miRSNPs from two different perspectives:

1. On the transcriptional level, which corresponds to polymorphisms or mutations affecting miRNA biogenesis (Section 6.1.1), and
2. the post-transcriptional level, which represents miRSNPs in miRNA target sites (Section 6.1.2).

The basis for our analyses are SNP-trait associations, which we extracted from A Catalog of Published Genome-Wide Association Studies\(^\text{130}\) [258, 309]. We extended the set of 3,035 distinct SNP:\[^\text{130}\] listed in the GWAS catalog to access the genetic architecture at the trait-associated loci by using LD data (Section 2.4.2). SNP:\[^\text{130}\] with a high LD (\(r^2 \geq 0.8\)) to a listed SNP were determined as in Chapter 5 using the SNAP tool [259]. This yielded a total set of 17,343 SNP:\[^\text{130}\] with reported trait association. In the following, we refer to this set as GWAS-SNPs.

### 6.1.1 GWAS-SNPs have no impact in microRNA biogenesis

Polymorphisms or mutations affecting miRNA biogenesis may affect the expression of several genes and have serious consequences (Figure 6.1). To analyze this, we mapped all GWAS-SNPs on miRNA coding regions retrieved from miRBase\(^\text{131}\), the central online repository for miRNAs [113–115]. Interestingly, this mapping revealed no significant results: none of the GWAS-SNPs was found to be located in any of the 719 miRNA coding regions. This suggests that there is no direct impact of SNP:\[^\text{130}\] on miRNA-mediated regulation on the transcriptional layer.

A potential explanation for this might be negative selection, which reduces the frequencies of disease causing variants characterized by early-onset morbidity and mortality [79]. So far, only few SNP:\[^\text{130}\] have been identified in small-scale analyses in the microRNA hairpin regions [305–308]. For these great impact on microRNA expression and microRNA targeting, which leads to great changes in cellular protein levels, has been described. MiRNA are master-regulators - comparable with high level TFs (Figure 4.1) - and a variant present in pri-, pre- and mature-miRNA can potentially influence the expression of hundreds of genes and pathways. Consequently, miRSNPs that are located in these genomic regions are not likely to persist in an organism due to evolutionary pressure. They are rather likely to be spontaneous and/or rare variants, and therefore not detectable in GWAS due to low predictive power (Figure 6.1 and Section 2.4.1). Another reason for these findings might be the history of and the course of action in GWAS. SNP arrays used in GWAS are build on pre-selected SNP sets (Section 2.4.2). An analysis based on GWAS data is

---

\(^{130}\) GWAS Catalog as of 08/26/2010

\(^{131}\) version 14.0 released in September 2009
thus highly unlikely to provide significant hits within miRNA coding regions. [19] Besides this localization bias, microRNAs are only short transcripts and for the distribution of their coding regions in the genome as well as of their target sites in the 3'-UTR of human transcripts no reliable patterns have been identified yet (...) which would allow for generalized fitting of SNPs on microRNA-related regions. Additionally, in contrast to, e.g., coding regions and the contained splice sites, a large-scale annotation of SNPs (...) located in microRNA coding regions or validated target sites is still not given yet. [19]

6.1.2 GWAS-SNPs tend to be located in the 3'-UTR

Next, we systematically analyzed the general distribution of all GWAS-SNPs to get an overview where SNPs are located, which impact they might have in general, and to see if a certain bias towards specific chromosomal regions exists. For this, we defined 13 so-called functional classes describing the chromosomal function of SNPs according to their location, i.e., if they are located within a gene or in an intergenic region. These classes can be divided into intergenic and intragenic categories. [19] The complete list and the corresponding translational effects are shown in Table 6.1.

<table>
<thead>
<tr>
<th>Functional Class</th>
<th>SNP Location</th>
<th>Translational Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>Exonic, translated</td>
<td>Protein amino acid change unknown</td>
</tr>
<tr>
<td>Nonsense</td>
<td>Exonic, translated</td>
<td>Amino acid changed to stop codon</td>
</tr>
<tr>
<td>Frameshift</td>
<td>Exonic, translated</td>
<td>Insertion or deletion interrupts reading frame</td>
</tr>
<tr>
<td>Coding-nonsynonymous</td>
<td>Exonic, translated</td>
<td>Protein amino acid change, but neither nonsense nor frameshift</td>
</tr>
<tr>
<td>coding-synonymous</td>
<td>Exonic, translated</td>
<td>No protein amino acid change</td>
</tr>
<tr>
<td>splice-5</td>
<td>Intronic (first two bases)</td>
<td>Not translated</td>
</tr>
<tr>
<td>splice-3</td>
<td>Intronic (last two bases)</td>
<td>Not translated</td>
</tr>
<tr>
<td>utr-5</td>
<td>Exonic, 5'-UTR</td>
<td>Not translated</td>
</tr>
<tr>
<td>utr-3</td>
<td>Exonic, 3'-UTR</td>
<td>Not translated</td>
</tr>
<tr>
<td>intron</td>
<td>Intronic</td>
<td>Not translated</td>
</tr>
<tr>
<td>near-gene-5</td>
<td>intergenic</td>
<td>Not translated</td>
</tr>
<tr>
<td>near-gene-3</td>
<td>intergenic</td>
<td>Not translated</td>
</tr>
<tr>
<td>intergenic</td>
<td>intergenic</td>
<td>Not translated</td>
</tr>
</tbody>
</table>

Table 6.1: Functional SNP classes. [19] Definition of the ten functional classes of SNPs located in a gene coding region (top of the table) and the three classes of intergenic SNPs (bottom of the table).
Based on this classification, we determined how SNPs are distributed over the single functional classes by calculating the fraction of GWAS SNPs and gold standard SNPs for each class, respectively. Detailed numbers are given in Table 6.2 and Figure 6.2 illustrates the distribution of the GWAS-SNPs and the gold standard SNPs over the single functional classes.\footnote{Already from this, a great deviation between both sets becomes obvious, especially for the intergenic class. To determine the statistical significance of these findings we calculated hypergeometric probability (Equation 6.1) and ORs (Equation 6.2) for each functional class as values on this reflect the tendency for over-/underrepresentation towards certain functional classes. Results indicate a significant under-representation of GWAS-SNPs in the intergenic regions ($OR = 0.57$), supporting our initial guess. Interestingly, a high amount ($\sim 45\%$) of the associated SNPs is located there. This is due to the great proportion of non-(protein-)coding DNA ($\sim 99\%$)\footnote{This set is defined as the gold standard.} and is speculated to be in consequence of still unknown regulatory elements.\footnote{Function classes for SNPs in the translated region (coding-nonsynonymous, coding-synonymous, frameshift, nonsense, coding) have been merged to the class 'coding'.} All other classes are significantly over-represented in the GWAS set (Table 6.2). A bias of GWAS-SNPs to the coding sequence (CDS) ($OR = 3.32$) and the 5'-region of the genes (5'-UTR $OR = 3.57$; outside the CDS within 2KB of the 5'-end: $OR = 3.69$) representing the promoter containing region became apparent. This goes along with recent findings where enhancers were uncovered to be affected by SNPs, see also Chapter 5. Strikingly, we found that GWAS-SNPs are also significantly over-represented in the 3' UTR ($P_{hyge} [X = 379] = 2.83e^{-071}$, $OR = 2.98$ [C.I. 95% : 2.69 – 3.30])\footnote{with confidence interval C.I.=95%}.}

\begin{align*}
P_{hyge} [X = x] &= \binom{K}{x} \binom{N-K}{n-x} \binom{N}{n} \quad (6.1) \\
OR &= \frac{x \cdot (N-n-(K-x))}{(K-x) \cdot (n-x)} \quad (6.2)
\end{align*}

X: functional class  
N: gold standard  
K: GWAS set  
n: total number of SNPs in N having functional class X  
x: total number of SNPs in K having functional class X
6.1 Accessing the role of miRSNPs in complex human traits

Figure 6.2: Distribution of the function classes. Distribution of the function class portions for the (...) GWAS-SNPs and the gold standard. The high amount of intergenic SNPs in the gold standard [(see Footnote [132] for a definition)] makes the interpretation of the other portions rather difficult based on the distributions only. Illustration adapted from Arnold [19].

<table>
<thead>
<tr>
<th>Functional Class</th>
<th>Gold standard SNPs (2,457,151)</th>
<th>GWAS-SNPs (17,343)</th>
<th>$P_{\text{hyge}} [X = x]$</th>
<th>$OR \ [C.I. \ 95%]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>coding</td>
<td>22,425</td>
<td>509</td>
<td>$1.71e^{-110}$</td>
<td>3.32 $[3.04 - 3.63]$</td>
</tr>
<tr>
<td>intergenic</td>
<td>1,468,098</td>
<td>7,976</td>
<td>$1.55e^{-294}$</td>
<td>0.57 $[0.55 - 0.59]$</td>
</tr>
<tr>
<td>intron</td>
<td>907,014</td>
<td>7,836</td>
<td>$&lt; e^{-300}$</td>
<td>1.41 $[1.37 - 1.46]$</td>
</tr>
<tr>
<td>near-gene-3</td>
<td>20,655</td>
<td>134</td>
<td>0.02</td>
<td>0.92 $[0.77 - 1.09]$</td>
</tr>
<tr>
<td>near-gene-5</td>
<td>17,648</td>
<td>443</td>
<td>$2.93e^{-089}$</td>
<td>3.69 $[3.36 - 4.06]$</td>
</tr>
<tr>
<td>utr-3</td>
<td>18,526</td>
<td>379</td>
<td>$2.83e^{-071}$</td>
<td>2.98 $[2.69 - 3.30]$</td>
</tr>
<tr>
<td>utr-5</td>
<td>2,669</td>
<td>66</td>
<td>$1.07e^{-017}$</td>
<td>3.58 $[2.80 - 4.57]$</td>
</tr>
<tr>
<td>splice-5</td>
<td>12</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>splice-3</td>
<td>10</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6.2: Frequencies and ORs. Background frequencies, hypergeometric probability $P_{\text{hyge}} [X = x]$ and odds ratio $OR$ for the functional classes of the SNPs in the (...) GWAS set. The combined evaluation of the hypergeometric probability and the odds ratio show that intergenic SNPs are highly underrepresented, the near-gene-3 class is the only one which meets the expected portion - at least roughly. All other classes are overrepresented in the set of trait-associated SNPs. Illustration adapted from Arnold [19].
6.1.3 In-depth investigation of 3'-UTR associated GWAS-SNPs

The overrepresentation of GWAS-SNPs in the 3'-UTR is a first evidence that these findings are not random. But are the GWAS-SNPs located in the 3'-UTR without effect, or are they biologically relevant in terms of affecting functional regulatory sites? So far, mutations in the 3'-UTR are rarely investigated [316]. „Conclusions of the few studies on effects of mutations of the 3'-UTR on regulatory functions are drawn either solely based on RNA fold change predictions [317] (...) [316]. The underlying processes, however, which lead to mutationally driven expression alterations remain in most cases unresolved [19]. To investigate this and due to the attention miRSNPs have gained during the last years [302], we now focus on SNPs affecting the 3'-UTR as this is the main target of miRISCs. In total, we found 253 genes having 379 SNPs located in their 3'-UTRs. In the following section, they are investigated in more detail to determine if those candidates are miRSNPs.

6.1.3.1 3'-UTR SNPs feature strong effects

A general characterization of the SNPs is usually done by analyzing their MAFs, which gives information about how prevalent the less common allele is in a given population.

„Due to the character of GWA studies, the effect size and the frequency of the SNPs are linked (Section 2.4.1). Whereas a high MAF does not permit direct conclusions regarding the effect size (although it suggests a lower effect on trait development), a low MAF strongly indicates a high effect size [98]“ [19].

To investigate the effect size of SNPs located in the 3'-UTR, we determined their MAF distribution using bins of 10% range with a focus on the functional class "3'-UTR". To this end, we defined different SNP sets, namely: gold standard SNPs, GWAS-SNPs, gold standard SNPs with 3'-UTR association, GWAS-SNPs with 3'-UTR association, and GWAS-SNPs with other chromosomal function than 3'-UTR [19].

Subsequently, we used these bins to calculate the ORs (Equation 6.2), which give information about tendencies towards over- or underrepresentation (Table 6.3). We described earlier in Section 2.4 that SNPs that are identified via GWA studies represent relatively common alleles having a MAF of > 1% [80]. We observed that this is also reflected by their MAF distribution. GWAS-SNPs were found to have a commonly higher MAF – especially from the [0.2 – 0.3] interval upwards – compared to the gold standard SNPs, regardless of their chromosomal function (Figure 6.3). This shows that the overall MAF of 3'-UTR GWAS-SNPs does not differ significantly compared to the MAF of the total of trait-associated SNPs in other locations. But the direct comparison of 3'-UTR GWAS-SNPs to SNPs located in the protein coding sequence (gold standard) shows that 3'-UTR GWAS SNPs are under-represented at lower frequencies (e.g., MAF) between [0 – 20%]: $OR = 0.48 – 0.72$ and over-represented at higher frequencies (e.g., MAF) between [20 – 50%]: $OR = 1.21 – 1.58$. This suggests that mutations in the 3'-UTR

---

135 Bin intervals in ISO 80000-2 notation: [0%; 10%], [10%; 20%], [20; 30%], [30%; 40%] and [40%; 50%].
6.1 Accessing the role of miRSNPs in complex human traits

feature strong effects\textsuperscript{136} on trait development as they are both, significantly associated with the according trait and also highly prevalent \textsuperscript{[19]}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6_3}
\caption{Binned MAF distribution of gold standard and GWAS-SNPs. The gold standard distribution evinces the expected trend, namely many SNPs with lower frequency and decreasing SNP count with increasing MAF. The (...) GWAS set on the other hand shows a peak at the bin of MAF between 30\% and 40\%.\textsuperscript{[19]} Illustration adapted from Arnold \textsuperscript{[19]}.}
\end{figure}

6.1.3.2 GWAS-SNPs are associated with heterogeneous function

Next, to address the previous question if these findings are random or biologically meaningful, we investigated if a certain bias towards genes with specific functions exists. Therefore, we applied GO enrichment analysis\textsuperscript{137,138} to the 229 genes (90\%) having a SNP in the 3’-UTR and for those functional annotation is available.

The enrichment analysis revealed that the concerned genes are involved in a wide range of cellular processes. No significant bias towards a specific function class could be observed. Furthermore, the overrepresented (...) terms [are quite heterogeneous and] connected to multiple downstream effects indicating a major participation of the genes in the initiation and regulation of cellular pathways, suggesting a broad spectrum of potential effects which may in part account for the strong effects of the 3’-UTR SNP\textsuperscript{3}.\textsuperscript{[19]} Interestingly, this also correlates with the broad functional spectrum of miRNAs (Section 2.5). We can therefore exclude that GWAS-SNPs are biased to specific functions. For more detailed information see the diploma thesis of Arnold \textsuperscript{[19]}. In summary, results of the functional annotation of genes containing a 3’-UTR SNP support the hypothesis that trait-associated SNP\textsuperscript{3} in the 3’-UTR might affect miRNA\textsuperscript{3} mediated regulation and thus are likely to be miRSNPs.

\textsuperscript{136} Strongness of the effect of a SNP is inferred based on statistical calculations. In GWAS usually odds ratios or p-values (...) are computed which depend on the MAF. The magnitude to which those effects translate into disease incidence, however, cannot be assessed via analysis of the MAF without the causative variants being determined. “\textsuperscript{[19]}

\textsuperscript{137} Benjamini & Hochberg procedure for multiple testing correction; significance level: p-value ≤ 0.05

\textsuperscript{138} We used BiNGO \textsuperscript{[319]} with GO Slim annotation \textsuperscript{[320]}.
### Table 6.3: ORs and background frequencies for MAF bins in the compared SNP sets.

<table>
<thead>
<tr>
<th>Compared SNP sets</th>
<th>MAF bin</th>
<th>Occurrences in set 1</th>
<th>Occurrences in set 2</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) 3’-UTR SNPs</td>
<td>0.0-0.1</td>
<td>48</td>
<td>2,397</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>67</td>
<td>3,294</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>106</td>
<td>3,818</td>
<td>1.34</td>
</tr>
<tr>
<td>(2) not-3’-UTR SNPs</td>
<td>0.3-0.4</td>
<td>78</td>
<td>3,914</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>0.4-0.5</td>
<td>80</td>
<td>3,541</td>
<td>1.01</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>379</td>
<td>16,964</td>
<td></td>
</tr>
</tbody>
</table>

|                 | 0.0-0.1 | 48                   | 4,815                | 0.41   |
| (1) 3’-UTR SNPs | 0.1-0.2 | 67                   | 4,357                | 0.69   |
| vs.              |         |                      |                      |        |
| (2) GS 3’-UTR SNPs | 0.2-0.3 | 106                  | 3,507                | 1.68   |
|                  | 0.3-0.4 | 78                   | 3,042                | 1.33   |
|                  | 0.4-0.5 | 80                   | 2,805                | 1.51   |
| total            |         | 379                  | 18,526               |        |

|                 | 0.0-0.1 | 2,445                | 578,289              | 0.54   |
| (1) GWAS SNPs   | 0.1-0.2 | 3,361                | 577,199              | 0.80   |
| vs.              |         |                      |                      |        |
| (2) GS SNPs     | 0.2-0.3 | 3,924                | 492,771              | 1.19   |
|                 | 0.3-0.4 | 3,992                | 439,825              | 1.40   |
|                 | 0.4-0.5 | 3,621                | 411,944              | 1.34   |
| total           |         | 17,343               | 2,500,028            |        |

|                 | 0.0-0.1 | 48                   | 578,289              | 0.48   |
| (1) 3’-UTR SNPs | 0.1-0.2 | 67                   | 577,199              | 0.72   |
| vs.              |         |                      |                      |        |
| (2) GS SNPs     | 0.2-0.3 | 106                  | 492,771              | 1.58   |
|                 | 0.3-0.4 | 78                   | 439,825              | 1.21   |
|                 | 0.4-0.5 | 80                   | 411,944              | 1.36   |
| total           |         | 379                  | 2,500,028            |        |

*Set 1 and set 2 are denoted in parentheses in the compared SNP sets column. Odds ratios are always calculated with respect to set 1. The Odds ratios for not-3’-UTR SNPs vs. GS SNPs have been calculated, too, but as being identical to those of the (...) GWAS SNPs vs. GS SNPs are not listed separately. The MAFs of 3’-UTR SNPs are distributed similar to those of the remaining (...) GWAS SNPs except for a shift in the interval 0.2-0.4 towards lower frequencies. This phenomenon is also detectable when comparing the (...) GWAS SNPs and the 3’-UTR SNPs against the GS SNPs. Aside from that, the SNPs in the (...) GWAS set (including 3’-UTR SNPs) present a generally higher frequency than the GS SNPs.*

[19]
6.1 Accessing the role of miRSNPs in complex human traits

6.1.3.3 Localization bias towards microRNA target sites

So far, we have seen an obvious bias of GWAS associated SNPs towards the 3'-UTR and that those genes feature strong effects. Besides that affected genes are quite heterogeneous in their function, this is also true for associated diseases (data not shown here, see diploma thesis of Arnold [19] for details). To investigate if miRNA binding sites are affected meaning that the identified SNPs are indeed miRSNPs, experimental verified binding sites are beneficial. Very recently, approaches towards large-scale experimental verification of exact AGO binding sites have been developed [122, 123]. These CLIP methodologies allow to exactly determine the footprints of the miRISC and consequently to infer miRNA binding sites from that (Section 2.5.1).

We mapped all experimentally verified miRNA target interactions provided by Ellwanger et al. [124] to the 253 genes having a GWAS associated SNP within their 3'-UTR region. We found AGO footprints and therefore binding sites for the RNA-induced silencing complex (RISC) in the 3'-UTR of 44 genes (17%) containing in total 61 SNPs (16%). Most of these genes contain a single SNP within their 3'-UTR, but twelve genes have more than one SNP located in their 3'-UTR (e.g. ERCC4 with four SNPs).

![Figure 6.4: Footprint distance distribution.](image)

(a) Histogram of the distances of 3'-UTR SNPs to the closest RISC binding site in PAR-CLIP [122]. SNPs are binned in 500 nucleotide [nts] distance intervals. Distance 0 [nts] would be the center of the RISC footprint. The distribution shows an accumulation towards zero with Gaussian character [19].

(b) Normal probability plot of the data points. The tails of the distribution drift from the normal distribution, nevertheless, in general the data fits a Gaussian distribution. [This could be confirmed] (...) by statistical testing: the Anderson-Darling test (p = 0.651), the two-sided Shapiro-Wilk test (p = 0.654) as well as the Lilliefors goodness-of-fit test (p = 0.973) did not reject the hypothesis that the data could be normally distributed for the chosen significance level (α = 0.05). [19] Illustration adapted from Arnold [19].

139 ERCC4: excision repair cross-complementing rodent repair deficiency, complementation group 4
As miRSNPs have been reported to affect miRNA functioning also in the proximity of miRNA binding sites [302], we investigated the absolute distances of all 3'-UTR SNPs to their corresponding footprint regions (Figure 6.4). This analysis revealed a general proximity bias of the SNPs towards miRNA binding sites: 27 (43.5%) of the SNPs lie within a 500 nt range up- or downstream of the binding sites. One SNP (rs10923 in the 3'-UTR of SMC4) is even located directly within a RISC binding region and another SNP (rs12916 in the 3'-UTR of HMGCR) is positioned only six bases from the footprint region. The binned distance distribution histogram of the SNPs’ location relative to the miRNA binding sites has the character of a Gaussian distribution. By applying several statistical criteria we verified that the distance distribution fits a normal distribution with mean $\mu \approx -36$ nt and a standard deviation of $\sigma \approx 1,528$ nt [19].

As already mentioned before, the information on validated miRNA binding sites is far from being complete (Section 4.2.6). We expect, however, that with an increasing data set the normal distribution would look even more smoothly. Concluding, on the current data basis it seems that the 3'-UTR SNPs and miRNA binding sites can be directly correlated via their distance [19].

### 6.2 The impact of miRSNPs - a systems biological perspective on impaired microRNA-regulation

The analyses in the previous section revealed several clues that GWAS-SNPs might play an important role in the predisposition to complex human diseases. We found that common variants have a clear bias towards 3'-UTR regions and furthermore, that they tend to be located in the proximity of miRNA binding sites. This, together with the finding that the affected genes are associated with a heterogeneous spectrum of functions and diseases, indicates that the 3'-UTR SNPs identified by GWAS are potentially miRSNPs. Thereby, they have an influence on miRNA-mediated regulation and might play an important role in disease development as impaired miRNA regulation has been reported to be causally linked with diseases for many cases [139, 140, 316].

Results from our large-scale in silico analysis underpin the key role of miRSNPs and suggest many candidate genes for further investigations and finally experimental verification. To investigate if the supposed miRSNPs affect the miRNA-mediated regulation of the 44 identified candidate genes and how they lead to the pathogenic effect the 61 corresponding variants are associated with, a systems biological perspective on this is beneficial (see Page 133). The microscopic view of systems biology focuses on small-scale mechanistic components of a system (Section 1.2) and therefore allows for a deeper functional, molecular, and causal understanding. In other words, it represents a step towards closing

140 SMC4: structural maintenance of chromosomes 4
141 HMGCR: HMG-CoA reductase
142 0.5 kB bins
143 Anderson-Darling test ($p = 0.651$), two-sided Shapiro-Wilk test ($p = 0.654$), and the Lilliefors goodness-of-fit test ($p = 0.073$)
the gap between genotype and phenotype. For more details see also Chapter 7.

Here, we want to attribute pathological phenotypes to impaired miRNA-regulation caused by miRSNPs and give detailed explanation of this relation in a systems biological fashion. This is achieved by systematically exploring the impact of the supposed miRSNPs and the broader biological context of affected candidate genes. To this end, we integrated relevant information from different resources\textsuperscript{144} with computational predictions\textsuperscript{145} and information from the literature extracted via Text Mining\textsuperscript{176}. Together with knowledge of clinical symptoms and molecular markers for the corresponding disease we are then able to present coherent systems biological models capable to contextualize the impact of common variants on miRNA-mediated regulation (see also Section 7.5). Each model is able to give a causal explanation on how the miRSNP leads to the associated phenotype. In the end, this results in coherent novel hypotheses for disease development leading to experiments which initiate a novel round in the systems biological cycle\textsuperscript{8,9}.

As discussing all potential candidates is beyond the scope of this thesis, we present two representative small-scale models: SMC4 (Section 6.2.1) and HMGCR (Section 6.2.2). For further models see the diploma thesis of Arnold\textsuperscript{19}.

### 6.2.1 SMC4 - miRSNP links primary biliary cirrhosis and cancer

"The autoimmune disease primary biliary cirrhosis (PBC) is associated with the damaging of the small bile ducts and is mediated by auto-antibodies\textsuperscript{146} \[326–328\]. The autoimmune response caused by those antibodies leads to inflammation followed by aggregation of dead cells via apoptosis induction, (...) \[among other things\], by reactive molecules effecting DNA damage \[329, 330\]."\textsuperscript{19} Thereby scar tissue (i.e., cirrhosis) is built up, which in turn blocks the normal flow of blood through the organ and thus leads to a loss of its function.

The genetic background of PBC was focus of a recent GWAS accomplished by Hirschfield et al.\textsuperscript{331}. But the rationale of the study was restricted to the major histocompatibility complex and interleukins and causative variants have not been identified so far. "DNA repair genes such as PARP1\textsuperscript{147} and XRCC1\textsuperscript{148} are over-expressed in cirrhotic tissue and are \[-\] in this context \[-\] hypothesized to feature pathogenic effects\textsuperscript{332}. For full functioning of the PARP1:XRCC1 complex in single-strand break repair, an association with the Condensin I complex (...) \[has to be\] established\textsuperscript{333}.\textsuperscript{19} Interestingly, SMC4 is part of the Condensin I complex\textsuperscript{334}. One of the SNPs (rs4679904) that has

---

\textsuperscript{144} GAD (genetic association database) for gene-disease associations\textsuperscript{222}; INTACT\textsuperscript{323} and CORUM (comprehensive resource of mammalian protein complexes)\textsuperscript{324} for PPI; miRBase\textsuperscript{113–115}, PhenomiR\textsuperscript{139} and miR2Disease\textsuperscript{140} databases for miRNA-related information.

\textsuperscript{145} mRNA fold predictions using the RNAfold web service (Vienna RNA package version 2.0.0, as from 10/4/2010) at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi\textsuperscript{225}; miRNA target site prediction based on seed complementarity only (perfect Watson-Crick complementary seed sites covering the first or second 5' miRNA heptamer). \[19\]

\textsuperscript{146} An autoantibody is an antibody manufactured by the immune system that is directed against one or more of the individual’s own proteins\textsuperscript{1}.

\textsuperscript{147} PARP1 poly (ADP-ribose) polymerase 1

\textsuperscript{148} XRCC1 X-ray repair complementing defective repair in Chinese hamster cells 1
been reported in the study by Hirschfield et al. \cite{331} is in high LD ($r^2 = 0.86$) with the SNP rs10923 (major allele: A - minor allele: G) located in the 3’-UTR of SMC4.

Mishra and Bertino \cite{302} reported that a SNP can either abolish or weaken a miRNA target, or create a perfect sequence match to the seed of a miRNA that otherwise was not associated with the given miRNA \cite{304}. Strikingly, rs10923 is localized within an experimentally validated RISC binding site \cite{122} and lies directly in the seed complementary region of hsa-miR-299-5p, a miRNA that has been shown to be up-regulated in PBC patients \cite{335}. Therefore, rs10923 is quite likely to be a miRSNP: “the minor allele G of rs10923 disrupts the complementary region and thus the ability of the (...) [hsa-miR-299-5p] to bind the transcript.”\cite{19} We suggest that, by this process, rs10923 contributes to the phenomenon of DNA repair perturbation in cirrhosis \cite{332}. Beyond that, this may also indicate a rate-limiting character of SMC4 in the generation of the Condensin I-PARP1-XRCC1 complex. Furthermore, PBC patients have a higher risk to develop different types of cancer \cite{336-338}, representing a potential explanation for the deranged DNA repair functionality in the disease \cite{19}. “In cancer development and progression, up-regulated DNA damage response is associated with mutagenesis and resistance to radio- and chemotherapy \cite{330, 339-341}.”\cite{19}

**Summary** Taken together, these findings strongly indicate that the rs10923 mutation in the 3’-UTR of SMC4 represents a miRSNP which plays a critical role in PBC development via impaired regulation of hsa-miR-299-5p regulation. The miRSNP causes a more frequent association of the Condensin I-PARP1-XRCC1 complex contributing to disturbed DNA repair in cirrhosis tissue \cite{332}. The complete model is summarized in Figure 6.5 and now, experimental validation is required to verify it \cite{19}.

**Figure 6.5: Impact model of mutated SMC4 in primary biliary cirrhosis.** Inflammation follows the autoimmune response (...) leading to the activation of the MAPK [signaling] pathway via signal molecules as e.g. TNF-alpha or stress (...). Transcription factors activated as downstream effect of MAPK activation \cite{342} lead to over-expression of DNA repair genes (...) [e.g. PARP1, XRCC1, and SMC4] \cite{332, 343}. The in PBC over-expressed [hsa-]miR-299-5p (...) is hypothesized to target SMC4 at the seed complementary region where rs10923 is located. With the major allele [A], SMC4 is silenced (...), whereas the mutated SMC4-G cannot be bound by [hsa-]miR-299-5p (...) and therefore is translated without interference. This results in the more frequent association of the Condensin I-PARP1-XRCC1 complex (...) contributing to disturbed DNA repair in cirrhosis tissue \cite{332-341} Illustration taken from Arnold \cite{19}.
6.2 The impact of miRSNPs - a systems biological perspective on impaired microRNA-regulation
6.2.2 HMGCR - the role of microRNAs in LDL cholesterol related traits

"Increased plasma low density lipoprotein cholesterol levels are associated with many disease classes ranging from cardiac to neurological traits." [19] The treatment of high low density lipoprotein (LDL) levels plays therefore an important therapeutically role for the prevention of, e.g., cardiovascular diseases. "HMG-CoA reductase (HMGCR) is the rate-limiting enzyme of cholesterol synthesis in human [344, 345] and in this role has been identified as (...) [a] drug target(...) : the treatment of high LDL cholesterol with HMGCR inhibitors (statins) such as Atorvastatin[149] or Lovastatin[150] is a standard therapy in medicine [348]." [19] Recently, it became apparent that patients with one or more risk factors for cardiovascular diseases (CVDs) such as hypertension, T2D or a positive family history, are treated with statins even though they show no elevated LDL levels at all. Statins, however, display several severe side effects ranging from cognitive loss to hepatic dysfunction. Consequently, untargeted treatment of (healthy) people is under critical observation and possibilities to evaluate the potential risk of statin dosage would have not only great medical, but also immense financial impact. But is there a way to estimate if the statin therapy will be paved with success [19]?

6.2.2.1 MiRSNP leads to elevated LDL cholesterol levels

We located the SNP rs12916 (major allele: T - minor allele: C) in the 3'-UTR of the HMGCR gene. This SNP is strongly correlated with rs12654264 ($r^2 = 0.912$), a causative SNP that has been identified in the context of high LDL cholesterol levels [349, 350]. Interestingly, we found that rs12916 lies proximal (six bases) to an experimentally determined RISC binding region [122] within the 3'-UTR of HMG-CoA reductase (HMGCR). According to Mishra and Bertino [302], also the accessibility of the miRNA-RISC complex can be affected, making rs12916 to a potential miRSNP. The corresponding AGO footprint is associated with five targeting miRNAs: hsa-miR-25, hsa-miR-32, hsa-miR-92a, hsa-miR-96, and hsa-miR-183 [19].

6.2.2.2 Interleukin 3 as link between cholesterol control and microRNA involvement

"Two of the miRNAs targeting HMGCR, namely [hsa-]miR-25 and [hsa-]miR-92a, have been shown to be over-expressed (6.9 fold and 70.2 fold, respectively) in model cells treated with interleukin 3 [IL3], a downstream effect identified by miRNA microarray analysis [351]. This effect correlates with the finding of Akgün et al. [352] (...) [who] identified IL-3 among other cytokines as a lowering factor of plasma LDL cholesterol." [19] This suggests that hsa-miR-25 and hsa-miR-92a might play a role in keeping the LDL cholesterol levels at healthy concentrations [19].

149 The statin Atorvastatin is sold by Mascot health series under the trade name Atossh and used for lowering blood cholesterol.
150 The statin Lovastatin is used for lowering cholesterol (hypolipidemic agent) in patients with hypercholesterolemia. Lovastatin is a naturally occurring drug found in food such as oyster mushrooms [346] and red yeast rice [347].
6.2 The impact of miRSNPs - a systems biological perspective on impaired microRNA-regulation

6.2.2.3 The loss of control over LDL production

In healthy individuals (carriers of the major allele rs12916-T), the regulation of cholesterol synthesis by negative feed-back loops (e.g. LDL signaling [353, 354]) leads to the inhibition or degradation of HMGCR [19]. "[This] is supported by microRNA-mediated silencing of the transcript of the HMGCR gene. The IL-3 signaling pathway (...) controls the silencing process by regulating the expression profiles of the two microRNAs. In more detail, IL-3 works through the stimulation of the JAK-STAT pathway (...) [by] activating downstream kinases (...) [MAPK151 and PKB152,] which (...) [in turn] activate [several] TFs (...) [MYC E2F1 and CCND1153] or support translation [CCND1] 353.4 [19]. Whereas CCND1 induces the expression of hsa-miR-92a 356, MYC and E2F1 cause the expression of both, hsa-miR-25 and hsa-miR-92a 127, 357. Strikingly, we found experimentally verified target sites in the 3'-UTR of HMGCR 122 for both miRNAs. This suggests a negative control of HMGCR's translation into the enzyme resulting in a limited synthesis of cholesterol through lowering the rate-limiting catalysis of the reduction of HMG-CoA154 to mevalonate. In carriers of the minor allele C, the binding of the RNA-induced silencing complexes is impaired. This leads to a depletion of the miRNA silencing potential as the binding accessibility for complex is shortened or even disturbed. In the end, this gives rise to the development of LDL cholesterol related traits 19.

6.2.2.4 MicroRNA control of rate-limiting enzymes

Mishra and Bertino [302] described that miRSNPs near a miRNA binding site affect the accessibility of a miRNA-RISC complex. This is confirmed by the secondary structure analysis of the 3'-UTR RNA sequence of HMGCR using RNAfold 325. The prediction results show a change in RNA folding caused by rs12916, leading to an impairment of the accessibility of the seed complementary region and the RISC binding site, respectively (Figure 6.6) [19].

Besides, Larsson et al. [358] reported recently that mRNA turnover rates have an important influence on the changes exerted by small RNAs (i.e., small-interfering RNAs (siRNAs) or miRNAs) on mRNA levels. SiRNAs are being developed for therapeutic use, for instance, in lowering LDL-cholesterol by targeting, e.g., enzymes such as PCSK9155 [358, 359]. Interestingly, miRNA silencing potential has been confirmed also for HMGCR. HMG-CoA reductase is acting for the class of rate-limiting enzymes where it is assumed that protein levels can be controlled via miRNA-mediated regulation in a convenient fashion. The potential of miRNA lies especially in fast changes in protein rates, making them likely to act as regulatory switches. This has also been reported for

151 MAPK: mitogen-activated protein kinase
152 PKB: protein kinase B
153 MYC: v-myc myelocytomatosis viral oncogene homolog (avian); E2F1: E2F transcription factor 1; CCND1: cyclin D1
154 HMG-CoA: 3-hydroxy-3-methylglutaryl-coenzyme A
155 PCSK9: proprotein convertase subtilisin/kexin type 9

139
other enzymes: for instance, the loss of \textbf{miRNA} cluster \textit{hsa-miR-29a/b-1} in \textbf{AD} patients correlates with an increased BACE/β-secretase expression. Those \textbf{miRNA}s are involved in the down-regulation of \textbf{BACE1} (beta-site APP-cleaving enzyme 1), which is the enzyme responsible for the rate-limiting step of the Aβ-production and a prime drug target for therapeutic intervention \footnote{[360]}.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6_6.png}
\caption{Secondary structure of the 3’-UTR of HMGCR under the influence of \textit{rs12916}. (a) Prediction results of RNA folding change of the 3’-UTR of HMGCR caused by \textit{rs12916} as provided by RNAfold\textsuperscript{19}. The seed complementary region for \textit{hsa-miR-25} and \textit{hsa-miR-92a} in the unmutated RNA (i.e. with the major \textit{T} allele) is accessible for the silencing complex\textsuperscript{19}. (b) The mutation (i.e. the minor \textit{C} allele) alters the folding making the binding of the \textit{RISC} impossible. (\ldots) Nucleotides are colored according to the probability of the base pairings (\ldots)\textsuperscript{19}. Illustration taken from Arnold \textsuperscript{19}.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6_7.png}
\caption{HMGCR - The role of microRNAs in LDL Cholesterol-Related Traits. Interleukin 3 (IL-3) was shown to reduce \textbf{LDL} cholesterol synthesis \textsuperscript{[352]}. The cytokine initiates the JAK-STAT signaling pathway resulting in the activation of \textbf{PKB} MAPK and c-MYC ([synonymous to MYC]). MAPK was shown to stimulate the transcription (\ldots) as well as - co-operating with \textbf{PKB} - the translation (\ldots) of cyclin D1 \textsuperscript{[355]}. cyclin D1 induces expression of the \textit{hsa-miR-17/20} cluster (\ldots) \textsuperscript{[260]} and additionally active c-MYC (\ldots) stimulates the expression of both \textit{hsa-miR-92a} (\ldots) as well as \textit{hsa-miR-25} (\ldots) \textsuperscript{[357]}. Both microRNAs are - incorporated in \textbf{RISC} (\ldots) - experimentally verified inhibitors of the translation of \textbf{HMGCR} transcripts (\ldots), thus limiting the concentrations of the enzymatically active end-product (\ldots). By limiting the throughput of the reduction of HMG-CoA to mevalonate, the cholesterol synthesis (\ldots) is directly lowered leading to normal \textbf{LDL} cholesterol (\ldots) concentrations in the plasma. In the mutated form of \textbf{HMGCR} (T to C mutation (\ldots)), the \textbf{RISC} are not able to bind the transcript (\ldots) leading to an increase of translated \textbf{HMGCR} and thus to raised mevalonate (\ldots) and cholesterol (\ldots) synthesis due to higher enzyme performance [see lower red box]\textsuperscript{[19]}. A result of impaired \textbf{miRNA} mediated control of \textbf{HMGCR} are high levels of blood \textbf{LDL} cholesterol that are treated with \textbf{statins} \textsuperscript{[19]}. Illustration taken from Arnold \textsuperscript{[19]}.}
\end{figure}
6.2 The impact of miRSNPs - a systems biological perspective on impaired microRNA-regulation
6 Analysis of the role of microRNAs in disease progression

Summary  Taking together, these findings strongly suggest that rs12916 represents a miRSNP, which causes a disturbance or even disruption of the miRNA binding signal of hsa-miR-25 and hsa-miR-92a. It has been functionally demonstrated that a polymorphism located in a miRNA binding site could lead to drug-resistance/drug sensitivity [302]. The potential miRSNP rs12916 might thus be a new player in that field in the context of LDL cholesterol related traits. As mentioned before, the broad medication with statins has been described as unnecessary for the prevention of CVDs [361]. For carriers of the mutation a statin treatment might however present a life prolonging treatment. Now, this hypothesis has to be validated experimentally [19]. The detailed systems biological model is summarized in Figure 6.7.

6.3 Summary and future perspectives

Within the last years, miRSNPs have been reported as a new functional class of polymorphisms that plays an important role in disease development and progression. Their detection may provide clues to many unanswered fundamental questions associated with disease as they shed new light on missing causative variants. MiRSNPs are emerging as a powerful tool to study the pathobiology of a disease, and have tremendous potential to be used in disease prognosis and diagnosis. Understanding the role and functions of miRSNPs has a promising future in pharmacogenomics, molecular epidemiology, and individualized medicine [302–304].

In this Chapter, we systematically analyzed the role of miRSNPs in complex human diseases from both, a transcriptional and a post-transcriptional perspective. By investigating the impact of SNPs detected in GWAS we were able to determine the significance of miRSNPs from the standpoint of population genetics. We used experimentally verified miRNA binding sites as they represent a much more reliable basis than computational predictions. Interestingly, we found no SNPs affecting miRNA biogenesis suggesting that common variants have no direct impact on miRNA-mediated regulation on the transcriptional layer. Besides strong evolutionary pressure also the widely discussed bias of GWAS towards coding regions might account for that. In contrast, we found several indications that GWAS associated SNPs strongly impact miRNA regulation on the post-transcriptional level. Common SNPs tend to be over-represented in the 3'-UTR sequences and even show a proximity bias towards experimentally validated miRNA target sites. We found many potential candidates and contextualized the results from a systems biological perspective in form of small-scale models. These models confirmed that our large-scale approach represents an appropriate framework enabling the biological interpretation of GWAS results under the particular consideration of impaired miRNA regulation. We generated detailed small-scale models where miRSNPs give causative explanations for disease development and therefore bridge the gap between genotype and phenotype on the microscopic level. Now, their experimental examination is mandatory to validate the proposed hypotheses. Especially for the potential miRSNP in the 3'-UTR of HMGCR this
might pave the way for new therapeutic targets or diagnostic methods. In addition to the hypothetical design of a RNA-based therapeutic, the miRSNP could be used as pharmacogenetic marker[19]. The broad medication with statins has been described as unnecessary for the prevention of CVD[361]. However, for carriers of the mutation it might present a life prolonging treatment[19]. Using the miRSNP as pharmacogenetic marker would therefore be beneficial to differentiate between both groups.

"(...) [By] adding microRNAs to the picture of genetic variance causing trait development, novel aspects not only for the interrelations in pathogenic disturbance of cellular processes, but also for the coherence of different traits can be addressed. For instance, differential tissue-specific expression patterns of microRNAs in combination with genetic variants may shed light on the still unknown functions driving the same cellular pathways (...)[20] different effects in diseased and healthy individuals. Closing the gap between trait-associated mutations and impaired microRNA-mediated gene regulation thus may lead the way to a"[19] better and more comprehensive understanding of cellular functions.

The presented findings highlight, however, also demands on GWAS-related research. "To advance the potential of conclusion drawing in GWAS results interpretation, it may be invaluable to overcome the GWAS bias towards the CDS and extend approaches based on 3'-UTR mutations. Conversely, more comprehensive data on microRNA binding sites might reveal regulation modules which could enhance our understanding of microRNA regulation functioning. (...) Aside from that, determination of microRNA regulation via RISC binding to other mRNA locations than the 3'-UTR might (...) [give information] on the effect of e.g. synonymous trait-associated alleles. The combination, i.e. extensive measurements of microRNA-mediated gene regulation in patients with traits for which a plausible model of microRNA involvement can be created in the context of associated mutations in the 3'-UTR, may further provide new perspectives of disease progression."[19]

Further analysis of the interactions between microRNA and other regulatory elements present in 3'-UTRs will shed more light on the function of microRNA polymorphisms and will eventually establish 3'-UTR as a hotspot for pathology.

"But the implications of the presented findings go beyond [mere] GWAS and microRNA related research. Considering the advances made in the exploration of the (...) [still poorly understood elements] of the genome, the impact of the presented results are pointed out. Just recently, the approaches of the modENCODE project[362, 363] led to the validation of the hypothesized hierarchical structure [(see also Chapter 4)] of physical regulatory networks in eukaryotes[156] which are based on a sophisticated interplay of microRNAs and transcription factors. Thus, the important role of this class of non-coding RNAs in the regulatory machinery of the cell is brought out on a large-scale level. If those findings can be transferred from the studied model organisms to human, the analysis of impairment of the transcription factor-microRNA network balance by mutationally altered target site functioning may lead to a completely new definition of genetically predisposed

i.e. Drosophila melanogaster and Caenorhabditis elegans.
diseases on a RNA-mediated, regulatory basis. Also, microRNAs are only one class of non-coding RNAs (...) [that] have been proven to feature regulatory power. Interference by, e.g., piwi-associated RNAs, siRNAs or (...) [long ncRNAs that] are all incorporated in protein-containing complexes targeting specific genes (especially, their 3'-UTR) will have to be assessed in this context to gather further insights.\[19\]

As a start towards this direction, we will focus on the role of post-transcriptional regulation in shaping tissue-specific metabolic activity profiles in Chapter 7.
CHAPTER 7

Systems biology of metabotypes

Genome-wide association studies and their follow-up studies, such as fine-mapping and next generation sequencing (NGS), provide novel insights into the predisposition to common diseases (Section 2.4). In the two previous Chapters, both, impaired up- and downstream mechanisms have been shown to be relevant for disease development. On the transcriptional level, disrupted enhancer binding sites have been found to be associated with diseases (Chapter 5). On the post-transcriptional side miRSNPs disturb microRNA binding sites causing regulatory imbalance and consequently leading to diseases (Chapter 6).

Genome-wide association studies, however, often fail to identify causal variants. Additionally, besides genetic predisposition, also environmental factors such as lifestyle and diet are supposed to play a major role in the onset of diseases, which after all are not accounted for by GWAS. Type 2 diabetes and other diseases related to the metabolic syndrome are not only of great medical, but also of enormous economical interest as the number of affected people is increasing dramatically. Prognoses for the year 2030 expect 439 million people (∼7.7% of the world population) to have diabetes. Meanwhile, the International Diabetes Federation (IDF) is already speaking of the epidemic of the 21st century. For developing appropriate therapies for their prevention and/or cure, an in-depth understanding on disease development and its underlying causal mechanisms is essential. However, in many cases the detailed connections between genotype, phenotype, lifestyle, diet and (impaired) metabolism are not well understood yet.

The metabolome represents the collection of all metabolites and thus is not only the functional readout of an organism’s physiological state. It is rather the very end of the sum of all influences it is exposed to, such as nutrition, environmental factors, or treatments (Section 2.3). Such factors are widely accepted to have an influence on disease progression, but the link between them is in most cases still missing.

Systems biological modeling allows for a deeper functional understanding of causalities. As shown previously in Chapter 6, we successfully applied this modeling procedure to explain pathological phenotypes attributed to impaired miRNA regulation (Sec-

---

157 http://www.diabetesatlas.org
158 http://www.idf.org/diabetes-epidemic-out-control
In this Chapter, we present a systems-biological perspective on metabolic phenotypes. We systematically analyze the impact of environmental influences in form of a high fat diet (HFD) in the context of non-alcoholic fatty liver disease (NAFLD) induction and progression. Based on a constraint-based computational approach (Section 2.3) originally presented by Shlomi et al. [22], we are able to study normal and abnormal metabolism in response to the diet in a time-resolved and thus dynamic manner. Significant changes in activity of specific metabolic categories during the study period were then analyzed on a more detailed level. In the end, comprehensive systems biological models reflecting these observations are constructed. MiRNAs have been associated with a wide range of diseases, in particular also with metabolic disorders [364–367]. Consequently, the focus lies on using miRNA-mediated regulation as missing link that brings genotype and phenotype closer together. For a more details see also [21].

7.1 Non-alcoholic fatty liver disease and its cross-links

Non-alcoholic fatty liver disease is one of the most common causes of chronic liver disease in the western world [368, 369] and supposed to be significant risk factor for the development of T2D and vascular disease [370]. "Today, it is commonly accepted, that NAFLD is closely related to over-[-nutrition and hepatic insulin resistance [(IR)]. Impaired insulin sensitivity [161] of hepatic tissue leads to an [excessive release of free fatty acids (FFAs)] due to [increased] lipolysis. At the same time, other insulin signal pathways [, which] (...) regulate glucose uptake and protein metabolism, remain unaffected. A biological explanation for that mechanism becomes clear against the background of times of fasting and starvation. [Under (...) these conditions, stored lipids - mainly triacylglycerols (TAGs) - serve as the primary energy source and fatty acid concentration[s] in liver raise (...). "[A] striking characteristic of NAFLD is triacylglycerol (TAG) accumulation in hepatocytes (fatty liver, hepatic steatosis)." Hepatic insulin resistance diminishes glucose uptake of liver and thus preserves glucose as an energy source for other glucose requiring organs (muscle, brain). (...) [Simultaneously], proteins, (...) essential for survival, are stored and not (...) supplied for energy [371]. (...) [But] the over-abundance of food, especially in the western world, made such an mechanism become needless. Fat-rich, ketogenic diets and positive caloric imbalance disrupt(...) such mechanisms controlling the interaction between genotype and phenotype [21], and ultimatively lead to the development of diseases such as NAFLD.

The strong connection between NAFLD and hepatic IR indicates that NAFLD might be linked to other diseases as well [21]. "For instance, a strong correlation between NAFLD and T2D (...) [has been] proposed by several studies [370, 372, 373]." Type 2 diabetes(...) is characterized by elevated plasma glucose (hyperglycemia) and insulin concent-

159 Basis for this Chapter was the diploma thesis of Jörn Leonhardt [21].
160 insulin resistance (IR) is a condition where insulin becomes less effective at lowering blood sugars.
161 Insulin sensitivity is known as the tissue responsiveness to insulin.
trations (hyperinsulinemia) due to a systemic impairment of insulin stimulated uptake of glucose. Although raised insulin secretion compensates this effect, ever less glucose is removed from the circulation on the long term [21]. „For a long time, T2D(...) has been seen as a disease of glucose homeostasis. (...) [Recently], it has been shown that lipid accumulation in hepatocytes is strongly correlated with systemic T2D [376, 377]. (...) The causal relation[s] between T2D and NAFLD [however,] remain(...) unclear. Whether T2D(...) promotes the accumulation of TAGs in liver or increased lipid concentrations (and their metabolic intermediates) result in systemic IR is not known [yet] [372, 378–381]. “[21]

“As insulin plays an important role in many processes and tissues [374], a connection of NAFLD to other metabolic disorders is [quite] likely [382]. Thus, gaining deeper insights in the mechanisms of hepatic metabolism and the onset of diseases like NAFLD is of particular interest.” [21]

7.2 Network-based prediction of metabolic behavior in [NAFLD]

To understand the underlying mechanisms of NAFLD, a detailed knowledge of closely linked disorders including IR, T2D or the metabolic syndrome is beneficial (Chapter 5), studying the functional, molecular, and causal relationship between genetic predisposition and environmental factors using animal models is essential.

For the study of NAFLD, dietary intervention studies in mice simulating human over-nutrition habits are typically used. By feeding a specific diet, NAFLD is induced in mice that in turn allows to study disease development and progression in a standardized and time-resolved fashion. In the following, details on the study design and the in silico approach used for analyzing changes in the metabolism due to dietary intervention are presented [21].

7.2.1 Study design and experimental setup

Here, a five-week safflower-oil rich diet\textsuperscript{162} using inbred male mice (C3HeB/FeJ strain\textsuperscript{163}) was implemented.\textsuperscript{164} The study is partitioned in two phases:

1. A three-week challenge period (d7, d14, d21) followed
2. by a two-week recovery phase (r7, r14).

In the challenge period, mice were fed with a safflower-oil rich diet, which we further refer to as high fat diet (HFD). After three weeks on HFD animals were nourished with standard chow again for the remaining two weeks within the recovery phase. For control purposes, littermate controls were fed with a standard chow during the complete study period. A schematic overview of the study is depicted in Figure 7.1.

\textsuperscript{162} The HFD consists of 27% safflower-oil.
\textsuperscript{163} Jackson Laboratory (http://jaxmice.jax.org/strain/000658.html)
\textsuperscript{164} Experimental data were provided by Melanie Kahle and Dr. Susanne Neschen, transcriptomics analysis was done by Barbara Fridrich and Dr. Martin Irmler (Institute of Experimental Genetics, Helmholtz Zentrum München).
Figure 7.1: Schematic overview of the HFD study design. Inbred male mice were fed differently during the study period. The challenge cohort was fed a HFD over a period of three weeks (blue). After three weeks, the diet was switched to standard chow (yellow) for the remaining two weeks. Littermate controls were fed a standard chow (yellow) over the full five weeks. After each week (challenge: d7, d14, d21; recovery: r7, r14) different samples were collected from challenged and control animals, respectively [21]. Illustration adapted from [21].

After each week (challenge: d7, d14, d21; recovery: r7, r14), different samples were collected from challenged and control animals, respectively. Besides physiological data in the form of insulin and glucose concentrations derived from plasma samples and TAG concentrations determined from liver tissue samples, hepatic gene expression profiles were measured for each time point [21].

7.2.2 In silico approach to model metabolic behavior

The metabolome represents the collection of all metabolites and thus is not only the functional readout of an organism’s physiological state, but also the very end of the sum of all influences it is exposed to, such as nutrition, environmental factors, or treatments [49]. This functional readout or metabolic behavior is measurable and represents the key to understanding of complex interrelations and by means of that allows for bridging the gap between genotype and phenotype.

As introduced in Section 2.3, CBMs are widely used for studying metabolic behavior. One very common approach to model cellular systemic behavior in the absence of detailed kinetic information is flux balance analysis (FBA). This kind of CBM-based approach assumes (...) that the biochemical network is regulated to maximize or minimize a (...) [defined] objective function [73, 383, 21] under certain conditions considering given bounds and balances, e.g. mass balance equation.

To analyze the impact of the HFD on NAFLD induction and progression in a time-resolved manner, we investigated changes in metabolic activity over the study period.
This was done according to the approach originally presented by Shlomi et al. [22], which represents an extension of FBA and facilitates the genome-wide study of normal and abnormal metabolism in a tissue-specific manner with a special focus on post-transcriptional regulation [21].

Predicting metabolic activity by FAA

Shlomi et al. [22] presented an approach to systematically predict tissue-specific metabolic behavior by integrating a genome-scale metabolic network with tissue-specific gene- and protein-expression data:

"Changes in gene- and protein-expression levels play a major role in controlling tissue-specific metabolic functions [384–386], and a strong correlation between gene expression and measured [387, 388] and predicted [389–391] metabolic fluxes is reported for microorganisms. To account for metabolic flux activity that is not reflected in the expression data (that is, post-transcriptional regulatory effects), (...) tissue-specific variations in enzyme-expression levels [were] not [treated] as the final determinants of enzyme activity, but as cues for the likelihood that the enzyme in question supports metabolic flux in its associated reaction(s). Network integration is then used to accumulate these cues into a global, consistent metabolic behavior, which reflects the outcome of putative post-transcriptional regulatory effects. [This approach that is further referred to as flux-activity analysis (FAA) extends the traditional FBA approach (Section 2.3.1).] [The] (...) method’s reliance on enzyme-expression data to infer tissue-specific metabolic flux eliminates the need for a priori [MLH] knowledge of tissue-specific objective functions and metabolites exchanged by the tissue with biofluids. Instead, the method provides predictions regarding tissue-specific metabolite uptake and secretion. To this end,(...) a discrete representation of significantly high or low enzyme-expression levels across tissues [was employed]. Network integration is then done by solving a constraint-based modeling optimization problem [(Section 2.3.1)] to find a steady-state [MLH] metabolic flux distribution (that is, an assignment of fluxes to all the reactions in the network) that, first, satisfies the stoichiometric and thermodynamic constraints embedded in the model and, second, maximizes the number of enzymes whose predicted flux activity is consistent with their measured expression level. In other words, the method aims to obtain a flux distribution where the number of flux-carrying reactions associated with highly expressed enzymes is maximized, and the number of flux-carrying reactions associated with lowly expressed genes is minimized.

The resulting predicted flux distribution is used to assign flux activity states to the genes, reflecting the presence and/or absence of nonzero flux through the enzymatic reactions they are associated with. For some of the genes, their flux activity state can be uniquely determined to be active or inactive, with associated confidence estimations. For others, their activity state cannot be uniquely determined because of potential alternative flux distributions with the same overall consistency with the expression data (mostly owing to isozymes or alternative pathways (...) [391, 392]. Because expression levels are not
enforced as exclusive determinants of metabolic flux, the flux activity states of genes may deviate from their expression states. Genes are thus considered to be post-transcriptionally up- or downregulated [(via miRNAs)] based on a difference between their measured expression level and their predicted flux activity state in a given tissue.\cite{22} An illustration of flux activity-state predictions is given in Figure 7.2. For more details on the method see the original publication by Shlomi et al.\cite{22}.

![Figure 7.2: Illustration of FAA according to Shlomi et al.\cite{22}. Circular nodes represent metabolites, whereas diamond nodes represent enzymes. White, red and green represent normal, significantly low and significantly high expression of the enzyme-encoding genes, respectively. Solid edges represent metabolic reactions. Broken edges associate enzymes with the reactions they catalyze. The predicted steady-state flux distribution, involving the activation of reactions, is shown as purple arrows.\cite{22} Based on the flux predictions, two enzymes are predicted to be post-transcriptionally regulated. As the highly expressed enzyme E7 is predicted not to support metabolic flux, it is hence considered to be downregulated. As the moderately expressed membrane transporter E4 is predicted to support metabolic flux, it is hence considered to be upregulated. Predicted fluxes through specific exchange reactions that cross the system boundaries represent the uptake and secretion of metabolites from the tissue. Of the five metabolites that can be exchanged with the tissue's surroundings (M1-2, M7-9), the method predicts the uptake of one metabolite (M1) and the secretion of two others (M7 and M8). Notably, the high-expression level of the membrane transporter of M1 indicates that it may be active, but it does not provide information regarding whether M1 is taken up or secreted from the tissue. In contrast, the integrated approach can determine the direction of flux for many reactions by propagating the known thermodynamic constraints on reaction reversibility and directionality throughout the network. In the current example, the direction of the activated pathway is inferred based on the irreversibility of enzymes E3 and E4.\cite{22} Illustration adapted from Shlomi et al.\cite{22}.
7.2 Network-based prediction of metabolic behavior in NAFLD

7.2.3 Metabolic activity in NAFLD

To combine the two layers of biological organization (metabolism and transcriptional regulation), we integrated the genome-scale mouse metabolic network provided by Selvarasu et al. [394] with treatment/time-specific expression data derived from the HFD (Figure 7.3). Expression values of genes from each time point were assigned to reactions in the metabolic model that are catalyzed by the corresponding enzyme (coded by the gene). Statistical significance of expression ratios between challenge cohort and control was determined using limma T-test. For expression profiles of days d7, d14, and d21, significantly altered expression values are those having a FDR < 10%, whereas the expression ratios of the top 50 genes (ordered by p-value) were used for days r7 and r14 as no genes with a FDR < 10% could be identified due to the similarity between challenge and control on these days. During the recovery period - the diet is switched to standard chow again - the expression values of both groups seem to converge [21].

Based on this, we generated five metabolic GEMs, one for each expression profile (time point). The resulting models differ in their reactions’ expression states and reflect the metabolic changes between challenge and control at the given time points (d7-d21, r7 and r14). Next, we applied the procedure of Shlomi et al. [22] to each of the GEMs with the aim to (i) systematically determine changes in metabolic activity behavior in response to the HFD and (ii) to manifest that those changes are actually attended by the NAFLD progression. Metabolic fluxes through the biochemical networks were predicted so that they achieved high consistency with the transcriptional status of the enzymes. As a result, for each of the five different physiological states the status of the metabolism was determined representing the metabolic behavior at the given time point. These time-specific GEMs serve as the basis for all further analysis [21].

7.2.4 Interpretation of metabolic activity profiles

The FAA of the genome-scale CBM in the context of NAFLD revealed predictions on profiles of metabolic activity corresponding to the system’s transcriptional state during the HFD period. Based on these activity profiles, we now investigate the potential impact of the HFD on NAFLD progression from different perspectives. For this we use different granularity scales as this top-down procedure has turned out to be successful in the previous Chapters. The workflow used is illustrated in Figure 7.3.

"First, on a large-scale [MLH] perspective, global trends in the changes of metabolic flux through the network due to the challenge [are] (...) studied. Regarding the underlying mRNA (...) expression profiles, the metabolic flux predictions set the information about the transcriptional state of the system into their metabolic context. As a result, [we]..."
(...), localize (...) biochemical subnetworks (...) showing significantly altered metabolic fluxes compared to the control, caused by the differential regulation of specific enzymes. These results provide (...) valuable information about a system-wide metabolic response to the physiological state and the role of certain metabolic subnets in the adaption to the challenge ([Section 7.3]). Second, on a mid-scale [MLH] level, (...) we study metabolic profiles of certain subnetworks that show(...) significant changes in their global trends. By focusing on smaller parts of the network, changes in the metabolic flux through single processes and enzymes [can] be observed ([Section 7.4]). Third, we generate novel hypotheses that give causal explanations for the effects observed on a large- and mid-scale. By integrating previous observations with information retrieved from the literature [176], qualitative small-scale models can be constructed in a systems biological fashion. They provide potential mechanistic explanations for the changes in hepatic metabolism and the progress of NAFLD from a microscopic perspective and represent the basis for new experiments and a refinement of the models (Section 7.5) [21].
7.2 Network-based prediction of metabolic behavior in [NAFLD]

Figure 7.3: In silico workflow of flux activity based analysis of mouse metabolism.  
(a) Treatment/time specific expression profiles are gathered from the experimental study (...). [These] data (...) [are] integrated with the genome-scale metabolic reconstruction of mouse metabolism (...)."[21] (b) "A large-scale constraint-based metabolic model is defined (...) and [FAA] is performed (...)."[21]. (c) "[On a large-scale,] context specific activity states of biochemical reactions refer to active or inactive metabolic pathways. Results are compared to physiological data and metabolomics measurements. "[21] (d) Metabolic categories with most significant changes are analyzed in more detail. (e) Small-scale qualitative models are generated to set the results into a systems biological context. (f) They lead to new hypothesis that have to be tested experimentally. This step represents a new iteration in the systems biological cycle [8, 9, 21].
7.3 Large-scale analysis – global trends of disease specific metabolism

A first impression on how HFD impacts on the mice metabolism is deducible from the predictions of the FAA (Section 7.2.3). To analyze metabolic changes over the study period, we compared predicted pathway activity and inactivity profiles of challenged mice and their corresponding control groups [21].

7.3.1 Metabolic activity decreases due to HFD

Pathways are classified in metabolic categories according to the Brite hierarchy from KEGG [395] (see also Figure 2.4). In Figure 7.4 we show the relative fraction of active and inactive reactions of each metabolic category, respectively. For single categories a similar trend in pathway activity can be observed, whereas the predictions on the inactivity reveal a considerable difference.

Interestingly, compared to other categories, lipid metabolism shows the strongest change in pathway activity and inactivity over the complete study period. This is a first indication that lipid metabolism is strongly deregulated due to the ketogenic diet. Besides, only slight differences in pathway activity were found for energy metabolism and metabolism of complex lipids. However, these are way stronger in pathway inactivity between the second and fourth week (d14-r7).

Taken together, these results indicate that hepatic metabolism in mice undergoes unique changes in response to the safflower-oil rich diet: metabolic activity decreases, whereas inactivity increases over the HFD period [21].

7.3.2 Reduced metabolic activity is accompanied by reduced lipid accumulation in hepatocytes

As mentioned before, lipid accumulation in hepatocytes is a typical characteristic of NAFLD [377]. Therefore, we compared TAG concentrations in liver between case and control animals in order to investigate whether predictions on metabolic (in-)activity correlate with the actual physiological state of the system.

Measured TAG concentrations show great divergences between both cohorts (Figure 7.5). After the first week (d7), animals that are challenged with the HFD show significantly higher hepatic TAG levels than controls (50% > 750 mg/g vs. 75% < 500 mg/g) [21].

169 If not especially mentioned, we refer to the challenge cohort.
170 amino acid metabolism, carbohydrate metabolism, energy metabolism, lipid metabolism, membrane transport, metabolism of cofactors and vitamins, metabolism of complex carbohydrates, metabolism of complex lipids, metabolism of other amino acids, nucleotide metabolism
171 The fatty acid structure represents an important building block of structurally more complex lipids such as phosphoglycerides or glycolipids.
172 Hepatic metabolism comprises all chemical reactions that occur within the liver and represents the source of many substances which are essential for an organisms’ health and survival.
Figure 7.4: Active and inactive reactions in main categories. (a) Values are the share of active reactions on each day (d7-d21, r7 and r14) in each category (compared to all reactions in the category, 100%). During the first and second week of the study, the number of reactions predicted active drops from approximately 12% (d7) to values around 3% (d14). After the third week, overall metabolic activity is predicted to be 1% (d21). During the recovery phase (r7 and r14), the values raise again to around 3%. In contrast, the ratios of inactive reactions amounts to 4% after the first, and 7% after the second week. A peak of 19% is reached after the third week of the study. During the recovery phase, the partition of inactive reactions amounts to 7%. Illustration adapted from [21].
Although the hepatic TAG levels remain elevated in the case group until the third week (d21) of the study, they are drastically reduced compared to the first two weeks: median TAG levels are 700 mg/g and 400 mg/g, respectively. Additionally, we observed that TAG level distributions in the challenged group begin to reduce. This results in a reasonable overlap of both cohorts and further assimilation of the TAG levels in both groups is even reached in the recovery phase. Here, additional concentration decrease in challenged animals as a consequence of the discontinued HFD [21].

A potential explanation for the elevated TAG concentrations during the first two weeks is an activation of metabolic processes involved in the storage of lipids in the liver. As shown before, the FAA predictions revealed a significant reduction of activity, but moderate levels of inactivity after the second week (Section 7.3.1). This suggests that processes relevant for lipid accumulation are still active after the second week. Interestingly, though still on HFD hepatic TAG concentrations decrease between the second and third week while involved metabolic processes are predicted to be inactive. The reduced TAG levels in the presence of lipid abundance is likely due to the down-regulation of processes involved in the synthesis of TAGs or mechanisms relevant in the uptake of necessary substrates by hepatocytes [21].

Both, the assimilation of the TAG concentrations and the predictions from the FAA imply that metabolic (in-)activity return slowly to the physiological conditions during the recovery phase [21].

Figure 7.5: Hepatic TAG levels in challenge and control. „Shown are the distributions of liver TAG levels measured after each week in animals from the challenge group (red) and their littermate controls (green). Black bars indicate the 50% quantile, whisker depict distributions' extrema. Lower and upper ends of the boxes indicate 25% and 75% quantiles, respectively.‘’ [21] Illustration taken from [21].
7.3 Large-scale analysis – global trends of disease specific metabolism

7.3.3 Relation between progression in NAFLD and metabolic flux predictions

Previous results indicate a direct relation of the observed changes in metabolic flux (in-)activity and NAFLD progression. Whether this is the result of raised lipid levels or due to processes leading to TAG accumulation becomes not clear from this large-scale perspective. To address this, we then focused on selected metabolic categories in order to analyze observed metabolic responses to the HFD in more detail [21].

Lipid and carbohydrate metabolism are intuitively connected to lipid accumulation in liver, which is strongly associated with hepatic IR. We expect a measurable impact of the HFD on both categories. Additionally, lipid and carbohydrate metabolism associated processes have not only several substrates in common, but are in general strongly interconnected (see Figure 2.4). Furthermore, much is known about the mutual regulation of glucose metabolism and lipid metabolism through insulin [396, 397].

7.3.3.1 Lipid metabolism and metabolism of complex lipids

Metabolic processes involved in lipid metabolism consist of two categories: lipid metabolism and metabolism of complex lipids. Whereas lipid metabolism covers reactions related to synthesis, degradation and transport of fatty acids and simple lipids, the metabolism of complex lipids comprises processes involved in the synthesis and degradation of TAG and other complex lipids.

Lipid metabolism showed the strongest change in its pathway activity and inactivity profiles over the complete study period (Section 7.3.1). For processes involved in the metabolism of complex lipids we observed significantly different trends (Figure 7.4). Together, this indicates that lipid metabolism responses to the HFD with an increased activity. A reasonable explanation for this is the abundant substrate supply for processes involved in fatty acid degradation [21]. Also stimulation of anabolic processes such as fatty acid synthesis, elongation or desaturation might account for enhanced lipid metabolism activity.

By producing even more substrate for lipid synthesis, processes involved in the biosynthesis of fatty acids might even intensify the effect of the HFD at the beginning of the study [21]. Lipid metabolism involved processes seem to respond more directly to the HFD than processes of the other categories. This explains the drastic decrease of metabolic activity during week two and three. In contrast to lipid metabolism, metabolism of complex lipids showed no characteristic response to the HFD rather the constant activity of processes involved in the generation of complex lipids might give reason for the accumulated TAG levels (Section 7.3.2) [21].
7.3.3.2 Carbohydrate metabolism

The metabolism of carbohydrates\(^{173}\) covers processes involved in the formation, breakdown and interconversion of carbohydrates. Although reactions related to carbohydrate metabolism show notable changes during the study, they differ significantly from those observed in lipid metabolism (Figure 7.4)\(^{21}\).

After the first week (d7), the ratio of active reactions is twice lower than those involved in lipid metabolism. The same accounts for inactive reactions after the third week (d21). Furthermore, the observed pathway inactivity shift takes place with one week delay compared to the corresponding shift in lipid metabolism. Lipid metabolism shows a rather direct response to the HFD induced TAG accumulation. In contrast, it seems that carbohydrate metabolism is affected later. The observed changes in carbohydrate metabolism can not be fully explained from this large-scale perspective as too many biochemical processes are involved. But it already gives a rough picture which categories are affected at all. Since causal relations to the observed patterns in lipid metabolism are not obvious, we will analyze this category in more detail\(^{21}\).

7.3.4 Affected enzymes show strong bias to lipogenic processes and disorders in hepatic lipid homeostasis

So far, we found first evidences that the deregulation of lipid metabolism leads to an accumulation of lipids as a consequence of the HFD. Subsequently, inactivation of related processes seems to improve progression of NAFLD. The underlying biochemical reactions are highly interconnected (Figure 2.4) and consequently, changes propagate within the metabolic network that are not considered yet. These are not obvious from the global (in-)activity trends. To identify affected connections, we performed an enrichment analysis\(^{174}\) of (in-)active enzymes (compared to the control) regarding metabolic disorders and biological processes using GeneCodis\(^{224, 225}\).

**MeSH enrichment** We mapped disease terms\(^{175}\) to all genes that have been predicted active or inactive. Then, we performed an enrichment analysis for each time point (d7-r7). Results from this showed a strong bias to diseases connected to lipid metabolism or liver in the first two weeks of the study. This is congruent with the significant changes in lipid metabolism activity identified by the FAA. Furthermore, these results also confirm that processes involved in lipid metabolism undergo particular responses due to the HFD. Identified regulatory changes however, are not limited to lipid metabolism.

---

\(^{173}\) Glycolysis/Gluconeogenesis; Citrate cycle (TCA cycle); Pentose phosphate pathway; Pentose and glucuronate interconversions; Fructose and mannose metabolism; Galactose metabolism; Ascorbate and aldarate metabolism; Starch and sucrose metabolism; Amino sugar and nucleotide sugar metabolism; Pyruvate metabolism; Glyoxylate and dicarboxylate metabolism; Propanoate metabolism; Butanoate metabolism; C5-Branched dibasic acid metabolism; Inositol phosphate metabolism

\(^{174}\) We used hypergeometric test; FDR; significance level 0.05

After the third week (d21), we found enrichment in diseases involved in glucose metabolism giving evidence towards a connection to IR. In the recovery phase, affected genes are enriched in more specific terms such as necrotic DNA degradation or kidney failure, indicating already initiation of induced IR [21]. For more details see the diploma thesis of Leonhardt [21].

GO enrichment Likewise, we applied GO enrichment analysis to all predicted (in-)active genes at each time point. To this end, the genes were tagged with their respective GO terms and analyzed for overrepresentation of certain functional terms. After the first two weeks, almost all enriched terms can be related to processes involved in lipid metabolism. This corresponds to our previous findings on pathway activity. After week three, we found that affected genes are enriched in processes related to transport (sterol transport and cholesterol transport), indicating first changes in molecular rearrangement and lipid exchange between tissues by lipoproteins [397]. Additionally, immune response related terms were found to be enriched after the third week. Such mechanisms due to dietary treatment have been reported in several studies, especially in the context of dietary induced IR [398–400]. Interestingly, we observed in Chapter 5 that autoimmune disorders are highly interlinked and also related to, e.g., obesity. Finally, terms related to carbohydrate and amino acid metabolism were found to be overrepresented during the recovery phase. A potential explanation for this might be that the switch to a standard diet after the third week has a more balanced ratio of lipids and carbohydrates than the HFD. Consequently, metabolic processes involved in carbohydrate metabolism might be induced [21]. An enrichment in Glucose transport indicates improved IR during the recovery phase, which would lead to enhanced glucose transporter expression and activity [397]. As before, the enrichment becomes more unspecific during the recovery phase [21]. For more details see the diploma thesis of Leonhardt [21].

7.4 Mid-scale analysis – changes in hepatic lipid metabolism

„(...) Deregulation of hepatic lipid metabolism is an important factor in the progress of NAFLD and its consequences [377, 401, 402].” From a macroscopic perspective, we found evidence that HFD seems to have a measurable impact on NAFLD development (Section 7.3). Using FAA predictions, we were able to identify notable changes in the metabolism of the challenged cohort due to the HFD. Especially processes involved in lipid metabolism showed striking variations underpinning its important role. This was even strengthened as genes predicted to be (in-)active tend to be enriched in processes and disorders related to lipid metabolism. Furthermore, also TAG concentration levels in the liver correlate with the observed metabolic changes. These findings suggest lipid metabolism as a critical player. To get a more comprehensive picture of the incidents in and the behavior of relevant processes, we now focus on the systematic analysis of reactions involved in hepatic lipid metabolism from a mid-scale perspective.
Lipid metabolism can be divided into anabolic processes involved in the synthesis\textsuperscript{176} of fatty acids and lipids (e.g. TAGs, phospholipids, or cholesterol), and catabolic processes that generate energy and primary metabolites\textsuperscript{177} by their degradation\textsuperscript{178} (Figure 2.4).

In the following, we discuss the results of the FAA and their relevance in the context of NAFLD for both anabolic (lipogenesis in Section 7.4.1) and catabolic processes ($\beta$-oxidation in Section 7.4.2). Whereas lipogenesis\textsuperscript{179} is the central source of TAGs in hepatocytes, hepatic fatty acid oxidation represents the counterpart to fatty acid biosynthesis\textsuperscript{21}.

7.4.1 Metabolic trends in lipogenic processes

Fatty acids, stored as TAGs in an organism, represent an important source of energy. One major source of hepatic TAGs is the generation of lipids from lipid- and non-lipid-precursors. Besides TAG-synthesis, lipogenesis comprises also processes for the generation of other complex lipids such as phospho- and sphingolipids and sterols. The two major sources of fatty acids used in lipogenesis are (i) \textit{(de novo)} fatty acid biosynthesis, which represents the synthesis of fatty acids from non-lipid precursors such as glucose via acetyl-CoA, and (ii) the elongation of existing fatty acids, which are either ingested or provided by lipolysis of lipids in adipocytes\textsuperscript{21}.

7.4.1.1 Prolonged feeding of safflower-oil impairs lipogenic processes

As described previously, we found enhanced lipid metabolism activity as a consequence of the HFD (Section 7.3.3.1). This becomes even more apparent when analyzing the individual lipogenic processes\textsuperscript{180} (Figures 7.6).

Almost 70\% of the reactions involved in fatty acid biosynthesis and fatty acid elongation were predicted to be active after the first week (d7), whereas they are significantly decreased during the second week (d14). Especially fatty acid biosynthesis showed a low activity. At the same time, inactivity profiles are considerably increased. Together, this suggests that the amount of fatty acids provided by \textit{(de novo)} fatty acid synthesis is rather low. Around 25\% of the reactions related to fatty acid elongation were predicted to be active, however, no change in metabolic inactivity is observable until the end of the second week (d14)\textsuperscript{21}. „[A] (...) delayed down-regulation of enzymes that are involved in fatty acid elongation [might account for this]. (...) [This is likely due to] feedback control of fatty acid biosynthesis through (...) [raised] fatty acid levels in the cell\textsuperscript{403}.“\textsuperscript{21} Likewise

\textsuperscript{176} fatty acid biosynthesis, fatty acid elongation and fatty acid esterification (belongs to glycerolipid metabolism)

\textsuperscript{177} Metabolite that is directly involved in normal growth, development, and reproduction.

\textsuperscript{178} fatty acid oxidation (also $\beta$-oxidation) and fatty acid metabolism

\textsuperscript{179} Lipogenesis is the process by which acetyl-CoA is converted to fats.

\textsuperscript{180} Lipid Metabolism: bile acid biosynthesis, fatty acid biosynthesis, fatty acid biosynthesis (path 2), fatty acid biosynthesis mitochondrial, fatty acid elongation, fatty acid elongation in mitochondria, fatty acid metabolism, fatty acid oxidation, phospholipid biosynthesis, sterol biosynthesis; Metabolism of Complex Lipids: folate biosynthesis, glycerolipid metabolism, inositol phosphate metabolism, phospholipid degradation, shingoglycolipid metabolism
7.4 Mid-scale analysis – changes in hepatic lipid metabolism

Figure 7.6: Active and inactive predictions of reactions involved in the processes of lipid metabolism and metabolism of complex lipids. Values reflect the ratio of (a) active and (b) inactive reactions involved in the processes of lipid metabolism after each week of the study (d7-d21, r7 and r14). Values reflect the ratio of (c) active and (d) inactive reactions in the processes involved in the metabolism of complex lipids after each week of the study (d7-d21, r7 and r14). Predictions are separated into corresponding Brite subcategories. Illustration adapted from [21].

applies after the third week (d21) where both anabolic processes (fatty acid biosynthesis and fatty acid elongation) showed low activity and high inactivity implying that fatty acid production through synthesis and elongation is significantly impaired at this time point [21]. Interestingly, while the ratios of inactive reactions in fatty acid biosynthesis remained high (..., ∼60%) during the recovery phase (r7, r14), ratios of reactions involved in fatty acid elongation (..., ) noticeably decreased again (Figure 7.6), [21]
Processes involved in the esterification\textsuperscript{181} of fatty acids in order to build TAGs (see glycerolipid metabolism) showed constant activity with only a moderate increase after the third week (d21). Inactive reactions show a similar profile with an one week shift forwards indicating that affected metabolic processes are not substantially different between cohort and control groups (Figure 7.6) \cite{21}.

7.4.1.2 Reduced lipogenic activity results in lower hepatic TAG levels

To correlate predicted metabolic states with observed physiological states, we compared – as before – the (in-)activity profiles of lipogenic processes to TAG concentrations measured in hepatocytes (Section 7.3.3). High activity rates of both, fatty acid biosynthesis and fatty acid elongation, indicated that abundant substrate for the generation of TAGs - by esterification to G3P (glycerol 3-phosphate) - is provided by the hepatic fatty acid pool. Elevated hepatic TAG levels after the first week confirmed this (Figure 7.5) \cite{21}. Substrates for fatty acid elongation originate most likely from the HFD as the diet is rich in linoleic acid (C18:2) \cite{21}, a common precursor for elongation and esterification \cite{397}.

\textquotedblleft After the second week, fatty acid biosynthesis seem[s] (...) to be quite inactive, while fatty acid elongation show[s] (...) still certain activity. [But] (...) elongation of existent fatty acids and those ingested by the diet seem(....) to be sufficient for active lipogenesis, as [indicated] (...) by elevated TAG levels at this time point. After the third week, both pathways involved in fatty acid supply show(....) significant (...) [and] glycerolipid metabolism certain inactivity. Simultaneously, TAG (...) levels begin to reduce	extquotedblright \cite{21}.

7.4.1.3 Glycerolipid synthesis slows down due to substrate shortage or regulatory switches

So far, a causal relation between (in-)activity profiles of hepatic lipid metabolism and the observed progression in hepatic lipid accumulation is not obvious. A potential explanation is that reduced fatty acid biosynthesis and elongation lead to decreased esterification of fatty acids \cite{21}. \textquotedblleft Providing less substrate (fatty acids) to the hepatic fatty acid pool [might] (...) result in a slower TAG (...) synthesis. [Furthermore,] (...) other substrates involved in the esterification of TAGs, especially G3P (...) from glycolysis cleaved TAGs, [might also account] (...) for slower glycerolipid synthesis (...) \cite{397}.

\textquotedblleft However, during the recovery phase, activity and inactivity ratios in glycerolipid metabolism returned to initial values. During [this period] (...) the two other anabolic categories showed rather inactive behavior. Besides, fatty acids ingested by the diet should be significantly less than in the challenge phase. Thus, a reduction of esterification activity due to fatty acid shortage becomes more unlikely. Another reason might be that regulatory mechanisms directly suppress processes involved in glycerolipid metabolism in response to the HFD. Especially the increased inactivity of most metabolic processes suggests a global switch in metabolic regulation.

\textsuperscript{181} Esterification is the chemical process for making esters, which are compounds of the chemical structure R-COOR’, where R and R’ are either alkyl or aryl groups. \url{http://science.jrank.org/pages/2573/Esterification.html}
Interestingly, glycerolipid metabolism showed certain activity during the recovery phase while TAG concentrations were diminishing. This might be due to an enhanced secretion of lipids by very low density lipoprotein (VLDL) particles. But plasma TAG concentrations after the third week showed only a slight increase compared to the first two weeks and were unlikely to account for the drastic reduction in hepatic TAG during the second and third week [21].

7.4.2 Metabolic trends in hepatic fatty acid oxidation

Another factor that contributes to TAG accumulation in hepatocytes is impaired fatty acid oxidation. Fatty acid oxidation (also β-oxidation) represents the counterpart to fatty acid biosynthesis and comprises metabolic processes involved in the step-by-step breakdown of fatty acids to acetyl-Coenzyme A (acetyl-CoA). Genes involved in β-oxidation are summarized in the category fatty acid metabolism (see also Figure 2.4).

7.4.2.1 Prolonged HFD leads to reduced catabolic activity

The (in-)activity profile of β-oxidation predicted by the FAA shows significant changes similar to those of fatty acid elongation during the study period (see fatty acid metabolism in Figure 7.6). Initially, β-oxidation seems to be highly induced by the HFD. Consequently, reduced levels of fatty acids are expected.

Fatty acids are - together with malonyl-Coenzyme A (Malonyl-CoA) - essential substrates for fatty acid elongation. In turn, one anticipates elevated levels of acetyl-CoA - substrate for fatty acid biosynthesis - as a result of stimulated fatty acid oxidation. This conceivably explains the high activity of fatty acid biosynthesis after the first week. Abundant lipids and consequently fatty acids are assumed to be provided by the diet, thus enough substrate for fatty acid oxidation and elongation should be present.

After the second week, the metabolic activity in fatty acid oxidation was significantly reduced, while inactivity raised. Although fatty acid oxidation is not the only source of acetyl-CoA [397] this is a potential connection to the reduced activity in fatty acid biosynthesis after the second week. Reduced cleavage of fatty acids by fatty acid oxidation also improves the availability of fatty acids in the cytosol for fatty acid elongation and glycerolipid synthesis. Predictions after the third week suggest low metabolic flux through this process. This fits with the overall metabolic inactivity predicted at this time point. During the recovery phase (r7, r14), activeness of β-oxidation gains again to some extent, while the amount of inactive reactions remains elevated [21].

7.4.2.2 HFD causes a deregulation of fatty acid oxidation

Interestingly, (in-)activation of fatty acid biosynthesis and β-oxidation related processes occur simultaneously in response to the HFD. Both are contrarily regulated under

---

182 propionyl-Coenzyme A (Propionyl-Coa) in the case of odd-numbered fatty acids
physiological conditions\footnote{Conditions of the external or internal milieu that may occur in nature for that organism or cell system.} and thus prevent useless cycles of fatty acid synthesis and degradation [403]. Assimilated regulation might, however, be beneficial in cases of abundant lipid availability as enhanced $\beta$-oxidation would antagonize the effects of TAG accumulation due to increased lipogenic activity \cite{21}.

Finally, the analysis of the step-by-step down-regulation of $\beta$-oxidation during the first three weeks of the study revealed interesting results. Especially the trend in inactivity shows a smooth progress, probably due to a differentiated control of various enzymes involved in fatty acid oxidation. To get a dynamic impression of enzymatic inactivity during the study period, we mapped their corresponding (in)-activity states to the metabolic subnetwork of $\beta$-oxidation (Figure 7.7). In the beginning (d7), none of the enzymes was predicted to be inactive switching to a high amount of inactive enzymes after two weeks (d14) that is even increased after three weeks (d21) (Table 7.1) \cite{21}.

The difference between the second and third week lies in the activity of enzymes responsible for the dehydrogenation of fatty acids. Whereas after the second week Acadm\footnote{Acadm: acyl-Coenzyme A dehydrogenase, medium chain} – important for degrading medium-chain fatty acids – is already predicted to be inactive, its counterpart Acadl\footnote{Acadl: acyl-Coenzyme A dehydrogenase, long-chain} – important for degrading long-chain fatty acids – is still active until the end of the third week. Acadm represents a crucial enzyme in the degradation and processing of medium-chain fatty acids. Its inactivity would therefore impair the complete degradation of fatty acids to acetyl-CoA. This has been reported, e.g., for medium-chain acetyl-CoA dehydrogenase deficiency, which is caused by mutations in the Acadm gene leading to altered the enzyme’s structure, reducing or abolishing its activity or in the worst case to an unstable enzyme that cannot function at all. With a shortage (deficiency) of functional Acadm enzyme, medium-chain fatty acids cannot be degraded and processed. As a result, these fats cannot be converted into energy leading to characteristic symptoms of this disorder, such as lack of energy and low blood sugar. Levels of medium-chain fatty acids or partially degraded fatty acids may build up in tissues leading to a damage of the liver and brain and thus causing more serious complications \cite{21}.

But what is happening then to incompletely oxidized fatty acids? Either they are exported to the cytoplasm where they might enter lipogenesis resulting in a re-esterification to TAGs or they are exported directly to plasma without esterification or binding to lipoproteins \cite{21}.
Figure 7.7: Inactive enzymes involved in fatty acid oxidation after each of the first three weeks of the study. The mapping of all inactive enzymes (red) to the mitochondrial fatty acid oxidation pathway shows the step-by-step inactivation of fatty acid biosynthesis over the first three weeks of the study: (a) d7, (b) d14, (c) d21. Illustrations adapted from KEGG [21, 395].
7 Systems biology of metabotypes

7.5 Small-scale analysis – from observations to systems biological models

The step-by-step analysis from the large- and mid-scale perspectives revealed that especially processes involved in hepatic lipogenesis and hepatic fatty acid oxidation show characteristic responses to the HFD. Observations such as the simultaneous activation and inactivation of fatty acid biosynthesis and fatty acid oxidation underscore the expectation that the regulatory mechanisms of hepatic metabolism is affected by the diet. But a causal relationship between observed metabolic states and corresponding physiological conditions can not be established so far [21]. To find causal and molecular explanations for the observed changes in lipid metabolism, the underlying biological processes have to be analyzed on a systems-level, likewise in Chapter 6.

Here, we combine results from the large- and mid-scale analyses with additional information from the literature [176] in order to construct small-scale systems biological models. This procedure turned out to be successful in generating plausible hypotheses, which facilitate a more detailed understanding of the underlying mechanisms in the context of a disease. The hypotheses in turn allow a subsequent experimental verification or falsification, thereby initiating the next round in the systems biological cycle (Section 6.2).

In this Section, we present representative hypotheses, which link the observed metabolic changes and physiological data in the context of NAFLD in a systems biological manner. For this purpose, we combine existing knowledge with a detailed analysis of how the interplay of involved genes and their potential post-transcriptional regulation impacts the development of NAFLD in response to the HFD. As discussing all potential hypotheses is beyond the scope of this thesis, we concentrate on three representative models where we narrow the gap between genotype and NAFLD by incorporating miRNA-mediated regulation [21].

7.5.1 Downregulation of Ppap2a improves hepatic steatosis and IR

Elevated plasma insulin levels (hyperinsulinemia) are one of the typical symptoms for impaired insulin sensitivity. The pancreas compensates for the reduced insulin mediated response of tissues by increasing insulin secretion [397]. Lipid accumulation in hepatocytes is a typical characteristic of NAFLD [377] that is correlated to hepatic and systemic IR. A causal relation between the observed metabolic states of lipogenic processes, hepatic TAG accumulation and insulin sensitivity is, however, still missing.

Here, we present a mechanistic explanation for a potential causal relation between the predicted activity of the enzyme Ppap2a [86] improved hepatic TAG accumulation and reduced plasma insulin concentrations [21].

---

186 Ppap2a: phosphatidic acid phosphatase type 2a
7.5.1.1 **HFD** leads to elevated plasma insulin levels

Measured insulin concentrations in the plasma showed significantly elevated values during the first three weeks (Figure 7.8). These observations strongly propose - together with the results from previous studies [370, 377, 404] – an evolving IR as a consequence of the HFD. Additionally, a correlation with the progression of NAFLD - reflected by increased hepatic TAG concentrations - becomes clear (Figure 7.5) [21].

7.5.1.2 Activity trends in lipogenesis

From the macroscopic analyses we found indications that lipogenic processes become inactivated during the HFD period of three weeks (Figure 7.6). Processes involved in the synthesis and elongation of fatty acids showed a strong reduction in activity after two weeks. In contrast, processes involved in the esterification of fatty acids to TAGs (glycerolipid metabolism) showed steady activity, which is almost identical to that of the control groups. Only after the third week, we found a considerable increase in enzymes that have been predicted to be inactive. Simultaneously, hepatic TAG concentrations showed considerably reduced values, though plasma insulin levels are raised (Figure 7.8) [21].

![Figure 7.8: Distributions of plasma insulin concentrations in challenge and control. Red boxes depict values of animals of the challenge cohort, green boxes those of the control. Shown are the 25% (lower bound), 50% (black bar) and 75% quantiles (upper bound) of the distributions. Whiskers depict extrema. Illustration taken from [21].](image)

7.5.1.3 Second messenger **DAG** impairs insulin signaling

Insulin action is mediated through the insulin receptor by phosphorylating specific insulin receptor substrates. Once insulin binds to the receptor, the phosphorylation of the
7 Systems biology of metabotypes

Protein kinase Irs2 is induced \[403\]. Activated Irs2 in turn stimulates the activation of other protein kinases including Akt2 \[403\], which has been shown to be a crucial activator and inactivator of IR involved in the regulation of hepatic lipid and glucose metabolism \[403\]. Furthermore, a direct relation between Akt2 and NAFLD has been proposed by Leavens et al. \[405\]. It is known that the lipid second messenger diacylglycerol (DAG), which is the precursor of TAG, plays an important role in this relationship as DAG is supposed to regulate the activity of protein kinase Prkce \[406\]. Raised DAG concentrations lead to the activation of Prkce that in turn prevents Irs2 to become phosphorylated. As a consequence, Irs2 cannot get activated in response to insulin \[407\]. This goes along [perfectly] with the insulin levels observed in the plasma \[21\] of HFD challenged animals (Figure 7.8).

That DAG plays a key role in the progress of NAFLD and IR has been confirmed by Kantartzis et al. \[381\] who showed that over-expression of the enzyme Dgat2 \[381\] leads to the dissociation of TAG accumulation and IR. Dgat2 is involved in the final phase of TAG esterification and catalyzes the step from DAG to TAG. Thus, over-expression due to elevated protein concentrations stimulates the rate of Dgat2 activity which in turn leads to reduced DAG concentrations \[21\]. „Besides, they showed (…) that stimulated rates of DAG cleavage abolished their negative effect on hepatic insulin signaling, while, at the same time, hepatic lipid accumulation was not decreased.“ \[21\]

7.5.1.4 Ppap2a regulates hepatic DAG and TAG concentrations

Ppap2a is a crucial enzyme, which is involved in the esterification of long-chain fatty acids to glycerol 3-phosphate (G3P) in order to build TAGs: it converts phosphatidic acid (PA) to DAG \[397, 402\]. Predictions from the FAA revealed no significant change in metabolic (in-)activity for Ppap2a during the first two weeks of the study. However, according to the predictions, Ppap2a becomes inactivated after the third week. This matches our previous observations on hepatic TAG levels that showed a significant decrease during the second and third week (Figure 7.5 \[21\]. „During the recovery phase, Ppap2a showed neither raised nor reduced activity (…). But what leads to the reduced TAG levels during the recovery phase?“ \[21\]

Fatty acids ingested from the diet should be significantly reduced at this time point. Furthermore, since fatty acid biosynthesis and fatty acid elongation showed trends towards inactivity, endogenous fatty acid supply should also be low. As a result, compared to the first two weeks of the study, a reduced DAG synthesis is caused by substrate shortage. Subsequently, we expect that there should be no increased induction of TAG synthesis due to elevated DAG concentration \[21\].

Interestingly, although no significant change in expression levels of Ppap2a has been ob-

---

187 Irs2: insulin receptor substrate 2
188 Akt2: thymoma viral proto-oncogene 2
189 Prkce: protein kinase c, epsilon
190 Dgat2: diacylglycerol O-acyltransferase 2
served, the FAA predicted a null flux through that enzyme after the third week. According to Shlomi et al. [22], genes with null fluxes indicate potential targets of post-transcriptional regulation, which potentially explains the difference between a gene’s observed transcriptional states and the corresponding (in-)activity predictions (Section 7.2.2) [21].

7.5.1.5 Ppap2a mediated improvement of insulin sensitivity and hepatic steatosis via post-transcriptional regulation

Our results indicate that Ppap2a has effects on hepatic TAG concentrations and DAG signal propagation. Inactivation of its enzymatic activity leads to either a drastic reduction of hepatic TAG concentrations or a long term improvement in insulin sensitivity [21]. But up to now, no causal explanation for the mechanism underlying its inactivation due to the prolonged HFD exists. We found neither evidence for a repression of Ppap2a on the transcriptional level nor information on this in the literature [21]. Consequently, it seems to be obvious that other mechanisms on the post-translational or -transcriptional level might account for the observed inactivation of phosphatidic acid phosphatase type 2a (Ppap2a) [21]. For instance, the particular composition of the safflower-oil rich diet - rich in C18:2 PUFA - might have an impact on Ppap2a activity. Furthermore, a constituent or metabolic intermediate might allosterically bind to the enzyme and impair its function [21]. Protein modification through phosphorylation or acetylation of specific residues are also conceivable [21]. Hirschey et al. [408] showed that so-called sirtuins (acetylases) represent powerful regulators of enzyme activity in the adaption to environmental stimuli. Interestingly, miRNA are predicted to control the activity of ~ 50% of all protein-coding genes and changes in miRNA expression have been associated with the pathogenesis of many human diseases such as cancer and metabolic disorders [107, 108]. Additionally, we showed that polymorphisms in the 3'-UTR of enzymes lead to impaired miRNA-mediated regulation and consequently have a significant impact on the enzyme’s turnover rates (Section 6.2.2.1). Fine regulation of gene expression helps to achieve a constant metabolic adjustment which is required for the maintenance of homeostasis during environmental flux. miRNAs are key players in this regulatory milieu (Section 2.5), they have been implicated in regulating gene expression within several metabolically active tissues including the endocrine pancreas, liver, and adipose tissue [109, 110]. According to the idea presented by Shlomi et al. [22], Ppap2a is a potential candidate for miRNA mediated regulation which would close the gap between the FAA predictions and the experimental data [21].

191 For literature search we used Text Mining [176].
192 poly-unsaturated fatty acids [PUFA]
Summary  Our results yielded the hypothesis that - regardless of the mechanism behind the inactivation of the enzyme \( \text{Ppap2a} \) represents a crucial factor in the connection between diet induced NAFLD and hepatic IR (Figure 7.9). The enzyme \( \text{Ppap2a} \) is involved in the esterification of long-chain fatty acids to \( \text{G3P} \) where it catalyzes the reaction from lysophosphatidic acid \( \text{LPA} \) to \( \text{DAG} \), the precursor of \( \text{TAG} \). According to our data we hypothesize that after week three of the study a so far unknown post-transcriptional regulator leads to inactivation of \( \text{Ppap2a} \). This has the consequence that \( \text{DAG} \) concentrations reduce resulting in lowered \( \text{TAG} \) concentrations due to substrate shortage. The impairment of \( \text{DAG} \) signaling leads to a reduced inhibition of insulin mediated phosphorylation with the result that insulin sensitivity in hepatocytes starts to improve. To test this hypothesis and to examine whether \( \text{miRNA} \) mediated control of \( \text{Ppap2a} \) accounts for its inactivation, further experiments are essential [21].

![Figure 7.9: Deregulation of Ppap2a reduces liver TAG concentrations and improves insulin sensitivity. In [21] it was hypothesized that \( \text{Ppap2a} \) catalyzes the reaction from lysophosphatidic acid \( \text{LPA} \) to diacylglycerol, which is a precursor of triacylglycerol. After the third week, a so far not known post-transcriptional regulator leads to inactivation of \( \text{Ppap2a} \). Such a regulation could be due to ligand binding, residue modification or \( \text{miRNA} \) mediated mRNA cleavage. As a consequence, \( \text{DAG} \) concentrations reduce. One immediate effect of this reduction is a lowered \( \text{TAG} \) concentration due to substrate shortage. Also \( \text{DAG} \) signaling is impaired, which leads to reduced inhibition of insulin mediated phosphorylation. In turn, insulin sensitivity in hepatocytes starts to improve. Red = induction/activation/increase, green = repression/inactivation/decrease. Illustration taken from [21].](image-url)
7.5 Small-scale analysis – from observations to systems biological models

7.5.2 Ppara and Fgf21 mediated regulation of hepatic fatty acid oxidation

Next, we present another systems biological model, which provides a potential mechanism explaining the role of Ppara and Fgf21 deregulation on hepatic fatty acid oxidation (Figure 7.10) (see also [21]). Again, mRNA mediated control allows to narrow down the gap.

7.5.2.1 Nuclear receptor Ppara is activated by fatty acids

The nuclear receptor (NR) Ppara, which gets activated by fatty acids [411–414], is required for the transcription of numerous genes involved in fatty acid transport and oxidation [415], including Acox1 [415] and Cpt1 [416, 417]. Cpt1 is a rate-limiting enzyme that regulates fatty acid oxidation in mitochondria by facilitating the transport of fatty acids across the outer mitochondrial membrane [397]. High Cpt1 levels have been associated with greater lipid - free fatty acids (FFAs) - oxidation and less conversion of FFAs into TAGs [418]. Furthermore, mice lacking Ppara have been shown to accumulate hepatic TAGs due to reduced fatty acid cleavage [419–421]. Besides, Ppara is a known activator of Fgf21 gene expression [422].

7.5.2.2 Fgf21 overexpression is strongly associated to stimulated fatty acid oxidation

Fgf21, which is predominantly expressed in liver [423], is known to stimulate glucose uptake in adipocytes [424]. Common polymorphisms in the Fgf21 signaling pathway have been associated with metabolic risk [425] and complete lack of Fgf21 due to knock-down experiments has been reported to lead to a decreased expression of several genes involved in fatty acid oxidation, namely Acadl, Acadm, and Hadh [426]. Furthermore, the repression of the hepatic fatty acid transporter Cd36 [427] has been reported [426, 427]. Besides, also an association between Fgf21 and Cpt1a (synonymous for Cpt1), and Acox1 respectively exists [422].

Interestingly, from the expression data we found Fgf21 to be increased in mice on the HFD, however at the same time, corresponding mRNA levels of Cpt1a and Acox1 were not increased in the presence of Fgf21. This raises the question whether a post-transcriptional regulation might account for that [21]?

7.5.2.3 Fatty acid oxidation is dissociated from Fgf21 expression

Previously, we observed a continuous reduction in β-oxidation activity with a clear switch from activity to inactivity after the second week (Section 7.4.2). Interestingly, for Fgf21 a two-fold increase in gene expression was measured for the same time. This observation

---

193 Ppara: peroxisome proliferator-activated receptor, alpha
194 Fgf21: fibroblast growth factor 21
195 Acox1: acyl-Coenzyme A oxidase 1
196 Cpt1: carnitine palmitoyltransferase 1
197 Hadh: hydroxyacyl-Coenzyme A dehydrogenase
198 Cd36: CD36 antigen
is consistent with a higher Ppara activity due to enhanced fatty acid binding. But for Cpt1a, which is a direct target of Fgf21, we found no enhanced expression after the second week [21].

Enhanced Fgf21 expression seems not to be sufficient for a stimulation of fatty acid oxidation as consequence to the HFD [21]. This is contrary to previous studies where a stimulated fatty acid oxidation under fatty acid abundance has been observed [422, 426, 428]. Cpt1a showed no elevated expression levels after the second week, though Fgf21 was still induced suggesting an alternative transcriptional regulation (miRNA-mediated regulation) of either Cpt1a and Fgf21 besides Ppara [21]. So, how does this fit together?

### 7.5.2.4 Fgf21 action gets impaired due to the diet

None of the genes involved in fatty acid oxidation showed significant reduction in gene transcription. This indicates that other factors than transcriptional control lead to the diminished action of Fgf21 on fatty acid stimulation during the first and second week [21].

On the post-translational level, direct control through diet induced ligands (e.g., fatty acids [429]), which allosterically control Fgf21 activity might finally lead to its inactivation after the second week [21]. Besides, reversible phosphorylation or acetylation of certain residues in Fgf21 – mediated by diet sensitive enzymes – might account for its observed activity profile [408]. Post-transcriptionally, miRNA-mediated cleavage of Fgf21 mRNA might also explain the observed changes in Fgf21 activity though mRNA levels did not change [21].

### 7.5.2.5 Ppara mediated induction of Cpt1 transcription is abolished by alternate factors

Interestingly, although both genes are direct targets of Ppara, we observed a significant reduction in the expression of Cpt1 after the second week, while Fgf21 transcripts showed two-fold elevation. Plenty of substrate, which is required for an activation of Ppara...
7.5 Small-scale analysis – from observations to systems biological models

is assumed to be supplied by the HFD, especially linoleic acid (C18:2). This suggests that Ppara is active during the challenge period and its activity would explain elevated Fgf21 levels, though not the reduced expression of Cpt1. An alternative mechanism accounting for the observed reduced transcription levels of Cpt1 is miRNA-mediated post-transcriptional regulation [21]. The importance of miRNA-mediated regulation in hepatic lipid metabolism has been confirmed in several studies [409, 410, 430] (see also Section 7.5.1). Strikingly, Iliopoulos et al. [431] showed that mmu-miR-370 directly targets the 3’-UTR of Cpt1 and thus stimulates its cleavage.

Summary Finally, these findings can be combined in a comprehensive systems biological model, providing a coherent explanation for the reduced β-oxidation activity as a consequence to the HFD (Figure 7.10). The hypothesis in [21] states that increased binding of fatty acids to Ppara induces Fgf21 and Cpt1a transcription followed by a subsequent activation of targets of Fgf21 such as Acadl, Acadm, Hadh, and Acox1. Prolonged HFD leads to the accumulation of specific ligands that allosterically bind to Fgf21 and inhibit its stimulating effects. As a result of the abolished stimulating effects of Ppara on Fgf21, the ratio of active and inactive Fgf21 inverts leading to decreased β-oxidation. Potential down-regulation of Cpt1a by mmu-miR-370 further suppresses enhanced fatty acid import to mitochondria and antagonizes the stimulating effects of Ppara on Cpt1a expression which has been activated by fatty acids [21]. Now, it has to be experimentally verified whether mmu-miR-370 is expressed in hepatic tissue at all [21].
In [? ] it was hypothesized that "Diet induced \textit{Ppara} delocalization to the nucleus stimulates \textit{Fgf21} and \textit{Cpt1} expression. Overexpression of \textit{Fgf21} induces transcription of several genes. Among these genes are \textit{Acadm}, \textit{Acadl}, \textit{Hadh} and \textit{Cd36}. A role in post-transcriptional stimulation of \textit{Cpt1a} activity is also suggested. [Our] (...) findings suggest that post-transcriptional or post-translational mechanisms disturb \textit{Fgf21} function after the second week, which results in diminished fatty acid oxidation. miRNA mediated control of \textit{Cpt1a} through \textit{[mmu-miR-370]} (...) which leads to impaired fatty acid import to mitochondria, might intensify this effect. Combined action of these mechanisms might explain the deregulation of combined \textit{Ppara} and \textit{Fgf21} action due to the safflower-oil rich diet. Red = induction/activation/increase, green = repression/inactivation/decrease." [21] Illustration taken from [21].
7.5.3 MicroRNA mediated control of Foxo1 improves the effects of IR

7.5.3.1 Foxo1 mediates the metabolic response to insulin signals

Foxo1 is a known master-regulator of signaling pathways and induces expression of many genes involved in energy homeostasis. The phosphorylation of Foxo1 in response to insulin leads to its nuclear exclusion and inactivation \[432\]. "This insulin dependent responsiveness makes it an important regulator of several processes involved in energy homeostasis. For instance, in hepatic tissue Foxo1 mediates the expression of genes involved glucose metabolism. Under fasting conditions, when plasma insulin concentrations are low, Foxo1 is localized to the nucleus [where it activates] (...) the expression of various genes involved in gluconeogenetic processes \[433–435\]." \[? \] Its importance in lipid metabolism has also been reported where Foxo1 induces the expression of apolipoprotein, leading to elevated plasma TAG levels (hypertriglyceridemia) \[436\]. Furthermore, the stimulation of β-oxidation by Foxo1 activity has been proposed by Bastie et al. \[437\]. Active Foxo1 increased the membrane content of the fatty acid transporter Cd36 which controls the rate of fatty acid uptake from the plasma and utilization. By this, Cd36 mediates the effects of Foxo1 on exogenous fatty acid oxidation.

In addition to Foxo1 induced increased fatty acid uptake, Foxo1 has been shown as a potential repressor of Acaca’s expression \[437\], which is responsible for the conversion of acetyl-CoA to Malonyl-CoA. As described previously, Malonyl-CoA represents an important inhibitor of Cpt1 (Section 7.5.2) which is the essential enzyme mediating the transfer of long chain fatty acyl-CoAs into mitochondria \[397\]. Reduced concentrations of Malonyl-CoA in response to Foxo1 activity is subsequently assumed to result in higher fatty acid oxidation rates \[21\]. Consequently, fatty acid oxidation is stimulated by active Foxo1 in two ways: (i) enhanced exogenous fatty acid uptake by Cd36 and (ii) impaired feedback inhibition of Cpt1a through Malonyl-CoA (by repression of Acaca) \[21\].

7.5.3.2 Insulin induced lipid synthesis is mediated through Srebf1

By controlling enzymes that are essential for catalyzing fatty acid biosynthesis and glycerolipid synthesis related reactions (namely: Acaca, Fasn, Gpam and Scd1\[442\]), Srebf1\[497\] plays a critical role in the process of insulin-mediated regulation of fatty acid and triglyceride synthesis in liver \[438\]. "Raised insulin concentrations in (...) plasma lead to [an] enhanced proteolytic processing of the membrane bound Srebf1 precursor [and thus allow] (...) it to enter the nucleus \[439\]. Besides proteolytic activation, insulin also induces Srebf1 transcription." \[21\]. The underlying mechanism, however, is still not known \[440\].
7.5.3.3 Expected effects of insulin resistance on Srebf1 and Foxo1 mediated processes

The distribution of insulin concentrations showed a drastic reduction of insulin sensitivity over the first three weeks of the study (Figure 7.8) [397, 403]. Impaired insulin signaling might thus lead to a diminished Srebf1 mediated lipogenic activity and raised activity of processes involved in glucose production and fatty acid oxidation are caused by Foxo1 that stimulates both processes [21]. „Under conditions of impaired insulin signaling, loss of insulin mediated repression leads to steady active Foxo1.“ [21]

7.5.3.4 Hepatic lipid synthesis shows typical signs of impaired insulin response

„Typical signs of lowered insulin sensitivity or IR are (...) reduced fatty acid and lipid synthesis, whereas fatty acid oxidation is stimulated [397].“ [21] To study the effects of the observed reduced insulin sensitivity on processes involved in fatty acid and lipid synthesis, we looked in detail to the (in-)activity prediction of these processes by the FAA. We found that the amount of active reactions was raised after the first week (Figure 7.6). Intact insulin signaling at the beginning of the study would account for that [21]. „[But] (...) during the first and third week, no further stimulation of lipogenic processes was predicted by the FAA (...) . [This is congruent with the arising IR] Due to the impaired insulin signaling, lipogenesis lacks of stimulation through Srebf1 mediated expression.“ [21]

7.5.3.5 Gluconeogenesis is not affected by IR

„A typical sign of insulin resistance (IR) is unrestrained production of glucose by gluconeogenic processes. (...) Foxo1 plays a crucial role in the regulation of these processes in response to insulin. Under physiological conditions, raised insulin concentrations [result in] (...) the inactivation of Foxo1 which in turn leads to reduced rates of gluconeogenic processes by diminished transcription of essential enzymes. Insulin resistance leads to an impaired inactivation of Foxo1 and unrestrained stimulation of gluconeogenesis further worsens plasma glucose levels (hyperglycemia) [403].“ [21]

Interestingly, results from the FAA predictions showed that there is no stimulation of processes involved in gluconeogenesis during the first and third week of the study (Figure 7.4). Measured plasma glucose levels, which should be elevated under condition of active gluconeogenesis, confirmed this: no significant changes between both groups were found, especially after the first and third week (Figure 7.11). Although glucose levels after the second week seemed to be higher in the challenge cohort, the 50% quantiles are almost equal in both cohorts and most of the animals have similar plasma glucose concentrations [21].

So far our observations, together with the predictions from the FAA, „suggest that gluconeogenesis is not stimulated by impaired insulin signaling. This is in contrast to the expected stimulation of gluconeogenic processes due to impaired insulin signaling“ [21] (Figure 7.12).
7.5.3.6 MicroRNA mmu-miR-139 compensates impaired insulin mediated Foxo1 inactivation

"(...) [T]he typical effects of insulin resistance on hepatic glucose and lipid metabolism are (...) disrupted in the (...) HFD. Srebf1 mediated control of lipid metabolism (...) showed the common characteristics [due to] (...) impaired insulin signaling: fatty acid and lipid synthesis [were not stimulated] (...) due to elevated insulin concentrations."[21] But we observed no stimulation of gluconeogenesis as expected [21]. Neither FAA predictions nor plasma glucose concentrations indicate[d signs for] a stimulation of endogenous glucose synthesis. As Foxo1 is the central regulator of gluconeogenic processes in response to insulin, (...) deregulation of insulin mediated control of Foxo1 [is likely to account] (...) for the observed characteristics."[21] Besides post-translational regulation through insulin mediated phosphorylation, acetylation and deacetylation of specific residues of Foxo1 have been shown to impact its activity [442]. "However, these (...) mechanisms would improve Foxo1 activity due to environmental stress [443] and (...) [consequently are not able to] explain the lack of stimulation observed from FAA results and physiological data."[21] Strikingly, Hennessy et al. [444] recently reported miRNA-mediated control of Foxo1. They showed that mmu-miR-139 binds to the 3'-UTR of Foxo1 in liver and as a consequence Foxo1 translation is reduced [444]. MiRNA-mediated down-regulation of Foxo1 by mmu-miR-139 is thus quite likely to „compensate for the usual effects of impaired hepatic insulin response (unrestrained glucose production) “[21].
7 Systems biology of metabotypes

Figure 7.12: Insulin acts as a inhibitor of Foxo1. „Under physiological conditions (a), insulin inhibits Foxo1 activation. Consequently, gluconeogenesis and fatty acid oxidation are tightly regulated. Diminished insulin sensitivity (b) leads to impaired inhibition of Foxo1 which in turn leads to raised endogenous glucose production and deregulation of lipid metabolism. In the case of insulin resistance (c), Foxo1 inhibition through insulin is completely disturbed. Severe states as hyperglycemia (high blood glucose) and ketoacidosis through uncontrolled fatty acid oxidation and ketogenesis are the results [441].“ [21]. Illustration adapted from Puigserver and Rodgers [441].

Summary  Figure 7.13 depicts the potential mechanistic explanation behind the missing stimulation through Foxo1 in the presence of raised insulin concentrations. The hypothesis in [21] states that „Arising IR has the common effects on Srebf1 mediated control of lipogenic processes. Impaired insulin signaling during week one and three leads to reduced activation of Srebf1. As a result, (...) fatty acid and lipid synthesis show no stimulation (...) [as well]. In contrast, the arising insulin resistance does not show effects on the control of Foxo1. Reduced insulin sensitivity is usually reflected by raised Foxo1 activity, which in turn leads to a stimulation of glucose production (gluconeogenesis). Unrestrained gluconeogenesis by insulin leads to raised blood glucose levels, which worsen the effects of insulin resistance. An explanation for this observation could be the compensatory effect of post-transcriptional down-regulation of Foxo1 by (...) [mmu-miR-139]. Raised levels of this miRNA lead to the inactivation of Foxo1 by reducing its protein level, therefore accounting for the lowered insulin sensitivity. However, it remains to be studied under which conditions such a mechanism compensates for the impaired insulin signaling." [21]
Summary and future perspectives

Non-alcoholic fatty liver disease is one of the most common causes of chronic liver disease and known to be an important risk factor for the development of T2D. It is widely known that NAFLD is closely related to over-nutrition and hepatic IR. But the underlying mechanisms are not clear yet and instead of a qualitative and static picture, one is interested especially in a dynamic perspective of NAFLD development and progression.

The best basis for a quantitative analysis towards this direction are metabolic networks, which represent the best described networks in biological systems so far. They are rather complete and a number of system-wide models is available. In this Chapter, we systematically analyzed the impact of a safflower-oil rich diet in mice in the context of metabolic disorders. Based on the CBM approach originally presented by Shlomi et al. [22], we studied the progression of diet induced NAFLD in a systems-biological fashion. We showed that this approach facilitates, on the basis of mRNA expression data and the underlying biochemical reactions in the genome-scale metabolic network, a qualitative description of changes in the metabolic behavior due to the HFD in a time resolved manner. This is achieved by interpreting predicted metabolic flux activity.
states in a step-by-step analysis from top to down. Several levels of complexity provide valuable information about metabolic changes in hepatic metabolism during the study period incorporating potential miRNA-mediated regulation [21]. Studying the role of post-transcriptional regulation in the context of tissue-specific metabolism, Shlomi et al. [22] found that great parts of metabolic enzymes might rely on post-transcriptional regulation as their expression levels and the predicted metabolic flux activity states are contradictory. The post-transcriptional up- or downregulation in turn accounts for the observed phenomena.

"First, we (...) correlated metabolic behavior predictions with physiological parameters [on a large-scale]. [We were] (...) able to connect changes in metabolic activity and the progression of the phenotype in a global fashion. Results from the analysis [on this granularity level] (...) underscored the central role of processes involved in fatty acid and lipid metabolism in the adaption to a HFD (...) [that has also been observed in previous studies [402, 407, 446]]. While at the beginning of the study the metabolism was stimulated in challenged animals, the metabolism shifted to a rather inactive state during the second and the third week of the study (...)." [21]

To understand the connection between the metabolic trends in lipid metabolism observed from this global perspective and the progression in NAFLD, we analyzed the metabolic states provided by the FAA in fatty acid metabolism and lipid metabolism in more detail. Thereby, we confirmed the characteristic role of processes involved in fatty acid and lipid metabolism in the progress of NAFLD. Various processes were identified that showed a certain degree of deregulation, e.g., fatty acid biosynthesis, fatty acid elongation, and fatty acid esterification. Although some of the observed effects can be explained by general mechanisms such as hormonal control through insulin signaling, for the bigger part that is suggested to be under post-transcriptional control relevant information is not manifested in the mRNA expression data [21].

For a detailed, systems-level understanding of these changes comprehensive small-scale models are indispensable (Section 7.5). Therefore, we generated qualitative systems biological models regarding the diet-induced lipid accumulation in the liver and the observed elevated insulin concentrations [on a small-scale] (...). The models [are based on the FAA results and] were enriched with additional information from literature (...). [We proposed] hypotheses about the causal connections between the predicted metabolic states and the observed progression in the phenotype (...). These small-scale models underscore the important role of miRNA-mediated post-transcriptional regulation in the adaption to the HFD e.g., that miRNA-mediated control of Foxo1 improves the effects of insulin resistance. Experiments to validate these statements are required and organized for the near future [21].

The integration of miRNA expression data will provide evidence about the role of post-transcriptional regulation of certain enzymes. Additionally, incorporation of data about post-translational protein modifications (such as acetylation and phosphorylation) would be useful [21]. "Gaining a deeper understanding of the dynamic activity of certain pro-
teins during the study will certainly explain some of the observed effects. The raised insulin concentrations [that have] (...) been observed during the study [period] and their relation to the metabolic states will be further investigated with the help of data from euglycemic clamp studies. In addition to that, data from quantitative metabolomic techniques will be correlated to the predicted metabolic states and enable a more differentiated look at the progression of the phenotype “[21]. Furthermore, tissue-specific models should be considered in future studies as the various tissues present in mammalian organisms vastly differ in their metabolic activity [21]. For instance[, the complete [β-oxidation] (...) is a(...) usual process in hepatic tissue, whereas adipose tissue lacks of the necessary enzymes and is hence not capable to these reactions. (...)To provide adequate flux predictions[,] (...) tissue-specific characteristics need to be incorporated into underlying metabolic models. A first step towards such tissue-specific models was recently published by Jerby et al. [447] [21].

“Taken together, [we studied] (...) how a CBM approach - like FAA - is a valuable basis for the study of complex phenotypes such as hepatic steatosis and NAFLD. The qualitative analysis of the metabolic behavior is a suitable way to understand the progression of a certain phenotype. [We] (...) successfully showed[,] how on the basis of these results[,] new hypotheses can be generated. Furthermore, the presented flux-activity analysis (FAA) approach is a generic framework and future studies can be analyzed in a similar fashion. The opportunity to integrate additional experimental data will enable the understanding of the development of a certain phenotype in more detail. Refined models can be generated and the new hypotheses derived from these models will result in new experiments.” [21]
CHAPTER 8

Conclusion and outlook

"Complex is perhaps the most common adjective used to describe biological phenomena. In every cell, complex networks of interactions occur between thousands of metabolites, proteins and DNA. (...) So where is the simplicity?" [448]

Networks of interactions between thousands of molecules seem to defy comprehension - even more in perturbed situations such as in diseased states - but shared principles of design may simplify the picture [448]. Based on these simplifying principles the field of network medicine arose, which strongly contributed in uncovering the interrelations between components in complex diseases [3]. They also laid the foundation towards systems biology where the challenge is to integrate complex and highly diverse information into a conceptual framework or network – one that is holistic, quantitative, and predictive, but at the same time simplex. The ultimate goal in systems biology is to find mechanistic explanations of the relationship between a systems components, their interactions and its phenotype.

In order to make sense of biology on all complexity levels, biological networks are an appropriate tool. However, they are apparently incomprehensible and thus simplicity in biology is necessary to discover the underlying general principles.

The goal of this thesis was therefore to reveal novel insights into complex human diseases from both, a macro- and a microscopic perspective. As the first step for reaching this goal is the necessity of algorithms capable to analyze complex biological networks, I started with the development of network analysis methods and data integration. By this, disease relations could be uncovered from a macroscopic perspective. Going into more detail, I concentrated on the microscopic perspective. This focus led to novel insights into the interrelations on a local level, causative explanations and novel hypotheses, and accordingly, towards a better understanding of diseases on a systems-level. To simplify matters and to contribute to a better understanding, each Chapter of this thesis encapsulated one of the consecutive steps towards capturing the whole, and therefore was intended to be as self-contained as possible. In the following, the main scientific contributions of this thesis are summarized.
A major challenge in systems biology is how to extract meaningful biological insights from large, heterogeneous data sets. The key point for this purpose are particularly sophisticated data integration and analysis methodologies that will infer the underlying dynamics and mechanisms of human diseases. Data integration facilitates the construction of large biological networks that are $k$-partite as they are complex and highly diverse. A key question is how to analyze and interpret the internal organization of these networks. Consequently, there is a major need for methods capable to dissect the internal characteristics, such as the topology of the cellular networks involved. I addressed this issue in Chapters 3 and 4.

First, we presented a novel fuzzy clustering algorithm based on NMF which allows to detect overlapping clusters in $k$-partite graphs in Chapter 3. We demonstrated that the algorithm is not only efficient and stable, but it is also able – using a real-world example of a gene-disease-protein complex graph – to identify biologically meaningful clusters on different resolution levels. Thereby, we got insights into the global network organization from a macroscopic perspective. Focusing on the small-scale architecture, we exemplified how overlapping clusters allow for reclassification, annotation, or even detection of misclassified elements on a local level.

Next, I focused on the hierarchical structure of regulatory networks, which is a central organizing principle of complex networks. It goes beyond simple clustering by explicitly including organization at all scales in a network simultaneously [13] and provides not only insights into the correlation among, e.g., genes and biological phenomena, rather it reveals the interdependencies of regulatory relationships [31]. In Chapter 4, I presented HiNO, an algorithm for inferring the hierarchical organization from regulatory networks. HiNO significantly improves existing methods and allows for resolving the global organization of GRN by analyzing the general properties of their hierarchical structure. Strikingly, I retrieved novel insights into, e.g., disease associations from a microscopic perspective by systematically investigating the individual regulatory layers of the human GRN.

Though both methods work differently, they represent the first steps towards a better understanding of network organization that is essential in the era of systems biology. They both decompose networks and thereby, they allow to disclose their global organization from a static and descriptive perspective. Subsequently, by focusing on disease relationships and disease causations in more detail, I broke the global organization step-wise down into smaller pieces. This allowed us to analyze the system from the microscopic view as shown in Chapters 5, 6 and 7.

As a start, the focus is shifted from the overall network structure and topology to the single interactions within a network. Based on the diseasome concept originally presented by Goh et al. [17], we investigated the role of pleiotropic genes in disease networks and how diseases themselves are cross-connected in Chapter 5. For this, we developed a concept that allows for a systematic multi-scale analysis of common human diseases in order to elucidate disease relationships, key genetic loci, and pathways. Thereby, we found many biologically suspected, as well as unexpected disease-disease relationships, which also led
to the conclusion that the term pleiotropy in the context of GWAS has to be used with care. However, our results yielded many candidate loci representing potentials to experimentally explore them and thus the genomic architecture of complex diseases in more detail.

Despite efforts to understand the interrelations between different diseases based on cellular network characteristics, and in order to understand the genotype–phenotype relationships in diseases, the modeling of disease-causing mutation effects on molecular networks is required [6]. We focused on this in Chapter 6 where we investigated the impact of impaired miRNA mediated regulation from the standpoint of population genetics. By systematically analyzing the global impact of miRSNPs, we looked on disease causation from a macroscopic perspective. We identified many candidates where a miRSNP represents a potential causal mechanism for disease development. Based on this, we were able to demonstrate that miRSNPs not only tend to be common variants, but most notably also have the potential to be used in disease prognosis and diagnosis. By contextualizing them in a systems biological manner, we ended up with small-scale models giving potential explanations about disease causations from a microscopic perspective. These models represent hypotheses that now have to be tested experimentally and therefore herald a new round in the systems biological cycle. Altogether, they led to the conclusion that a systems biological interpretation of GWAS data is essential in order to get a comprehensive understanding of a complex disease.

A similar top-down procedure was applied in Chapter 7 where we focused on disease causations, which are not solely genetically predisposed, but rather due to environmental changes. Therefore, we analyzed the impact of a HFD in the context of NAFLD induction and progression. Based on the predictions of a CBM approach, we were able to quantitatively study normal and abnormal metabolic behavior in response to the HFD in a time-resolved manner. By using a multi-scale analysis approach, we were able to identify metabolic categories where the most significant changes in response to the HFD occurred. Based on this macroscopic perspective, we were able to deduce comprehensive systems biological models giving potential causative explanations for our observations from a microscopic view. Important to note is that, as before, also here post-transcriptional regulatory mechanisms represent the missing piece in the puzzle. Again, identified models led to novel insights and hypotheses, which are currently being tested experimentally.

In this thesis, I analyzed complex human diseases and their interrelations in a systems biological fashion using both, the macro- and the microscopic dimension. In other words, I worked on simplifying complex systems based on multi-scale network analyses. In a nutshell, starting from a general, static, and descriptive view of the global organization, by using heterogeneous experimental data, I headed towards a more specific, dynamic, predictive, and hypothesis-generating direction. The former covers methods that allow to structure networks and thus to understand complex systems by decomposing them into interrelated components. The latter represents the individual connections between those components. Figuratively spoken, when piecing a puzzle together one usually begins with assembling the frame and then goes to individual parts that can be merged according to
similar forms and/or colors. In the end, both together represent (i) pieces in the puzzle and (ii) strategies to solve it. Consequently, they allow to narrow down the gap between genotype and phenotype.

The network analysis methods, systems biological models, and concepts to retrieve them, which I presented in this thesis pave the way towards a better understanding of complex human diseases. Using different resolution levels turned out to be a successful way in "approaching" and, after all, understanding a disease on systems-level. We revealed several qualitative models representing novel insights in diseases. For instance, we presented that impaired miRNA regulation might lead to elevated LDL cholesterol levels, or that miRNAs might account for lowered insulin sensitivity. These models reveal novel insights and allow a better understanding of diseases in a qualitative manner on a small-scale. To capture the whole, meaning to completely understand a complex disease, its onset, progression, the predisposition to and the probability to develop it at all, these models are far too small and most important, they are missing any dynamics. However, understanding diseases entirely on a systems-level is far out of reach as at the moment, their total complexity and the parameter space that has to be considered exceed our imagination. Besides, we cannot capture and less than ever cope with it, neither conceptionally nor technically. Nevertheless, this thesis is a start towards this direction and we are confident that with ever faster developing high-throughput technologies we will achieve the final goal in the future. Beyond suggestions to extend and improve the proposed methods/concepts in the individual Chapters, first of all, especially the integration of more, particularly quantitative data, is necessary. This enables, e.g., mathematical modeling and consequently a more dynamic perspective on human disease. The outcome is not merely a more refined picture, but offers a new level of mechanistic understanding.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acaca</td>
<td>acetyl-Coenzyme A carboxylase alpha</td>
</tr>
<tr>
<td>Acadm</td>
<td>acyl-Coenzyme A dehydrogenase, medium chain</td>
</tr>
<tr>
<td>Acadl</td>
<td>acyl-Coenzyme A dehydrogenase, long-chain</td>
</tr>
<tr>
<td>acetyl-CoA</td>
<td>acetyl-Coenzyme A</td>
</tr>
<tr>
<td>Acox1</td>
<td>acyl-Coenzyme A oxidase 1</td>
</tr>
<tr>
<td>acyl-CoA</td>
<td>acyl-Coenzyme A</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer’s disease</td>
</tr>
<tr>
<td>ADHD</td>
<td>attention-deficit/hyperactivity disorder</td>
</tr>
<tr>
<td>AGO</td>
<td>argonaute</td>
</tr>
<tr>
<td>Akt2</td>
<td>thymoma viral proto-oncogene 2</td>
</tr>
<tr>
<td>ALS</td>
<td>amyothropic lateral sclerosis</td>
</tr>
<tr>
<td>apoE</td>
<td>apolipoprotein E</td>
</tr>
<tr>
<td>AS</td>
<td>ankylosing spondylitis</td>
</tr>
<tr>
<td>BACE1</td>
<td>beta-site APP-cleaving enzyme 1</td>
</tr>
<tr>
<td>BA</td>
<td>Barabási-Albert</td>
</tr>
<tr>
<td>BCL3</td>
<td>B-cell leukemia/lymphoma 3</td>
</tr>
<tr>
<td>BFS</td>
<td>breadth first search</td>
</tr>
<tr>
<td>BiGG</td>
<td>Biochemical Genetic and Genomic knowledge base of large-scale metabolic reconstructions</td>
</tr>
<tr>
<td>CAD</td>
<td>coronary artery disease</td>
</tr>
<tr>
<td>CBFB</td>
<td>core binding factor beta</td>
</tr>
<tr>
<td>CBM</td>
<td>constraint-based model</td>
</tr>
<tr>
<td>CBM</td>
<td>constraint-based modeling</td>
</tr>
<tr>
<td>CCND1</td>
<td>cyclin D1</td>
</tr>
<tr>
<td>CD</td>
<td>Crohn’s disease</td>
</tr>
<tr>
<td>Cd36</td>
<td>CD36 antigen</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>cyclin-dependent kinase inhibitor 2A</td>
</tr>
<tr>
<td>CDKN2B</td>
<td>cyclin-dependent kinase inhibitor 2B</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>CeD</td>
<td>celiac disease</td>
</tr>
<tr>
<td>CLIP</td>
<td>in vivo Cross-Linking and Immuno-Precipitation</td>
</tr>
<tr>
<td>CLP</td>
<td>common lymphocyte progenitor</td>
</tr>
</tbody>
</table>
Malonyl-CoA malonyl-Coenzyme A
MAPK mitogen-activated protein kinase
MAPKK mitogen-activated protein kinase kinase
MECP2 methyl CpG binding protein 2
MEP megakaryocyte/erythroid progenitor
MeSH Medical Subject Heading
metabotype metabolic phenotype
MGD Mouse Genome Database
MGI Mouse Genome Informatics
MGRN mouse gene regulatory network
MHC major histocompatibility complex
MILP mixed-integer LP
miRNA microRNA
miRISC miRNA-induced silencing complex
mRNA messenger RNA
MS multiple sclerosis
MTRN mouse transcriptional regulatory network
MYC v-myc myelocytomatosis viral oncogene homolog (avian)
OMIM Online Mendelian Inheritance in Man
OR odds ratio
PA phosphatidic acid
PAR-CLIP photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation
PARP1 poly (ADP-ribose) polymerase 1
PAX6 paired box gene 6
PBC primary biliary cirrhosis
PCSK9 proprotein convertase subtilisin/kexin type 9
PD Parkinson’s disease
PKB protein kinase B
PL postnatal lethal
Ppap2a phosphatidic acid phosphatase type 2a
Ppara peroxisome proliferator-activated receptor, alpha
PPI protein-protein interaction
Prkce protein kinase c, epsilon
Propionyl-Coa propionyl-Coenzyme A
PSCA prostate stem cell antigen
PTPN22 protein tyrosine phosphatase, non-receptor type 22 (lymphoid)
PUFA poly-unsaturated fatty acids
QTLs quantitative trait loci
RA rheumatoid arthritis
RBP RNA-binding protein

190
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>RELN</td>
<td>reelin</td>
</tr>
<tr>
<td>RGRN</td>
<td>rat gene regulatory network</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RTRN</td>
<td>rat transcriptional regulatory network</td>
</tr>
<tr>
<td>RUNX1</td>
<td>runt-related transcription factor 1</td>
</tr>
<tr>
<td>Scd1</td>
<td>stearoyl-Coenzyme A desaturase 1</td>
</tr>
<tr>
<td>siRNA</td>
<td>small-interfering RNA</td>
</tr>
<tr>
<td>SLE</td>
<td>systemic lupus erythematosus</td>
</tr>
<tr>
<td>SMAD3</td>
<td>SMAD homolog 3 (Drosophila)</td>
</tr>
<tr>
<td>SMC4</td>
<td>structural maintenance of chromosomes 4</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>ST-HSC</td>
<td>short term hematopoietic stem cells</td>
</tr>
<tr>
<td>STAT1</td>
<td>signal transducer and activator of transcription 1</td>
</tr>
<tr>
<td>STAT4</td>
<td>signal transducer and activator of transcription 4</td>
</tr>
<tr>
<td>Srebf1</td>
<td>sterol regulatory element binding transcription factor 1</td>
</tr>
<tr>
<td>SVD</td>
<td>singular value decomposition</td>
</tr>
<tr>
<td>T1D</td>
<td>type 1 diabetes</td>
</tr>
<tr>
<td>T2D</td>
<td>type 2 diabetes</td>
</tr>
<tr>
<td>TAG</td>
<td>triacylglycerol</td>
</tr>
<tr>
<td>TCF7L2</td>
<td>transcription factor 7-like 2 (T-cell specific, HMG-box)</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>TNF</td>
<td>tumor necrosis factor</td>
</tr>
<tr>
<td>TNFAIP3</td>
<td>tumor necrosis factor, alpha-induced protein 3</td>
</tr>
<tr>
<td>TRBP</td>
<td>(TAR) RNA-binding protein</td>
</tr>
<tr>
<td>TRN</td>
<td>transcriptional regulatory network</td>
</tr>
<tr>
<td>TP53</td>
<td>tumor protein p53</td>
</tr>
<tr>
<td>TP63</td>
<td>tumor protein p63</td>
</tr>
<tr>
<td>UBE3A</td>
<td>ubiquitin protein ligase E3A</td>
</tr>
<tr>
<td>UC</td>
<td>ulcerative colitis</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>VLDL</td>
<td>very low density lipoprotein</td>
</tr>
<tr>
<td>WNT3</td>
<td>wingless-type MMTV integration site family, member 3</td>
</tr>
<tr>
<td>XRCC1</td>
<td>X-ray repair complementing defective repair in Chinese hamster cells 1</td>
</tr>
<tr>
<td>YGRN</td>
<td>yeast gene regulatory network</td>
</tr>
<tr>
<td>YTRN</td>
<td>yeast transcriptional regulatory network</td>
</tr>
<tr>
<td>ZBTB32</td>
<td>zinc finger and BTB domain containing 32</td>
</tr>
<tr>
<td>ZMIZ1</td>
<td>Zinc finger MIZ domain-containing protein 1</td>
</tr>
</tbody>
</table>
## List of Figures

1.1 A simple complexity pyramid .............................................. 2
1.2 Illustration of thesis outline ............................................ 5

2.1 Example graph .................................................................. 10
2.2 Network models ............................................................... 13
2.3 Illustration of the concept of overlapping communities .......... 17
2.4 Global map of metabolic pathways ..................................... 22
2.5 The conceptional basis of constraint-based modeling ............. 27
2.6 DNA sequence variation in the human genome ...................... 31
2.7 The Genome-wide association study .................................... 33
2.8 MicroRNA biogenesis ....................................................... 35
2.9 MicroRNA regulation ........................................................ 37

3.1 Illustration of the fuzzy clustering approach ....................... 42
3.2 Fuzzy clustering algorithm ............................................... 45
3.3 Illustration of the cluster decomposition of a bipartite toy example . 49
3.4 Performance on toy models .............................................. 48
3.5 Adjacency matrix for cluster stability analysis .................... 49
3.6 Cluster stability analysis .................................................. 50
3.7 Cluster structure - cost function profile ............................. 51
3.8 Cluster structure - membership behavior ........................... 52
3.9 Cluster structure - details on Figure 3.7c ............................. 53
3.10 Illustration of the gene-disease-protein complex graph ......... 54
3.11 Decomposition of a gene-disease-protein complex network .... 56
3.12 Illustration of large-scale cluster structures in the gene-disease-protein complex network ........................................ 57
3.13 The degrees of membership from the large-scale clustering of the gene-disease-protein complex graph ........................................ 58
3.14 Evaluation of the backbone of the gene-disease-protein complex network ......................................................... 60
3.15 The small-scale fuzzy clustering in the neighborhood of MECP2 .. 62
3.16 The small-scale hard clustering in the neighborhood of MECP2 .. 63

4.1 Hierarchical structure clarifies regulatory interdependencies .... 67
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2</td>
<td>Illustration of the hierarchy extraction approach</td>
<td>69</td>
</tr>
<tr>
<td>4.3</td>
<td>Pseudo code of HiNO</td>
<td>70</td>
</tr>
<tr>
<td>4.4</td>
<td>From the GRN to the hierarchical structure</td>
<td>72</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison BFS-method vs. HiNO using the TRN of E. coli</td>
<td>74</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison BFS-method vs. HiNO using the TRN of S. cerevisiae</td>
<td>75</td>
</tr>
<tr>
<td>4.7</td>
<td>Level and target distribution of the HTRN</td>
<td>78</td>
</tr>
<tr>
<td>4.8</td>
<td>Distribution of embryonic and postnatal lethal phenotypes</td>
<td>81</td>
</tr>
<tr>
<td>4.9</td>
<td>Distribution of HTRN associated functions and pathways</td>
<td>83</td>
</tr>
<tr>
<td>4.10</td>
<td>Abstract overlay of the HTRN with the hierarchy in hematopoiesis</td>
<td>85</td>
</tr>
<tr>
<td>4.11</td>
<td>MicroRNA association of the HTRN</td>
<td>88</td>
</tr>
<tr>
<td>4.12</td>
<td>Clustergram of miRNA regulatory profiles</td>
<td>90</td>
</tr>
<tr>
<td>4.13</td>
<td>Distribution of gene expression levels</td>
<td>91</td>
</tr>
<tr>
<td>4.14</td>
<td>Distribution of gene expression levels in single hierarchies</td>
<td>92</td>
</tr>
<tr>
<td>4.15</td>
<td>Heat-map of TF expression in 32 human organs and tissues</td>
<td>94</td>
</tr>
<tr>
<td>4.16</td>
<td>Tissue association of the HTRN</td>
<td>95</td>
</tr>
<tr>
<td>4.17</td>
<td>Distribution of associated diseases</td>
<td>96</td>
</tr>
<tr>
<td>5.1</td>
<td>Data acquisition and processing procedure</td>
<td>103</td>
</tr>
<tr>
<td>5.2</td>
<td>Illustration of the bipartite disease-locus association network</td>
<td>104</td>
</tr>
<tr>
<td>5.3</td>
<td>Trait and loci association distribution</td>
<td>106</td>
</tr>
<tr>
<td>5.4</td>
<td>Study distribution</td>
<td>108</td>
</tr>
<tr>
<td>5.5</td>
<td>Disease class distribution</td>
<td>109</td>
</tr>
<tr>
<td>5.6</td>
<td>The human GWAS diseasome</td>
<td>111</td>
</tr>
<tr>
<td>5.7</td>
<td>Illustration of the disease-centric projection of the human GWAS disease network</td>
<td>112</td>
</tr>
<tr>
<td>5.8</td>
<td>38 top traits</td>
<td>114</td>
</tr>
<tr>
<td>5.9</td>
<td>Illustration of the human GWAS disease-locus association network of the 38 top traits</td>
<td>116</td>
</tr>
<tr>
<td>5.10</td>
<td>Diseasome of widely studied traits</td>
<td>118</td>
</tr>
<tr>
<td>5.11</td>
<td>Heat-map of trait associations</td>
<td>119</td>
</tr>
<tr>
<td>5.12</td>
<td>Pathway association of 38 top traits</td>
<td>120</td>
</tr>
<tr>
<td>6.1</td>
<td>Potential effects of SNPs on microRNA processing and function</td>
<td>125</td>
</tr>
<tr>
<td>6.2</td>
<td>Distribution of the function classes</td>
<td>129</td>
</tr>
<tr>
<td>6.3</td>
<td>Binned MAF distribution</td>
<td>131</td>
</tr>
<tr>
<td>6.4</td>
<td>Footprint distance distribution</td>
<td>133</td>
</tr>
<tr>
<td>6.5</td>
<td>SMC4 - miRSNP links PBC and cancer</td>
<td>135</td>
</tr>
<tr>
<td>6.6</td>
<td>Secondary structure of the 3’-UTR of HMGCR</td>
<td>137</td>
</tr>
<tr>
<td>6.7</td>
<td>HMGCR - The role of microRNAs in LDL Cholesterol-Related Traits</td>
<td>139</td>
</tr>
<tr>
<td>7.1</td>
<td>Study design</td>
<td>148</td>
</tr>
<tr>
<td>7.2</td>
<td>Illustration of FAA according to Shlomi et al.</td>
<td>150</td>
</tr>
<tr>
<td>7.3</td>
<td><em>In silico</em> approach to model metabolic behavior</td>
<td>153</td>
</tr>
</tbody>
</table>
7.4 Active and inactive reactions in main categories

7.5 Hepatic TAG levels

7.6 Metabolic flux through lipid metabolism and metabolism of complex lipids

7.7 Fatty acid oxidation over the study period

7.8 Plasma insulin levels

7.9 Deregulation of Ppap2a reduces liver TAG concentrations and improves insulin sensitivity

7.10 Mechanisms for the disruption of the common control of fatty acid oxidation through Ppara and Fgf21

7.11 Plasma glucose levels

7.12 Insulin action on Foxo1

7.13 mmu-miR-139 compensates the effects of impaired insulin action on Foxo1
List of Tables

3.1 Parameters for the simulated data models ........................................ 47
4.1 Detailed numbers of GRNs and TRNs ............................................... 71
4.2 Level distribution of the TRNs ....................................................... 72
4.3 Level distribution of the TRNs ....................................................... 77
4.4 Level distribution of the HTRN and comparison with random networks .... 79
5.1 MeSH disease classes ................................................................. 108
5.2 SNPs that link multiple phenotypes are also highly correlated. ............... 113
5.3 SNPs with antagonistic effects in distinct traits .................................. 119
6.1 Functional SNP classes ............................................................. 127
6.2 Frequencies and ORs ................................................................. 129
6.3 ORs and background frequencies for MAF bins in the compared SNP sets .... 132
7.1 Prediction and expression states of genes involved in fatty acid oxidation .... 172
Bibliography


