TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

The impact of microRNAs on signaling pathways: From general perspectives to a computational model of the JAK-STAT pathway

Andreas Kowarsch

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. D. Frischmann

Prüfer der Dissertation: 1. Univ.-Prof. Dr. H.-W. Mewes
2. Univ.-Prof. Dr. Dr. F. Theis

Die Dissertation wurde am 20.07.2011 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 15.12.2011 angenommen.

Acknowledgements

Here, I would like to thank the following people that accompanied and supported me during the last three years.

First of all, I would like to thank my supervisors Prof. Dr. Hans-Werner Mewes, Prof. Dr. Ursula Klingmüller and Prof. Dr. Dr. Fabian Theis for being members of my thesis committee and their valuable comments and discussions. I would like to thank Prof. Dr. Dr. Fabian Theis for giving me the opportunity to become a member of his group. Thanks for all your support and making various exciting projects and collaborations possible. Especially, I would like to thank you for giving me the opportunity to make a research visit to Dr. Kevin Janes.

The entire CMB group, for our stimulating working atmosphere, discussions, entertaining coffee breaks, nice conference trips - special thanks to Daniel and Jan, and of course for all our bbq evenings. All people at the IBIS, for the great atmosphere at the Institute, many constructive discussion rounds and successful collaborations.

Sebastian, Ruth, Svantje and Dr. Ursula Klingmüller of the DKFZ, for very successful collaborations and entertaining evenings in Heidelberg.

Carsten, for being my room mate for three years. Thanks for all your support, scientific advices and of course for your helpful comments and proof-reading of almost every manuscript.

Daniel, for your support during several projects, fights against statistical tests and especially for our legendary time in Boston.

Florian, for the very pleasant collaboration in the GraDe project and many competitive table soccer games. Thanks for your support during the nights before the final submission.

My parents and brother, for all your support, not only during the last three years and your appreciation that I had definitely too little time the last years.

And finally, especially Julia, without your support and help I would not have been able to successfully finish this thesis. Thanks for being appreciative that I spent too much of our leisure time to work on this thesis.

Abstract

In this thesis we analyze regulatory motifs of microRNA-mediated regulation with a special focus on signaling pathways. MicroRNAs are a large class of post-transcriptional regulators that bind to the 3'-untranslated region of messenger RNAs and alter their expression. They play a critical role in many cellular processes and have been linked to the control of signaling pathways. With the identification of thousands of microRNAs, the future challenge is to unveil their biological functions.

Within this thesis, we first study the regulatory role of microRNAs from a general perspective. Using more than three hundred studies, we show that the integration of data only from cell culture studies is in conflict to conclusions drawn from patient studies. In addition, we link disease-associated microRNAs to signaling pathways to unveil their regulatory pattern. On a global scale, we identify a core set of signaling pathways, which are enriched by microRNA target genes across human diseases. More locally, we identify significantly different patterns for disease-associated microRNAs and genes based on their cellular location and process role.

Second, we introduce a novel measurement for the detection of functional microRNA-pathway associations. Usually, the impact of microRNA control is assessed by identifying pathways with enriched targets. However, this approach ignores the features provided by biological networks. We propose a novel measure that takes the network topology into account. Defining the proximity score, we identify novel microRNA-pathway associations that differ from those usually inferred with the enrichment score. Finally, we integrate our novel approach into the miTALOS web server that provides novel features for the identification of microRNA-pathway associations.

Third, we analyze the regulatory role of microRNAs on the dynamic of phosphorylation processes. We initially study a linear signaling cascade to analyze the differences between a system with and without microRNA regulation. Then, we use a full model of the gp130-STAT pathway and show that pre-induced microRNAs change the overall ratio of pSTAT3/STAT3 expression.

Finally, we present a novel framework for the analysis of multi-scale biological data. Matrix factorization techniques are well suited for the analysis of multi-layered temporal responses induced by signaling pathways. We extend this technique by integrating prior knowledge to link genes and microRNAs along an underlying network. We are able to define a graph-delayed correlation function and to use this framework as constraint to the matrix factorization task to set up our graph-decorrelation algorithm GraDe.

Zusammenfassung

In dieser Dissertation werden regulatorische Motive von microRNAs mit einem speziellen Fokus auf Signalwege untersucht. MicroRNAs bilden eine Klasse von posttranskriptionalen Regulatoren, die an den 3'-untranslatierten Bereich der messenger RNAs binden und die Expression beeinflussen. Sie spielen eine zentrale Rolle in vielen zellulären Prozessen und werden mit der Kontrolle von Signalwegen in Verbindung gebracht. Mit der Identifikation von Tausenden von microRNAs ist die Erforschung der biologischen Funktionen eine zukünftige Herausforderung.

Zuerst untersuchen wir die regulatorische Rolle von microRNAs aus einem generellen Blickwinkel. Durch die Verwendung von mehr als dreihundert Studien zeigen wir, dass die Ergebnisse von Zellkulturstudien im Widerspruch zu Patientenstudien stehen. Des Weiteren verknüpfen wir krankheitsrelevante microRNAs und Signalwege, um die regulatorische Rolle von microRNAs zu untersuchen. Global gesehen identifizieren wir eine Kernmenge an Signalwegen, die eine signifikante Anhäufung an microRNA-Zielgenen zeigt. Detailliert betrachtet finden wir unterschiedliche Muster für krankheitsrelevante microRNAs und Gene bezogen auf deren Position und Rolle in Prozessen innerhalb der Signalwege.

Zweitens stellen wir eine neue Methode zur Identifizierung von microRNA-Signalweg-Verbindungen vor. Normalerweise werden microRNA-Signalweg-Verbindungen durch eine signifikante Anhäufung von microRNA-Zielgenen identifiziert. Der Nachteil dieser Methode ist, dass sie keine Informationen der biologischen Netzwerke verwendet. Daher stellen wir eine neue Methode vor, die die Netzwerkstruktur berücksichtigt. Durch die Definition eines Proximity Wertes finden wir eine neue Klasse von microRNA-Signalweg-Verbindungen, die sich von den durch die Standard Enrichment Methode identifizierten Verbindungen unterscheidet. Schließlich stellen wir den miTALOS Webservice vor, in den wir neben anderen Besonderheiten auch die Proximity Methode integriert haben.

Drittens analysieren wir den Einfluss von microRNAs auf die Dynamik von Phosphorylierungs-Prozessen. Wir untersuchen die Rolle von microRNAs in einer linearen Signalkaskade, erweitern dieses Model später zu dem gp130-STAT Signalweg und zeigen, dass die Reduktion von STAT3 vor der Stimulation zu einer Veränderung des pSTAT3/STAT3 Verhältnisses führt.

Schließlich stellen wir eine neue Methode zur Analyse von biologischen Daten vor. Matrixfaktorisierungs-Methoden sind ideal geeignet für die Analyse von mehrschichtigen und zeitlichen Antworten, die durch Signalwege erzeugt werden. Wir erweitern diese Methode und integrieren Vorwissen über Regulationsprozesse, um Gene und microRNAs durch das Netzwerk zu verbinden. Dadurch können wir eine Graph-Dekorrelationsfunktion erstellen und diese mit der Matrixfaktorisierung verbinden, was uns dann erlaubt unseren Graph-Dekorrelationsalgorithmus GraDe zu entwickeln.

Content

List of Figures

List of Tables

List of Abbreviations

AML	acute myeloid leukemia, page 52
BSS	blind source separation, page 38
BTO	Brenda Tissue Ontology, page 48
DNA	deoxyribonucleic acid, page 11
DO	Disease Ontology, page 47
EST	expressed sequence tag, page 24
FAB	French-American-British, page 53
GES	gene expression source, page 144
GO	Gene Ontology, page 48
ICA	Independent Component Analysis, page 39
IL-6	Interleukin-6, page 29
JAK	Janus kinase, page 29
KEGG	Kyoto Encyclopedia of Genes and Genomes, page 63
LOD	log-odds score, page 55
MCMC	Markov chain Monte Carlo, page 109
MeSH	Medical Subject Heading, page 47
miRNA	microRNA, page 11
mRNA	messenger RNA, page 11
NCA	Network Component Analysis, page 124
NCI PID	National Cancer Institute Pathway Interaction Database, page 63
ncRNA	non-coding, page 11
NGS	Next-generation sequencing, page 25
ODE	ordinary differential equation, page 36
OMIM	Online Mendelian Inheritance in Man, page 47
PCA	Principle Component Analysis, page 40
PCR	polymerase chain reaction, page 22
pri-miRNA	primary miRNA, page 16
Q-PCR	quantitative PCR, page 24
RNA	ribonucleic acid, page 11
rRNA	ribosomal RNA, page 14
siRNA	small interfering RNA, page 11
snoRNA	small nucleolar RNA, page 11
snRNA	small nuclear RNA, page 11
SOCS	suppressors of cytokine signaling, page 29
STAT	Signal Transducer and Activator of Transcription, page 29
Uniprot	Universal Protein Resource, page 85
UTR	untranslated region, page 15

1 Introduction

In 1993, Lee, Feinbaum, and Ambros discovered that the gene lin-4 encodes a pair of short RNA transcripts instead of proteins that regulate the timing of larval development by translational repression of lin-14 (156). They postulated a regulation of lin-14 based on sequence complementarity between lin-4 and a unique pattern within the 3'-untranslated region of the lin-14 messenger RNA (mRNA). Seven years later a second microRNA (miRNA), let-7 was discovered (219). Since these discoveries, thousands of miRNAs have been identified in organisms from viruses to human and define a new research field in the area of post-transcriptional regulation. Nowadays, miRNAs are commonly described as small endogenous RNA molecules ($\sim$21-25 nucleotides (nt)) that regulate gene expression by either targeting one or more mRNAs for translational repression or cleavage. Although the discovery of miRNAs was over ten years ago, we have just begun to understand the scope and diversity of these regulatory molecules. Within this thesis, we analyze and present regulatory motifs and mechanisms of miRNA regulation with a special focus on signal transduction pathways. Signaling pathways and miRNAs share common principles, which make the study of their interaction highly interesting. The effectiveness of signaling pathways relies on their capacity to control the expression of genes and therefore the resulting proteins in time and space. Within this thesis, we study these two mechanisms from a general perspective to a detailed model of a regulatory interaction of miRNAs and the gp130-STAT pathway. In Chapter 3 and 4, we analyze the regulatory pattern of disease-related miRNAs, followed by introducing a novel approach for the detection of functional miRNA-pathway associations in Chapter 5. Increasing the granularity of our analysis, we study the potential complexity of the miRNA-signaling network relationship by a quantitative mathematical model in Chapter 6. Finally, we present a novel approach for the large-scale analysis of biological data, which is beyond the simple exploratory clustering of 'omics' data. Our graph-decorrelation algorithm (GraDe) can

be applied to study the multi-layered and temporal responses of signaling pathways. Moreover, GraDe offers a natural way to integrate different kind of available data, e.g. to analyze combined miRNA-mRNA expression data.

1.1 Overview of this thesis

In this thesis, we analyze of regulatory motifs of miRNA-mediated regulation in signaling pathways. To unveil these regulatory motifs we combine various methods from different fields of science covering bioinformatics, mathematical biology and molecular biology.

Chapter 2 gives an overview of biological mechanisms, which are covered in this thesis. In the first part, we introduce the transcriptome and its regulation starting from transcriptional to miRNA regulation, followed by an overview of state of the art technologies to measure the transcriptome. Further, we introduce the concepts of cellular communication and signal transduction with a special focus on the gp130-Stat pathway, which is one of the central pathway covered in this thesis. We then give a brief overview about quantitative models in Systems Biology and finally an introduction about latent variable models.

In **Chapter 3**, we investigate the influence of differentially regulated genomic miRNAs on diseases from a larger-scale statistical point of view. Data for the analyses was retrieved from our novel manually curated database PhenomiR (226). Within this study, we focus on the differences in observed miRNA expression pattern between cell cultures and living organisms. Cell lines have been established in life sciences as easy to manipulate model systems for the study of all kind of cellular processes. However, studies using both in vitro and in vivo systems have shown that results, especially in cancer, do not always correlate. Previous studies observed differences in gene expression patterns between cell lines and their fresh-frozen tissue counterparts. With respect to the discrepancies between cell cultures and living organisms mentioned above, we questioned whether cell cultures are reliable disease models for the analysis of differential miRNA expression. Using PhenomiR data from more than one hundred cell culture studies and more than two hundred patient studies, we found that depending on disease type, integration of independent information from cell culture studies are in conflict to conclusions drawn from patient studies. These observations and the results of our study show that the potential of cell cultures to investigate miRNA expression

in diseases is limited. As a consequence, the suitability of cell cultures has to be verified for each disease and cell line before using such data as tool for the prognosis of diseases in human beings.

While creating the PhenomiR database, we identified individual studies, which did not investigate the impact of single deregulated miRNAs but miRNA genomic clusters. We then asked whether the impact of miRNA clusters on diseases is only restricted to a few examples or if miRNA clusters significantly correlate to the pathobiology of diseases? Using the comprehensive dataset from our PhenomiR database, we identify for the first time in a systematic analysis that deregulated miRNA clusters are significantly overrepresented in the majority of investigated diseases, compared to singular miRNA gene products. These experimental findings and the systematic correlation of miRNA cluster deregulation on human disease strongly support the idea that a coordinated regulatory effect is a general attribute of miRNA clusters. The pivotal role of miRNA clusters in miRNA-based gene silencing found in human diseases suggest that effective treatment of various diseases may require a combinatorial approach to target not singular miRNAs but rather miRNA clusters.

Chapter 4 studies the regulatory role of miRNAs on signal transduction pathways from a statistical point of view (145). Starting with a general analysis of the interaction of miRNAs and phenotypic observations in Chapter 3, we further focus in this thesis on the regulatory role of miRNAs in signal transduction pathways. Various studies indicate that miRNAs can function as tumor suppressors or even as oncogenes when aberrantly expressed. To study general aspects of these regulatory mechanisms, we set up a multipartite graph consisting of miRNAs, proteins, diseases, and signaling pathways in a tissue-specific manner. We analyze the impact of disease-associated miRNAs on human signaling pathways from two perspectives. On a global scale, we identify a core set of signaling pathways with enriched tissue-specific miRNA targets across diseases. The functions of these pathways reflect the affinity of miRNAs to regulate cellular processes associated with apoptosis, proliferation or development. To illustrate the robustness of our result, we compare our result with findings for different miRNA prediction tools as well as randomization procedures. Using this procedure, we provide evidence that cancer and non-cancer related miRNAs show no significant different patterns. To unveil the interaction and regulation of miRNAs on signaling pathways more locally, we compare the cellular location and process type of disease-associated miRNA targets and proteins. While disease-associated proteins are highly enriched in extracellular components of the pathway, miRNA targets are preferentially located

in the nucleus. Moreover, targets of disease-associated miRNAs preferentially exhibit an inhibitory effect within the pathways in contrast to disease proteins. This chapter provides systematical insights into the interaction of disease-associated miRNAs and signaling pathways and uncovers differences in cellular locations and process types of miRNA targets and disease-associated proteins. With this chapter, we apply the commonly used enrichment score to infer functional miRNA-pathway associations. This technique takes only the number of miRNA target genes into account to infer functional interactions. We improved this approach by integrating features provided by biological networks and present a new approach in the next chapter.

In **Chapter 5**, we introduce a novel measurement for the detection of functional miRNA-pathway associations (182). Usually, the functional impact of miRNA control is assessed by identifying pathways with enriched targets. The standard procedure quantifies the proportion of targets of a miRNA in the specified pathway among the expected proportion of targets, inferred from the overall set of proteins. The underlying assumption of enrichment analysis is that miRNAs are believed to have many targets, which are only weakly regulated. However, this does not apply to the case where it is functionally sufficient for a miRNA to regulate a small sub-part or even a single transcript, e.g. a bottleneck or a pathway hub in a pathway. Therefore, to predict the impact of a miRNA on a specific pathway, the mere knowledge of the target transcripts is mostly not sufficient: miRNAs are promiscuous regulators, with often hundreds of targets. The challenge for computational biology in this context is: How can one infer signaling pathways under miRNA control from a large number of miRNA-target transcript relationships? We therefore propose a novel measure that is independent of pathway size and which takes network topology into account. We calculate average distances between targets of a single miRNA in signaling networks. Via comparison with random targets, we define a proximity score and identify pathways with proximal and distal target patterns. We apply the proximity score to experimentally validated miRNA targets in network representations of KEGG signaling pathways and identify miRNA-pathway associations that differ from those inferred with the conventionally used enrichment score. A gene ontology analysis reveals that each target pattern corresponds to a specific function in cell signaling. In addition to the statistical analysis, we developed the miTALOS web resource (146), which provides insight into miRNA-mediated regulation of signaling pathways. As a novel feature, miTALOS considers the tissue-specific expression signatures of miRNAs and target transcripts to improve the analysis of miRNA regulation in biological pathways. MiTALOS identifies poten-

tial pathway regulation by (*i*) an enrichment analysis of miRNA target genes and (*ii*) by using our novel proximity score to evaluate the functional role of miRNAs in biological pathways by their network proximity. Moreover, miTALOS integrates five different miRNA target prediction tools and two different signaling pathway resources (KEGG and NCI). A graphical visualization of miRNA targets in both KEGG and NCI PID signaling pathways is provided to illustrate their respective pathway context. We perform a functional analysis on prostate cancer related miRNAs and are able to infer a model of miRNA-mediated regulation on tumor proliferation, mobility and anti-apoptotic behavior. MiTALOS provides novel features that accomplish a substantial support to systematically infer regulation of signaling pathways mediated by miRNAs.The application of concepts from graph theory to signal transduction allows the identification of novel miRNA-pathway associations to promote functional hypothesis on miRNA control. We surmise that the concept of proximity can serve as a powerful tool to identify patterns in signal transduction beyond miRNA regulation, like disease genes or drug targets.

So far, we analyzed the regulatory control of miRNAs on signal transduction pathways from a general perspective. In **Chapter 6**, we study the impact of miRNAs on pathway dynamics using two different dynamical pathway models (147). In a first model, we study a signaling cascade to analyze the differences between a system without miRNA regulation and a pathway cascade targeted by miRNAs. Altering the pathway readout from a steady state level, we are able to show that miRNAs decrease the time of dimming the signal, where the recovery time of the systems is unaltered. In further analysis, we adapt this signaling cascade to a full model of the gp130-STAT3 pathway by integrating receptor activation as well as negative feedback. We obtain time-series data of IL-6 stimulation in primary mouse hepatocytes, which were done by the group of Dr. Klingmüller at the DKFZ in Heidelberg. We study the phospho-dynamics as well as the impact of miRNAs on the pSTAT3/STAT3 ratio in the cytoplasm. We are able to show that a model integrating miRNA influence on the pathway results in reliable turnover rates compared to model without miRNAs. Using the miRNA model, we show that a pre-induced decrease of STAT3 changes the overall ratio of pSTAT3/STAT3 in the cell. Moreover, this effect results in a time shift of the maximal pSTAT3 concentration in the cytoplasm. Finally, the model reveals that induced miRNAs based on active pSTAT3 have no influence on the pathway dynamics in primary hepatocytes based on IL-6 stimulation. Our results presented in this chapter indicate that miRNA regulation as an additional layer of transcriptional control allows

the cell to alter the signal transduction and recovery in context specific manner.

In **Chapter 7**, we introduce a novel framework for the analysis of activated signaling pathways in large-scale biological data. External stimulations of cells by hormones, cytokines or growth factors activate signal transduction pathways that subsequently induce a re-arrangement of cellular gene expression. The analysis of such changes is complicated, as they consist of multi-layered temporal responses. While classical analyses based on clustering or gene set enrichment only partly reveal this information, matrix factorization techniques are well suited for a detailed temporal analysis. In signal processing, factorization techniques incorporating data properties like spatial and temporal correlation structure have shown to be robust and computationally efficient. However, such correlation-based methods have so far not been applied in bioinformatics, because large-scale biological data rarely imply a natural order that allows the definition of a delayed correlation function. We therefore develop the concept of graph-decorrelation. We encode prior knowledge like transcriptional regulation, protein interactions, miRNA regulation or metabolic pathways in a weighted directed graph. By linking features along this underlying graph, we introduce a partial ordering of the features (e.g. genes) and are thus able to define a graph-delayed correlation function. Using this framework as constraint to the matrix factorization task allows us to set up the fast and robust graph-decorrelation algorithm GraDe (144). The first section of this chapter gives an overview of the mathematical concept. We then demonstrate the applicability of GraDe by two toy examples. Furthermore, we introduce G-MA processes, which we used to evaluate the performance of GraDe. Finally, we present three biological studies with different experimental settings and goals to illustrate the versatility of GraDe: (*i*) The cytokine interleukin IL-6 mediates the production of acute phase proteins by hepatocytes and promotes liver regeneration (63). In order to unveil the multi-layered temporal gene responses in these processes, we measure gene expression in *IL-6* stimulated mouse hepatocytes by a time-course microarray experiment. The experiment were done by the group of Dr. Klingmüller at the DKFZ in Heidelberg. Applying GraDe with a literature based gene regulatory network, we are able to infer associated biological processes as well as the dynamic behavior of *IL-6* related gene expression. In addition, we find that the estimated factors are robust against the high number of false positives contained in large-scale biological databases. (*ii*) We apply Grade to microarray data from a stem cell differentiation experiment (16). In contrast to other factorization techniques, it finds a structured and detailed separation of known biological processes. (*iii*) Finally, we apply GraDe to combined mRNA

and miRNA data, which were measured in collaboration with the group of Dr. Götz at the Helmholtz Zentrum München. We use information about gene regulation and miRNA target genes to link mRNA and miRNA in a regulatory network. Applying GraDe, we are able to identify a core regulatory network of the differentiation process in glutamatergic neurons from high-throughput data (175).

In **Chapter 8**, we summarize and conclude the individual projects described in the thesis. Furthermore, we give an outlook on possible further projects.

1.2 Main scientific contributions

The main scientific contributions of this thesis have been published in the following peer reviewed publications, which are sorted by the corresponding chapters. Some of these projects lead to further collaborations and resulting publications are also cited within this thesis.

Chapter 3

- Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ: *PhenomiR: a knowledgebase for miRNA expression in diseases and biological processes.* Genome Biology 2010, 11(1):R6

- Ruepp A, Kowarsch A, Theis FJ: PhenomiR: *MicroRNAs in human diseases and biological processes.* MicroRNA Expression Profiling: Methods and Protocols. Series: Methods in Molecular Biology, Springer, 2011

Chapter 4

- Kowarsch A, Marr C, Schmidl D, Ruepp A, Theis FJ: *Tissue-specific target analysis of disease-associated miRNAs in human signaling pathways.* PLoS one 5(6), 2010

Chapter 5

- Kowarsch A, Preusse M, Marr C, Theis FJ: *miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse miRNAs.* RNA 17(5), 2011

- Marr C, Kowarsch A, Preusse M, Backofen R., Theis FJ: *Beyond enrichment: Measuring miRNA-pathway associations in signaling networks.* submitted

Chapter 6

- Kowarsch A, Schmidl D, Braun S, Bohl S, Merkle R, Klingmüller U, Theis FJ: *MicroRNA-mediated regulation has an impact on the dynamic behavior of the JAK-STAT pathway.* Manuscript in preparation

Chapter 7

- Kowarsch A, Blöchl F, Bohl S, Saile M, Gretz N, Klingmüller U, Theis FJ: *Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation.* BMC Bioinformatics, 11(1), 2010

- Blöchl F, Kowarsch A, Theis FJ: *Second-order source separation based on prior knowledge realized in a graph model.* In Proc. LCA/ICA 2010, Lecture Notes of Computer Science, Springer, 2010

- Lutter D, Walcher T, Lerch M, Röh S, Kowarsch A, Götz M, Ninkovic J, Theis FJ: *A Bayesian approach to infer boolean models for neuronal progenitor cell differentiation*. Manuscript in preparation

1.3 Further scientific projects and collaborations

Besides the main scientific contributions, which are described in this thesis, I have worked on various projects and collaborations that lead to further peer-reviewed publications that are not described in this thesis:

- Jungraithmayr TC, Hofer K, Cochat P, Chernin G, Cortina G, Fargue S, Grimm P, Knueppel T, Kowarsch A, Neuhaus T, Pagel P, Pfeiffer KP, Schäfer F, Schönermarck U, Seeman T, Toenshoff B, Weber S, Winn MP, Zschocke J, Zimmerhackl LB: *Screening for NPHS2 Mutations May Help Predict FSGS Recurrence after Transplantation.* Journal of the American Society of Nephrology 22, 2011

- Raia V, Schilling M, Böhm M, Hahn B, Kowarsch A, Raue A, Sticht C, Bohl S, Saile M, Möller P, Gretz N, Timmer J, Theis FJ, Lehmann WD, Lichter P, Klingmüller U: *Dynamic Mathematical Modeling of IL13-induced Signaling in*

Hodgkin and Primary Mediastinal B-cell Lymphoma Allows Prediction of Therapeutic Targets. Cancer Research 70(22), 2010

- Konopka W, Kiryk A, Novak M , Herwerth M, Parkitna JR, Wawrzyniak M, Kowarsch A, Michaluk P, Dzwonek J, Arnsperger T, Wilczynski GM, Merkenschlager M, Theis FJ, Köhr G, Kaczmarek L, Schütz G: *miRNA loss enhances learning and memory in mice.* Journal of Neuroscience 30(44), 2010

- Balluff B, Elsner M, Kowarsch A, Rauser S, Meding S, Schuhmacher C, Feith M, Herrmann K, Röcken C, Schmid RM, Höfler H, Walch A, Eberl MP: *Classification of HER2/neu status in gastric cancer using a breast-cancer derived proteome classifier.* Journal of Proteome Research 9(12), 2010

- Kowarsch A, Fuchs A, Frishman D, Pagel P: *Correlated mutations: a hallmark of phenotypic amino acid substitution.* PLoS Computational Biology 6(9), 2010

2 Background

2.1 The transcriptome: Transcription and non-coding RNAs

Transcription specifies the process of creating a complementary ribonucleic acid (RNA) copy of a specific deoxyribonucleic acid (DNA) sequence. The resulting transcriptome is the total set of transcripts in a cell. In contrast to the genome, which is roughly fixed for a cell, the transcriptome varies with environmental conditions. The description of the transcriptome is essentially limited to the characterization of the transcription products of annotated genes. These products are mainly coding messenger RNAs (mRNAs) and known stable non-coding RNAs (ncRNAs), such as tRNAs, small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). However, the level of complexity began to increase, with the discovery of interfering RNAs such as small interfering RNAs (siRNAs) and microRNAs (miRNAs). In the following section, we give an overview of the transcription process and its corresponding control mechanisms.

2.1.1 Transcription

Whenever a specific RNA is need within a cell, a complementary RNA copy of a specific DNA sequence is generated. This process is called transcription and the DNA helix is used as a template. In living organisms, DNA usually exists of two long strands in the shape of a double helix. One strand is called "sense" if its sequence is the same as that of the complementary RNA copy. The sequence on the opposite strand is called "anti-sense". In both eukaryotes and prokaryotes, the function of produced anti-sense RNA sequences is unknown. The DNA helix double strand is organized to a complex called chromatin. The packaging of the DNA into condensed chromatin makes it inaccessible to DNA transcription, replication, recombination or repair. The chromatin is

built of DNA, histone and non-histone proteins, subdivided into nucleosomes. The nucleosome itself is composed out of an octamer of four core histones (H2A, H2B, H3, H4) that are wrapped around 147 base pairs of DNA. The accessibility of a specific DNA region is controlled by the structure of the chromatin. Histone modifications and nucleosome remodeling have an enormous influence on the transcription of the DNA. Histones can be modified through at least eight different ways (143), which all have influence on transcriptional activity. In general the chromatin structure controls the transcription activity on a basal level. Therefore, it controls the accessibility of the DNA, thus coding and noncoding RNAs. These two principles of the design of the DNA are fundamental to understand the whole process of transcription.

The transcription process starts by binding of the RNA polymerase to the "open" DNA on a specific region on the DNA defined as core promoter sequence. Promoter regions are found, in eukaryotes, between -30 and -90 base pairs upstream from the start of a transcript that can be either a coding mRNA or a ncRNA. For the binding of a RNA polymerase, various specific transcription factors are needed. This is a complex process in eukaryotes as the eukaryotic RNA polymerase does not directly recognize the core promoter sequences. Rather a collection of transcription factor proteins mediates the binding of RNA polymerase and the initiation of transcription. Beside the regulatory role of the chromatin structure, this is a first layer of a regulatory mechanism. General transcription factor (GTF) proteins are required by the RNA polymerase to bind to the DNA and to initiate the transcription. Further transcription factor proteins can promote or block the recruitment of RNA polymerase either alone or with proteins in a complex, catalyze the acetylation or deacetylation of DNA or recruit co-activator or co-repressor proteins. Approximately 5-10% of all mammalian proteins are involved as regulators of gene transcription (283). After RNA polymerase binding, the DNA is unwound by breaking hydrogen bonds between complimentary nucleotides. The template strand of the DNA is then used for RNA synthesis. In further proceeds, RNA polymerase traverses the sense strand and uses base pairing complementarity with the DNA template to create a RNA copy. As the RNA polymerase traverses the DNA from 3' to 5' the newly formed RNA is an exact copy of the 5' to 3' coding strand, except that thymines are replaced with uracils and the RNA molecules are composed of a ribose (5-carbon) sugar where DNA has deoxyribose. An important difference between DNA replication is that transcription produces multiple copies of the DNA template by simultaneous binding of multiple RNA polymerases. Finally, the transcription is terminated, the new transcript is cleaved and the 3'-end undergoes

the process of polyadenylation by adding around 250 adenosine residues to the end to form the poly(A)-tail.

2.1.2 mRNA

The chemical copy of the specific DNA segment is a mRNA molecule, which in turn encoded a protein product. The genetic information is encodes in the sequence of nucleotides of the purine bases adenine and guanine, and the pyrimidines thymine and cytosine. In a first progression step called transcription, RNA polymerase makes a copy of specific DNA segments to mRNA. This process is similar in eukaryotes and prokaryotes. One notable difference is that prokaryotic RNA polymerase is associated with mRNA processing enzymes during transcription so that processing can quickly proceed after the start of transcription. The unprocessed product of the transcription is called pre-mRNA, which is rapidly progressed containing different kind of modification such as 5' and 3' end as well as the splicing process. The 5'-end of precursor messenger RNA is modified by adding a RNA 7-methylguanosine cap (5'-cap) shortly after the start of transcription. A second step called polyadenylation occurs also immediately after transcription. The mRNA fragment is cleaved through an endonuclease complex associated with RNA polymerase. After the cleavage, around 250 adenosine residues are added to the free 3'-end at the cleavage site generating the poly(A)-tail. Polyadenylation is important for transcription termination, export of the mRNA from the nucleus, and translation.

The human genome project encoded a smaller than expected number of genes. This surprising observation has renewed interest in alternative pre-mRNA fragments. These alternative fragments are generated by the splicing process, a mechanism to generate a complex proteome from a limited number of coding transcripts. In fact, alternative splicing affects the expression of 60% of human genes. In order to understand this important process, one has to go one step back and has a closer look on the structure of already modified pre-mRNA. The coding sequences of eukaryotic genes are in many cases separated into small pieces of coding segments, called exons, which are separated by a non-coding sequences, so-called introns. Alternative splicing is the process, which reconnects the exons of a pre-RNA in multiple ways. This process leads to different mRNAs, which in turn will be translated into different protein isoforms and therefore greatly increases the diversity of proteins that can be encoded by the genome. In humans, approximately 95% of multi-exonic genes are alternatively spliced (206).

2.1.3 non-coding RNA

RNAs can be split into two subtypes: mRNAs, which are translated in protein and ncRNAs, which are functional RNA molecules that are not translated into proteins. The term *non-coding RNAs* include several highly functionally important RNAs such as long ncRNAs, miRNAs, piRNAs, ribosomal RNA (rRNA), siRNAs, snoRNAs, and tRNAs. Until now, the number of ncRNAs, which are present within the human genome is unknown. Table 2.1 summarized the most important ncRNAs but there exist thousands of longer transcripts and most of whose functions are unknown. These ncRNAs can be summarize as a hidden layer of internal signals, which control various levels of gene expression.

Name	Function
long ncRNA	non-protein coding transcripts longer than 200 nucleotides
miRNA	putative translational regulatory gene family
piRNA	Piwi-interacting RNA linked to transcriptional gene silencing of retrotransposons
rRNA	RNA component of the ribosome
siRNA	short interfering RNA involved in the RNA interference (RNAi) pathway
snoRNA	most known snoRNAs are involved in rRNA modification
tRNA	transfers a specific active amino acid to a growing polypeptide chain

Table 2.1: **Overview of non-coding RNA molecules.** Overview of the most important ncRNAs in eukaryotes.

The biological role of ncRNAs is diverse. The functions range from highly conserved across most cellular life to more ncRNAs specific to one or a few closely related species. Most of the highly conserved ncRNAs are involved in the protein translation process. First discovered in prokaryotes, ncRNAs mainly regulate mRNA translation and its stability. More than 60 ncRNAs involved in these functions were identified during the last years in *Escherichia coli* (185). Commonly, ncRNAs, which are co-expressed with mRNAs, will be cleavaged after transcription. In prokaryotes, ncRNAs are not the major class of genomic output, where in general 80-95% of transcripts are protein-coding sequences. In higher organisms, the proteome is much more stable as

well as the number of protein-coding genes, which varies by less than 30% between Caenorhabditis elegans (which has only approximately 10^3 cells) and humans (approximately 10^{14} cells). Eukaryotes have a far more developed RNA processing and signaling system than prokaryotes, which leads to much more sophisticated pathways of gene regulation and complex genetic mechanisms such as post-transcriptional control, DNA methylation, chromatin modification or imprinting.

Infrastructural ncRNAs have been studied for a long time and have well-described functions. This class of ncRNAs includes tRNAs, rRNAs, spliceosomal uRNAs or snRNAs and the common snoRNAs. Both translation and splicing process requires ncRNAs not only for recognition of RNA substrates, but also for its own process. Regulatory RNAs function in most mediated by base-pairing with complementary sequences in other RNAs and DNA forming RNA:RNA and RNA:DNA complexes. One of the most well characterized examples of regulatory RNA sequences is the untranslated regions (UTRs) of mRNAs by forming secondary structures that bind regulatory proteins or other small RNA molecules, such as miRNAs.

MicroRNA biogenesis

MiRNAs are endogenous, non-protein coding, approximately 22-nucleotide (nt) long RNA molecules that have recently emerged as post-transcriptional regulators. It has been reported, that miRNAs control diverse aspects of biology, including developmental timing, differentiation, proliferation, cell death, and metabolism (70; 138). Analyzing the human genome, more than 60% of human protein coding genes are predicted to contain miRNA binding sites within their 3'-untranslated regions (UTRs) (237). Up to now, more than 1400 human miRNAs were detected (82). Their diversity and number suggests that a vast number of normal and pathological outcomes may be controlled, at least in part, through miRNA-mediated repression. In this section, we would like to introduce miRNA biogenesis, function and post-transcriptional control of mRNA transcripts mediated by miRNAs.

MiRNAs are located in diverse regions of the whole genome including both protein coding and non-coding transcription areas. Approximately 50% of miRNAs are derived from non-coding RNA transcripts, while approximately 40% are located within the introns of protein coding genes, called intronic miRNAs. Most of the miRNAs are transcribed by RNA polymerase (RNA pol) II and will be modified right after the transcription by a 5'-cap and a poly-A-tail at the 3'-end, which is similar to the mRNA

transcription (158). Co-expression of mRNA and miRNA is achieved by embedding the miRNA sequence within the intron of a coding mRNA transcript. Rarely, embedded miRNAs can also be located with the 3'-UTR of an mRNA and will be transcribed by read-through transcription (196). Figure 2.1 illustrates the miRNA biogenesis and gives an overview about the important steps in the miRNA biogenesis.

The splicing of intronic primary miRNA (pri-miRNA) mediated by Drosha has no effect on the host gene splicing or stability. This indicates that protein-coding transcripts could give independently rise to both miRNA and mRNA transcripts (196). A recent work suggests that the presence of a miRNA within an intron may enhance processing of the miRNA by extending the time of pri-miRNA association with chromatin (211). The location of a miRNA in an intron does not necessarily lead to coexpression of the host gene and miRNA. Microarray analysis of mRNA and miRNAs reveal that only approximately 50% of intronic miRNAs are strongly co-expressed with their corresponding host gene (11). The difference in host gene and miRNA expression could occur through differential miRNA processing or through alternative promoter usage (47). Various studies indicate that 25-33% of intronic miRNAs are transcribed from independent promoters leading to an independent expression pattern (205). Finally, many miRNAs are encoded in the genome as clusters, which can range from 2 to 19 miRNA hairpins encoded in tandem in close proximity.

One third of miRNAs are encoded in the genome as clusters within a range of 5kb. Analysis of polycistronic miRNA transcripts indicates that clustered miRNAs can derive from a single transcript (11). In Chapter 3, we study the role of miRNA cluster in human disease and show their pivotal role in miRNA-based gene silencing in human diseases. The majority of miRNA genes are transcribed by the same RNA polymerase as protein coding genes. Moreover, the epigenetic control of miRNAs is similar to protein coding genes. The transcription of the pri-miRNA undergoes two cleavage steps to generate the mature miRNA. The first cleavage is mediated by the RNase III enzyme Drosha within the nucleus. In general, the transcribed pri-miRNA could be several thousand nucleotides long. Drosha cleaves at the base of the stem to generate a 60-100 nt hairpin RNA. A subset of miRNAs has been identified, which by-passes the Drosha cleavage and defines the set of 'mirtrons'. These miRNAs are embedded within short introns and the ends of the pre-miRNA hairpin are determined by the 5' and 3' splice sites of the intron (225). Spliced pri-miRNAs are translocated from the nucleus into the cytoplasm through interaction with exportin-5. Export to the cytoplasm pre-miRNA is released from exportin-5. After translocation into the cytoplasm, the

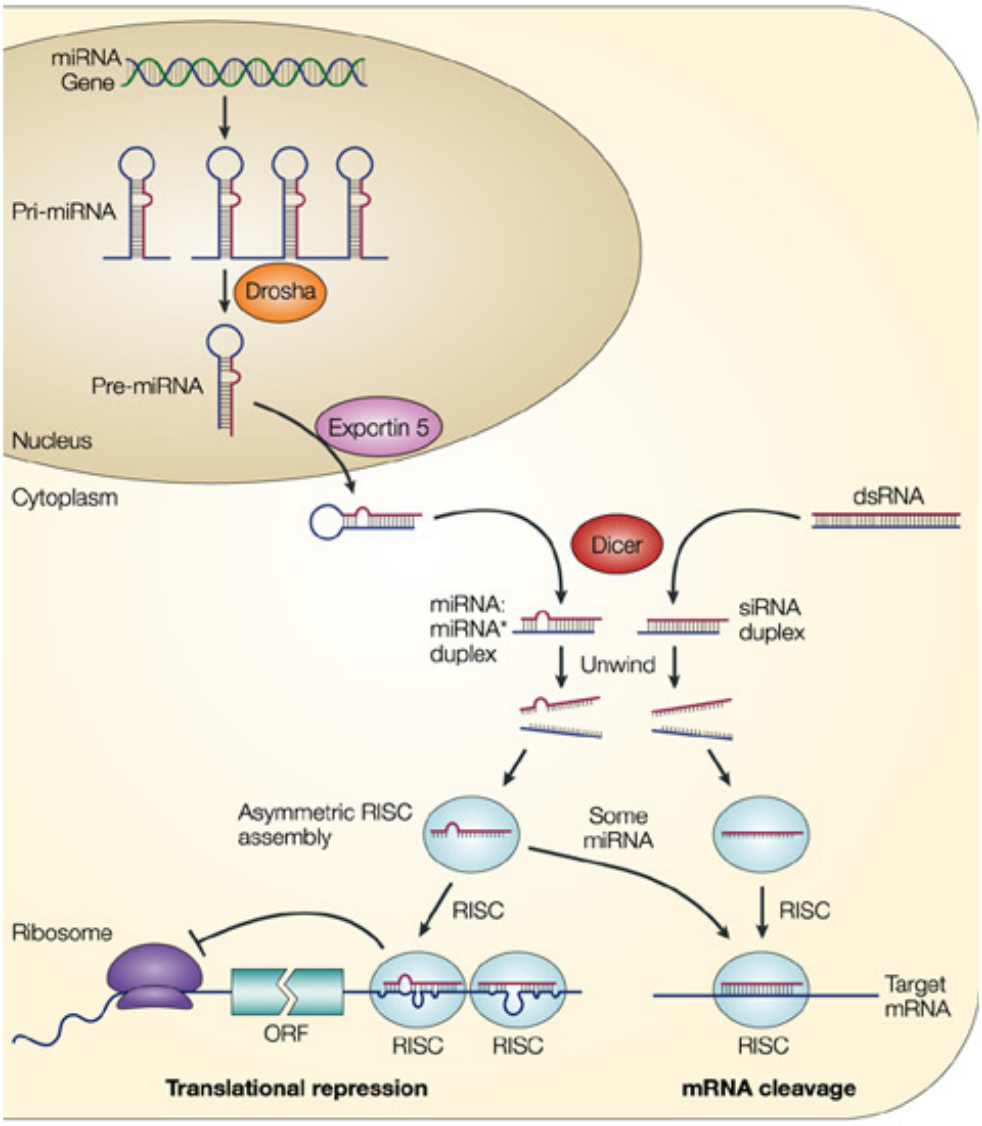

Figure 2.1: **The current model for the biogenesis and post-transcriptional suppression of microRNAs and small interfering RNAs.** The nascent primary microRNA (pri-miRNA) transcripts are first processed into 70-nt pre-miRNAs by Drosha inside the nucleus. Pre-miRNAs are transported to the cytoplasm by Exportin 5 and are processed into miRNA:miRNA* duplexes by Dicer. Dicer also processes long dsRNA molecules into siRNA duplexes. Only one strand of the miRNA:miRNA* duplex or the siRNA duplex is preferentially assembled into the RNA-induced silencing complex (RISC), which subsequently acts on its target by translational repression or mRNA cleavage, depending, at least in part, on the level of complementarity between the small RNA and its target. ORF, open reading frame. Figure adopted from (92).

pre-miRNA is cleaved near the terminal loop by the RNase type III protein Dicer generating an approximately 22-nt double-stranded miRNA. Cleavage by Dicer results in an unstable double-stranded miRNA containing an active strand (miRNA) and the passenger (miRNA*) strand. RLC is then responsible for unwinding the double-stranded miRNA, which leads to a degradation of the passenger strand. The single-stranded miRNA is then loaded into the RISC complex. Similarly to the control miRNA expression levels, the miRNA-mediated repression is highly regulated by diverse factors. Following maturation, the mature miRNA is loaded onto Ago proteins and associates with additional proteins to form the RISC complex. Regulation of Dicer, RISC or Ago protein expression results in an auto-regulation of expressed mature miRNA transcripts. Finally, the impact of expressed miRNAs can also be altered indirectly by modification of the mRNA target site. For example, mRNA binding factors, which interfere with the interaction of miRNA with target sites, have been reported.

2.1.4 Post-transcriptional control

Gene expression is regulated at multiple levels and within cells these differences in post-transcriptional control have to be coordinated. A first layer of transcriptional regulation is provided by chromatin arrangement and due to the activity of transcription factors, which lead to differentially transcribed genes. Transcript turnover and translational control are two integral parts of gene expression. Moreover, modifications of the mRNA transcript are powerful mechanisms to influence the gene expression: (*i*) The capping process changes the 5'-end of the mRNA transcript to a 3'-end by a 5'-5' linkage. This process protects the mRNA from 5' exonuclease, which degrades RNA. In addition, the 5'-cap modulates the ribosomal binding for the initiation of the translation. (*ii*) Transcribed mRNA (pre-mRNA) contains of exons that will make up the mRNA product, but which are interrupted by non-coding introns in the DNA and in the initial pre-mRNA transcript. The splicing process removes the introns and joins the exons, which is mediated by the spliceosome, a macromolecular ribonucleoprotein complex that assembles on the pre-mRNA in a series of complexes (117). (*iii*) Poly-adenylation of the 3'-tail of the pre-mRNA. The addition of poly-(A)-tail acts as a buffer to the 3' exonuclease in order to increase the half-life of mRNA. (*iv*) RNA-editing is the molecular process in which the information content in an RNA molecule is altered through a chemical change in the base sequence. RNA-editing is observed in tRNA, rRNA, and mRNA molecules of eukaryotes but not prokaryotes. In this section,

we will briefly discuss the main mechanisms of post-transcriptional control with strong impact on the composition of the transcriptome and gene expression.

RNA transport and localization

Localization of mRNAs is a widespread post-transcriptional mechanism for protein synthesis. This process is involved in the generation of cell polarity, asymmetric segregation of cell fate determinants and germ cell specification. The key elements during RNA localization are actin and microtubule filaments. Active transport of mRNAs along cytoskeletal filaments is the major localization mechanism in most cells (265). Although we do not know what fraction of the transcriptome is controlled by localization, recent reports from yeast suggest that in this simple unicellular eukaryote localized mRNAs make up >1% of expressed genes (180). In general, localized RNAs are characterized by signals that can generally be found in the 3'-untranslated region of mRNAs. These signals are recognized by RNA-binding proteins that are part of the localization machinery. Both the protein sequence and the structure of the RNA signal are important for the localization process, but so far no consensus signals have been identified (110). Similar to the diversity of the signals, different mechanisms that have been proposed for mRNA localization including mRNA transport, retention and site-specific degradation/protection (137). In general, every RNA is exported from the nucleus and has to pass the nuclear membrane via the nuclear pore complexes. So far, this process is poorly understood, but the nuclear RNA export is highly selective and is mainly mediated by a protein family termed exportins (karyopherins). These exportins depend on the activity of a small co-factor, which are able to recognize and process only completely processed mRNAs. The main recognition signals are cap-binding, poly-A-tail and further binding of transport proteins.

For miRNAs, the nucleus export process is well studied. Drosha-processed pre-miRNAs are transported by a heterotrimer of Exportin5 and Ran. After passing the nuclear pore complexes, the Ran-GTP complex is hydrolyzed to Ran-GDP, which in turn leases the pre-miRNAs. Unbound pre-miRNA in the cytoplasm binds to Dicer, an RNase III enzyme, which cuts the double stranded pre-miRNA generating a 22 nucleotide miRNA duplex. One strand is incorporated into RISC, whereas the other miRNA-strand is degraded (26).

RNA degradation or turnover

The mRNA turnover process is an important aspect of mRNA physiology. The mRNA turnover plays a main role in controlling the gene expression both is setting the basal level of gene expression and as a site of regulatory responses (209). Second, the mRNA turnover process recognizes and degrades aberrant mRNAs, thereby increasing the quality control of mRNA biogenesis (179). The stability of human mRNA is estimated to control the mRNA level of about 5-10% of all genes (18). Studies in several organisms indicate that the majority of mRNA transcripts are stable (216; 291). Within these studies, it appeared that the half-life of each mRNA transcript is related to its physiological role. For example, housekeeping genes mostly have long mRNA half-lives. In contrast, transcripts that are only required for specific process in the cell, e.g. cell cycle, during development, growth, differentiation or in response to external stimuli, often have short half-lives. So far, two pathways of mRNA decay have been identified in eukaryotic cells. The mRNA degradation usually starts with the deadenylation of the poly-(A)-tail at the 3'-end of the mRNA. After deadenylation, decapping enzymes remove the 5'-cap of the mRNA and therefore expose the transcript for digestion by a 5'$\rightarrow$3' exonuclease. Alternatively, mRNAs can also be degraded in a 3'$\rightarrow$5' direction mediated by the exosomes. Finally, mRNA degradation is initiated by endonuclease cleavage either by sequence-specific endonuclease or in response to miRNAs.

Whereas the framework of miRNA biogenesis is established, factors involved in miRNA degradation remain unknown (217). Recent studies from plants and nematode worms indicate that miRNA degradation is mediated by both 5'$\rightarrow$3' and 3'$\rightarrow$5' exonucleases (119). In addition, modifications of the miRNA 3'-end mediated by methylation or non-templated nucleotide addition have a severe impact on the stability of the miRNA. For detailed review about the molecular mechanism of miRNA degradation see (119; 217).

Transcriptional control mediated by microRNAs

The discovery of the abundance of miRNAs raised the question how they regulate their mRNA target transcripts. Initial evidences for miRNA target recognition came from the observation of the first miRNA lin-4 that has a sequence complementarity to multiple conserved sites within the lin-14 mRNA (157). Based on this and further findings the miRNA-mRNA interaction is given by a short perfect match complemented by im-

perfect matches in close vicinity. This region is defined as the seed sequence and is considered to be a 6-8 nucleotides (nt) long substring within the first 8 nt at the 5'-end of the miRNA (163). This feature regards to be the most important feature for target recognition by miRNAs in mammalians (10). A central feature of current miRNA target prediction approaches is the evolutionary conservation of miRNA target sites. But there is evidence that non-conserved miRNA target patterns are highly common (8). Several new features were detected and tested for their suitability of miRNA target prediction. In order to mention the most favored features: Extensive previous work revealed that 7-8 nt at the 5'-end of the miRNA are very important for target recognition. A recent work proposed that the majority of functional target sites are formed by less specific seeds of only 6 nt (56). The structural accessibility of miRNA target site, and the number of multiple target sites in close proximity have also been reported to be highly predictive for functional miRNA target sites. So far mostly miRNA target recognition is based on prediction methods but the upcoming of new large-scale technologies for the exact detection of miRNA targets will identify the important features for target recognition. Two recent works provided a technology to generating a miRNA:mRNA interaction map that contains the absolute chromosomal positions of miRNA seed positions within mRNA transcripts (37; 89). Moreover, these technologies will also improve the understanding of the interaction between miRNAs and mRNAs.

2.1.5 Measuring the transcriptome

The study of the transcriptome can be described as the measurement of the activity of thousands of genes or miRNAs at once, to create a global picture of cellular function. These profiles are often obtained using high-throughput techniques based on DNA microarray technology. Beside this well-established technique, a new technology known as next-generation sequencing was introduced in the 21^{st} century. In this section, we would like to give an overview about the principles of these technologies and will discuss the applicability as well as the main issues and restrictions.

Gene expression using microarrays

The main principle of the microarray technology is the use of a labeled sample, called "target", which is immobilized on a solid surface on an array and hybridized in parallel to a large number of DNA sequences. Using this principle, tens of thousands of

transcript can be detected and quantified simultaneously. During the last decade, this technology was developed further. Although many different microarray supplier are on the market, the most commonly used microarray technologies can be divided into two groups: complementary DNA (cDNA) and oligonucleotide microarrays. Figure 2.2 gives a schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays. Microarray probes for cDNA arrays are usually generated by polymerase chain reaction (PCR) from cDNA libraries or clone collections and are printed onto glass slides or nylon membranes. Probes are then organized in so-called spots, complementary to nucleotide sequences of known transcripts. Typical cDNA array consist of more than 30,000 cDNAs on the surface of a conventional microscope slide (234). Probes for oligonucleotide arrays are synthesized in situ, either by photolithography onto silicon wafers or by ink-jet technology. Moreover, pre-synthesized oligonucleotides can also be directly printed onto the glass slides. A first advantage of the oligonucleotide arrays is that there is no time-consuming handling of cDNA resources. Second, the much shorter oligonucleotide probes can be designed to cover a unique part of a given transcript, which allows a better detection of closely related genes or splice variants. Third, oligonucleotide probes have less specific hybridization and reduced sensitivity. Another important difference between high-density oligonucleotide arrays and cDNA arrays is the target preparation. For both types, mRNA from cells or tissues is extracted and translated into cDNA. Followed by a labeling step, the cDNA is hybridized on the surface of the array and can be detected by phospho-imaging or fluorescence scanning. Based on the high reproducibility of oligonucleotide arrays, an accurate comparison of signals between separate arrays is guaranteed. For cDNA arrays, the process of girding is not accurate enough to allow precise comparison between different cDNA arrays. One of the biggest differences between cDNA and oligonucleotide microarrays is the use of either one target cDNA or in case of the cDNA arrays mRNAs from two different cell populations or tissues. In this case, mRNAs from two different cell populations or tissues are labeled each in a different fluorescent dye (either Cy3 or Cy5), mixed and hybridized to the same array. This procedure results in a competitive binding of the target to the same probe. After hybridization and washing, the array is scanned using two different wavelengths corresponding to the two dyes and the intensity of the same spot in both channels is compared. This procedure results in a ratio of transcript levels for each gene. Due to different signal strength of the two dyes, one cDNA array experiment normally includes two arrays in which the dyes are swapped for the two

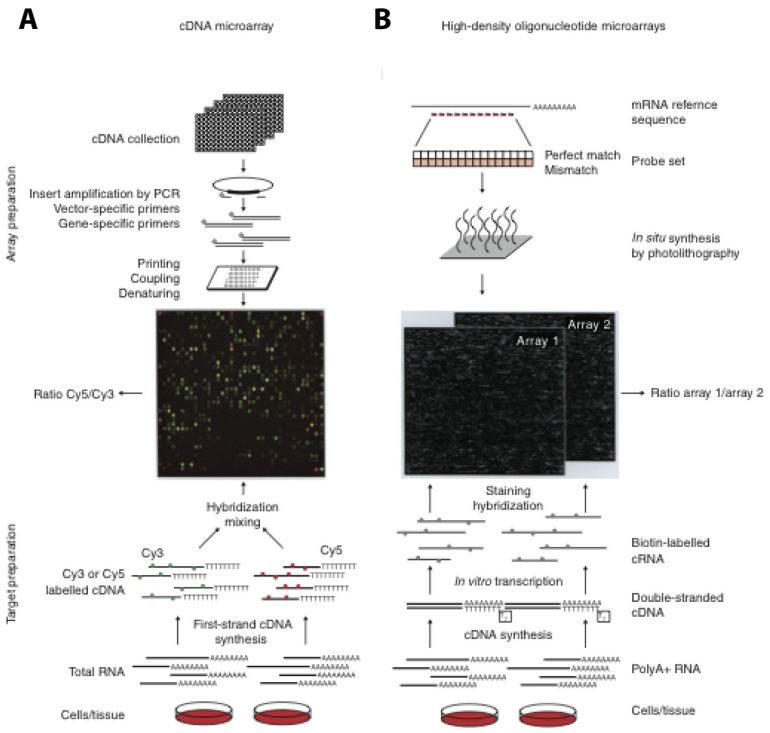

Figure 2.2: **Schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays.** (**A**) cDNA microarrays. Array preparation: inserts from cDNA collections or libraries are amplified using either vector-specific or gene-specific primers. PCR products are printed at specified sites on glass slides. Through the use of chemical linkers, selective covalent attachment of the coding strand to the glass surface can be achieved. Target preparation: RNA from two different tissues or cell populations is used to synthesize single-stranded cDNA in the presence of nucleotides labeled with two different fluorescent dyes (e.g. Cy3 and Cy5). Both samples are mixed in a small volume of hybridization buffer and hybridized to the array surface resulting in competitive binding of differentially labeled cDNAs to the corresponding array elements. High-resolution confocal fluorescence scanning of the array with two different wavelengths corresponding to the dyes used provides relative signal intensities and ratios of mRNA abundance. (**B**) High-density oligonucleotide microarrays. Array preparation: sequences of 16-20 short oligonucleotides are chosen from the mRNA reference sequence of each gene, often representing the most unique part of the transcript in the 5'-untranslated region. Light-directed, in situ oligonucleotide synthesis is used to generate high-density probe arrays containing over 300,000 individual elements. Target preparation: polyA+ RNA from different tissues or cell populations is used to generate double-stranded cDNA carrying a transcriptional start site for T7 DNA polymerase. During in vitro transcription, biotin-labeled nucleotides are incorporated into the synthesized cRNA molecules. Each target sample is hybridized to a separate probe array and target binding is detected by staining with a fluorescent dye coupled to streptavidin. Signal intensities of probe array element sets on different arrays are used to calculate relative mRNA abundance for the genes represented on the array. Figure adopted from (234).

mRNA target probes. A whole dye-swap experiment will compensate the dye effect as the Cy5 fluorescent dye typically leaves less signal than the Cy3.

The biggest limitation of the microarray technology comes from the fact that only genes can be detected and measured, which are represented on the array. However, not all genes or transcripts are known yet or sequences are wrongly identified during genome annotation. Moreover, the high throughput technologies lack in accuracy compared to single expression methods such as quantitative PCR (Q-PCR). One of the main imprecisenesses of the microarray technology is caused by cross-hybridization of mRNA transcript to a "wrong" probe. Furthermore, probes designed from genomic expressed sequence tags (ESTs) may be incorrectly associated with a transcript of a specific gene. Since a particular probe is mainly designed to match parts of the coding sequence, alternative splice forms of a single gene can not be determined. Finally, microarray methods detect only the expression level of mRNA transcripts. As these are subjects to post-transcriptional regulatory mechanisms the measured mRNA level does not give information about corresponding protein expression level.

Measuring microRNA expression

In contrast to mRNA profiling technologies, measuring miRNA expression has to take into account the difference between mature miRNAs and their precursors, and finally should also distinguish between miRNAs that differ by as little as a single nucleotide (246). For microarrays, either synthetic oligonucleotides or cDNA fragments are used as capture probes. An ideal probe should have high specificity and high affinity. It has been shown that the sequences of mature miRNAs have unequal melting temperatures and therefore sample labeling and probe design is a major challenge for miRNAs. For example, the melting temperatures of miRNAs vary between $45\,^{\circ}$C and $74\,^{\circ}$C. In case of a hybridization temperature (e.g. $55\,^{\circ}$C) set for the entire array, probes with a lower melting temperature will yield lower signals, whereas probes with higher melting temperatures will produce impaired nucleotide discrimination and lower specificity. Castoldi and co-workers (34) developed a locked nucleic acid modified capture probe, which can elevate thermal duplex stability. Another approach is 20-O-(2-methoxyethyl)-modified oligoribonucleotides that can detect newly identified validated or predicted miRNA candidates. In addition, probes for miRNAs have to be complementary to the sense and antisense strands of miRNAs. Moreover, different control probes are also required, which include exogenous and endogenous positive

controls and negative controls. These control probes are used for normalization and to provide reference points for quality control.

Beside array technologies, a variety of methods and tools have been developed for miRNA expression profiling. As hybridization of miRNA samples onto an array require a large amount of total RNA, a PCR based approach was developed to address the issue in case of access to only a small and limited amount of RNA. The PCR technique starts with a reverse transcriptase (RT) reaction. The resulting small RNA fractionation will be polyadenylated, followed by a standard RT protocol. This method has two advantages, as the total amount of RNA is much less compared to the array technology and second it can be applied for a rapid detection of miRNAs and precursors. Compared to the array technology the costs are much higher.

Over the past four years, massively parallel DNA sequencing platforms have become widely available and making the sequencing capacity of a genome center available for even small research institutes. This new technology, called Next-generation sequencing, is rapidly evolving and is a new effective approach for the comprehensive analysis of genomes, transcriptomes and interactomes. For further information and an overview of profiling strategies, we would like to recommend these two reviews: (140; 294). In the following section we would like to introduce this new technology and the applicability of expression profiling of mRNAs and miRNAs.

Next-generation sequencing of RNA

Next-generation sequencing (NGS) technologies include a number of methods that can be broadly summarized by template preparation, sequencing, imaging, and data analysis. In the following, we will discuss these steps for different sequencing platforms. For template preparation, in general NGS methods randomly break genomic DNA into smaller pieces from which either fragment templates or mate-pair templates are created. A common principle among NGS technologies is that the template is immobilized to a solid surface. The adventure of this immobilization is that thousands of sequencing reactions can be performed simultaneously. As fluorescent detection approaches are not designed to detect single events, NGS technologies require an amplification of the templates. The two most common methods are emulsion PCR (emPCR) (52) and solid-phase amplification (64). The advantage of emPCR is the amplification of the template in a cell-free system, which avoids the loss of genomic sequence. After the successful amplification and enrichment of emPCR beads, millions of reads

can be immobilized based on the platform either in a polyacrylamide gel on a standard microscope slide (Polonator) (241), chemically crosslinked to an amino- coated glass surface (Life/APG; Polonator) (130) or deposited into individual PicoTiterPlate wells (Roche/454) (154). Such an amplification can produce 100-200 million spatially separated template clusters (Illumina/Solexa) and provides free ends, which can be hybridized to a universal sequencing primer. Beside the clonally amplified methods, some NGS approaches use single-molecule templates, which are more straightforward and requires less starting material (<1 μg) than the amplified methods (3-20 μg).

The sequencing step is fundamentally different between clonally amplified and single-molecule templates. Clonal amplification results in a population of identical templates and each of them have undergone the sequencing process. During the imaging process the observed signal is a consensus signal of all probes of an identical template for a particular sequence cycle. The addition of multiple probes in a given cycle can result in leading-strand dephasing, which increases fluorescence noise, causing base-calling errors and shorter reads (58). For NGS methods based on single-molecule templates, dephasing is not an issue, therefore requirement for cycle efficiency is relaxed. For these techniques, deletion errors may occur owing to quenching effects or no signal will be detected due to incorporation of probes. The sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition. In most cases, the sequencing relies on synthesis. The enzyme driving synthesis can be either a polymerase or a ligase. Data acquisition can be done by imaging of the full array at each cycle (e.g., of fluorescently labeled nucleotides incorporated by a polymerase).

Global advantages of next-generation sequencing strategies relative to array-based sequencing is the much higher degree of parallelism than conventional capillary-based sequencing. As the size of one single sequencing spot can be around 1μm, millions of sequencing reads can potentially be obtained in parallel by imaging the surface area. Moreover, the planar surface can be enzymatically manipulated by a single reagent volume, in practice, reagent volumes in microliter-scale are essentially amortized over the full set of sequencing features on the array. This results in a dramatically lower costs for DNA sequencing. But nevertheless, NGS techniques also have several disadvantages. The most prominent, which is valid for all new platforms, is the read-length that is currently much shorter than conventional sequencing reads. Second, the raw accuracy of base-calls generated reads are at least tenfold less accurate than base-calls generated by Sanger sequencing. These limitations open an interesting research field

for Bioinformaticians to develop new algorithms, which address these issues. Table 2.2 gives an overview of current NGS technologies and their features. For further information about NGS technologies, we would like to refer to following reviews (192; 240).

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	cons	Biological applications
Roche/454 Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo- polymer repeats	Bacterial and insect genome de novo assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics
Illumina/ Solexa	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9II	18‡ 35,	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Life/APG SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14II	30‡,50	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Polonator	MP only/ emPCR	Non- cleavable probe SBL	26	5II	12II	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery
Helicos Bio-Sciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods
Pacific Biosciences	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks

Table 2.2: **Comparison of next-generation sequencing platforms.** *Average read-lengths. ‡Fragment run. II Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection. Table adopted from (192).

2.2 Cellular communication and signal transduction

The ability to respond to environmental stimuli belongs to the basic properties of life. Thus, organisms ranging from simple prokaryotes to humans are cognitive systems (181). The processing of complex amounts of information requires a regulatory network of switching devices based on the laws of mathematical logic, which are able to adapt and learn (3). In an organism, such networks can be found at any level of

detail. First, an organism is organized in a network of (specific) cells. Second, macromolecules between and within these cells form a second regulatory network. Third, these networks define smaller sub-networks of interacting proteins that define the lowest level.

Within the organism, signal processing is mainly mediated by proteins. Proteins are ideal candidates, due to their structural flexibility and capability to interact with each other. These properties make them efficient switching elements of a complex information-processing network. By interacting with each other they establish "neural networks" of logical gates that are in principle able to process any kind of logical information (181). The processing of information within these networks is based on protein communication mediated by short and long-range interactions. Short-range effects are mediated by direct contacts using specific binding domains, while long-range interactions are transmitted by signaling molecules, including proteins, small peptides, amino acids, steroids, retinoids, lipid derivatives, and even gases. Signaling molecules can be exchanged either via direct cell-cell interaction or through the interplay between signaling cells secreting specific molecules. A group of secreted signal molecules are cytokines (secreted glycoproteins), which are not able to cross the plasma membrane and therefore bind to receptor proteins that are located integral in the membrane. Based on this principle, extracellular received signals are converted into a series of biochemical reactions that in turn builds a signal transduction pathway. The most important biochemical process, which is responsible for signal maintenance within these signaling cascades is alteration of the phosphorylation state of certain proteins. The binding of the ligand on the extracellular region induces conformational changes within the receptor, thus activating either an intrinsic kinase domain of the receptor or a receptor-associated kinase. Upon this event, a series of proteins is modified, e.g. by phosphorylation, transporting the signal to a certain cellular compartment. In case of a transport to the cell nucleus, such proteins can modify gene expression by either initiation or suppression of the transcription process. With this process, metabolic processes, growth, differentiation, protein synthesis, or secretion processes are controlled and coordinated.

JAK-STAT signaling

In this thesis, we study the interaction of miRNAs and signaling pathways. Within Chapter 6 and 7, we special focus on the analysis of the JAK-STAT pathway. Here,

we introduce the biological background of this pathway. The pathway is triggered by several ligands and their cognate receptors, for example, growth hormones, such as epidermal growth factor (EGF), tyrosine kinases (RTK), such as the EGF receptor (EGF-R), SRC or ABL as well as G-protein-coupled receptors (GPCR) (24). Different cytokine families including the INF or Interleukin-6 (IL-6) are the major activators of JAK-STAT signaling cascade (232). After receptor re-organization the pathway is activated upon cytokine stimulation and phosphorylate tyrosine residues within the cytoplasmic domain of the receptor providing docking sites for Janus kinase (JAK) and Signal Transducer and Activator of Transcription (STAT) proteins. Binding to the receptor, STAT proteins become tyrosine-phosphorylated by JAK proteins. Activated STAT proteins dissociate from the receptor, dimerize and then are translocated to the nucleus where they can act as transcription factors to modulate gene expression (134). Figure 2.3 illustrates the main JAK-STAT cascade activated by the IL-6 via the gp130 receptor.

The JAK protein family includes JAK1-3 and the tyrosine kinase 2. JAK1 is the dominant kinase in IL-6-induced signaling, as cells lacking JAK1 are highly impaired in signal transmission. Prior to cytokine binding to gp130, JAK proteins are already associated at the receptor. Ligand-induced receptor oligomerization triggers auto- and trans-phosphorylation of JAK proteins, which in turn also involve other kinases such as SRC and RTKs. Activated JAK induces signal transduction by phosphorylation of STAT proteins and receptor tyrosine residues. Studies revealed that the JAK kinases do not seem to exhibit specificity for a particular STAT protein (46; 139). The specificity for STAT phosphorylation seems to be determined by specific docking sites for STAT proteins at the receptor site (218). STAT protein transduces the signal from the membrane to the nucleus and directly alters gene expression as transcription factor. The STAT protein family includes STAT1-6, whereas STAT5 is differentiated into STAT5A and 5B. IL-6 activates JAK-STAT cascade, mainly triggers STAT3 and to minor extent STAT1. For the activation of STAT3 by JAK1, STAT has to associate with the gp130 receptor unit. These protein-protein-interactions as well as other functions of STAT proteins are mediated by different domains, which are conserved within the STAT family. As uncontrolled signaling activity mediated by cytokines can lead to the development of cancer, signal termination at multiple control points is an essential part of signal transduction. In JAK-STAT signaling, three major groups of proteins can suppress the signaling cascade: protein tyrosine phosphatases (PTPs), protein inhibitors of activated STATs (PIAS) and suppressors of cytokine signaling (SOCS). In contrast

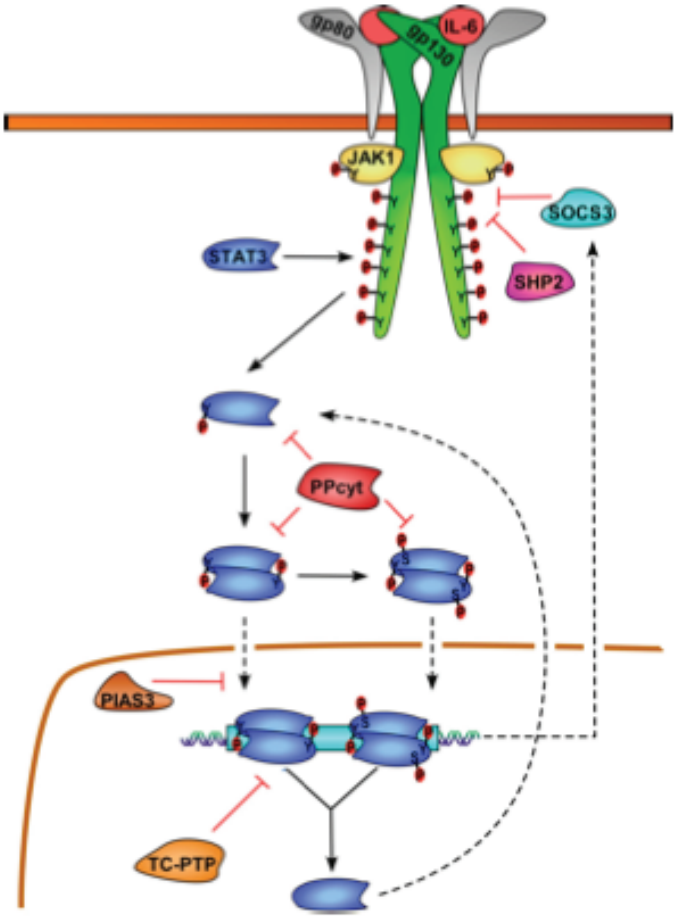

Figure 2.3: **IL-6-JAK1-STAT3 signaling pathway.** IL-6-induced receptor oligomerization and conformational changes of the receptor induce transactivation of associated JAK1 by auto-tyrosine phosphorylation. Activated JAK1 phosphorylates tyrosine residues of the cytoplasmic domain of gp130, providing binding sites for STAT3. STAT3 molecules become tyrosine-phosphorylated by JAK1, dissociate from the receptor and dimerize. STAT3 dimers are potentially phosphorylated on serine residues and translocate to the nucleus where target gene expression is altered. After dephosphorylation, the dimers dissociate and monomeric STAT3 re-enters the cytoplasm. Induction of SOCS represents a negative feedback loop for STAT3 activity. In addition to SOCS proteins, downregulation of gp130 signaling is mediated by recruitment of protein tyrosine phosphatases including SHP-2. Figure modified from (176).

to the constitutively expressed phosphatases, such as PIAS family, the SOCS protein family is a transcriptionally induced negative feedback. The SOCS protein family includes SOCS1-7 and the cytokine inducible SH2 domain-containing protein (CIS).

After signal induction upon cytokine stimulation, the expression of SOCS/CIS genes is rapidly induced. The inhibitory function of SOCS proteins is mediated via different protein-protein interactions. Binding of the SH2 domain to phosphotyrosine residues of the receptor inhibits the signal transduction by a competitive binding with STATs to the receptor leading to inhibition of STAT phosphorylation. The CIS protein prevents the activation of STAT5 by competitively binding to the Epo receptor (296). A direct protein interaction of SOCS with JAK negatively regulates JAK-STAT signaling by preventing STAT phosphorylation (198).

MicroRNA control in signaling pathways

As previously described, miRNAs are integral elements in the post-transcriptional control of gene expression. Within this thesis, we analyze the regulatory role of these elements within signal pathways. A specific question hereby is the effect of miRNA-mediated regulation of the mRNA expression level on the pathway and therefore protein level. Signaling pathways and miRNAs share common features, which make the study of their interaction highly interesting. The effectiveness of signaling pathways relies on their capacity to control the expression of genes and therefore the resulting proteins in time and space. Two main principles are used here to achieve this result: context-dependent transcriptional activation and repression. Repression mediated by miRNAs lead to a target gene expression, which is turned on exclusively in the presence of signal but kept repressed in its absence. This is mainly achieved at the transcriptional level, which is similar to the same responsive element on a gene promoter switching from repression to signal-dependent activation. An interesting question hereby is, whether the transcriptional control is sufficient to explain tight signaling regulation. One can argue that a cell is only challenging between an on or off situation. Much more frequently, cells have to distinguish between a real signal and stimulation that is too weak or transient. Here, miRNAs could be crucial for signal interpretation: by dampening positive mediators of signaling cascades, miRNAs may raise the threshold for pathway activation, restricting it only to appropriate zones of competence (106). Although the main function of miRNAs is the repression of gene expression, miRNAs can also act as signal activators. By targeting signal suppressors, context-specific miRNA activation can lead to a pathway activation (278).

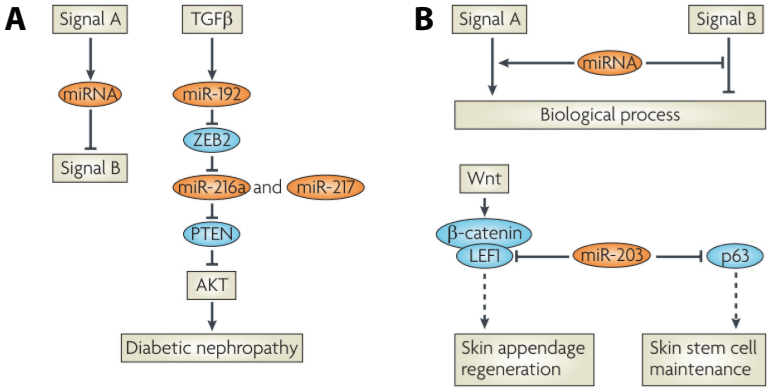

Figure 2.4: **MicroRNAs in signaling crosstalk and coordination.** (**A**) miRNAs can serve as mediators of crosstalk between signaling pathways. On the left, signal A induces the expression of a miRNA to negatively regulate signal B. On the right is a different example, with miRNAs enabling positive crosstalk between transforming growth factor-β (TGFβ) and AKT signaling. In glomerular mesangial cells, TGFβ induces the expression of miR-192, which represses the transcription factor zinc finger E-box-binding homeobox 2 (ZEB2). This results in the derepression of miR-216a and miR-217, enabling them to inhibit phosphatase and tensin homologue (PTEN), which leads to enhanced AKT activation. In these cells, this pathway triggers cell survival, extracellular matrix deposition and hypertrophy, all classic features of diabetic nephropathy. (**B**) miRNAs as signaling coordinators. A single miRNA can act simultaneously on two signaling pathways to coordinate their biological effects in a tissue or cell (upper diagram), as exemplified by miR-203-mediated regulation of skin tissue homeostasis (lower diagram). By antagonizing both WNT signaling (at the level of the transcriptional cofactor lymphoid enhancer-binding factor 1 (LEF1)) and p63 activity, miR-203 may have a general role in skin regeneration and self-renewal. Figure adopted from (106).

Beside the repression or activation of signal transduction, miRNAs also have the regulatory capacity to interconnect distinct signaling pathways. Figure 2.4 summarize these functions of miRNAs. The robustness of biological systems is a highly studied research field and miRNAs have been proposed to contribute to robustness by several mechanisms: balancing and buffering. For example, miRNAs can target both an activator and an inhibitor of the same pathway. Within the Nodal pathway, miR-430 targets both the Nodal homologue Squint and its inhibitor, Lefty (40). In case, miRNAs target the agonist-antagonist pair, this regulation leads to a reduction of their expression level (dampening effect), or regulates their relative levels to achieve an optimal signaling strength (balancing effect). Moreover, a buffering effect can be attained by limiting undesired signaling fluctuations.

Signaling pathways are highly interconnected and the flow of information may be controlled by many feedback loops (106). Recent works predict that miRNAs are crucial elements for the regulation of these networks (3; 99; 272). Within this thesis, we focus in Chapter 5 on this topic and present a novel approach to detect functional miRNA-pathway associations by taking the network topology into account (146; 182). We already described that miRNAs can act as repressor or activator of signals. In case of a coherent feedback loop in which the miRNA and its target are oppositely regulated by the same signal, miRNAs may act as stabilizer to fine tune the expression pattern of its target genes and repress it in cells where the expression is not desired.

Another interesting type of feedback pattern is defined by a miRNA that is negatively regulated by its own target (positive feedback). This double negative configuration is similar to toggle switches and can be used to convert a transient signal into a longer-lasting cellular response. Recent studies have reported several examples of miRNAs in toggle switches (85).

We introduced the biological interaction of miRNAs and signal transduction pathways. Using this knowledge, we will study in Chapter 6 the miRNA-mediated regulation of signaling pathway by analyzing the impact on the phosphorylation dynamic. Illustrating the complexity of the miRNA-signaling network relationship, it is difficult to escape the prediction that any reductionist approach will greatly benefit from the guidance of quantitative mathematical modeling. The involvement of miRNAs in feed-forward and feedback motifs makes miRNAs ideal reagents in the hands of systems biologists to offer insights into the physical properties of signaling pathways, something that could not be reached by intuition alone (106).

2.3 Models in Systems Biology

System-level understanding, the approach defined as systems biology, requires a shift in the notion of "what to look for" in biology (135). The pure identification of all genes and proteins in an organism is like generating a list of all parts. While such a list provides a catalog of the individual components, by itself it is not sufficient to understand the complexity underlying the organism. We need to know how these parts are assembled to form and regulate complex biological networks. A few years ago, a lot of effort was put into the generation of gene-regulatory networks and their biochemical interactions. Such networks provide limited knowledge of how changes to one gene

of a system may affect another gene, but can only provide limited understanding of a particular network. A system-level understanding of a biological system can be derived from insight into four key properties, which are presented by (135):

(i) System structures. These include the network of gene interactions and biochemical pathways, as well as the mechanisms by which such interactions modulate the physical properties of intracellular and multicellular structures.

(ii) System dynamics. How a system behaves over time under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as phase portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors. Bifurcation analysis traces time-varying change(s) in the state of the system in a multidimensional space where each dimension represents a particular concentration of the biochemical factor involved.

(iii) The control method. Mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease.

(iv) The design method. Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error.

So how can we systematically study complex biological processes? Mathematical modeling and computer simulations may help us to understand the dynamic and complexity of a system. Moreover, we can define a prediction by using these models, which can help us to find properties of the system. How can we define a model?

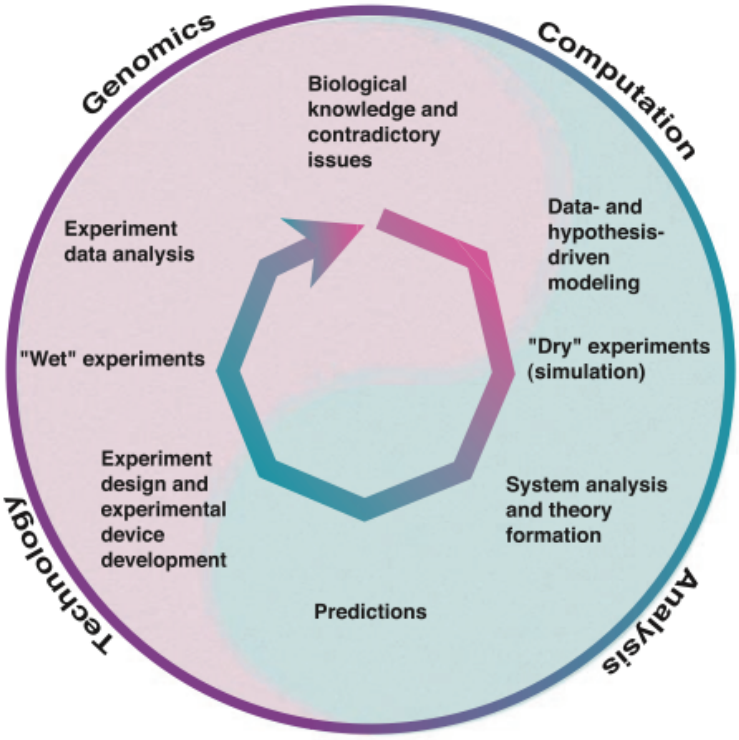

Figure 2.5: **Hypothesis-driven research in systems biology.** A cycle of research begins with the selection of contradictory issues of biological significance and the creation of a model representing the phenomenon. Models can be created either automatically or manually. The model represents a computable set of assumptions and hypotheses that need to be tested or supported experimentally. Computational "dry" experiments, such as simulation, on models reveal computational adequacy of the assumptions and hypotheses embedded in each model. Inadequate models would expose inconsistencies with established experimental facts, and thus need to be rejected or modified. Models that pass this test become subjects of a thorough system analysis where a number of predictions may be made. A set of predictions that can distinguish a correct model among competing models is selected for "wet" experiments. Successful experiments are those that eliminate inadequate models. Models that survive this cycle are deemed to be consistent with existing experimental evidence. While this is an idealized process of systems biology research, the hope is that advancement of research in computational science, analytical methods, technologies for measurements, and genomics will gradually transform biological research to fit this cycle for a more systematic and hypothesis-driven science. Figure adopted from (135).

In a broad sense, a model is an abstract representation of a (biological) process that may explain properties of this process. Biological and biochemical reaction networks can

be represented by a graphical sketch, but the same network can also be described by a system of differential equations, which then captures the dynamic of the underlying system. Systems biology models are often based on physical properties, for instance thermodynamic dynamic or Michaelis-Menten kinetic. One has to keep in mind, that the mathematical models are tailored systems to describe the underlying system as close as possible. For model generation, one has to consider the trade-off between a highly complex system, which in-depth describes a biological process but has many parameters to estimate, whereas a simpler model could be sophisticated enough to describe the process. George Box coined the phrase of "essentially, all models are wrong, but some are useful" (22).

2.3.1 Quantitative modeling

The era of the discovery and analysis of large interaction networks developed the concept of systems biology, which is now being applied to study complex biological networks. Starting with concepts from the graph theory, systems biology approaches tried to understand biochemical networks by determining the connection information within these networks. While cells are highly dynamical systems, new approaches to study these systems involve the creation of models that take into account the move and function in time. Now, the era of large-scale dynamical models of biochemical networks arise and to solve current questions in biology. Therefore, the most common way is the analysis of a dynamical network by describing a model with a set of differential equations. This approach has been extensively used to study protein and/or gene interactions, as well as small molecules like ncRNAs. One of the advantages of continuous modeling via differential equations is the precision and realistic description of molecular interactions. A deterministic modeling approach tries to describe a quantitative model with ordinary differential equations (ODEs):

$$\frac{dx_i}{dt} = \sum_j f_j(x), i = 1 \cdots N_x, j = 1 \cdots N_f \tag{2.1}$$

with x the vector of state variables, $f_j(x)$ defines the rate equations, N_x the number of total states, and N_f the number of rate equations. Rate equations $f_j(x)$ can be linear and nonlinear functions of the state vector x. In order to model pathway dynamic, one has to describe enzyme dynamic over time. One of the simplest and best understand models for enzyme kinetics is Michaelis-Menten kinetic. The Michaelis-Menten

equation describes the rates of irreversible enzymatic reactions by reaction rates of the concentration of the substrate.

$$x_i(t) = \frac{V_{max}[S]}{K_m + [S]} \tag{2.2}$$

with $x_i(t)$ defining the current reaction rate at time point t, V_{max} the maximum reaction rate, K_m the inverse of enzyme affinity and $[S]$ the substrate concentration. We apply this formula to an enzymatic reaction, which is assumed to be irreversible, and the product does not bind to the enzyme.

$$E + S \rightleftarrows ES \rightarrow E + P \tag{2.3}$$

where k_1 defines the reaction $E + S \rightarrow ES$ and $k_{-1} = E + S \leftarrow ES$ and k_2 the reaction $ES \rightarrow E + P$. Using this reaction, we obtain for $v_0 = k_2[ES]$, $V_{max} = k_2 E_0$ and $K_m = k_{-1} + k_2/k_1$.

The Hill equation is an extension of Michaelis-Menten kinetics. It describes the fraction of macromolecules saturated by ligand as a function of the ligand concentration. It can be used to determine the degree of cooperativeness of the ligand binding to the enzyme or receptor molecule.

$$\Theta = \frac{[L]^n}{(K_A)^n + [L]^n} \tag{2.4}$$

with $[L]$ describing the free ligand concentration, K_A is the dissociation constant and n defines the Hill coefficient. A Hill coefficient of 1 indicates completely independent binding, $n > 1$ indicate positive cooperativity, while $n < 1$ indicate negative cooperativity. The Hill coefficient was originally devised by the cooperative binding of oxygen to Hemoglobin (97).

Beside these two possibilities to describe the dynamic of protein or gene interaction, the law of mass action is a widely used principle to describe complex biological systems. The law of mass action explains the behaviors of solutions in dynamic equilibrium. Mass action can be described with two features: (*i*) the composition of a reaction mixture at equilibrium and (*ii*) the kinetic. Guldberg and Waage (84) introduce this concept by studying kinetic data and propose that chemical equilibrium is a dynamic process in which rates of reaction for forward and backward reactions must be equal.

In Chapter 6, we will use a quantitative model to study the impact of miRNAs on the signal dynamic by setting up two models based on ODEs. For modeling the protein, gene and miRNA reaction, we use Michaelis-Menten and mass action kinetics.

2.4 Latent variable models

In the context of blind source separation (BSS), latent variable models attracted much attention in the signal processing community over the last decades. These techniques, try to recover latent variables from underlying sources of observed mixture models. Blind source separation defines a set of algorithms, which separates a set of signals from a set of mixed signals, without information about the source signals or the mixing process. Blind signal separation relies on the assumption that the source signals do not correlate with each other. Signals may be either statistically independent or decorrelated. BSS thus separates source signals into a set of signals, such that the regularity of each signal is maximized and the regularity between the signals is minimized.

In order to illustrate the concept of BSS, a classic example is the "cocktail-party problem": where a number of n people are talking simultaneously in a room and therefore, acting like $k(k \leq n)$ sound sources. The clutter of conversations are recorded by m microphones, which are spread throughout the whole room. Due to the different distances between speakers and each microphone, every recorded signal is then a weighted mixture of the conversations. Several approaches have been proposed for the solution of this problem. Successful approaches are principal component analysis and independent component analysis, which work well when the records are free of delays or echoes. To extract both the individual speakers (source signals) and the mixing process from the recorded (mixed) signals, we assume that the mixing process should be instantaneous and linear. The recorded signals $x_1(t), \dots x_m(t)$ by time t can be formulated by a linear combination of the weighted source signals at time t by $s_1(t), \dots s_n(t)$:

$$\begin{aligned} x_1(t) =& a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1n}s_n(t) \\ &\vdots \\ x_m(t) =& a_{m1}s_1(t) + a_{2m}s_2(t) + \cdots + a_{mn}s_n(t). \end{aligned} \tag{2.5}$$

where a_{ij} denotes the mixing coefficients by weighting speaker s_j in signal x_j. Now, we can aggregate the m recorded signals into a mixture matrix $\mathbf{X} \in \mathbb{R}^{m \times l}$. Finally,

we write the upper equations in matrix representation, which results in the well know linear mixing model:

$$\mathbf{X} = \mathbf{AS}. \tag{2.6}$$

First, we assume that the mixing matrix $\mathbf{A}$ has to be full rank. Moreover, we center all data ($\mathbf{X}$) to mean zero. It is obvious that the matrix decomposition (2.6) has an infinite number of solutions, so further assumptions have to be made.

In the following, we introduce matrix factorization techniques that are used in this thesis. These techniques have a multitude of relevant applications in biological and biomedical signal processing (17; 104; 267).

2.4.1 Independent component analysis

One common technique is the independent component analysis (ICA). This approach assumes that the underlying sources are statistically independent. Statistical independence of the resulting sources is the strongest constraint compared to other requirement such as decorrelation, which we will discuss later on. The first step of almost all ICA algorithms is the data whitening. The whitening process is a decorrelation method that converts the covariance matrix of samples into the identity matrix I. This process transforms the input matrix closer towards white noise. This can be expressed by:

$$\mathbf{X}_w = \Lambda^{-\frac{1}{2}}\Phi^{\mathbf{T}} \tag{2.7}$$

where Φ is the matrix with the eigenvectors of the covariance matrix and Λ is the diagonal matrix of corresponding eigenvalues. In general the matrix decomposition has an infinite number of solutions, but in case the mixing matrix has full column rank and at least one of the sources has a Gaussian distribution, we achieve a unique solution of the matrix decomposition (28; 268). The latter implies that the number of mixtures is at least as large as the number of sources. A unique solution means in this context a unique modulo scaling and permutation.

In general, ICA identifies independent components by maximizing the statistical independence of the estimated components. The independence can be defined by: (*i*) Minimization of Mutual Information and (*ii*) Maximization of non-Gaussianity. The family of Minimization-of-Mutual information algorithms uses measures like Kullback-Leibler Divergence (151) and maximum-entropy, whereas the Non-Gaussianity family of ICA algorithms uses kurtosis and negentropy. Inferred from the central limit

theorem, maximizing non-gaussianity is a way to reveal the independent underlying sources. This property can be quantified by the fourth-order cumulants, the kurtosis, which is a measure of the peakedness of the probability distribution of a real-valued random variable:

$$k = \frac{\mathbf{E}(x-\mu)^4}{\sigma^4} \tag{2.8}$$

The second measure is the negentropy, which is used as a measure of distance to normality. In case a signal has a normal distribution, negentropy is always nonnegative and vanishes if and only if the signal is Gaussian. Negentropy is defined as

$$J(x) = S(\varphi_x) - S(x) \tag{2.9}$$

with $S(\varphi_x)$ being the differential entropy of the Gaussian density with the same mean and variance as x and $S(x)$ is the differential entropy of x. In the ICA, negentropy can be understood as the information that can be saved when representing x in an efficient way. Robust approximations of negentropy instead of kurtosis enhance the statistical properties of the resulting estimator. The two widely used ICA algorithms are based on non-Gaussianity. JADE is based on the approximation of the joint diagonalization of the fourth-order cumulants (31). The second approach, called FastICA is based on a fixed-point iteration scheme that maximizes negentropy (103).

2.4.2 Principal component analysis

Beside the strongest constraint of statistical independence sources, a second approach based on the decorrelation requirement is called principle component analysis (PCA) (212). PCA is based on a unique decomposition of correlated signals into a number of uncorrelated random variables. This technique transforms multivariate data into a new orthogonal basis, where the first new basis vector refers to the direction with the largest data variance. The first principal component (PC) is defined by a vector y_1 such that the linear combination $s1 := y_1^T\mathbf{X}$ has maximum variance. As we are only interested in the direction of $y1$, we require $|y_1| = 1$. PC1 is then defined as

$$\mathbf{C}_{s_1} = \mathbf{C}_{y_1^T\mathbf{X}} = y_1^T\mathbf{C}_{Xy_1} = y_1^T\mathbf{V}_T\mathbf{D}\mathbf{V}_{y_1} \tag{2.10}$$

with $\mathbf{V}_T\mathbf{D}\mathbf{V}$ being the eigenvalue decomposition of the data covariance. With $\mathbf{V}$ being orthogonal, also $|\mathbf{V}_{y_1}| = 1$ and we obtain the maximum variance when y_1 is the

unit vector. s_1 is the eigenvector to the largest eigenvalue D_{11} of $\mathbf{CX}$. We obtain the second PC by getting the orthogonal vector to PC1, which has then the second largest variance. Analogously, we identify all n PCs. The orthogonality of eigenvectors implies then the decorrelation of the different principal components at the end. In contrast to ICA, the matrix decomposition into decorrelated PCs is unique except for scaling. PCA calculates the first two moments (mean and variance) of the data by taking the data variance as a measure for information content. Therefore, PCA is often used for dimensionality reduction. Given a feature space $x_i \in \mathbb{R}^{\mathbf{N}}$, we find a mapping $y = f(x) : \mathbb{R}^{\mathbf{N}} \rightarrow \mathbb{R}^{\mathbf{M}}$ with $\mathbf{M} < \mathbf{N}$ such that the transformed feature vector $y_i \in \mathbb{R}^{\mathbf{M}}$ preserves most of the information in $\mathbb{R}^{\mathbf{N}}$. An optimal mapping $y = f(x)$ will be one that results in no increase in the minimum probability of error and is in most the cases a non-linear function.

2.4.3 Second-order methods using structural information

Finding repeating patterns in a biological data set is of frequent interest. In this section, we present the mathematical background of a novel approach, which is later on described in Chapter 7. In signal processing, various matrix factorization techniques have been developed that employ intrinsic properties of data such as repeating patterns (12; 270; 271). These methods are based on delayed correlations that can be defined as data having temporal or spatial structure. Time-delayed correlations quantify the similarity of a signal with itself after a time shift τ. For instance, the *time-delayed correlation matrix* of a centered, wide-sense stationary multivariate random process $\mathbf{x}(t)$ is defined as

$$\left(\mathbf{C}_{\mathbf{x}}(\tau)\right)_{ij} := E\left(\mathbf{x}_i(t+\tau)\mathbf{x}_j(t)^{\top}\right). \tag{2.11}$$

where E denotes expectation. Here, off-diagonal elements detect time-shifted correlations between different data dimensions. Given l features, e.g. genes and/or miRNAs, aggregated in a data matrix $\mathbf{X}$, the cross-correlation matrix can be easily estimated with the unbiased variance estimator. For $\tau = 0$ this measure reduces to the common cross-correlation.

$$\mathbf{C}_{\mathbf{X}} = \frac{1}{l-1}\mathbf{X}\mathbf{X}^{\top}. \tag{2.12}$$

These approaches solve the matrix factorization problem by considering only independent random variables, where the samples in particular have no intrinsic order. However, the experimentally generated quantitative data sets we face in bioinformatics

rarely imply a natural order which allows the definition of a generic kind of delayed correlation. In the following, we introduce a technique that use the temporal structure of the sources instead of taking into account higher-order moments. In order to deal with biological data, we generalize this concept by introducing prior knowledge that links features (e.g. genes and/or miRNAs) along a pre-defined underlying network. Integrated information may be of qualitative (e.g. interaction) as well as quantitative nature (e.g. interaction strength, reaction rates).

We will then use these information obtained in a time-delayed covariance matrix as constraint to the blind-source-separation problem. We try to find a factorization such that all time-delayed cross-covariances vanish. Therefore, $\overline{\mathbf{C}}_{\mathbf{S}}(\tau)$ has to be diagonal for all τ. Using this assumption, the time-delayed correlation matrices of the observations is defined as

$$\overline{\mathbf{C}}_{\mathbf{X}}^{G}(\tau) = \begin{cases} \mathbf{A}\overline{\mathbf{C}}_{\mathbf{S}}^{G}(\tau)\mathbf{A}^{\top} + \sigma^2\mathbf{I}, & \tau = 0 \\ \mathbf{A}\overline{\mathbf{C}}_{\mathbf{S}}^{G}(\tau)\mathbf{A}^{\top} & \tau \neq 0 \end{cases} . \tag{2.13}$$

A full identification of $\mathbf{A}$ and the covariance matrix is not possible because the linear mixing model for the input data matrix $\mathbf{X} \in \mathbb{R}^{m \times l}$ is given by

$$\mathbf{X} = \mathbf{AS} + \boldsymbol{\varepsilon} \, . \tag{2.14}$$

The linear mixing model defines them only up to scaling and permutation of columns. We can take advantage of the scaling indeterminacy by requiring our sources to have unit variance, i.e. $\overline{\mathbf{C}}_{\mathbf{S}}^{G}(0) = \mathbf{I}$. As we assumed white data $\tilde{\mathbf{X}}$, we see that $\mathbf{AA}^{\top} = \mathbf{I}$, i.e. $\mathbf{A}$ is orthogonal. Since the sources are assumed to be uncorrelated and with Equation (2.13) we obtain

$$\overline{\mathbf{C}}_{\mathbf{X}}(0) = \mathbf{AA}^{\top}. \tag{2.15}$$

After whitening our observations, the factorization in Equation (2.13) represents an eigenvalue decomposition of the symmetric matrix $\overline{\mathbf{C}}_{\mathbf{X}}^{G}(\tau)$. The reason why we focused on the symmetrized correlation matrix is, because then the eigenvalue decomposition is well defined and also simple to compute. If we assume that $\overline{\mathbf{C}}_{\mathbf{S}}^{G}(\tau)$ has pairwise different eigenvalues, the spectral theorem guarantees that $\mathbf{A}$ is uniquely decomposition determined by $\mathbf{X}$ except for permutation. In addition to this result, we see that the unmixing matrix $\mathbf{U}$ for a fixed choice of τ can be easily obtained by calculating the eigenvalue decomposition of $\overline{\mathbf{C}}_{\mathbf{X}}(\tau)$.

We can subsume this procedure to the AMUSE (Algorithm for Multiple Unknown

Signals Extraction) algorithm (195; 271). AMUSE can be summarized by:

(i) Data transformation: $y := Cx$ where C defines the whitening transformation

(ii) Select a τ and compute the eigenvalue/eigenvector decomposition

Applying AMUSE, it is often the case that the eigenvalue decomposition turns out to be the most problematic step, but choosing a different τ may often solve this issue. The performance of AMUSE is known to be relatively sensitive to additive noise and the numerical estimation by a finite amount of samples may lead to a badly estimated autocorrelation matrix (269). In order to improve the performance of AMUSE, more than on time lag can be considered, which is used in SOBI (12), TDSEP (301), or TFBSS (65).

In Chapter 7, we extend and modify the procedure of AMUSE to biological data. Hereby, we use prior knowledge to introduce a temporal structure to the data. Applying our graph-decorrelation algorithm (GraDe), we are able to extract biological 'signals', which can be interpreted as active signal pathways or processes. Furthermore, we show that our novel approach is able to handle multi-scale data.

3 Deregulated microRNAs from multiple disease studies

3.1 Background

MicroRNAs (miRNAs) are approximately 22-nucleotide endogenous RNAs predicted to regulate the expression of most mammalian genes (68). Since the discovery of miRNAs in Caenorhabditis elegans, the influence of these regulatory RNAs on cellular processes has been established in a large variety of metazoa (83). Accordingly, individual studies as well as large-scale endeavors have detected a growing number of miRNAs (14; 152), up to 695 in human according to miRBase release 12.0 (82). A proteome study that investigated the influence of the abundance of a single miRNA on cells found that the mode of regulation occurs through modulation of protein expression rather than as a binary off-switch (8).

However, the potential of deregulated miRNA expression to cause severe impairments has already been demonstrated in the early days of miRNA research (227). In 2004, it was shown that deregulated miRNA expression is associated with human diseases such as lung cancer (261). One year later, Lu et al. (172) analyzed miRNA expression in cancer types and observed that miRNA profiling is a more reliable indicator for cancer than mRNA expression profiles. In the meantime, additional studies have demonstrated that miRNAs are significant indicators for specific diseases and can, for example, be used to create decision trees differentiating cancer types solely by miRNA expression profiles (223; 277). In recent years, deregulated expression of miRNA has also been found to be associated with human diseases such as cardiomyopathy, muscular disorders and neurodegenerative diseases (55; 93; 105). The samples used for these studies stem from biopsies of patients or cell cultures, which are used as easily tractable experimental models. Besides diseases, miRNAs are also known to

have functional roles in eukaryotic organisms. MiRNA-mediated gene silencing was shown to be involved in a number of cellular processes, such as cell growth, larval development and B-cell differentiation (8; 23; 36).

Due to the increasing amount of data in miRNA research, several resources have been established, covering topics such as experimentally validated miRNA targets (Tarbase (208)), and prediction of miRNA targets (TargetScan (163), PITA (127), PicTar (149)) or serving as miRNA repositories (miRBase (82)).

In order to provide a comprehensive overview of differentially regulated miRNA expression data in diseases and general biological processes, we generated the PhenomiR database. We aim at high data quality by manual annotation by experienced biocurators. PhenomiR provides an in-depth annotation of the studies, not only including information like the mode of miRNA expression (up or down) and the miRNA detection method, but also data such as the quantitative fold-change of miRNA expression, the sample size and the origin of the samples (patients or cell culture) analyzed (Figure 3.1), which are not available from any existing resource. This comprehensive repository allows for the first time a large-scale statistical analysis of aspects such as genomic localization of deregulated miRNAs or the influence of sample origin. Using PhenomiR data from cell culture studies and patient studies, we found that, depending on the disease type, independent information from cell culture studies is in conflict with conclusions drawn from patient studies. Furthermore, a systematic analysis of 94 diseases shows for the first time that deregulated miRNA clusters are significantly overrepresented in the majority of investigated diseases (approximately 90%) compared to singular miRNA gene products.

3.2 Results and Discussion

3.2.1 Database contents

In recent years, a wealth of studies published in the scientific literature has investigated deregulation of miRNA expression in diseases and other biological processes. PhenomiR provides a repository that offers all the scattered information about miRNA expression in a structured and uniform format. This allows users to perform individual queries for specific miRNAs and diseases as well as to use the complete dataset for large-scale statistical analyses. All information in PhenomiR is extracted from published experiments and has been manually curated. The literature reference for each

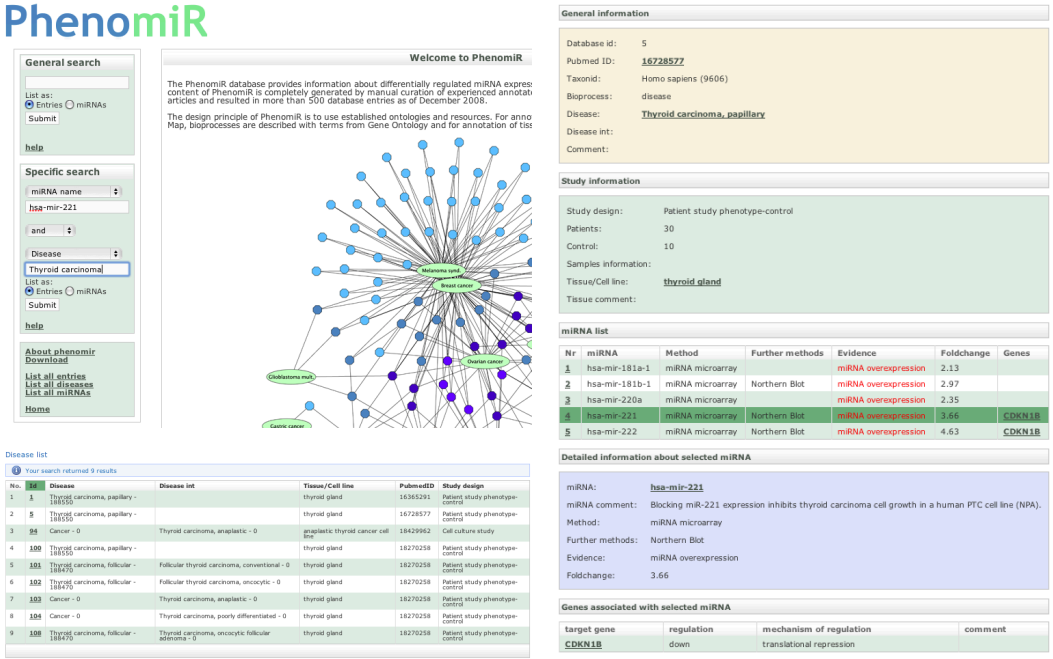

Figure 3.1: **Overview of the PhenomiR web page, the search options, search results and a database entry.**

database entry is annotated as a PubMed identifier and is hyper-linked to PubMed in the web frontend. Each individual entry of the database refers to an instance of a publication describing a specific disease or bioprocess (Figure 3.1). Currently, PhenomiR documents data from 296 articles that describe 542 studies. This dataset includes 11,029 data points, each representing one deregulated miRNA in an experiment.

A design principle of PhenomiR is to use well-established ontologies and resources. As miRBase is the primary resource for miRNA annotation and nomenclature, we use the miRBase identifiers and nomenclature for annotation of miRNAs. In order to enable convenient analysis of the dataset, miRNA designations from previous nomenclature releases were mapped to miRBase release 12.0. For annotation of diseases we use information from the Online Mendelian Inheritance in Man (OMIM) Morbid Map (4). The OMIM Morbid Map is an alphabetical list of diseases described in OMIM, including their corresponding cytogenetic locations. In contrast to disease vocabularies like Disease Ontology (DO) or Medical Subject Heading (MeSH) disease categories, the widely popular OMIM classification scheme contains additional information about the disease, such as clinical features, population genetics and genes that are experimentally shown to be involved in the respective disease. If no appropriate OMIM disease term

is available for the annotation of a disease (currently the case for 20.7% of studies), we introduce additional terms like 'dermatomyositis' and 'thyroid carcinoma, medullary'. In addition to the OMIM terms, PhenomiR annotates Morbid Map-associated higher-level disease classes, such as cancer or cardiovascular, which were introduced by Goh et al. (77). In this system, each annotated disease from the Morbid Map is associated with one of 22 disease classes. miRNA expression analyses of biological processes are predominantly performed for developmental processes and responses to conditions like folate starvation. For the annotation of biological processes we assign terms from Gene Ontology (GO) (7). Cell lines or tissues (Figure 3.2) that were used as samples in the analyses are annotated using the Brenda Tissue Ontology (BTO) (35).

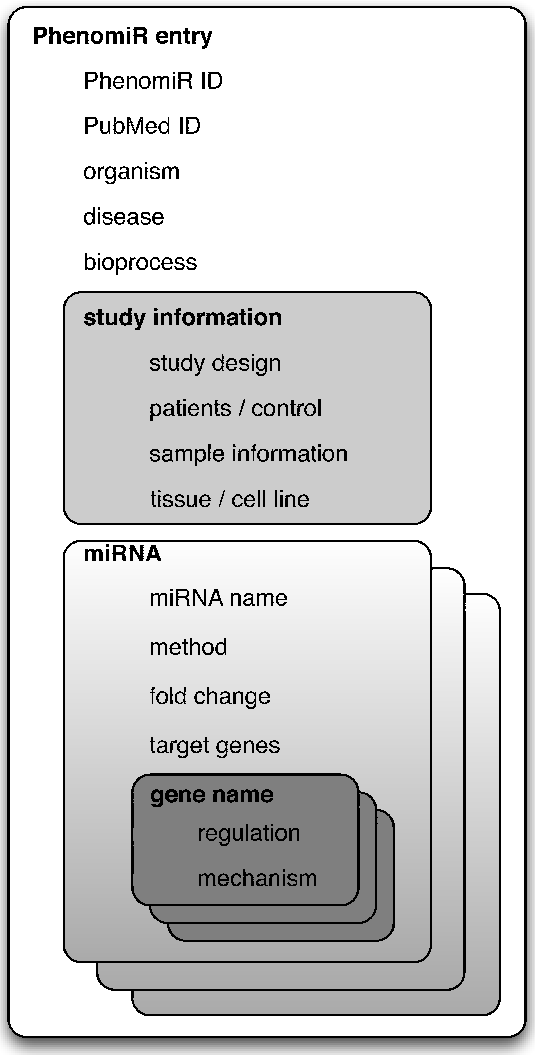

Figure 3.2: **Overview of a PhenomiR entry structure.**

In addition to the sample information, PhenomiR provides the experimental methods used for miRNA expression analyses: to a large extent, expression studies of miRNAs have been performed with microarrays (29%) of all miRNA phenotype correlations. Other methods, such as RT-PCR (47%) and Northern blot (10%), are also used to re-confirm the results for selected miRNAs (Figure 3.3A). Information about differential expression of miRNAs in PhenomiR is given as the qualitative attributes 'miRNA over-expression' or 'miRNA downregulation'. In most articles (75%) authors also publish quantitative results. This information allows discrimination between marginally and significantly deregulated miRNAs. If such information is available, quantitative data (as fold-change) are additionally annotated in PhenomiR. Data content from miRNA expression studies curated in PhenomiR show a high heterogeneity in the amplitude of fold-change and the available measurements. Studies like that of Nikiforova et al. (201) present only few values considered to be significant by the authors, whereas in an analysis of melanoma and neural system tumor syndrome 222 values are presented (298). Accordingly, the extent of maximum miRNA deregulation lies in a range from 1.42-fold in a renal cell carcinoma study (78) to 5,997-fold for acute lymphoblastic leukemia (233). Therefore, we do not set arbitrary thresholds for the numbers of deregulated miRNAs or the fold-changes but present the data as they are provided by the scientists, leaving possible filtering and thresholding or weighting to any later analysis.

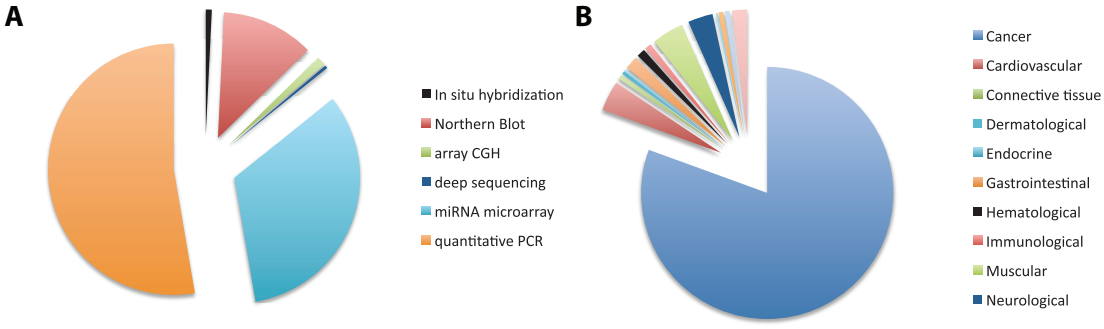

Figure 3.3: **Fraction of annotated miRNA detection methods and diseases in PhenomiR.** (**A**) Distribution of detection methods for all disease entries in PhenomiR. (**B**) Distribution of disease entries in PhenomiR.

In case studies containing analyses of putative target genes from significantly deregulated miRNAs, verified target genes are annotated. The annotation includes gene name and identifier of the target gene, the effect on gene product expression (up- or down-regulation) as well as the mechanism of regulation, for example, transcriptional

repression or translational inhibition (Figure 3.1).

A survey about the PhenomiR dataset reveals that cancers are by far the most thoroughly investigated diseases (81%) followed by muscular (4.3%) and cardiovascular (4.1%) disorders (Figure 3.3B). The largest number of cancer studies is devoted to leukemia (16.7%), colorectal cancer (10.6%) and breast cancer (9.5%).

PhenomiR is the largest publicly available resource of miRNA deregulation in diseases and biological processes, providing 11,029 data points (miR2Disease: 2,663 data points) and 572 miRNAs (miR2Disease: 347 miRNAs) as of September 2009. Out of 542 PhenomiR entries, 90 provide information about miRNA expression in biological processes such as cardiac muscle development or eye development, which are not available from any other existing database. In comparison to resources like miR2Disease (111) and HMDD (173), PhenomiR provides comprehensive experiment information such as fold-change of miRNA dysregulation, cohort information and study design. Moreover, we particularly focused on the thorough use of ontologies, which are invaluable for in-depth statistical analysis and further exploitation of the data as shown in the analyses below. Especially in publications presenting all data on deregulated miRNAs, fold-change information allows a threshold to be set in order to separate marginally from significantly deregulated miRNAs. Cohort information specifies the number of patients analyzed in a study and thus determines the statistical significance of the data. Data from cell culture studies and patient studies are identified by the study design information. Without this information the first data analysis shown below would not have been possible. Finally, to our knowledge, manually annotated data about differential miRNA regulation in bioprocesses are not found in any other publicly available database. PhenomiR is freely accessible and the data can be downloaded as tab-delimited text files. New content releases for PhenomiR will appear every half year.

3.2.2 Search options and predefined datasets

In order to obtain an overview of the PhenomiR dataset, the web page links to three lists that display: all entries; all diseases; and all annotated miRNAs (Figure 3.1). In addition, statistical information about the number of database entries, most frequently annotated miRNAs, and so on are provided on the front page. Currently, 567 different miRNAs were found to be deregulated in at least one entry.

For queries, PhenomiR offers two search options, a 'General search' as well as a

'Specific search' (Figure 3.1). The 'General search' performs simultaneous queries across several attributes like 'miRNA name', 'disease' or 'gene name'. This is optimized for searches where comprehensiveness rather than specificity is required. The results can be displayed either as respective entries or associated miRNAs. The 'Specific search' allows the selection of individual annotated attributes shown in a pull-down menu. Additionally, specific searches can be combined by using the logic operators AND, OR and NOT. As in the 'General search', results can be displayed as a list of database entries. Another way to depict the results is to generate a list of all miRNAs found in any of the corresponding studies. Results of both search options are linked to the respective entries.

To demonstrate the additional value of the comprehensive annotation in PhenomiR, we investigated the influence of differentially regulated genomic miRNAs on diseases from a large-scale statistical point of view.

3.2.3 Differences between disease-associated microRNA expression in patients and cell lines

Cell lines have been established in life sciences as easy to manipulate model systems for the study of cellular processes. However, studies using both *in vitro* and *in vivo* systems have shown that the results from each - for example, in cancer - do not always correlate. In previous studies differences in gene expression patterns between cell lines and their fresh-frozen tissue counterparts have been observed (124). Accordingly, analysis of DNA copy number alterations between cell lines and fresh tissue revealed recurring deviations in cell lines (80). Cell cultures are also frequently used to investigate differential miRNA expression in cellular systems. In PhenomiR, we have collected 119 *in vitro* studies of miRNA expression in various diseases revealing implications for the prognosis of diseases. With respect to the discrepancies between cell cultures and living organisms mentioned above, we asked whether cell cultures are reliable disease models for the analysis of differential miRNA expression.

In order to analyze the concordance of *in vivo* and *in vitro* data, we extracted disease information from PhenomiR with sufficient miRNA annotation for both study designs. We first compared the consistency of miRNA annotation within each disease for both *in vivo* and *in vitro* experiments. This was done by means of an intra-consistency score, defined as the fraction of miRNAs showing a concordant expression pattern within a disease annotated by at least two experiments. In a second step, we computed the

cross-consistency score between *in vivo* and *in vitro* data as the fraction of miRNAs showing the same expression pattern between these two study designs. Figure 3.4 shows the obtained consistency scores for 15 diseases that had sufficient data coverage in PhenomiR. Only 6 out of 15 diseases (glioblastoma, ovarian cancer, hepatocellular carcinoma, colorectal cancer, gastric cancer and chronic myeloid leukemia) show the expected high cross-consistency (73 to 100%) between *in vivo* and *in vitro* experiments. On the other hand, we found six diseases (pancreatic cancer, non-Hodgkin lymphoma, neural system tumor, lung cancer, breast cancer and prostate cancer) with only moderate cross-consistency scores (51 to 61%). Analyzing the corresponding *in vivo* and *in vitro* data, we obtained high intra-consistency scores, which indicate a high homogeneity within these experiments. However, the resulting cross-consistency scores are rather low, implying limited relevance of *in vitro* experiments for those diseases. Finally, we also found three diseases (squamous cell carcinoma, (AML) and cervical cancer) with low cross-consistency (24 to 38%), revealing severe discrepancies between cell culture experiments and patient studies. High intra-consistency scores corroborate the significance of this observation and exclude the possibility that the results stem from different experimental conditions.

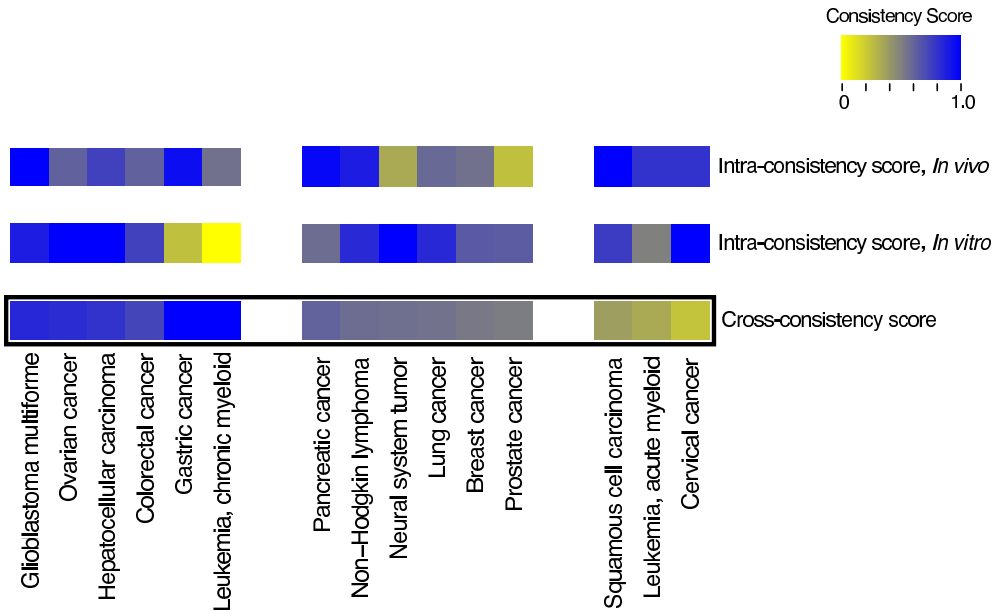

Figure 3.4: **Comparison of consistencies in expression profiles between *in vivo* and *in vitro* experiments.**

These findings could possibly arise as an artifact of the selection of cell cultures or subtypes of diseases investigated in the published studies. Indeed, this is not the case,

as can be seen from the example of AML. AML is, according to the (FAB) classification system, divided into eight subtypes, M0 through to M7, based on the type of cell from which the leukemia developed and its degree of maturity. All AML cell culture studies have been performed with NB-4 cells and HL-60 cells, which are both from the M3 subtype (promyelocytic or acute promyelocytic leukemia). In contrast, patient studies from different AML subtypes are annotated in PhenomiR. For our analysis, data from the respective patient studies has been pooled in order to obtain a statistically sufficient amount of data. However, comparison of cell culture studies with data from Saumet et al. (228), which also analyzed patients with AML from the M3 subtype, confirm that our findings hold true for similar AML subtypes. In cervical cancer the two patient studies did not classify specimen according to the World Health Organization classification or the Bethesda System. However, the three different cell lines that were used for *in vitro* studies exhibited consistent results (100%). In conclusion, these two examples show that our findings are not distorted by the origin of the samples that were used in the studies.

Recent studies have shown that miRNA expression profiles have a high prognostic potential in disease classification and that it is even possible to build decision trees in order to differentiate cancer tissue origins (223; 277). However, our large-scale analysis including data from more than 413 surveys has shown that data from *in vitro* and *in vivo* studies correlate for diseases like pancreatic cancer and ovarian cancer, but display significant inconsistencies in squamous cell carcinoma and cervical cancer. Discrepancies between experimental results from organisms and cell cultures could occur for two reasons. Most notably, the cell line immortalization process has been implicated as a source of cytogenetic changes (112). In addition, multiple growth passages, to which commercially available cell lines are routinely subjected, have been shown to be associated with random genomic instability (190). These observations and the results of our study show that the potential of cell cultures in the investigation of miRNA expression in diseases is limited. As a consequence, the suitability of cell cultures has to be verified for each disease and cell line before using such data as a tool for the prognosis of diseases in human.

3.2.4 MicroRNA clusters are significantly overrepresented in most investigated diseases

While creating the PhenomiR database we found individual studies that investigated the impact of not only deregulated single miRNAs but also miRNA clusters, such as miR-17-92, miR-106b-25 and miR-222-221, on diseases, especially cancer (131; 191). Using the comprehensive dataset from our PhenomiR database we asked whether the impact of miRNA clusters on diseases is restricted to only a few examples or whether miRNA clusters significantly correlate with the pathobiology of diseases. According to release 12.0 of miRBase, 695 miRNA genes have been detected in the human genome so far. Analysis of the genomic distribution of miRNAs shows that it is strongly biased towards neighborhoods on chromosomes. Given a maximum distance of 5 kb, about 34% of human miRNAs appear as miRNA clusters of at least two members, leading to 62 miRNA clusters. Microarray profiling of miRNAs has shown that neighboring miRNAs within a distance of up to 50 kb are frequently co-expressed (11). It can be assumed that miRNA clusters are not only often jointly expressed but also act in a concerted way on interrelated cellular functions (131).

First, we systematically analyzed the homogeneity of expression patterns within miRNA clusters. In order to determine the concordance of expression, we excluded those clusters from further analysis having less than half of all miRNAs annotated in PhenomiR, leading to 47 remaining clusters. The clusters are denoted as exhibiting a homogeneous expression pattern if all annotated miRNAs are either up- or downregulated. In total, disease-associated clusters revealed homogeneous expression patterns for 77% of all annotated diseases, which confirms the hypothesis of co-expression of miRNA clusters. For example, cluster mir-221-222 shows a consistent expression pattern in 93% of the diseases.

As some of the investigated diseases show an extremely unidirectional expression pattern - that is, almost all annotated miRNAs are either upregulated or downregulated - we might find homogeneous patterns even by chance. In order to take this effect into account, we created a null model by randomly linking miRNA expression patterns (10,000 times within each disease). In total, 23 clusters (50%, P-value < 0.05) showed a significantly higher homogeneity pattern in all annotated diseases compared to that expected by chance. These clusters exhibit a homogeneous expression pattern in at least 87% of all annotated diseases.

To investigate the association of miRNA clusters with human diseases, we estimated

the enrichment of miRNA clusters in disease-associated miRNAs. Analysis from articles restricted to only a few miRNAs could introduce an overestimation of disease association with miRNA clusters. In order to avoid this bias, we chose only data from patient studies using miRNA microarrays, since microarrays are standardized tools that aim to cover the most comprehensive dataset of known miRNAs. For the estimation of cluster enrichment we used a (LOD): we calculated the fraction of disease-associated miRNAs within a cluster for each disease and divided this number by the background frequency of 34% (Figure 3.5). We found enrichment for 46 out of 52 (88.5%) diseases (P-value = 6.1^{-3}). Within these 46 diseases, miRNAs located in clusters are, on average, 1.4 times (LOD = 0.58) enriched compared to random. However, it may be argued that polycistronic miRNA loci are more likely associated with diseases because multiple combinations of miRNAs could possibly generate a phenotype, that is, only one miRNA of a cluster may act as the causative while the others act as 'bystanders'. In order to address this question, we considered miRNA clusters as single loci and calculated the enrichment of polycistronic miRNA loci for each disease (see Materials and methods). We found that polycistronic loci are, on average, 3.5 times more disease-associated than expected by chance, whereas differentially expressed single miRNA loci are not enriched in diseases. In conclusion, both analyses show a significant enrichment of clustered miRNAs in diseases regardless of whether the single miRNA members are used for the analysis or the cluster is viewed as one locus. Thus, it is highly unlikely that only one miRNA of a cluster is associated with a disease. Indeed, experimental analyses show that different miRNAs from miRNA clusters act synergistically (131; 141). miRNA cluster members are, on average, 1.4 times (LOD = 0.58) enriched compared to random. Green points depict enriched diseases (at least LOD = 0.15 for lung cancer). Black points indicate diseases without enrichment compared to random and red points depict disorders with few deregulated cluster members. These results show that deregulation of miRNA clusters in diseases is obtained not just in a few examples but appears to occur systematically in the vast majority of human diseases investigated in our analysis. Although studies of miRNA expression in diseases are dominated by various types of cancer (Figure 3.3), comparable results are found if cancer and non-cancer diseases are examined separately. The lower deregulation of single miRNAs is probably due to the fact that miRNAs do not act as binary off-switches but rather modulate protein expression (8). For instance, the response to altered miR-223 expression on the human proteome indicates that, for most interactions, miRNAs act as rheostats to make fine-scale adjustments to protein

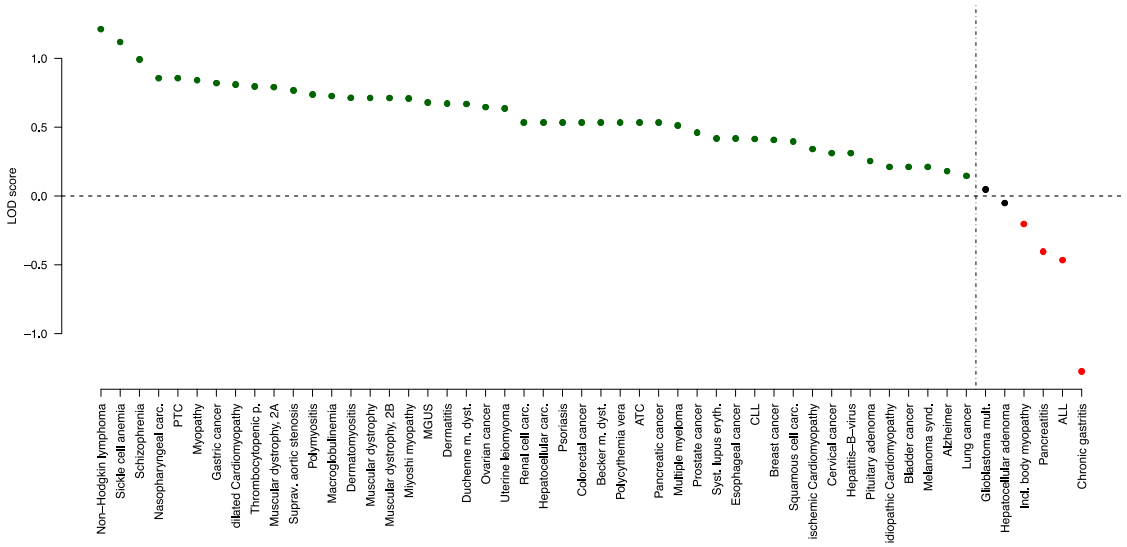

Figure 3.5: **miRNA cluster enrichment in human diseases.** For each disease the log-odds (LOD) score is plotted. There is an enrichment of miRNA cluster members for 46 diseases (88.5%).

output (8). As shown above, deregulation of miRNA clusters affects expression of several miRNAs at the same time. A concerted action of several miRNAs on a common target or pathway has a much higher potential to influence cellular processes. In fact, for several specific clusters such a synergistic regulatory effect has been shown. Overexpression of miR-200 miRNA clusters in NMuMG cells hindered epithelial-mesenchymal transition by enhancing E-cadherin expression through direct targeting of ZEB1 and ZEB2, which encode transcriptional repressors of E-cadherin (141). In gastric cancer, two miRNA clusters, miR-106b-93-25 and miR-222-221, were found to suppress different cell-cycle inhibitors (191). Such results are further corroborated by *in silico* analyses of target genes of members of miRNA clusters (290). These experimental findings and the systematic correlation of miRNA cluster deregulation with human disease shown here strongly support the idea that a coordinated regulatory effect is a general attribute of miRNA clusters. The pivotal role of miRNA clusters in miRNA-based gene silencing found in human diseases suggests that effective treatment of various diseases may require a combinatorial approach to target not singular miRNAs but rather miRNA clusters.

3.3 Materials and Methods

3.3.1 Comparison between *in vivo* and *in vitro* experiments

To evaluate the expression consistency of miRNAs *in vivo*, we first allocated all *in vivo* expression profiles in PhenomiR to the corresponding diseases. Within each disease, these entries were grouped by miRNAs and all groups containing less than two entries were discarded. Subsequently, we checked for consistent expression profiles - that is, all entries for a specific miRNA must show the same expression (either down- or up-regulation) to be counted as consistent. The intra-consistency score for*in vivo* or *in vitro* experiments is defined for every disorder to be the fraction of miRNAs with consistent expression patterns throughout all allocated entries of a group and all miRNAs involved in the disease. For the estimation of the cross-consistency score we grouped all miRNAs with consistent expression profiles in both study designs (*in vivo* and *in vitro*) for that specific disease. Additionally, we added those miRNAs to the groups that contained only one entry in the *in vivo* and *in vitro* experiments, respectively. The cross-consistency score for comparison of *in vivo* and *in vitro* experiments was then calculated as the fraction of miRNAs showing a consistent expression pattern compared to the total number of miRNAs for each disease.

3.3.2 Human microRNA cluster data

A miRNA cluster is defined as set of miRNAs in which each member has at least one other member of the same set within 5 kb according to chromosomal locations. Chromosomal positions for all human miRNAs were obtained from miRBase (release 12.0). In total, we obtained 62 human miRNA clusters containing, in sum, 240 of 695 (34%) human miRNAs in miRBase.

3.3.3 Analysis of homogeneous expression patterns within microRNA clusters

For the systematic analysis of coexpression of miRNA clusters, we considered all miRNAs associated with the particular diseases. miRNAs not belonging to any cluster and miRNAs of clusters of which at least half the members are not associated with the appropriate disease were discarded. For clusters containing only two members, both

miRNAs had to be present. In total, we obtained 47 unique clusters. We defined a cluster to be homogeneous if all present members (which is at least half of all members) show the same expression pattern (either all up- or all downregulated). For each unique cluster we thereafter computed the homogeneous-fraction, that is, the fraction of co-expression throughout all obtained disease entries, and calculated a P-value for this fraction by the following sampling approach: for every disease entry the expression of all its associated miRNAs was distributed randomly within these miRNAs for 10,000 times, keeping the distribution of up- and downregulated miRNAs constant for each step. For each sampling step the homogeneous fraction over all disease entries was computed, which yields the P-value as the number of sampled homogeneous fractions exceeding the original homogeneous fractions divided by 10,000.

3.3.4 Enrichment analysis of microRNA clusters in human diseases

For this part of the study only data from microarray experiments were taken into account in order to avoid a bias introduced by expression experiments investigating only a few miRNAs by, for example, RT-PCR. To measure the enrichment of cluster miRNAs compared to single miRNAs in human diseases, we set up a sampling algorithm based on log-odds (LOD) scores: for each disease, d, we calculated the number of cluster miRNAs, x_d, and the number of non-cluster miRNAs, y_d. The LOD score for disease d is then computed by:

$$LOD_d = log_2 \frac{x_d/(x_d + y_d)}{x_{overall}/(x_{overall} + y_{overall})} \tag{3.1}$$

where $x_{overall}$ denotes the number of the 240 human cluster miRNAs and $y_{overall}$ denotes the number of the 455 human miRNAs not contained in any cluster as obtained from miRBase (release 12.0). Note that $x_{overall}$ and $y_{overall}$ take into account all known human miRNAs, not just those annotated in PhenomiR. It can be easily seen that the LOD score for the enrichment of miRNAs not contained in any cluster computes to -LOD_d, where d is again the disease index. A positive LOD score indicates enrichment for cluster miRNAs compared to non-cluster miRNAs in a specific disease. For evaluation of the hypothesis of enrichment of cluster miRNAs throughout all human diseases we randomly shuffled the genomic position of all miRNAs in each disease 100,000 times and computed the fraction of cases where the number of sampled positive LOD scores was at least as high as the number of positive LOD scores obtained

from the data. In addition, we considered miRNA clusters as single loci and calculated the enrichment of polycistronic miRNA loci by a LOD score: for each disease d we calculated the number of polycistronic miRNA loci, x_d, and the number of single miRNAs, y_d. $x_{overall}$ denotes the number of the 62 human polycistronic miRNA loci and $y_{overall}$ denotes the number of the 455 human miRNAs not contained in any cluster as obtained from miRBase (Release 12.0).

3.4 Conclusions and Outlook

Within this study, we focus on the differences in miRNA expression pattern between cell cultures and *in vivo* organisms. Cell lines are well suited systems for studying all kinds of cellular processes. However, studies using both *in vitro* and *in vivo* systems have shown that results, especially in cancer, do not correlate. Comparing data from more than three hundred studies, we show that depending on the disease type, integration of independent data from cell culture studies is in conflict to conclusions drawn from patient studies. Therefore, our result provides evidence that cell culture studies are not well suited to investigate the expression profiles of disease-associated miRNAs. In addition, we show in a systematic analysis that deregulated miRNA clusters are significantly overrepresented in the majority of investigated diseases, compared to singular miRNA gene products. The systematic correlation of miRNA cluster deregulation on human disease strongly supports the pivotal role of miRNA clusters in human diseases as effective treatment of various diseases. In the next chapter, we further study the regulatory role of disease-associated miRNAs. Therefore, we link miRNA expression data obtained from patient studies to signaling pathways. We set up a multipartite graph consisting of miRNAs, proteins, diseases, and signaling pathways. Using this framework, we analyze first in a global and later on in a more local manner the regulatory motifs of miRNA-mediated regulation in signal transduction pathways.

4 The role of disease-associated microRNAs in signaling pathways

4.1 Background

While there is evidence (48; 202; 259) that miRNA expression and maturation is induced by signaling pathways, miRNAs also emerge as regulators of signaling proteins. In zebrafish, miR-9 has been shown to regulate several components of the FGF signaling pathway, and thus controls neurogenesis in the midbrain-hindbrain domain during late embryonic development (160). In another recent example in fruit fly (125), miR-8 has been identified to target both a transmembrane protein and a transcription factor of the WNT signaling pathway. Ricarte-Filho et al. (220) showed that the RET-pathway is mediated by let-7, which inhibits the activation of the RET/PTC-RAS-BRAF-ERK cascade exemplifying the direct influence of a single miRNA on a submodule of a signaling pathway. Given the generally large number of miRNA targets (162) it is natural to assume that many miRNAs regulate not only a single important pathway protein, but rather coordinate protein levels on a pathway-wide scale. Altered miRNA levels might then result in inaccurate target protein levels, consequently fallacious signal transduction, and potentially a disease phenotype.

From this perspective, it is intriguing to observe that medical sciences increasingly focus on the impact of miRNA-mediated regulatory control on diseases, especially in cancer: miRNAs are intensively used as diagnostic and prognostic disease markers (27), and even appear in first clinical trials (42). Given the linkages between signaling pathways and miRNA regulation on the one hand, and miRNAs and disease phenotypes on the other, we aim to unveil the connection between phenotypes, pathways and miRNA-mediated regulation.

Here, we analyzed the tissue-specific regulatory patterns of disease-associated miR-

NAs in signaling pathways on different scales. Globally, we investigated the enrichment of disease-associated miRNAs on different pathways, and more locally, on the cellular location and process type of target proteins. We used manually annotated data from hundreds of patient studies to estimate the impact of disease-associated miRNAs on signaling pathways. We identified a core set of pathways, homogeneously enriched throughout nearly all diseases. Most of these pathways have been associated with cell growth, proliferation, and apoptosis. However, deregulation of signaling pathways can be induced by diverse factors. Point mutation of central signaling cascade proteins (115) have a severe impact on the information flow as well as any change in the expression pattern of *cis* or *trans* regulators. We thus compared the cellular localization and process type of signaling proteins that are miRNA targets with proteins that have been identified as disease-associated. In the following, we show that in contrast to disease proteins, miRNA targets are significantly enriched as inhibitors within the nucleus.

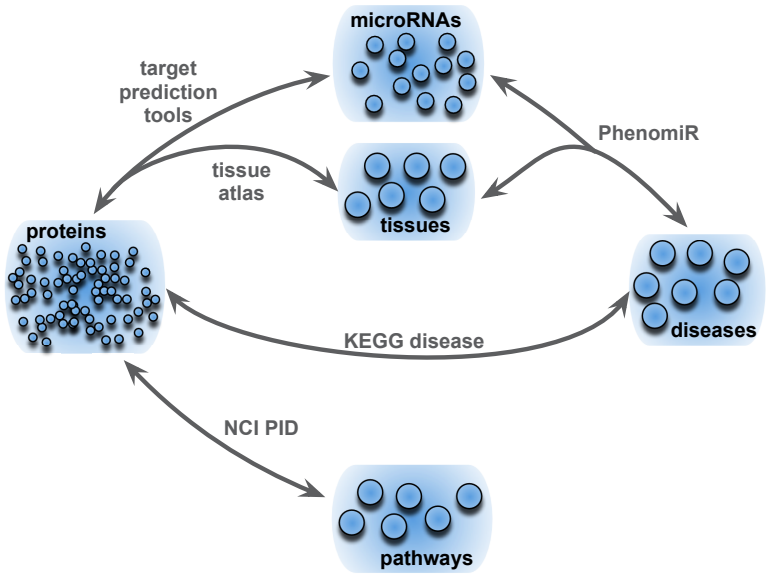

Figure 4.1: **Illustration of the interactions between diseases, tissue, annotated disease-associated miRNAs, proteins, and human signaling pathways** The multipartite graphs consists of five sets of nodes and links between them, established by different data resources: 165 miRNAs from the PhenomiR database with annotated deregulation in 63 diseases, 4907 target transcripts, predicted by TargetScanS and filtered by the tissue atlas, 79 signaling pathways with constitutive proteins as given by the NCI PID database, and finally the subset of disease proteins as provided by the KEGG DISEASE database.

4.2 Results

We captured the different entities of our investigation in a multipartite graph. The graph consists of five sets of nodes representing the entities miRNAs, proteins, tissue, diseases, and pathways and links between but not within the set of nodes. Links are given by a prediction tool and four databases. miRNAs (as provided by the miRBase database (81)) are linked to diseases and corresponding tissue via the PhenomiR database (226), a manually curated database containing disease-associated miRNAs in human disorders. miRNA target transcripts are determined by TargetScanS (162) a prediction tool that shows a high performance on different miRNA target data sets (91). In addition, we used the tissue atlas provided by Su et al. (257) to filter potential miRNA targets for a specific disease and a given tissue. We unified the set of mRNA transcripts and corresponding proteins to a set of nodes denoted simply as proteins. This set is linked to signaling pathways via the National Cancer Institute Pathway Interaction Database (NCI PID) (230), containing 79 human pathways together with its constituting components. Finally, disease proteins are identified by their Kyoto Encyclopedia of Genes and Genomes (KEGG) DISEASE annotation (121) (see Methods for a detailed description of the materials used). Figure 4.1 summarizes the entities and connections used. Notably, similar results were obtained with other miRNA prediction tools and a different set of disease genes, as provided by OMIM (90).

4.2.1 MicroRNAs induce a core set of signaling pathways across diseases and tissues

We first analyzed the connection between diseases and signaling pathways, mediated by disease-associated miRNAs. In order to project the properties of the multipartite graph onto a disease-pathway correlation, we calculated the enrichment of disease-associated miRNA targets in a particular pathway. We used the tissue annotation in PhenomiR to filter for expressed miRNA targets, as given by the tissue atlas of Su et al. (257). For a particular disease and a specific pathway, we computed the log odds ratio (LOD score) by dividing the relative number of associated miRNA targets in this pathway and tissue with the expected number, based on the relative number of associated miRNA targets in all signaling pathways given a specific tissue. Disease-pathway interactions with no targets (white fields in the heatmap Figure 4.2A) were excluded from further analyses (see Methods for a detailed description). We obtained

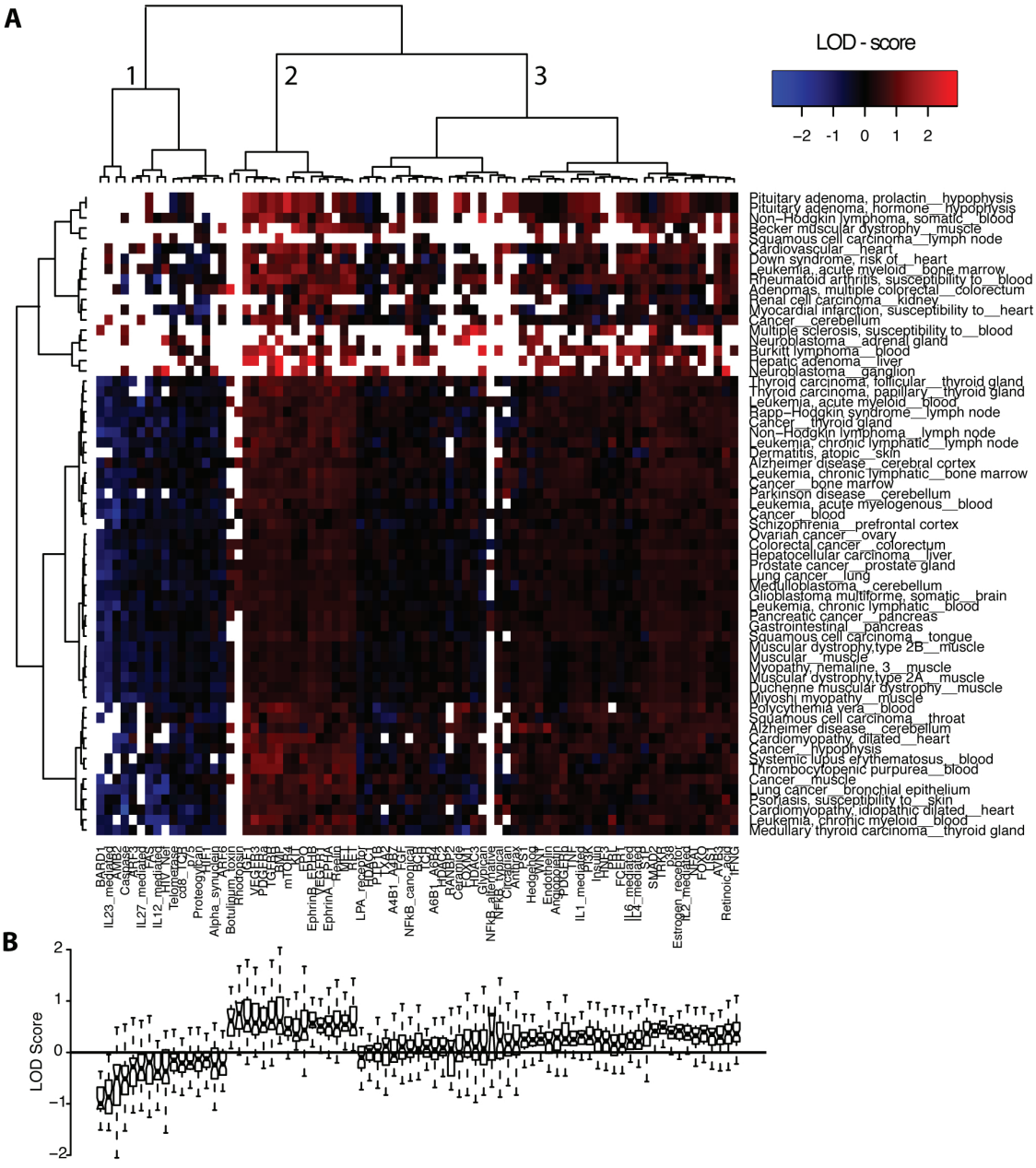

Figure 4.2: **Impact of disease-associated miRNAs on signaling pathways** Enrichment for a particular disease and pathway was calculated by a LOD score. A positive score indicates an enrichment of miRNA targets for a disease-pathway interaction. Negative scores indicate depletion. **A**: Heatmap of miRNA target enrichment for a particular disease and pathway. Pathways and diseases are ordered by hierarchical clustering using Manhattan distance and ward clustering. **B**: Boxplot of disease-pathway associations ordered according to hierarchical clustering along the pathways. Red fields indicate an enrichments and blue a depletion. White fields indicate that no miRNA targets were found for this disease-pathway association.

a matrix of LOD scores, where each entry indicates the enrichment or depletion of tissue-specific targets of disease-associated miRNAs in the respective signaling pathway. We ordered this matrix according to a hierarchical clustering along the disease axis and pathway axis, respectively. Two features of the resulting heatmap are remarkable: First, dividing the hierarchical clustering of the signaling pathways into 3 major sub-clusters, we found one cluster (cluster 2; mean LOD = 0.55 , variance = 0.008) showing a high enrichment throughout all diseases (see Figure 4.2A). We define this cluster as the core set of signaling pathways highly enriched with disease-associated miRNA targets. The remaining clusters show a high variance (cluster 3; mean LOD = 0.21, variance = 0.02) and a common depletion of miRNA targets (cluster 1; mean LOD = -0.36, variance = 0.07). Second, the 63 diseases split into two clusters with high and low miRNA-pathway associations. Within the larger of the two clusters, the enrichment of miRNA targets is homogenous. Moreover we performed a multi-scale bootstrap resampling approach (relative sample sizes of bootstrap replication of 20%) (245) to test whether clusters 1 - 3 are robust against variation in the data. We can reject the hypothesis that the clusters do not exist with a significance level $\alpha < 0.05$ indicating that the clusters 1 - 3 may stably be observed by increasing the number of observations. All signaling pathways located in the core set are given in Table 4.1. The functions of these pathways reflect the affinity of miRNAs to regulate cellular processes associated with apoptosis, proliferation or development, as we will outline with three examples. (*i*) The PDGFa pathway, for example, promotes cell migration, proliferation, and survival (94; 165; 224; 286). PDGF expression has been demonstrated in a number of different solid tumors, from glioblastomas to prostate carcinomas. Its biological function varies from autocrine stimulation of cell growth to subtler paracrine interactions involving adjacent stroma or vasculature (75). (*ii*) It was recently reported that let-7 has an influence on the RET-pathway by effecting the cell growth and differentiation of papillary thyroid cancer (220). Ricarte-Filho et al. (220) concluded that let-7 inhibited the activation of the RET/PTC-RAS-BRAF-ERK cascade exemplifying the direct influence of a single miRNA on a submodule of a signaling pathway. (*iii*) The Reelin pathway has been directly correlated with tumor aggressiveness (214; 236; 279). Evangelisti et al. (59) linked this pathway for the first time to cancer by showing the inhibition of Reelin by miR-124a.

The pathways with the highest negative enrichments based on disease-associated miRNA targets, are the IL-23 mediated pathway (playing a pivotal role in autoimmunity (188)) and BRAD1, which is associated with cell survival and cell death (107).

Pathway	Median LOD	miRNA	Z-score$_{Targets}$	Z-score$_{Pathways}$
Rhodopsin	0.76	miR-154	8.69	6.58
Botulinum	0.61	miR-29b	8.58	8.10
TGFBR	0.61	miR-216a	12.20	7.10
BMP	0.60	miR-224	9.37	7.93
IGF1	0.59	miR-375	9.39	8.12
VEGFR3	0.57	miR-422a	8.29	7.89
EphrinB/EPHB	0.57	miR-422a	11.44	8.06
PDGFa	0.56	miR-383	7.20	7.59
MET	0.55	miR-422a	10.96	7.61
EphrinA/EPHA	0.53	miR-136	8.31	8.15
RET	0.52	miR-422a	9.04	7.24
VEGFR1	0.51	miR-422a	11.72	7.82
REELIN	0.51	miR-197	7.76	6.86
TRKR	0.49	miR-335	12.94	7.88
mTOR4	0.47	miR-375	7.23	7.44
EPO	0.43	miR-134	6.75	8.00

Table 4.1: **Core set of signaling pathways with highly enriched miRNA targets.** The Median LOD score is calculated over all diseases for a particular pathway. miRNA is the most enriched single miRNA within the corresponding pathway. Z-score$_{Targets}$ was calculated by comparing the median LOD score with the obtained score by a random sampling of miRNA targets. Z-score$_{Pathway}$ was calculated by comparing the median LOD score with the obtained score by a random sampling of pathway proteins.

Although we found a core set of pathways across diseases, differences between disorders can arise due to different expression levels of the respective miRNAs. The PDGFa pathway for example shows high enrichments across diseases independent of the miRNA prediction tool. We found miR-144 to be highly enriched in the PDGFa pathway. Analyzing the expression profile, we found miR-144 down-regulated in cancer, but up-regulated in Parkinson disease and idiopathic Myelofibrosis. Predicted targets of miR-144 are SRF, a transcription factor activated by PDGFa, and FOS that is thought to have an important role in signal transduction, cell proliferation and differentiation (21; 25; 203). This finding shows that although different diseases are associated with the same signaling pathway, differences in the effects of the stimulated pathways can be induced by complementary expression profiles of miRNAs.

As the PhenomiR data set is dominated by cancer-related diseases (60%), we divided the set of diseases into a subset of cancer and non-cancer related miRNAs to

study differences between both groups. We found 14 out of 16 pathways of the global core set also in the cancer-specific core set. The core set for the non-cancer related pathways contains 12 pathways that were also found by the global data set, but we also identify also two non-cancer specific pathway enrichments such as the KIT pathway and the NFκB pathway, that is involved in the expression of genes associated with development, cell death, and immune response (19; 38; 76; 213).

4.2.2 Robustness analysis of the core set of signaling pathways

In order to ensure that our results are not artifacts of the chosen prediction tool, we analyzed the data with four other prediction tools: PicTar (149), Miranda (114), TargetSpy (255), and RNA22 (194). Different features like conservation of the seed region or binding energies are taken into account to predict miRNA-transcript interactions in each tool. Based on these differences the overlap between the target sets from different tools is generally rather low (238). We define for each tool the core set of signaling pathways, which are highly enriched by miRNA targets and compare these list with our core set listed in Table 4.1. The result shows that the signaling pathways in our core set are mostly consistent with different prediction tools. We found 8 out of 16 pathways within the core set of at least 3 different prediction tools.

In order to test the significance of these pathways, we performed a randomization approach, by comparing the median LOD score of these pathways with the median scores obtained by two random samplings. We first sampled 10.000 times pathway proteins keeping the pathway size constant, second, we generated 10.000 times a random miRNA predictor by sampling for each miRNA the corresponding targets. Finally, we calculated a z-score to estimate the significance of each pathway within the core set. We obtained high z-scores for the pathways within the core set independent of the sampling approach (see Table 4.1). The mean z-score for all pathways is 12.51 (Z-score$_{Targets}$) and 7.65 (Z-score$_{Pathways}$), respectively.

The enrichment of miRNA targets is summarized in the boxplot in Figure 4.2B, where the distribution of LOD scores for each pathway is shown. The median LOD scores and their variance for the set of signaling pathways are significantly negatively correlated (Pearson correlation coefficient C$_P$ = -0.37, $p = 7 \cdot 10^{-3}$). In contrast to depleted pathways, highly enriched pathways are homogeneously targeted by miRNAs across diseases. This indicates that disease-associated miRNAs in human disorders target a core set of signaling pathways irrespective of the specific disease and tissue.

We ensure that the LOD scores are not trivially biased by the pathway size (C_P = -0.032, $p = 0.83$). We noticed that the pathway enrichment is significantly negatively correlated with the number of miRNAs with targets in this pathway (C_P = -0.31, $p = 0.0010$), with up to 159 targeting miRNAs in the SMAD2 pathway.

4.2.3 Interaction of disease-associated proteins and microRNA targets

Much effort has been invested in understanding the mechanisms underlying the complex network of factors contributing to human diseases. Databases like OMIM (90), KEGG DISEASE (121), or HGMD (253) link dysfunctional proteins and genetic mutations to human disorders. In order to focus on already confirmed gene-disease interactions, we used the KEGG DISEASE database to study similarities and differences to miRNA targets in signaling pathways. In the following, we analyzed 23 diseases that are both annotated in KEGG DISEASE and PhenomiR (see Methods). In this subset, we analyzed 365 KEGG DISEASE proteins located in the NCI PID signaling pathways and identified 123 (33.7 %) proteins as miRNA targets. The current estimation for the amount of miRNA targets in the human genome lies between 30 - 35% (162; 68). This implies that there is no higher rate of miRNA targets in the set of disease proteins than expected. In order to study the interplay of disease proteins and miRNA targets, we compared their mapping to NCI PID pathways (see Figure 4.1). We found that typically, disease-affected proteins are widely distributed over pathways for a particular disease. Focusing on pathways showing a high fraction of disease-associated proteins, we found no correlation of miRNA target enrichment and the fraction of disease-affected signaling proteins. These findings imply that disease-affected proteins and disease-associated miRNA targets do not prefer a common set of signaling pathways. To elucidate those differences, we changed the scale of our investigation and compare the localization and process type of disease-associated miRNA targets and disease proteins.

4.2.4 MicroRNA targets are preferentially located in the nucleus in contrast to disease proteins

To question whether miRNA targets and KEGG DISEASE proteins differ with respect to their cellular location and process type annotation, we divided the set of signaling

proteins according to their NCI PID annotation into four groups: extracellular region, cell membrane, intracellular region, and nucleus. We then estimated the fraction of miRNA targets as well as disease proteins for each group and calculated the LOD enrichment scores (see Methods for a detailed description). Surprisingly, we found opposing patterns of cellular localization for disease-associated proteins and miRNA targets (see Figure 4.3A). Deregulated miRNAs preferentially target nuclear proteins (LOD = 0.57, $p = 0.020$), while disease-associated proteins in the nucleus are underrepresented (LOD = -0.41, $p = 0.032$). Therefore, miRNA targets are almost twice more frequently located in the nucleus as compared to disease proteins. Furthermore, proteins located in extracellular region are only weakly controlled (LOD = -0.81, $p = 4.9 \cdot 10^{-3}$) by miRNAs. Disease associated proteins showing again a complementary result compared to miRNA targets (LOD = 0.44, $p = 0.068$), being more than twice more frequently located in the extracellular region. Proteins located in the cell membrane or intracellular region show no significant differences and enrichments

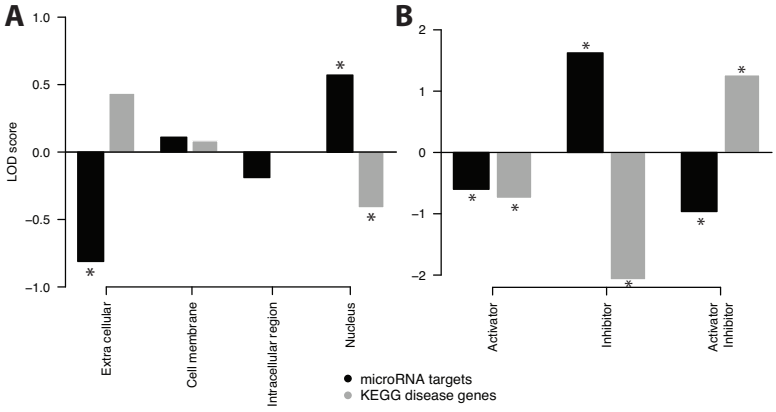

Figure 4.3: **Analysis of cellular location and process type distribution for miRNA targets and disease proteins A**: Signaling proteins are divided into four different cellular location groups (extracellular region, cell membrane, intracellular region, and nucleus) based on their NCI PID annotation. We calculated the enrichment of miRNA targets and disease proteins by a LOD score. We found an opposing patterns of cellular localization for disease-associated proteins and miRNA targets. **B**: Process type information obtained by the NCI PID database was used to divide signaling proteins into three different groups, activators, inhibitors, and ambivalent proteins (annotated as both activators and inhibitors). The result indicates again complementary patterns for miRNA targets and human disease proteins. * indicates significant enrichment obtained by Fisher's exact test ($p = 0.05$).

for miRNAs or disease-associations. Comparing these results with the subset of cancer-related miRNAs we obtained the similar finding of a preferred target location in the nucleus. This result shows that preferred location is not based on a disease-specific set but a common pattern, valid for cancer as well as non-cancer related miRNAs. We repeated the location analysis with different prediction tools and obtained similar results for miRNA targets. Analyzing miRNA targets located in the nucleus by Gene Ontology, we found 50% of those genes involved in transcriptional regulation. In addition, we used the OMIM database to select disease-associated genes and found again a opposite pattern of cellular localization for OMIM and miRNA targets.

4.2.5 In contrast to disease proteins, microRNA targets frequently exhibit an inhibitory effect

We sorted the set of signaling proteins into three different groups according to their process type annotation: activating proteins, inhibiting proteins and proteins that can act as either activators or inhibitors, further on denoted as ambivalent. We then counted the number of miRNA targets as well as disease proteins for each group in our signaling pathways and calculated the LOD score. The result shows again a complementary pattern: As shown in Figure 4.3B, targets of disease-associated miRNAs are preferentially inhibitors (LOD = 1.62, $p = 1.2 \cdot 10^{-4}$), whereas only 6 disease-associated proteins (LOD = -2.08, $p = 1.5 \cdot 10^{-5}$) show a inhibitory effect. miRNA targets are enriched almost 14 times more in inhibiting proteins compared to disease proteins showing a complementary focus. Ambivalent proteins show a strong under-representation for miRNA targets (LOD = -0.96, $p = 7.3 \cdot 10^{-5}$), whereas disease-affected proteins are significantly enriched (LOD = 1.26, $p = 3.6 \cdot 10^{-9}$). For activators, we found a significant under-representation for both disease proteins (LOD = -0.75, $p = 1.0 \cdot 10^{-4}$), and miRNA targets (LOD = -0.60, $p = 2.7 \cdot 10^{-3}$), respectively. Again, we found the same result for cancer and non-cancer related miRNA targets indicating a common pattern. Notably, the enrichment of process types of disease proteins remains for the OMIM data set.

4.3 Discussion

In order to study the role of disease-associated miRNAs in pathways, we applied a thorough statistical analysis to a multipartite graph consisting of miRNAs, proteins, dis-

eases, tissue and signaling pathways. We investigated enrichment of disease-associated miRNAs globally on different pathways by considering of tissue-specific transcript expression, and more locally, on the cellular location and process type of target proteins.

We found that the amount of regulatory control mediated by disease-associated miRNAs differs from pathway to pathway. For the majority of diseases, a homogeneous enrichment profile of miRNA targets throughout all pathways emerged. From our analysis of the constituting multipartite graph, we found that pathways are heterogeneously targeted by miRNAs. However, the core set of pathways appear to be homogeneously enriched by miRNA target genes throughout the majority of diseases, since many diseases are linked to a large number of miRNAs. So far, almost two third of the currently known miRNAs are linked via large-scale expression analysis to a phenotype. It is obvious that beside the phenotype responsible miRNAs, many miRNAs are detected as deregulated in human diseases but are not functionally linked to the phenotype.

What could be the biological function of a core set of globally enriched pathways? We showed that these pathways are targets of numerous deregulated miRNAs. One possible hypothesis is that these pathways could serve as disease sensors, transferring the information of erroneous cellular functions via deregulated miRNAs to important output proteins, like cell cycle checkpoints. From this perspective, it is intriguing that most top enriched pathways are associated with apoptotic, proliferation or developmental processes (116). Entries in the PhenomiR database obtained by patient studies are more than 60% cancer-related diseases. Alterations in the expression or function of genes controlling cell growth and differentiation are considered to be the major cause of cancer. Notably, degenerative disorders like Alzheimer or Parkinson disease shows a similar pathway profile compared to cancer-related phenotypes, although often with different direction of miRNA expression.

Presumably, the impact on signaling pathways for disease-associated proteins and miRNA targets differs. However, there might be an interaction between the disease-associated miRNAs and proteins to mediate deregulation of signaling pathways. It would be interesting to evaluate whether a given disease emerges due to protein deregulation caused by mutations with a successive deregulation of miRNAs, or due to deregulated miRNA levels, leading to pathogenic protein levels in turn. For a subset of miRNAs, located in the intron of a host gene, an examination of a common phenotypic effects is possible. Recently, we showed that intronic miRNAs support the regulatory effect of their host genes (174). Here, we find one disease-associated miRNA-target pair with a common phenotype: both the host gene PTK2 and its intronic miRNA miR-

151 are annotated with lung cancer in KEGG DISEASE and PhenomiR, respectively. In this case, the impact on the associated signaling pathways via correlated mir-151 and PTK2 deregulation is probably controlled by a single promoter. To unveil interactions between miRNAs and pathway proteins on a systems level, a much more precise knowledge of miRNA transcriptional regulation is needed.

We analyzed the subcellular location and process type behavior of disease-associated proteins and miRNA targets. Our result on the preferred cellular locations of miRNA targets shows an enrichment of proteins in the nucleus. This finding is in line with a study by Cui et al. (44), who obtained a similar result for the localization of miRNA targets on a much smaller set of signaling networks and miRNAs in mammalian hippocampal CA1 neurons. In addition, we found that disease-associated proteins often constitute the initial players of signaling networks and thus show an opposite pattern to miRNA targets. The deregulation of a single proteins at the cell surface receptor can have a severe impact on the whole signaling information flow stimulated by the receptor. For example, for growth factor receptors, the activation under normal conditions promotes cellular survival, whereas over-expression promotes tumor cell growth (1). Therefore, cell surface receptors are well suited as drug targets, as diminishing the signal through these receptors has the potential to normalize cellular behavior. The deregulation of a single protein in the intracellular region or the nucleus might influence only a subpart of the signaling network.

A large fraction (50%) of miRNA targets located in the nucleus are involved in transcriptional regulation. It was shown that transcription factors like MYC, JUN, or FOS, have a short mRNA lifetime based on their RNA stability (118; 293). Within these studies the importance of the 3' untranslated region for the mRNA stability was mentioned. Thus, miRNAs presumably tune RNA stability in a tissue or stage dependent manner. Deregulated miRNAs changing the stability of transcription factors of a signaling pathway may then lead to malfunction of different cellular processes (98). Motivated by the affinity of miRNAs to regulate with associated pathways apoptosis, proliferation or development (138), we suppose that the regulation of stability extends to proteins with short half-lives that are required only for limited time in, e.g. cell cycle, growth, or differentiation.

In a recent study, Legewie et al. (159) introduced a set of signal inhibitors with a short mRNA and protein lifetime that are transcriptionally induced upon stimulation. These rapid feedback inhibitors (RFIs) are thought to tune the signal transduction cascades, allow for swift feedback regulation and establish short latency phases af-

ter signaling induction. As we found an enrichment of inhibitory proteins targeted by miRNAs, the question arises, if RFI proteins are potential miRNA targets. Using the TargetScanS prediction tool we were able to confirm 18 out of 19 (95%) RFIs as miRNA targets ($p = 0.023$). We thus assume that the short mRNA lifetime of RFIs can be attributed to the degradation activity promoted by miRNA binding. Inhibiting proteins are preferentially located in the nucleus, whereas activating or ambivalent proteins are randomly distributed in the cellular regions. Interestingly, disease proteins showed a frequent association with ambivalent process type. We assume that for ambivalent proteins, deregulation of the expression levels imparts a more severe effect on signaling cascades as compared to activators or inhibitors alone.

4.4 Materials and Methods

In this section, we give a detailed overview about the resources and methods, which were used to interconnect the different entities shown in Figure 4.1.

4.4.1 Human signaling pathway data

Human signaling pathway data was obtained from the National Cancer Institute Pathway Interaction Database (NCI PID) (230), which is a manually curated collection of biomolecular interactions and key cellular processes assembled into signaling pathways. NCI PID holds 128 pathways including 47 sub-networks. We combined all subnetworks with their parent networks to the set of signaling pathways. In addition, we kept all pathways that have more than one predicted miRNA target gene, leading to a final data set of 79 human signaling pathways containing 1573 unique human proteins. The database also provides information on subcellular location terms from the Gene Ontology Consortium. We used this information to divide all subcellular locations into four different groups: extracellular region, cell membrane, intracellular region and nucleus. Finally, location information for 1083 proteins containing 135 extracellular region, 344 cell membrane, 373 intracellular region and 231 proteins located in the nucleus were obtained. In addition, we extracted process type information for each biological process, which can be input, output, positive or negative regulator. In total, there are 1120 interactions of which 765 are activating, 74 inhibiting and 281 proteins acting as activators as well as inhibitors.

4.4.2 Disease-associated microRNAs

Human disease-associated miRNAs were obtained from the PhenomiR database (226). PhenomiR is a manually curated collection of miRNA-disease associations, containing a total of 11029 miRNA expression-phenotype relations collected from 542 different experiments. We used patient study data only and obtained 486 disease-associated miRNAs in 83 different diseases including up to 5 subtypes per disorder. For each disease, we take only those miRNA into account, that have at least one target in the specific tissue annotated by PhenomiR and obtained finally 165 different miRNAs in 63 diseases-tissue combinations.

4.4.3 MicroRNA target prediction

Hausser et al. (91) analyzed different features of miRNA targets and showed within their work that TargetScanS has a good performance on different data sets. We used TargetScanS as the main prediction tool but to handle the issue of the unknown reliability of miRNA prediction tools we used several other prediction tools like PicTar, intersection of PicTar and TargetScanS, Miranda, RNA22, and TargetSpy to confirm our results. We used for each method default parameter settings.

4.4.4 MicroRNA targets filtered by tissue expression

As miRNA expression is tissue-specific annotated in PhenomiR, we used the tissue atlas provided by Su et al. (257) to filter potential miRNA targets in a specific tissue. The data was downloaded from the NCBI Gene Expression Omnibus (GEO), and the processed data was used. We mapped the predicted miRNA target transcripts on the tissue atlas and considered a transcript as expressed in a specific tissue, if either one replicate has a present call or both show at least a marginal call, similar to the work of McClintick et al. (186).

4.4.5 Human disease data

Human disease proteins were taken from the KEGG DISEASE database (121). It associates 5 neurodegenerative disorders, 5 infectious and metabolic disorders and 13 different cancer diseases. Finally, we obtained 909 proteins from 23 different diseases, which are also found in the PhenomiR database.

4.4.6 Pathway profile

Pathway profiles were calculated for all diseases annotated in PhenomiR passing the tissue filter. For each disease-pathway interaction we estimated the enrichment of miRNA targets of disease i in pathway j defined by a log odds ratio (LOD score):

$$\text{LOD}_{i,j} = \log_2\left(\frac{T_{i,j}}{P_j} \Big/ \frac{\sum_{k=1}^{n} T_{i,k}}{\sum_{k=1}^{n} P_k} \right)$$

where $T_{i,j}$ is the number of miRNA targets for all disease-associated miRNAs in disease i and pathway j; P_j is the number of proteins in pathway j; $\sum_{k=1}^{n} T_{i,k}$ is the number of miRNA targets for all disease-associated miRNAs in disease i over all pathways; $\sum_{k=1}^{n} P_k$: is the number of proteins over all pathways. We use these LOD scores to build up a heatmap using Manhattan distance function and ward clustering. A positive value indicates an enrichments and a negative a depletion. Whenever we identified no target for a particular disease-pathway interaction $T_{i,j} = 0$ and therefore the resulting LOD score $_{i,j}$ is $-\infty$. As commonly done, we excluded all cases with $T_{i,j} = 0$ for calculating the mean and quantiles for each pathway. In addition, these cases were also excluded from the clustering taking the reduced dimensions into account.

4.4.7 Cellular location analysis

We used the subcellular location annotation of the NPI PID database to estimate the miRNA target enrichment. The enrichment was calculated by the logarithm of base 2 of the odds ratio (LOD score) and its significants was obtained by Fisher's exact test.

4.4.8 Process type analysis

In addition to the subcellular location, the NPI database provides information about specific process types of proteins in signaling processes. We used this information to analyze the interaction between inhibiting as well as activating proteins in signaling processes. Within this analysis we calculated the enrichment of miRNA targets as well as KEGG DISEASE proteins for different process types. The enrichment was calculated by the logarithm of base 2 of the odds ratio (LOD score) and its significants was obtained by Fisher's exact test.

4.5 Conclusions and Outlook

The usage of hypergraphs for a proper representation of interconnected entities in systems biology has recently been acknowledged (136). Here, we applied a thorough statistical analysis not only to bipartite but to a multipartite graph consisting of miRNAs, proteins, diseases, and signaling pathways in a tissue-specific manner. Using this framework, we uncover the impact of disease-associated miRNAs on human signaling pathways. From a global perspective, we identify a core set of signaling pathways with enriched tissue-specific miRNA targets across diseases. In addition, we divide the set of disease-associated miRNAs into cancer and non-cancer sets and find no significant difference for both groups. The resulting core set reflects the affinity of miRNAs to affect central cellular processes. More locally, we show that disease-associated miRNAs and proteins prefer different cellular locations and process types. This chapter provides systematical insights into the interaction of disease-associated miRNAs and signaling pathways and uncovers differences in cellular locations and process types of disease-associated miRNAs and proteins. In the following chapter, we focus on the theoretical aspects of inferring functional miRNA-pathway associations. We will introduce a novel approach, which use the signal network structure to improve the inference of miRNA-pathway interactions.

5 Beyond enrichment: Measuring microRNA-pathway associations in signaling networks

5.1 Background

Cellular signaling pathways act as information processing devices: They integrate diverse inputs and compute, based on the states of the signaling molecules, a cellular output. This output often crucially depends on the precise levels of the involved signaling proteins. miRNAs, a large class of post-transcriptional regulators that predominantly decrease mRNA levels (86), have been shown to optimize and fine-tune protein abundances in signaling pathways (for a review, see (106)). High-throughput methods like the HITS-Clip protocol (37), target prediction tools like the TargetScan algorithm (162), and databases of validated targets (100; 207; 287) deliver evidence for miRNA-transcript interactions. Unfortunately, to predict the impact of a miRNA on a specific pathway, the mere knowledge of the target transcripts is mostly not sufficient: miRNAs are promiscuous regulators, with often hundreds of targets. The challenge for computational biology in this context is: How can one infer signaling pathways under miRNA control from a large number of miRNA-target transcript relationships? The standard procedure to filter potentially functional miRNA-pathway associations (MPAs) from this mass of interactions is to calculate an enrichment score (see (43; 145; 290) for examples). To that end, one divides the proportion of targets of a miRNA in the specified pathway by the expected proportion of targets, inferred from the overall set of proteins. However, this measure only works if a considerable number of targets can be found - for single or few targets in a pathway, it fails. From a functional perspective, it might suffice for a miRNA to regulate a small sub-part or even a single transcript in a path-

way. To assess these aspects of miRNA-mediated control on signal transduction, we here introduce the proximity score P, which takes the topology of the underlying signaling network into account. It is based on the shortest paths between miRNA targets in a signaling pathway and identifies proximal and distal regulatory patterns, where the targets are either more closely connected or more outspread than expected by chance.

Currently, only few resources are available that link miRNAs and biological pathways. MiRDB (280) is a miRNA target prediction web resource that also provides precompiled information about single miRNAs regulators. MiRGator (199) offers functional annotation of miRNAs targets as well as mapping of single miRNAs in pathways. DIANA-mirPath (207) integrates miRNA targets in KEGG (121) pathways and provides three different target prediction tools.

It has already been shown that many miRNAs exhibit temporal and tissue-specific expression patterns (153), regulating many transcripts to further define tissue-specific transcript profiles (61). However, miRNA prediction algorithms do not take expression profiling of both miRNA and mRNA levels into account. Therefore, functional analysis on the global set of predicted targets may lead to wrong miRNA-pathways associations. Based on the highly tissue-specific expression signatures of miRNAs and target transcripts, tissue-specific gene expression has to be considered to improve the analysis of miRNA regulation in biological pathways. We have therefore developed miTALOS, an interactive tool that integrates tissue and pathway filters to restrict the functional analysis as first publicly available resource. MiTALOS performs an enrichment and proximity analysis of predicted target genes in signaling pathways. The widely applied enrichment analysis uses the number of target genes in a specific signaling pathway to infer miRNA-pathway associations (89; 145; 290). As the enrichment analysis focuses on the whole signaling pathway as a set of genes without taking its topology into account, sub-cascade specific relations between miRNAs and pathways are ignored. In order to cover these interactions, miTALOS provides a second approach based on the network proximity of miRNA targets. As there is strong evidence that miRNAs can act in concert with each other in order to affect a signaling pathway (109), miTALOS addressed this aspect through the simultaneous analysis of multiple miRNAs or even predefined genomic miRNA clusters. In addition, target genes and miRNAs are linked to external databases to offer additional information. Finally, graphical visualization of the miRNA targets in a given pathway allows functional insights into miRNA dependent regulation of signaling pathways.

5.2 Results and Discussion

We consider a data set of miRNA targets in mouse brain, based on high-throughput sequencing of RNAs isolated by immunoprecipitation of crosslinked Ago-RNA complexes (Argonaute HITS-CLIP, see `http://ago.rockefeller.edu`, (37)). In contrast to target prediction tools with only a limited degree of accuracy (see (8; 237) for a experimental validation), here a miRNA-transcript relation is only present if miRNA and transcript are expressed, if the argonaute protein is bound to the transcript, and if the transcript contains the miRNA seed. Chi et al. (37) claim that crosslinking of the miRNA-Ago complex reduces false-positive rate compared to computational target prediction algorithms and conventional Ago-immunoprecipitation. They calculate a specificity of 93%, a false-positive rate of 13-27% and a false-negative rate of 15-25% for the HITS-CLIP approach, based on comparison to a genome wide systematic analysis of the seed sequences of miR-124 (122; 177; 281). Unlike databases with experimentally validated miRNA targets, HITS-CLIP delivers the complete target set of the 20 most abundantly expressed miRNAs in mouse brain, which is necessary for a global network analysis as conducted in this study. Mapped on the proteins in the signaling pathways provided by KEGG (121), each pathway is on average regulated by 16 out of 20 miRNAs with varying number of targets, resulting in a densely connected miRNA-pathway network.

We filter miRNA targets in KEGG signaling pathways with distinct regulatory patterns by introducing the proximity score P. It explicitly takes the topology of the signaling network into account by calculating the average shortest path length between all miRNA targets in a pathway. For each miRNA-pathway association (MPA), we calculate the z-score by considering 10^4 randomly generated samples as a null model. Here, for a miRNA, the same number of targets is chosen randomly from the pathway, giving rise to a null model mean μ and standard deviation σ, from which the proximity score is calculated as $P = (l - \mu)/\sigma$ (see Figure 5.1C). We further analyze MPAs with $|P| > 2$, as is conventionally done in network biology (see, e.g. (9; 193). The networks are inferred from the respective KEGG signaling pathways (see Figure 5.1A) as follows: Based on the KEGG Markup Language (KGML) pathway files of 119 signaling pathways, we convert each pathway to a network, where a node represents a protein, a protein complex, or a protein family, and each molecular interaction or relation is represented as a link. Since we are interested in the topological distance between targets, we take the emerging networks as undirected. To avoid artifacts from

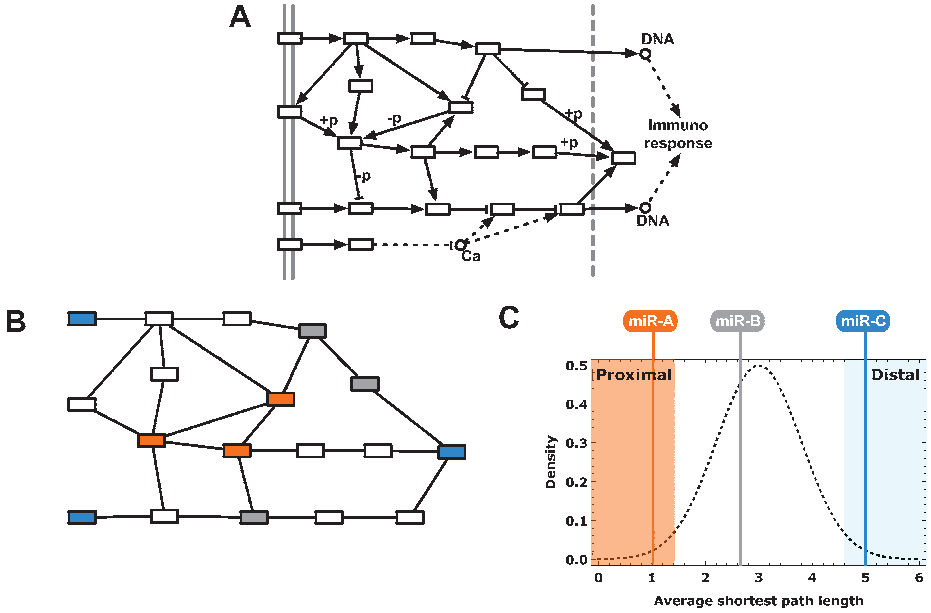

Figure 5.1: **Proximity score for miRNA target patterns in signaling pathways.** (**A**) An exemplifying signaling pathway. Rectangles represent pathway proteins and solid links stand for molecular interactions, like phosphorylation, (denoted with +p and -p), activation, or inhibition. We convert this pathway into an undirected protein-centered network (**B**), where unconnected fragments have been removed. The colored nodes (orange and blue) are targets of the illustrative miRNAs miR-x and miR-y with three targets each. While the number of targets is equal for the two miRNAs, the regulatory patterns in the network are considerably different: miR-x targets an interconnected core of the pathway, while the targets of miR-y are spread over the whole pathway. (**C**) The proximity score is calculated for each miRNA and as a z-score. Here, the average shortest path length between all targets (1.0 for miR-x, 4.7 for miR-y) is compared to a null model of 3 randomly chosen targets (dotted line). We call the targets of a miRNA proximal, if their z-score (defined as the deviation from the null model mean in units of the standard deviation) is below -2 (red shaded area), and distal, if the z-score is larger than 2 (blue shaded area).

fragmented graphs, we only consider the largest connected component of the networks and the targets therein (see Figure 5.1B).

5.2.1 MicroRNAs have a proximal and distal target pattern in signaling pathways

We identify 52 MPAs (out of a total of 20x119=2380 and 863 with at least 2 targets) with $|P| > 2$. They contain 18 out of the 20 most abundantly expressed miRNAs in

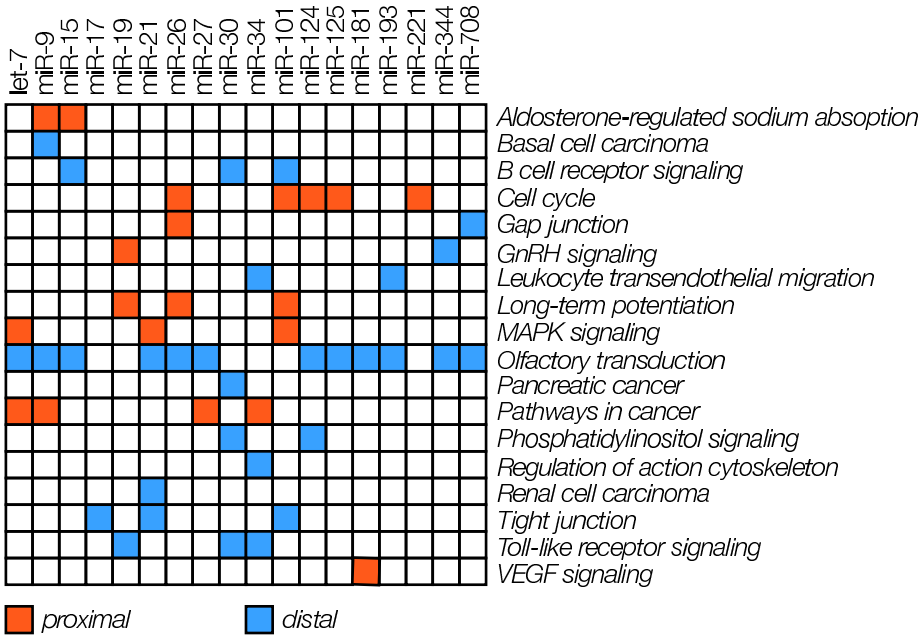

Figure 5.2: **miRNA-pathway associations (MPAs) with an absolute proximity score $|P| > 2$.** We calculate the proximity score P as the z-score of the average shortest path-length of miRNA targets in a signaling network, compared to randomly selected targets. MPAs with $P < -2$ are called proximal, those with $P > 2$ distal. We find that many miRNAs exhibit both proximal and distal target patterns.

mouse brain, and 18 out of 89 pathways with a largest connected component of at least 2 nodes (smaller components are inapt for the proximity score). Since proximal and distal MPAs represent the extremes of regulatory patterns the sets of the 21 proximal $(P < -2)$, and the 31 distal $(P > 2)$ MPAs are mutually exclusive (see Figure 5.2). However, two pathways (Gap junction and GnRH) appear in both distal and proximal MPAs, though with different miRNAs associated. Among the pathways with most proximal target patterns we find the MAPK and cell cycle pathways.

Within the MAPK pathway, miR-21 proximally targets the downstream cascades RAS-RAF1-MEK of the epidermal growth factor (EGF) receptor. Based on these findings we predict a pivotal role of miR-21 in MAPK-based signaling in brain. Indeed, Zhou et al. (300) recently showed that mmu-miR-21 effects the growth of glioblastoma cells by inhibiting the MAPK pathway via EGFR. The functional impact of the MAPK network in brain is backed by several studies indicating that this pathway is a key mediator of EGF (53) and brain-derived neurotrophic factor signaling (210; 295). The MAPK pathway is also essential for growth-factor-induced cell-cycle progression.

Growth-factors induce MAPK pathway, which in turn activates cyclin-dependent kinases (CDKs). CDKs are key regulatory proteins that are activated at specific phases of the cell cycle. CDK4 and CDK6 are essential for the transition from G1 phase to DNA replication. We find 4 proximal MPAs in the cell cycle pathway (see Figure 5.2). Among those, miR-124 targets CDK4, a member of the early cell cycle (G1 phase), and two members of the G2/M phase (CDC25 and YWHAQ). Silber et al. (249) showed that transfection of miR-124 inhibits proliferation of glioblastoma cells by inducing a G1 cell cycle arrest, which is in line with our prediction. This finding shows that alteration of miR-124 expression has a severe impact on cell cycle progression, further indicating that MPAs inferred with our proximity score have a significant functional impact. The pathway with most (12) distal targets patterns is 'Olfactory transduction'. A closer look at this this pathway reveals a unique network structure. From the 15 nodes comprising the network, one represents several hundred receptors (further on called the receptor node), involved in the olfactory system (49). The other 14 are individual proteins involved in downstream signaling. When targets are chosen randomly to build the null model for this pathway, it is very likely that all three are located on the receptor node. An MPA will have a proximity score P > 2 if already one target is not located on the receptor node. Thus, the network structure of the Olfactory transduction pathway does not allow a meaningful calculation of the proximity score. If this pathway is not taken into account, the number of MPAs with a distal target pattern decreases by one third.

5.2.2 Enriched targeted pathways represent only a small subclass

We compare the results of our proximity score with the well-established enrichment measure, which relies on the assumption that the sheer number of targets indicates if a pathway is controlled by a miRNA. Using the identical data sources as above, we find 25 enriched MPAs ($p < 0.05$, Bonferroni corrected, similar to (51)). Two of these 25 MPAs (miR-9 in MAPK, miR-26 in Long-term potentiation) also exhibit proximal target patterns (Figure 5.3A). To understand the differences in the associations, we compare the number of miRNA targets in proximal, distal, and enriched MPAs in Apparently, the two scores identify MPAs with very different target abundances: We find typically more than 10 targets in an enriched MPA (with a median of 14), while the median number of targets in proximal and distal MPAs is 7 and 3, respectively (see Figure 5.3B). Thus the established enrichment score misses, by construction, pathways

with only very few targets. The proximity score on the other hand fails for large target abundances, since differences between proximal and distal patterns then disappear. However, the majority of pathways show only few targets, which is covered by the proximity rather than the enrichment score.

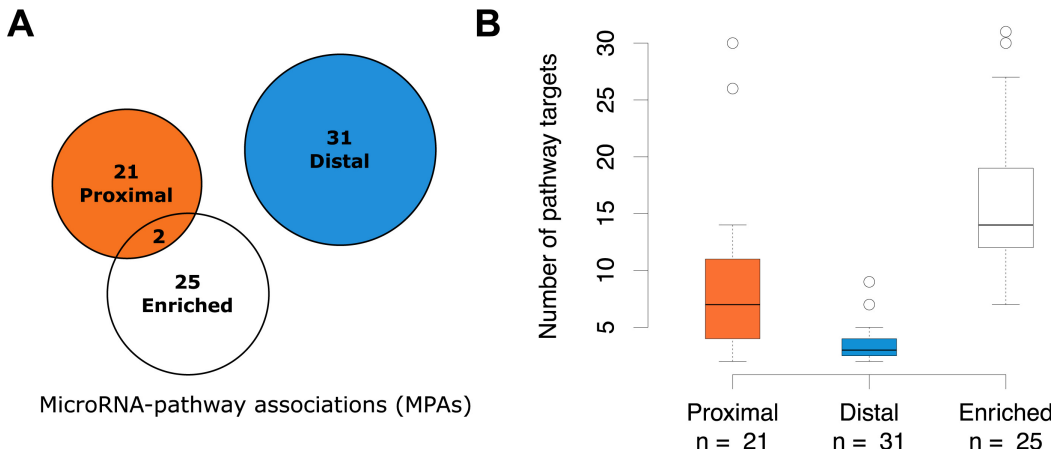

Figure 5.3: **Proximal, distal, and enriched MPAs.** (**A**) We find 21 proximal (P<-2), 31 distal (P>2), and 25 enriched MPAs. Two MPAs (miR-9 in MAPK, miR-26 in Long-term potentiation) are both exhibit enriched and proximal target patterns (**B**) Density of number of pathway targets in significantly targeted pathways. While the number of targets in enriched pathways is predominantly higher than 10 (white), proximally (red) and distally (blue) targeted pathways have mostly less than 10 targets. Thus, the proximity concept identifies MPAs with only few targets, as opposed to the enrichment concept, which per se favors many targets. However, 95% of all HITS-CLIP MPAs have less than 11 targets.

5.2.3 MicroRNA target pattern corresponds to a specific function in cell signaling

Both scores miss a specific target pattern: miRNAs with only a single target in a signaling pathway. From a functional perspective, the regulation of an important pathway protein might suffice to alter the dynamics of the associated pathway as a whole (see, e.g. (8) for the functional impact of single nodes on complex networks' dynamics). If we consider all 2380 MPAs in our data, we find 15% with only a single miRNA target. To study if these targets share specific properties, we conduct a gene ontology (GO) (7) analysis with DAVID (101) a tool that automatically clusters enriched and related GO terms (102). The two top scoring groups of GO process terms are shown in

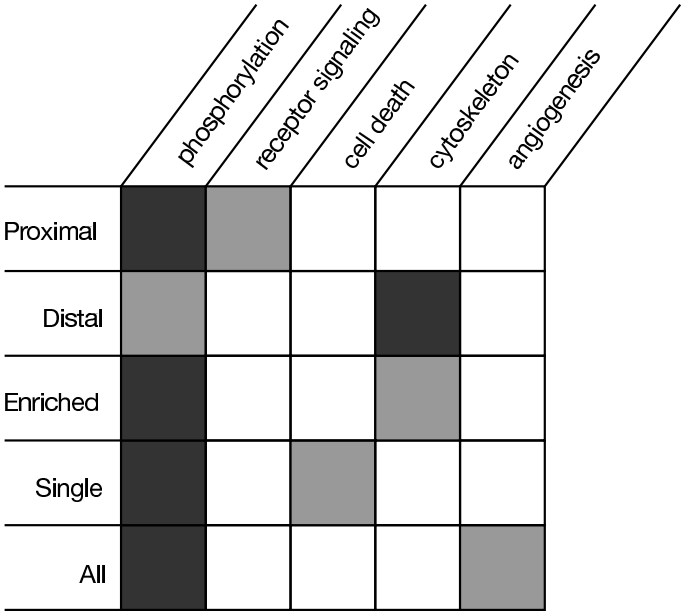

Figure 5.4: **GO analysis of proximal, distal, enriched and single targets of HITS-CLIP miRNAs.** We identify clusters of GO biological process terms associated with the targets, using the Functional Annotation Clustering (7) of the DAVID software (101). The two top scored clusters for each class of MPA are shown in dark and light grey respectively. While phosphorylation-associated functions appear in all identified target patterns, underlining the assumption that miRNAs target mostly intracellular components of signal transduction networks (44), each pattern also exhibits a specific biological function.

Figure 5.4 for proximal, distal, enriched, and single targets. A 'phosphorylation' cluster appears in all four groups, comprising the biological processes 'protein amino acid phosphorylation', 'phosphorylation', 'phosphate metabolic process', and 'phosphorus metabolic process'. This cluster shows up as the dominant cluster of all targets, which corroborates earlier findings (44; 145), where miRNAs have been shown to predominantly target intracellular components of signal transduction networks. However, each class also exhibits a specific cluster of targeted processes: Proximal miRNA targets appear predominantly in trans-membrane or cell surface receptor associated processes. Apparently, the control of signal transduction at the receptor level requires a coordinated regulation of a few proximate proteins by a specific miRNA. Single targets are highly associated with the regulation of cell death, the most important decision a cell can make. This is remarkable, since it stresses the idea pivotal pathway proteins can be under specific miRNA control. Finally, distal and enriched targets both participate

in the regulation of cytoskeleton organization.

5.2.4 miTALOS: Workflow of the functional analysis

The miTALOS web resource provides insight into the tissue-specific regulation of signaling pathways mediated by miRNAs. First, a single or multiple miRNAs can be analyzed by miTALOS, whereby the input data can also be selected from a list of predefined genomic miRNA clusters (Step 1; see Figure 5.5a). Several studies indicated that not only deregulated single miRNAs but also miRNA clusters, such as miR-17-92 (131), miR-106b-25 and, miR-222-221 (191), have a strong impact on signaling transduction and corresponding phenotypes. It is well established that many miRNAs are limited in their expression to certain stages in development, tissues, and cell types (10). To address this issue, miTALOS is the first resource that provides an additional tissue filter (Step 2). We mapped the predicted miRNA target transcripts on the tissue atlas data from (256) providing the expression patterns for 79 human and 61 mouse tissues.

MiTALOS offers two different resources of signaling pathways. All non-metabolic human and mouse pathways were integrated from KEGG (121). For the analysis of human miRNA regulation, we also included pathway information from the National Cancer Institute Pathway Interaction Database (NCI PID) (230). In addition to the tissue filter, miTALOS also provides a pathway filter to restrict the functional analysis. MiRNA target transcripts are obtained from five different prediction tools: TargetScanS (162), RNA22 (61), PicTar (149), PiTa (127), TargetSpy (255). Due to imperfect base pairing and the short length of binding sites, prediction of miRNA target genes often yields false positive target genes. It has already been shown that the intersection of the prediction tools can yield improved specificity with only a marginal decrease in sensitivity relative to any individual algorithm (238). To address this issue, miTALOS provides the ability to generate intersections from 2-5 prediction tools (Step 3; see Figure 5.5A). By default, miTALOS presents miRNA-pathway associations having a p-value < 0.05. An intuitive option menu allows the user to modify all parameters (enrichment, proximity cutoffs and p-values).

The result web page lists all identified miRNA-pathway associations (see Figure 5.5B). MiTALOS sorts all pathways by increasing enrichment p-value along with the names of each miRNA's target genes involved in either KEGG or NCI PID pathways. Target genes are linked to the Universal Protein Resource (Uniprot) (6). MiTALOS also links queried miRNAs to disease-associations obtained from the Phe-

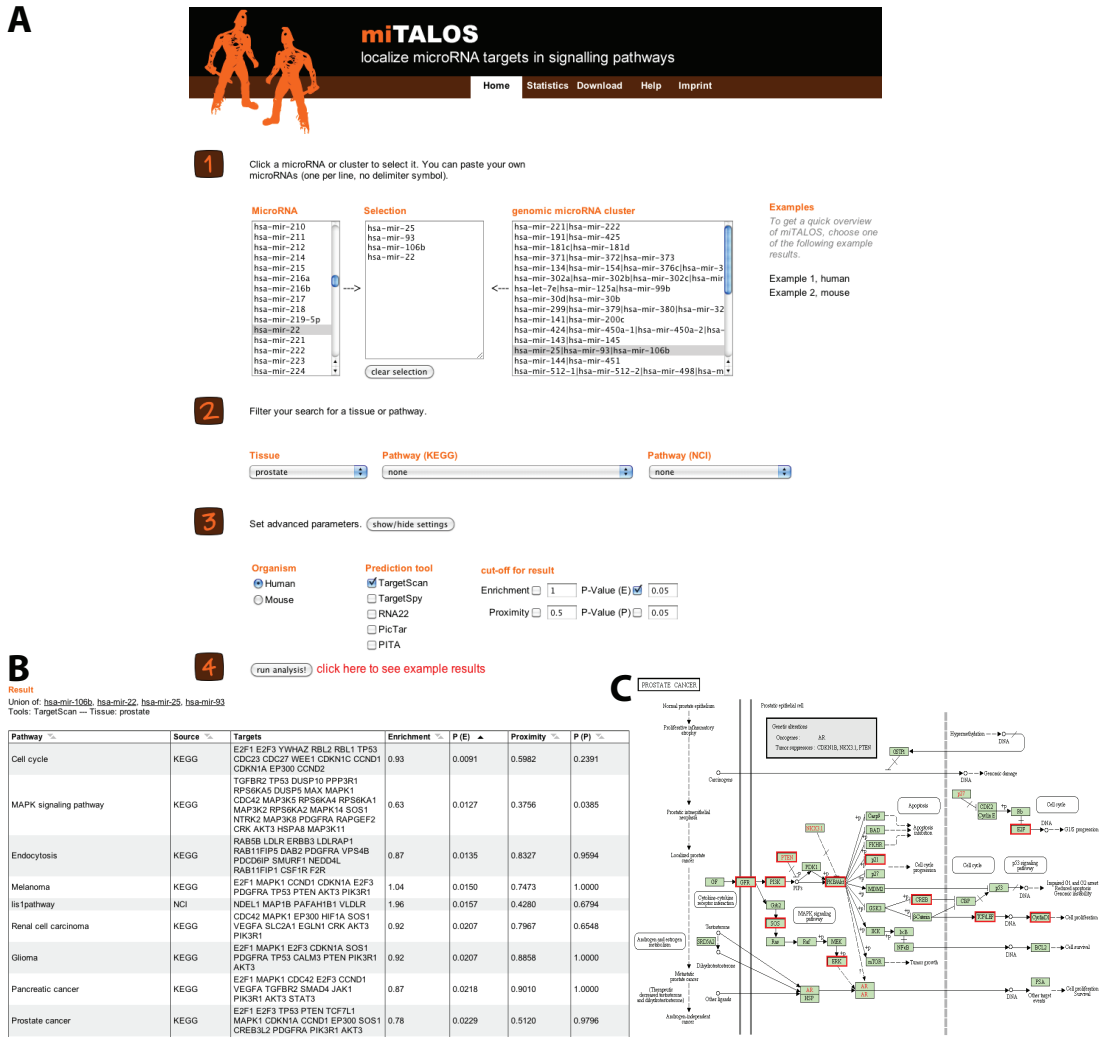

Figure 5.5: **Overview of the miTALOS web resource.** (**A**) After selecting a single or multiple miRNAs (which can also be chosen from a list of predefined genomic miRNA clusters) as input, the user can restrict the analysis to a specific tissue and/or pathway. In addition, miRNA prediction methods and output parameter such as p-value cutoffs can be defined. (**B**) The result page shows the identified miRNA-pathway associations. By default, miTALOS sorts all pathways by an increasing enrichment p-value along with the names of each miRNA's target genes involved in either KEGG or NCI PID pathways. Multiple sorting options and links to disease-association are provided. (**C**) MiRNA target genes in a given pathway are graphically annotated (highlighted in red boxes) in the pathway map.

nomiR database (226), which is a manually curated database of differentially regulated miRNA expression in diseases and other biological processes. The result page also provides multiple sorting options for the user. Finally, miRNA target transcripts located in a given pathway are graphically annotated onto the pathway map (see Figure

5.5C), which is also linked to the result page.

MiTALOS uses a MySQL database to store pathways from KEGG and NCI and tissue specific expression data from the tissue atlas. MiRNA predictions and random samplings for all pathways are precomputed and stored in separate database tables. The database structure allows to update pathways and miRNA targets independently. The integration of external resources such as miRBase, KEGG, NCI or target prediction tools is automated and will be updated whenever these resources have a new major release. Calculation of the enrichment and proximity score is performed on-the-fly and makes use of the precomputed data. We are thereby able to offer highly responsive computation of miRNA-pathway associations based on the most recent biological knowledge. MiTALOS is based on Java EE and running on a Tomcat servlet container. The business logic is implemented with servlets and additional Java libraries while Java ServerPages are used to present the data. AJAX capabilities are added with jQuery, a comprehensive JavaScript library.

5.2.5 Identification of microRNA-pathway associations

MiTALOS analyzes miRNA-mediated regulation of signaling pathways using two different approaches. A first approach uses the number of miRNA targets in a specific signaling pathway to calculate an enrichment score, as widely applied in various applications (89; 145; 168; 290). The enrichment score assumes that miRNAs target specific signaling pathways to influence specific functions of the cell by the sheer number of target transcripts. The enrichment score is then defined as the fraction of target genes compared to the expected number of target genes in a given pathway. The significance is obtained with Fisher's exact test (66), corrected by the Benjamini-Hochberg procedure (13) (see Material and Methods for a detailed description).

Signaling pathways often have different cascades activated by different stimuli. As the enrichment method focuses on the whole pathway, sub-cascade specific relations between miRNAs and signaling pathways are ignored. To cover these miRNA-pathway associations, we developed a second approach based on the network proximity of miRNA targets in signaling pathways. We assume that miRNAs target signaling cascades in a proximal manner. To reveal signaling pathways with proteins that function in a proximal manner and are targeted by the same miRNA, we introduce the proximity measure. We determine the distances between all pairs of targets in the corresponding signaling pathway. For each target, the minimal distance is chosen and the proximity

score is defined as the mean of all minimal distances (see Material and Methods for a detailed description).

5.2.6 Difference between microRNA enriched and proximal pathways

In contrast to the proximity measure, the enrichment measure relies on the assumption that miRNA control on a pathway is mediated by the number of targets. As there is evidence that miRNAs have a strong impact on the signal transduction, the degree of downregulation often tends to be quantitatively modest. A miRNA typically downregulates most of its target transcript by less than 50% (8). This consideration suggests that although many genes are predicted to be miRNA targets, only a fraction of these interactions will have an impact on biological responses and phenotypes (164). Analyzing the distribution of the number of targets in signaling pathways (KEGG pathways and TargetScanS), the result shows the highest miRNA target density for 1 to 8 targets per pathway (see Figure 5.6) reflecting 90% of all miRNA-pathway associations.

Applying the enrichment method, we obtained 265 significant miRNA-pathway associations (FDR < 0.01). We analyzed the number of targets in significantly enriched pathways and obtained mainly between 10 and 20 target transcripts per signaling pathway (see Figure 5.6). This result indicates that the enrichment approach mainly focuses on a small subpart of miRNA-pathways associations that only reflects 7% of the total associations. Therefore, it can be argued that the enrichment approach identifies signaling pathways having in general transcripts under miRNA control. Therefore, the basic hypothesis behind the enrichment concept might be unsound: Often, it suffices for a miRNA to regulate a small subpart or even a single transcript in order to influence the function of a whole pathway (125; 160; 220). This assumption is also affirmed by our finding reflecting that miRNAs target in general a small amount of pathway player (Figure 5.6).

In order to assess these signaling components of miRNA-mediated control, we introduced the proximity measure that calculates the proximity of miRNA targets in signaling pathways based on the distance of the miRNA targets in a pathway. Using this method, we obtained a set of 125 significantly proximal (FDR < 0.01) miRNA-pathway associations. This set of miRNA-pathway associations has the highest density for 3 to 7 targets per pathway (Figure 5.6). Apparently, the two measures identify significant miRNA-pathway associations with very different target abundances. Thus,

proximity and enrichment scores identify two alternative forms of miRNA control. The proximity method (mean number of miRNA targets per pathway (mtpp) = 5.90) is clearly shifted to a smaller number of target transcripts compared to the enrichment method (mtpp = 15.31; $p < 2.2^{-16}$). Comparing the proximity method with the distribution of all miRNA-pathway associations (mtpp = 5.40), shows that the proximity based approach focused on miRNA-pathway associations that are in general more common.

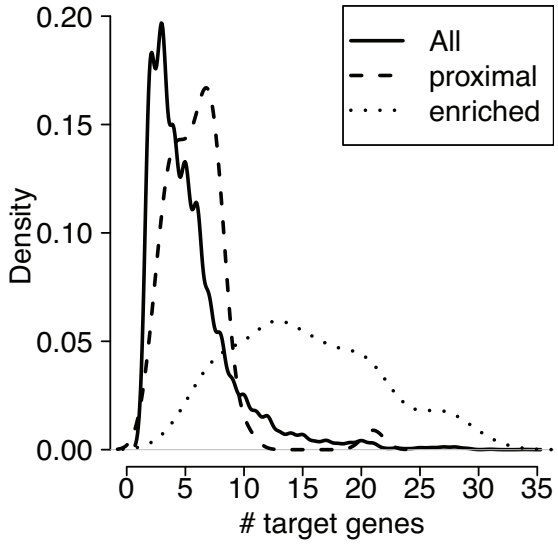

Figure 5.6: **Proximity vs. enrichment in signaling pathways.** Density of the number of targets in miRNA-pathway associations (solid line). While the number of targets in significantly enriched (FDR < 0.01) pathway is mainly between 10 and 20 (dotted line), significantly proximal (FDR < 0.01) targeted pathways have 3 to 7 targets (dashed line). Thus proximity and enrichment score identify two alternative forms of miRNA control.

5.2.7 Case study: microRNAs in prostate cancer

Recent studies have supported that miRNA mutations or deregulation are associated with various human cancers indicating that miRNAs can function as tumor suppressors and oncogenes (189; 297). Prostate cancer is one of the most significant cancers and second leading cause of cancer death among American men, exceeded only by lung cancer (200; 252). In order to unveil the impact and interaction of miRNAs with the important and altered signaling pathways in prostate cancer, we performed a functional analysis with miTALOS using miR-106b-93-25, miR-22, TargetScanS, and the prostate expression profile as tissue filter. A putative oncogenic function was proposed

Pathway	Genes	TS		TS & PT		PT & R	
		E	P	E	P	E	P
LIS1	MAP1B, NDEL1, PAFAH1B1, VLDLR	**1.96**	0.43	**2.55**	0.44		
Cell cycle	CCND1, CCND2, CDC23, CDC27, CDKN1A, CDKN1C, E2F1, E2F3, EP300, RBL1, RBL2, TP53, WEE1, YWHAZ	**0.93**	0.60	**1.35**	0.69	**1.66**	0.67
Endocytosis	CSF1R, DAB2, ERBB3, F2R, LDLR, LDLRAP1, NEDD4L, PDCD6IP, PDGFRA, RAB11FIP1, RAB11FIP5, RAB5B, SMURF1, VPS4B	**0.87**	0.83	0.92	0.69	**2.08**	0.69
PI3K	CALM3, IMPA2, OCRL, PIK3R1, PIP4K2A, PIP4K2B, PIP5K1C, PTEN, SYNJ1	**0.84**	0.19	0.73	0.18	**1.99**	0.43
Prostate cancer	AKT3, CCND1, CDKN1A, CREB3L2, E2F1, E2F3, EP300, MAPK1, PDGFRA, PIK3R1, PTEN, SOS1, TCF7L1, TP53	**0.78**	0.51	**1.03**	0.61	**1.78**	0.50

Table 5.1 continued on next page

Pathway	Genes	TS		TS & PT		PT & R	
		E	P	E	P	E	P
MAPK	AKT3, CDC42, CRK, DUSP10, DUSP5, HSPA8, MAP3K11, MAP3K2, MAP3K5, MAP3K8, MAPK1, MAPK14, MAX, NTRK2, PDGFRA, PPP3R1, RAPGEF2, RPS6KA1, RPS6KA2, RPS6KA4, RPS6KA5, SOS1, TGFBR2, TP53	**0.63**	**0.38**	0.69	0.36	0.26	0.47
Neurotrophin	AKT3, CALM3, CDC42, CRK, IRS2, MAP3K5, MAPK1, MAPK14, NTRK2, NTRK3, PIK3R1, RPS6KA1, RPS6KA2, RPS6KA4, RPS6KA5, SOS1, TP53, YWHAZ	**0.60**	0.29	0.07	0.37	0.23	0.35

Table 5.1 continued on next page

Pathway	Genes	TS		TS & PT		PT & R	
		E	P	E	P	E	P
Actin cytoskeleton	ARPC5, CDC42, ERK, F2R, FGD1, GRLF1, ITGA10, ITGA5, ITGA6, MYH9, PDGFRA, PFN2, PIK3R1, PIP4K2A, PIP4K2B, PIP5K1C, PPP1R12A, SLC9A1, SOS1, TIAM1, VCL	**0.56**	0.52	-0.19	0.41	-0.03	0.50
Long-term	MAPK, PPP2CA	-1.28	**0.09**				
p53	CCNG2, CDKN1A, PTEN, TP53	-0.13	**0.13**	0.51	0.17	0.62	0.17
Circadian	CLOCK, NPAS2	2.32	**0.15**	**3.32**	**0.14**		
Wnt	PPP2CA, PPP3R1, NFAT5, FBXW11, FZD7	0.43	0.60	**0.90**	0.61	0.33	0.67
Nfat	EGR3, RNF128, KPNA2	0.85	0.46	1.43	0.43	**3.31**	0.43
Smad2	KPNA2, RBBP7	-0.45	0.32	0.55	0.32	**2.42**	0.32
Jak-STAT	CCND1, CCND2, STAM2, STAT3, AKT3, SPRY4	0.51	0.21	0.78	0.22	**1.72**	0.23
ErbB	AKT3, CDKN1A, ERBB3	-0.01	0.36	0.24	0.28	0.98	**0.17**

Table 5.1: **Enriched and proximal signaling pathways.** The table shows enriched ($p < 0.05$) and proximal ($p < 0.05$) pathways identified by miTALOS using different prediction tools and the prostate tissue filter. TS, TargetScanS; PT, PicTar; R, RNA22. KEGG disease pathways for tissues other than prostate are omitted. Genes listed target transcripts of miR-106b-25 cluster and miR-22. E shows the enrichment score, P the proximity score. Bold scores are significant ($p < 0.05$).

for the miR-106b-25 cluster and miR-22 in prostate cancer (5; 215). It was found that miR-22 operates as a proto-oncogene in combination with c-MYC (215) and plays an

important role in retardation of tumor cells (289). For cluster miR-106b-25, recent studies proposed an anti-apoptotic role in prostate cancer (69; 120).

We performed a functional analysis with miTALOS using the miR-106b-25 cluster, miR-22, prostate tissue filter, and TargetScanS (see Table 5.1). One feature of miTALOS is the ability to use intersections of miRNA prediction tools that can improve the target gene specificity. We therefore also applied miTALOS using the intersection of TargetScanS and PicTar, which shows a good performance and achieved just slightly less sensitivity than either program individually (238). Furthermore, we used the intersection of two prediction methods (PicTar and RNA22), which are based on different features, to illustrate the scope of miTALOS (for a complete list of identified miRNA-pathway associations see Table 5.1).

Using miTALOS, we obtained a significant enrichment ($p < 0.05$) of miRNA target genes in KEGG's prostate cancer pathway independently by the chosen prediction set. This pathway summarizes key molecular alterations in prostate-cancer in a combined pathway. The result shows that the queried miRNAs have a strong impact on critical components of the phenotype of prostate-cancer. In addition, miTALOS identifies an enrichment of target genes in actin cytoskeleton pathway indicating the association between the queried miRNAs and cell motility in prostate cancer. Cell motility is a critical determinant of prostate cancer metastasis (50). RHO/ROCK kinase induces reorganization of the actin cytoskeletal dynamics in several metastatic tumors (178). Zohrabian et al. (302) showed that a downregulation of ERK leads to increased cell migration. We found ERK and GRLFI targeted by miR-106b-25 indicating the influence of the prostate related miRNAs on the repression of ROCK and therefore the activation of cell migration (see Figure 5.7a). Furthermore, we identify an association between miR106b-25, miR-22 and the MAPK pathway. IL-6 activates prostate cancer cell proliferation via JAK-STAT (274) and MAPK (243) pathways (see Figure 5.7B). Downregulation of AKT and DUSP leads to an activation of the MKK/JNK cascade, which is involved in the tumor growth in prostate cancer (244). MiTALOS identifies inhibitors such as AKT, DUSPs, and MAPKs targeted by miR-106b-25 and miR-22. In addition, MAPK9 is a validated target gene of miR-93 (197). The result of miTALOS shows that central inhibitors of the MAPK related proliferation are under miRNA-mediated repression, which may facilitates tumor proliferation (see Figure 5.7B).

In addition, miTALOS links the queried miRNAs to the cell cycle and Phosphatidylinositol pathways. In prostate cancer, PI3K/AKT signaling cascade is activated to en-

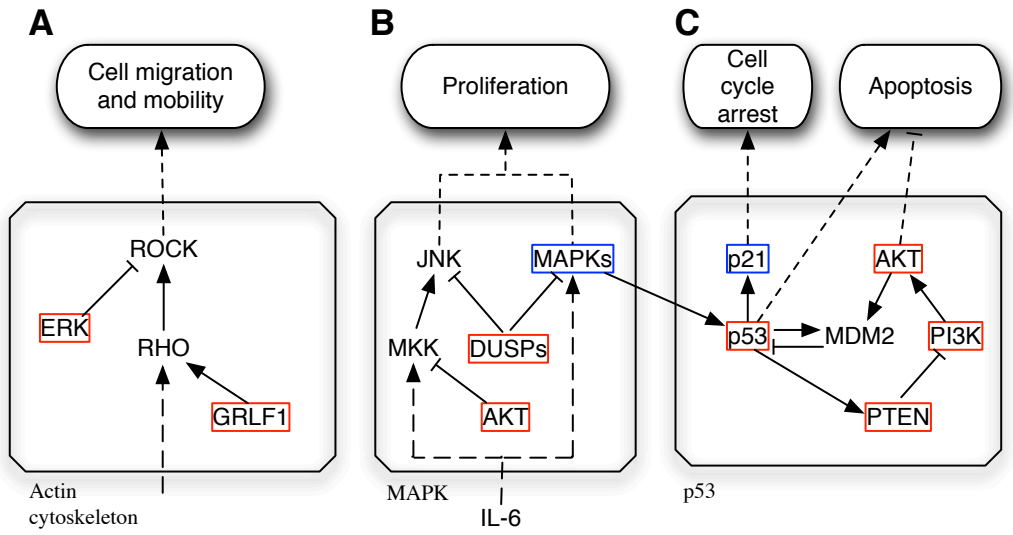

Figure 5.7: **Model for central prostate cancer related processes and their miRNA-mediated regulation.** Framed transcripts in red are predicted targets by miR-106b-25 cluster and/or miR-22. Framed transcripts in blue are validated miRNA target genes. Arrows indicates activation, dashed lines inducement, and blunted arrows inhibition. (**A**) RHO/ROCK (RHO kinase) signaling regulates actin cytoskeletal dynamics in several metastatic tumors (178). ERK/MAPK regulates the actin cytoskeleton and contraction required to drive cell motility, whereas a downregulation of ERK leads to cell migration (302). We found ERK and GRLF1 targeted by miR-106b-25 **(B)** IL-6 mediated cell proliferation via activation of the MAPK pathway. Downregulation of AKT and DUSP leads to an activation of MKK/JNK (62), which is required for the growth of prostate carcinoma (244). We found inhibitors such as AKT and DUSP targeted by miR-106b-25 and miR-22 indicating the oncomir character of the queried miRNAs. (**C**) Activation of the p53 pathway is induced by MAPK. The p53 pathway is actively involved in cell cycle arrests and p53-dependent apoptosis (30; 235). We found central players of cell cycle arrest targeted by miR-106b-25 and miR-22. (108) showed that p21 is a direct target of miR-106b and that its silencing plays a key role in cell cycle progression by modulating checkpoint functions.

sure cell survival and protection against apoptosis (247). It was shown that the Wnt signaling is involved in AKT activation (292), which inhibits angiogenesis and tumor growth (60). Moreover, there is *in vivo* evidence that the ErbB family receptors activate the PI3K/Akt/NF-kB pathway in prostate cancer cells (79). Using miTALOS, we identify these pathways, which are in crosstalk to the PI3K/AKT signaling cascade. In addition, we found PI3K and AKT targeted by the queried miRNAs in the related pathways. This result supports the activation of MAPK-related tumor growth and indicated the role of miR-106b-25 and miR-22 as oncomirs (see Figure 5.7B).

The p53 pathway was only found by our new proximity method ($p < 0.05$). P53 and cell cycle related pathways are active and involved in the lack of cell cycle checkpoint arrests and p53-dependent apoptosis (30; 235). DNA damage results in a sharp increase of p53 protein that, in turn, can enhance cell cycle arrest and apoptosis. We identify central players of cell cycle arrest targeted by miR-106b-25 and miR-22 independently of the prediction tool facilitating the inhibition of cell cycle arrest and apoptosis (see Figure 5.7C). In addition, Ivanovska et al. (108) showed that p21 is a direct target of miR-106b and plays a key role in cell cycle progression.

We summarized the miRNA-mediated regulation on tumor proliferation, mobility and anti-apoptotic behavior of the prostate cancer related miR-106b-25 and miR-22 in a model illustrated in Figure 5.7. The functional analysis and inferred model indicate that the global effect of the up-regulated miRNAs do not only depend on single central target genes but also on the interaction of multiple components in the signaling pathways. We were able to show that the features of miTALOS provide a substantial support to infer miRNA-mediated regulation of signaling pathways in systematical manner.

5.2.8 Functional microRNA-pathway associations

MiRNAs play a pivotal role in the regulation of signal transduction pathways. Different aspects of this control have been analyzed and discussed, e.g. miRNAs as noise-buffering devices in networks (96) or their influence on motif dynamics (288). However, the identification of regulatory miRNAs in pathways has, up to now, been only addressed with the enrichment score, which ignores the underlying topology and is biased to pathways with many targets. In contrast to the enrichment measure, the proximity score explicitly uses the network topology inscribed in the wiring of signaling pathway, with a bias to only few targets. The two measures discern different regulatory patterns and can thus be seen as complementary tools. The functional implications of proximal and distal patterns have to be discussed. For proximal patterns (see Figure 5.2), links between the observed MPAs and functional aspects are well documented in the literature. At the same time, our GO analysis reveals that proximal targets are predominantly related to receptor signal transduction (see Figure 5.4), indicating that this pattern indeed controls signal transduction pathways in a specific, pathway independent manner. Many distal MPAs however appear to be an artifact of the unique network structure of the Olfactory transduction pathway. Interestingly, it

has been shown with a Dicer knockout that miRNAs are dispensable for mature olfactory neurons (39). Our GO analysis (Figure 5.4) reveals that distal targets are mostly associated with the cytoskeleton, a biological process that is not directly related signal transduction, and the same functional cluster emerges for enriched MPAs. Taken together, we assume that the functional implications of distal MPAs, unlike proximal ones, are debatable. Finally, we also analyze pathways with single targets. Theoretical studies show that the regulation of single important proteins can have a dramatic effect on the network dynamics (see, e.g. (8)). From this perspective, it is interesting to realize that MPA with single targets are specifically linked to the 'regulation of apoptosis' (see Figure 5.4). Note that all these results are, due to the HITS CLIP resource, restricted to miRNA control in mouse brain. Our proximity score complements the computational toolbox to identify miRNA-pathway associations. It is an alternative measure to the established enrichment score and based on the hypothesis that a functional relationship is probable if the targets of a miRNA are proximal and linked in a pathway rather than randomly distributed and unrelated. These two measure present two different ways of inferring target-pathway associations. Notably, the enrichment measure is unable to identify pathways with a small number of targets. Vice versa, the proximity measure fails if a pathway contains a large proportion of targets. Refinements of our score would include not just the largest, but also smaller connected component of a signaling network. Moreover, one could account for the direction and the character of the interactions, thus discrimination between, e.g. concatenated activations and more complex target patterns. Finally, our score could be used to boost target prediction by filtering out proximal patterns.

5.3 Material and Methods

5.3.1 MicroRNA data

Human and mouse miRNAs were extracted from the miRBase database (82), which is a collection of published miRNA sequences and annotation. As there is strong evidence that miRNAs can act in concert with each other, miTALOS also provides a list of predefined miRNA clusters. MiRNA clusters are defined as a set of miRNAs, where each member is having at least one other member of the same cluster within 5kb distance according to chromosomal locations. Chromosomal positions of all human and mouse miRNAs were obtained from the miRBase database.

5.3.2 MicroRNA target prediction

The miTALOS web resource uses several target prediction methods to infer miRNA target transcripts: TargetScanS (162), RNA22 (61), PicTar (149), PiTa (127), TargetSpy (255). Sethupathy et al. (238) showed that the intersection of the prediction tools can yield improved specificity with only a marginal decrease in sensitivity relative to any individual algorithm. MiTALOS can handle this issue by generating intersections from at least two prediction methods. Hausser et al. (91) analyzed different features of miRNA targets and showed that TargetScanS has the best performance on different data sets. Therefore, TargetScanS was defined as the default prediction method for miTALOS and used for the case study.

5.3.3 Tissue expression profiles

MiRNA and their corresponding target transcripts show a highly tissue-specific expression pattern. We used the tissue atlas provided by (257) to filter potential miRNA targets in a specific tissue. The human and mouse data was downloaded from the NCBI Gene Expression Omnibus (GEO) and the processed data was used. We mapped the predicted miRNA target transcripts on the tissue atlas and considered a transcript as expressed in a specific tissue, if either one replicate has a present call or both show at least a marginal call, similar to the method used by (186).

5.3.4 Signaling pathways

For the functional analysis of miRNA-pathway associations, miTALOS offers two different resources of signaling pathway. All non-metabolic pathways for human and mouse were integrated from KEGG (121). The KEGG Pathway database is a collection of manually curated pathway maps for various genomes. For the analysis of human miRNA regulation, we also included signaling pathway information from NCI PID (230). NCI PID is a manual collection of biomolecular interactions and key cellular processes assembled into signaling pathways. The database is curated by Nature Publishing Group editors and reviewed by experts in the field.

5.3.5 Enrichment score

The identification of miRNA-pathway associations by miTALOS is obtained using two different approaches. A first approach use the number of target genes in a specific pathway to calculate an enrichment score. Here, we assume that miRNAs target specific pathways to influence specific functions of the cell by the sheer number of target transcript. Calculating the enrichment of targets T_{Pi} in a pathway i with P_i proteins leads to an enrichment score E, which has been used in previous studies:

$$E = \frac{T_{Pi}/P_i}{T_P/P}.$$

where T_{Pi} is the number of targets in pathway i, P_i is the number of all proteins in pathway i, T_P is the number of all targets in all pathways and P the number of all protein in the KEGG or NCI PID pathways. The significance is obtained by Fisher's exact test (66), corrected by the Benjamini-Hochberg procedure (13).

5.3.6 Proximity score

We assume that some miRNAs target signaling cascades in a proximal manner. To reveal pathways with proteins that function in a proximal manner and are targeted by the same miRNA, we introduce a proximity measure P. Let us consider a pathway i with P_i proteins and T_{Pi} targets of a specific miRNA, we can determine the distances d_{xy} between all $T_{Pi}(T_{Pi}-1)/2$ pairs of targets x, y $(x \neq y)$ in the corresponding signaling pathway. To condense this set of distances for a miRNA pathway pair into a real number in $[0, 1]$, we calculate the minimal distance for each target x. The proximity score P is then defined as the mean of all minimal distances $< d_{xy} >$ as the power of base α. The proximity score P is the defined as:

$$P = 1 - < \alpha^{-d_{xy}} >_{xy} .$$

The base α can be chosen appropriately to ensure a reasonable separation of the distances occurring in the network. Based on the observed distance scores, we chose α = 1.1. In order to obtain significant miRNA-pathway associations we perform random sampling. For each pathway i and specific number of targets, we randomly choose 10.000 times miRNA targets and calculate the corresponding proximity scores P. These samplings are then used to calculate the p-values by counting the number of

proximity scores that are less than the original score divided by 10.000. Final p-values were then corrected by the Benjamini-Hochberg procedure.

5.4 Conclusions and Outlook

Within this Chapter, we link miRNAs and signaling pathways with a novel proximity measure going beyond the common enrichment approach by incorporation the topology of the underlying network. Applying the proximity score to a global set of experimentally validated miRNA targets, we identify miRNA-pathway associations that differ from those inferred with the conventionally used enrichment score. This finding indicates the existence of additional subclasses of miRNA pathway associations in addition to the enrichment of miRNA target pattern. A gene ontology analysis reveals that proximal target patterns correspond to a specific function in cell signaling. Summarizing, the application of concepts from graph theory to signal transduction allows the identification of novel miRNA-pathway associations. Furthermore, we presented the miTALOS web server that provides novel features for the functional analysis of miRNA-mediated regulation in biological pathways. MiTALOS offers two different methods and pathway resources to identify signaling pathways altered by the expression of miRNAs. The two measures provide significant miRNA-pathway associations for two alternative forms of miRNA control. As miRNAs and their target genes show highly tissue-specific expression signatures, miTALOS provides a tissue filter. This is a novel feature in contrast to already existing resources, where the functional analysis is corrupted by targets that are not expressed in the tissue under consideration. In a functional analysis of prostate cancer related miRNAs, we showed the benefit of the novel features to identify biological meaningful miRNA-pathway associations. Given the increasing amount of evidence that miRNAs have an important impact on signaling pathway regulation, miTALOS provides a substantial support to infer systematical insights of miRNA-mediated regulation. More generally, we think that the concept of proximity can serve as a powerful tool to identify patterns in networks beyond miRNA regulation in signal transduction. For drug targets in metabolic networks or disease genes in signaling pathways, our tool might generate useful hypothesis beyond the commonly used enrichment method. In the following chapter, we study the regulatory role of miRNAs not from a large-scale point of view but rather analyze the impact of miRNAs on the the dynamic of phosphorylation processes in signal pathways.

6 Mathematical models of microRNA-mediated regulation in signaling pathways

6.1 Background

Inhibition of miRNA biogenesis clearly reveals that miRNAs are essential for diverse cellular processes ranging from proliferative, differentiation to apoptosis (138). Various studies showed that miRNA expression and maturation is induced by signaling pathways (48; 202; 259), more importantly it was shown that miRNAs emerge as regulators of signaling proteins. In zebrafish, miR-9 has been shown to regulate several components of the FGF signaling pathway, and thus controls neurogenesis in the midbrain-hindbrain domain during late embryonic development (160). In another recent example, Ivanovska and colleagues (109) showed that miR-106b has a severe impact on the tumor cell proliferation by targeting p21. In fruit fly, miR-8 target both a transmembrane protein and a transcription factor of the WNT signaling pathway and hereby antagonizes the pathway at multiple levels (125). The work of Ricarte-Filho et al. (220) indicated that let-7 inhibits the activation of the RET/PTC-RAS-BRAF-ERK cascade exemplifying direct influence of a single miRNA on a pathway cascade.

Despite these studies, the functions of miRNAs within cells remain largely uncharted. Uncovering the function of individual miRNAs is challenging: First, each miRNA has numerous targets that have diverse functions. Second, miRNAs have a strong impact on the signal transduction, the degree of downregulation often tends to be quantitatively modest (106). A miRNA typically down-regulates most of its target transcripts by less than 50% (8). This consideration suggests that although many genes are predicted to be miRNA targets, only a fraction of these interactions may have an

impact on biological responses and phenotypes (164).

In this chapter, we analyze the impact of miRNAs on pathway dynamics studying two different models. In a first model, we use a signaling cascade to study the differences between a system without miRNA regulation to a pathway cascade targeted by miRNAs. Altering the pathway readout from a steady state level, we are able to show that miRNAs decrease the time of dimming the signal, where the recovery time of the systems is unaltered. In further analysis, we adapt this cascade to a full model of the gp130-STAT3 pathway by integrating receptor activation as well as negative feedback. We observe time-resolved data after IL-6 stimulation in primary mouse hepatocytes. Using this time-series data, we study the phospho-dynamics as well as the impact of miRNAs on the pSTAT3/STAT3 ratio in cytoplasm. Therefore, we analyze the effect of known miRNA regulation on JAK1 and STAT3. We are able to show that a model integrating miRNA influence on the pathway results in reliable turnover rates compared to model without miRNAs. Using the miRNA model, we show that a pre-induced decrease of STAT3 changes the overall ratio of pSTAT3/STAT3 in the cell. Moreover, this effect results in a time shift of the maximal pSTAT3 concentration in cytoplasm. Finally, the model reveals that induced miRNAs based on active pSTAT3 have no influence on the pathway dynamics in primary hepatocytes based on IL-6 stimulation.

6.2 Results and Discussion

6.2.1 Dynamical modeling of microRNA-mediated regulation on protein phosphorylation

The main objective of this section is to study the post-transcriptional regulation of signaling cascades by miRNAs. In order to get a first understanding of the miRNA regulation on signaling transduction pathways, we consider a signaling cascade. Stimulation of a receptor leads to the consecutive activation of a protein kinase (see Figure 6.1). The signal output of this activation is the phosphorylated kinase, which in turn activates the readout protein that elicits a cellular response (e.g., activation of a transcription factor). The signal is terminated by an inhibitor, which inhibits the activation in a non-competitive manner. This general scheme is representative of many signaling pathways, for example, growth factors such as PDGF, EGF or IL-6, which lead to activation of many cytokine receptor systems, regulate growth, survival and

differentiation. Signals can be terminated by protein-tyrosine, serine-threonine phosphatases or degradation. In contrast to expressed phosphatases, proteins such as SOCS family are transcriptionally induced negative regulators. The SOCS protein family includes SOCS1-7 and the cytokine inducible SH2 domain-containing protein (CIS). After cytokine stimulation, the expression of SOCS/CIS genes is rapidly induced via the gp130-STAT3 pathway displaying a classical example of negative feedback loops (275).

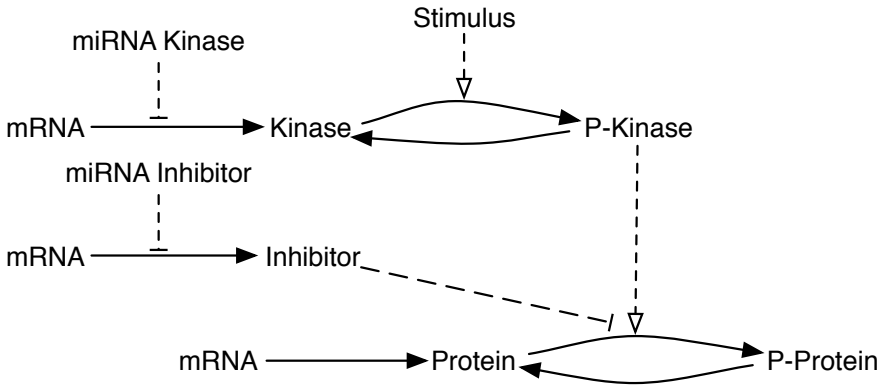

Figure 6.1: **General model of miRNA regulation in a phosphorylation cascade**. A stimulation of a receptor leads to the consecutive activation of a down-stream protein kinase, which in turn can phosphorylate a protein that elicits a cellular response, e.g. activation of a transcription factor. The signal is terminated by an inhibitor, which inhibits the activation in a non-competitive manner. MiRNAs inhibit the translation process of the kinase and inhibitor by decreasing the mRNA levels. In order to study the differences in gene and miRNA-mediated regulation, we shutdown the pathway signal either by upregulating the miRNA Kinase (Kin down miRNA) or downregulating the miRNA Inhibitor (Inh down miRNA). Altering the mRNA levels, we shut down the signal by increasing the inhibitor (Ind down) or decreasing the kinase (Kin down).

The schematic representation (Figure 6.1) is transferred into an ordinary differential equation (ODE) model following mass action kinetics mostly. For model simplification, we use Michaelis-Menten kinetics for all phosphorylation reactions (see Material and Methods for a detailed description). We randomly select turnover rates for all mRNAs and proteins between 1h and 24h, as well as between 8h and 24h for miRNAs regulating the kinase and inhibitors to cover the common range of signaling-related mRNAs, proteins and miRNAs (15; 159; 187). Obtaining a set of turnover rates, we first determine the steady-state of the system, in a next step we then shut down the pathway signal to 50% either by altering the kinase or the inhibitor concentration. Fi-

nally, we recover the system by turning off the introduced regulation on the kinase or inhibitor. To compare both scenarios either of gene regulation or miRNA regulation we calculate the time to shutdown the signal to 75% (shutdown time). Based on the resulting new steady-state level of 50% we calculate then the recovery time of the signal to reach again a signal strength of 75% (see Figure 6.2).

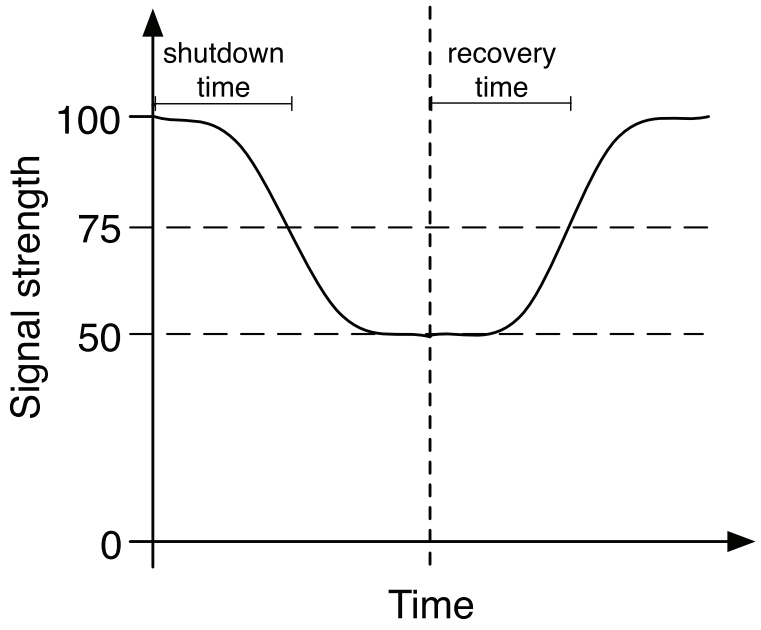

Figure 6.2: **Illustration of the pathway shutdown and recovery.** Starting from a steady-state level at 100% of the pathway signal, we reduce the signal strength to a new steady state level at 50% by either alter the mRNA or miRNA expression of the kinase or inhibitor. The shutdown time is defined by the time until the signal is reduced to 75%. Recover the pathway signal from its new steady-state level at 50%, we measure then the time until the signal reach again 75% of its original strength and define this as recovery time.

MicroRNAs align the pathway shutdown speed

To analyze the impact of miRNAs on the pathway signal, we compare the signaling cascade regulated by miRNAs with the same cascade by changing the gene production of the kinase and inhibitor. To study the influence of the turnover rate, we sample 10.000 different turnover rates for mRNA and proteins. After determination of the initial steady-state of the signal cascade for each set of turnover rates, we either alter the kinase or inhibitor by changing the gene production or miRNA impact to shutdown the pathway signal to 50% (see Figure 6.2). Figure 6.3 shows the obtained shutdown

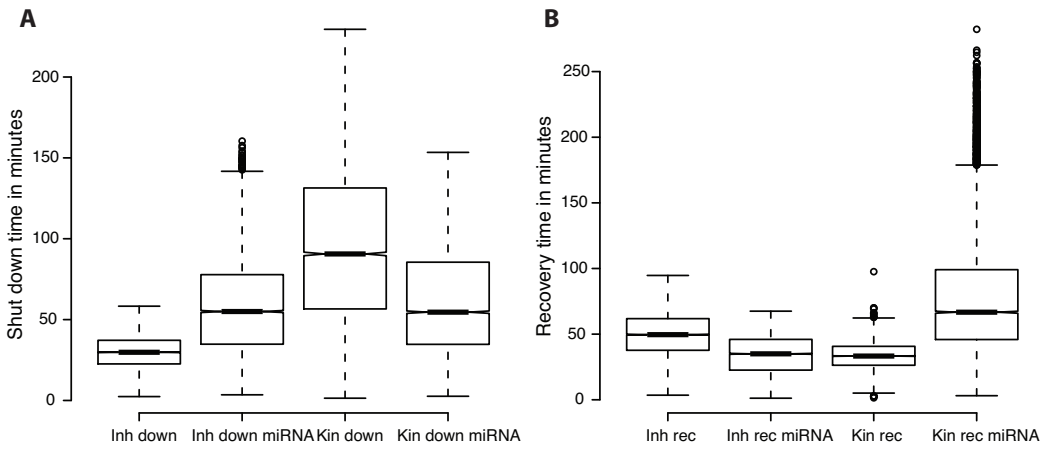

Figure 6.3: **Shutdown and recovery time of a pathway signal**. (**A**) Shutdown time of the pathway signal either by regulating the inhibitor via mRNA expression (Inh down) or miRNA (Inh down miRNA) or the kinase via mRNA expression (Kin down) or miRNA (Kin down miRNA). (**B**) Recovery time of the pathway signal either by regulating the inhibitor via gene regulation (Inh rec) or miRNA (Inh rec miRNA) or the kinase via gene regulation (Kin rec) or miRNA (Kin rec miRNA).

time for the four different regulation mechanisms. The result shows that the shutdown time for the pathway by altering the inhibitor (Inh down) is much faster compared to the kinase (Kin down). This result is in line with a previous work (159), which shows the same effect for a single protein phosphorylation. On average, the shutdown is three times slower via the kinase compared to the inhibitor. This strong difference could be explained by the different regulation. To shutdown the signal via the kinase, the kinase has to be degraded whereas the inhibitor concentration has to be increased. Therefore, the time difference can be explained by an in general faster production process.

Comparing the differences for kinase and inhibitor regulated by miRNA, we show that both mechanisms result in similar shutdown speed (average difference less than four minutes). To shut down the signal via the inhibitor either the inhibitor concentration has to be increased or the miRNA concentration has to be decreased. Interesting, the fast production speed compared to the very slow miRNA degradation has no effect on the shutdown time. These results show that altering the pathway signal via the kinase or the inhibitor has an extreme effect on the response time of the pathway. Moreover, we are able to show that altering the pathway by miRNA-mediated regulation clear this time difference. One explanation could be that differences in the steady-state level are responsible for a fast shutdown even via miRNA degradation. Another explanation could be some dependencies in the turnover rates. Analyzing the

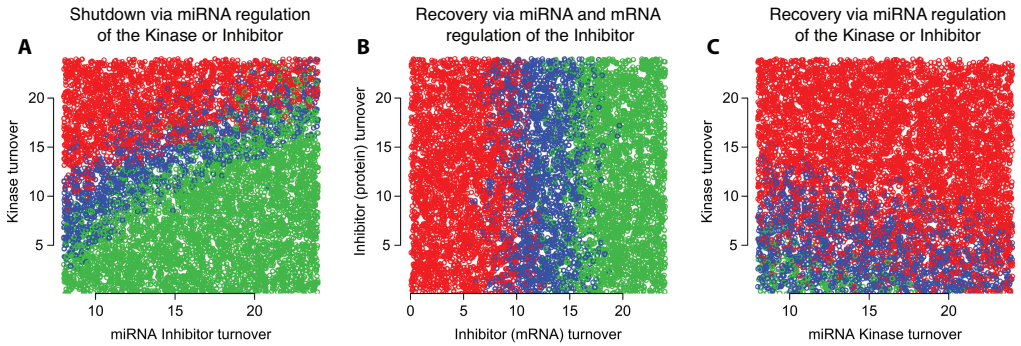

Figure 6.4: **Correlations between parameter**. (**A**) Correlation between miRNA turnover of the Inhibitor and the Kinase. The blue color indicates a time difference in the shutdown via miRNA regulation of the inhibitor or the kinase by less than 10 minutes, whereas red shows a faster shutdown via the inhibitor and green via the kinase, respectively. (**B**) Correlation between the protein and mRNA turnover rate of the inhibitor. The blue color indicates a difference in the recovery time of the system via the miRNA or gene regulation of the inhibitor between 10 and 20 minutes. Red indicates less than 10 and green more than 20 minutes, respectively. (**C**) Correlation between the miRNA and protein turnover rate of the kinase. The blue color shows a difference in the recovery time of less than 15 minutes between gene or miRNA regulation, whereas red indicates a time delay via gene regulation of more than 15 minutes and green a delay via miRNA, respectively. The turnover rate is given in hours.

turnover rates in respect to the time differences between a shutdown via miRNA regulation of the kinase or inhibitor, we identify a correlation between the turnover rate of the inhibitor's miRNA and kinase protein (see Figure 6.4A). The blue color indicates a time difference of less than 10 minutes, whereas a faster shutdown via the inhibitor is indicated in red and green via the kinase, respectively. We obtain a similar correlation between the turnover rate of the inhibitor's miRNA and other turnover rates indicating that a fast miRNA turnover allows a fast shutdown via the inhibitor.

Pathway recovery is unaffected by microRNA regulation

After reducing the pathway signal to 50%, we measure the recovery time of the pathway to reach again 75% of the steady state level. Figure 6.3B shows the average recovery time for the four different regulation mechanisms. One result is that compared to the shutdown behavior, the recovery time showed a more homogenous pattern. We obtain almost no time difference between a system altered either by miRNA regulation of the inhibitor or a gene regulation of the kinase. Moreover, the time average differ-

ence in the recovery time between miRNA or gene regulation of the inhibitor is around 15 minutes. Figure 6.4B shows the turnover rate for the mRNA and protein turnover of the inhibitor. We are able to show that a fast degradation of the mRNA leads to a faster recovery time via gene regulation, whereas a turnover rate of > 15 hours indicates a faster regulation via miRNAs.

We obtain the largest time difference of the recovery time between miRNA regulation of the inhibitor and kinase (see Figure 6.3B). Analyzing the turnover parameter, we identify a strong dependency of the protein kinase turnover rate. Figure 6.4C shows the correlation for this parameter with the kinase miRNA turnover, which is representative for the other turnover rates. Blue indicates a difference in the recovery time of less than 10 minutes, whereas red indicates a delay via miRNA regulation of the kinase and green via the inhibitor, respectively. A fast protein kinase turnover results in no time differences in the recovery time, whereas a more stable protein leads to a time delay in the recovery.

In summary, we are able to show that alteration in the concentration of signal transduction proteins and transcripts via gene or miRNA regulation has a severe impact on the signal maintenance and shutdown. Changes in the intracellular pool of signal transduction activators or inhibitors lead either to a fast signal shutdown via the inhibitor or a slow signal repression via the activator. We are able to show that an increase in miRNA concentration also leads to a shutdown but we are not able to identify any difference in the shutdown time. This result indicates that a context dependent regulation of either the signal transduction activator or inhibitor has a strong impact on changes in the signal transduction. Moreover, we show that gene or miRNA regulation lead to similar recovery times. We identify specific turnover rate ranges, which lead to a context specific increase or delay in the recovery time. This analysis shows that miRNA regulation an addition layer of transcriptional control allows the cell to respond in context specific manner. In the next section, we extended this signaling cascade to a whole pathway model for further analysis of these mechanisms in the context of signal dynamic and maintenance.

6.2.2 Mathematical modeling of gp130-STAT3 signaling including microRNA regulation

The gp130-STAT3 pathway is triggered by several different ligands and their corresponding receptors. Beside, growth hormones, the major activators of the gp130-

STAT3 pathway are members of different cytokine families including INF, IL-6 or EPO (232). In primary mouse hepatocytes, IL-6 mediates two major responses. First, hepatocytes start to produce acute phase proteins upon infection-associated inflammation. These proteins include complement factors to destroy or inhibit growth of microbes. Second, IL-6 promotes liver regeneration and protects against liver injury [19]. Upon binding to its cell surface receptor, IL-6 activates the receptor associated Janus tyrosine kinase (JAK) 1 and signal transducer and activator of transcription (STAT) 3. The latent transcription factor STAT3 is phosphorylated and translocated to the nucleus after activation. Subsequently STAT3 alters gene expression and activates the negative feedback protein SOCS, which leads to a shutdown of the signal at receptor level.

We translate the gp130-STAT3 pathway, schematized in Figure 6.5, into an ODE model using mass action kinetics mostly. For model simplification, we used Michaelis-Menten kinetics for all phosphorylation and dephosphorylation processes (see Material and Methods for a detailed description). We introduce miRNA regulation into this system by modeling the mRNA:miRNA interaction for JAK1, STAT3 and SOCS3. Free miRNA in the cytoplasm can bind to free mRNA and therefore form the mRNA:miRNA complex. Due to technical limitation to resolve the effect of each single miRNA on these mRNAs, we model the miRNA regulation by combining miRNAs to a whole miRNA regulation process.

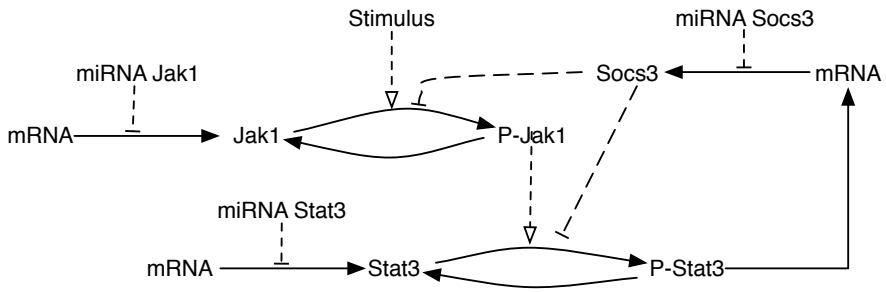

Figure 6.5: **Model of the gp130-STAT3 pathway**. Stimulation of the gp130 receptor with IL-6 activates JAK1. The active JAK1 in form of phosphorylated JAK1 (pJAK1) leads to an activation of STAT3. The active transcription factor STAT3 (pSTAT3) is transported into the nucleus and altered gene expression and activates the negative feedback protein SOCS3. MiRNAs inhibit the translation process of the JAK1, STAT3, and SOCS3 by decreasing the corresponding mRNA levels.

Primary mouse hepatocytes were stimulated with 1nm lL-6 and protein and phosphoprotein concentration for gp130, JAK1, STAT3 and SOCS3 were measured by quantitative immunoblotting, normalized using calibrator or normalizer proteins as described

in (231). These measurements were done in the group of Dr. Klingmümller at the DKFZ in Heidelberg. MRNA concentrations are measured by RT-PCR for mJAK1, mSTAT3 and mSOCS3. Finally, we used Illumina next-generation sequencing to measure the miRNA concentration in hepatocytes before and after IL-6 stimulation. These data were used to calibrate the model and estimate kinetic parameters. Parameters distributions were estimated using a Markov chain Monte Carlo (MCMC) approach (222). For each measurement, we assume a normal-distributed error, whereas deviations of the measured data were fitted within the model (see Material and Method).

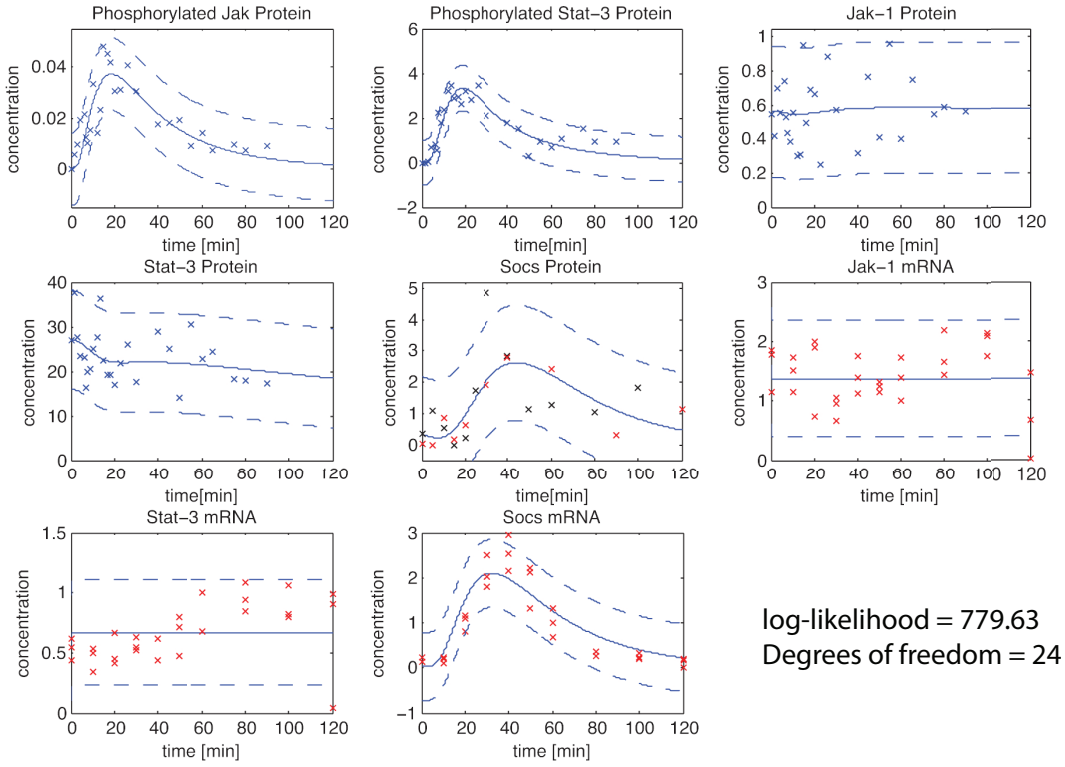

Figure 6.6: **Best fit of the model without miRNA influence**. Predicted time-course of the gp130-STAT pathway proteins and mRNAs obtained by the pathway model without miRNA influence. The solid line illustrate the best fit and the dashed line show the two-times standard-deviation of the estimated experimental error.

Influence of microRNAs on mRNA turnover rates

To analyze the effect of miRNAs on the signal dynamics in gp130-STAT3 signaling, we first fit a model without miRNA regulation to the data. In this model, degradation

of mRNA is only modeled by a single turnover rate, which summarize all degradation processes, e.g. decay or even miRNA effects. We estimated parameters using an MCMC approach to fit the model to the experimental data with independent random initial parameter guesses. Figure 6.6 shows the best fit obtained by this model. The gp130-STAT3 model without miRNAs is sufficient in describing the data but estimated turnover rates for mJAK1, mSTAT3 are less than 1 minute, which not supports previous work (87). For mSOCS3 we obtain a turnover rate of 7.61 minutes, which is in line based with previous work (285) (see Table 6.1 for all turnover rates obtained by both models).

Protein/gene	Model	Half-life (Model)	Half-life (Literature)
PJAK1	without	0.86 h	∼3.2 h (248)
	with miRNA	8.4 h	
JAK1	without	0.39 h	∼3.2 h (248)
	with miRNA	4.3 h	
mJAK1	without	3.3^{-5} h	∼10.3 h (239)
	with miRNA	11.5 h	
PSTAT3	without	2.6 h	∼4.0 h (129; 248)
	with miRNA	3.4 h	
STAT3	without	0.6 h	∼4.0 h (129; 248)
	with miRNA	3.0 h	
mSTAT3	without	3.6^{-5} h	∼7.1 h (239)
	with miRNA	2.4 h	
SOCS3	without	0.14 h	∼0.25 h (285)
	with miRNA	0.16 h	
mSOCS3	without	0.15 h	∼0.25 h (285)
	with miRNA	0.24 h	

Table 6.1: **Estimated half-life of proteins and mRNAs based on the two gp130-STAT3 models.** Protein/gene defines the corresponding protein or gene within the gp130-STAT3 pathway. The capital P defines the phosphorylated protein and the m in lower case the mRNA transcript. *Without* defines the model without miRNA regulation and *with miRNA* the miRNA-extended model. Half-life shows the turnover rates in hours.

The resulting kinetic parameters indicate that a single turnover rate is not able to capture the degradation processes in a biological manner. One explanation could be that the time-course behavior favors a rapid degradation process to well fit the model to the data. On the other side, a more complex turnover using miRNA and mRNA degradation could be able to capture this issue. Therefore, we integrated miRNA-

mediated mRNA degradations into the model. MiRNA influence is modeled by a production, turnover, and mRNA:miRNA forming process. Moreover, we model an mRNA:miRNA complex, which decrease free mRNA and miRNA in the cytoplasm. We do not model mRNA:miRNA release, as this process is stable and on a large time-scale compared to the observed time points (169; 183). Obviously, mRNA transcripts have several different miRNA target sites and could be therefore regulated by different miRNA simultaneously. As we do not know the exact number of miRNAs regulating the different gp130-STAT3 transcript, we summarize different miRNAs for a specific transcript into a single miRNA term.

The miRNA-extended model was fitted to the same time-series data as the model without miRNA regulation. Figure 6.7 shows the best-fit for the miRNA-extended model. To compare both models (without miRNA and miRNA-extend), we use the log-likelihood ratio test (D=0.042, $p > 0.99$). The result shows that the smaller model without miRNA regulation is not significant better than our miRNA-extended model, Furthermore, we are able to capture the dynamic moments of pJAK1, pSTAT3 and SOCS3, as well as the expression pattern of JAK1 and STAT3 using the extended model. The model indicates that IL-6 promotes a rapid increase of pJAK1 and pSTAT3. JAK1 and mRNA expression of JAK1 (mJAK1) show no increase in expression, whereas we obtain an increase in expression for mSTAT3, but a constant level for STAT3. Phosphorylated STAT3 is transported into the nucleus and activates mRNA expression of SOCS3 (mSOCS3). The model fits the delayed activation pattern of mSOCS3, which results in a similar pattern for SOCS3. This transmission of activation is supported by a rapid protein and mRNA turnover of < 10 minutes, which we obtain by our model (see Table 6.1). The obtained turnover rates for JAK1 (> 260 min) and STAT3 (> 180 min), respectively for the miRNA-extended model. These predicted turnover rates are in line with previous work (129; 248) and show the improvement of the miRNA-extended model. As we model the miRNA-mediated regulation as the total miRNA influence, we are not able to infer the parameter directly from miRNA data. To obtain miRNA expression, we fit the initial expression of miRNAs within each run. The model predicts a constant expression pattern for JAK1-related miRNAs and a slight decrease of STAT3-related miRNAs. For mSOCS3, the model predicts no miRNA regulation by setting the mRNA:miRNA binding parameter c_8 to 0.

MiRNA expression profiles confirm model predictions

In order to confirm the predicted miRNA expression levels, Illumina next-generation sequencing was used to obtain a time-course profile of expressed miRNAs. The obtained expression profiles confirm three findings of the miRNA-extended model: (*i*) mJAK1 and mSTAT3 are under miRNA regulation, (*ii*) experimentally verified miRNA-target interactions show the predicted time-course expression pattern, predicted by the miRNA-extended model and (*iii*) we found no expressed miRNA targeting mSOCS3. In hepatocytes the turnover rate of mSOCS3 is < 10 minutes. Moreover, the typical activation time of IL-6 is less than 1 hour. Therefore, one can argue that neither a gene regulation via changes in transcription factor expression nor a miRNA expression change is as slow to have a severe impact on SOCS3 expression during this time. Although, a rapid turnover seems to be the optimal regulation process for a rapid feedback loop such as SOCS3 to shut down the intracellular signal.

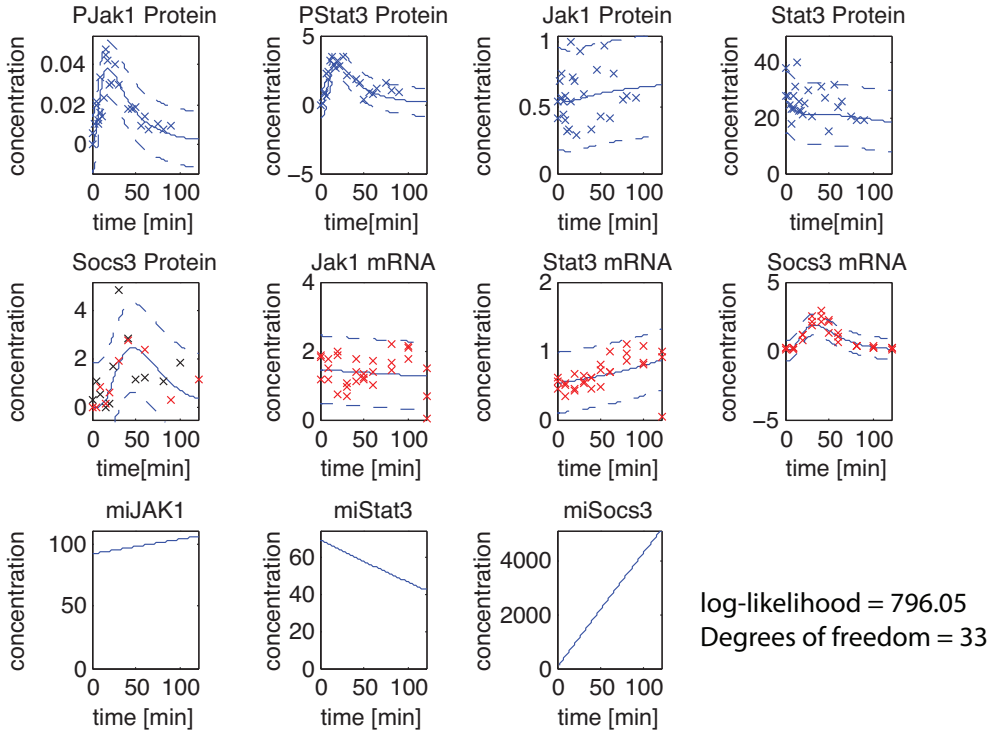

Figure 6.7: **Best fit of the miRNA-extended model**. Predicted time-course of the gp130-STAT pathway proteins, mRNAs and miRNAs obtained by the miRNA-extended model. The solid line illustrate the best fit and the dashed line show the two-times standard-deviation of the estimated experimental error.

Using the set of highly expressed miRNAs, we identified several experimentally validated interactions with STAT3. The miRNA expression profiles for STAT3 related miRNAs show a constant expression pattern with a decrease after 1h, which is in line with our model predictions. Based on these profiles, we can exclude a negative feedback mechanism accomplished by miRNAs, but rather miRNAs could be responsible to prevent expression over-bursts of mSTAT3. This mechanism was recently described as a common function for miRNAs (204). JAK1 shows a similar expression profile as STAT3. Using the set of experimentally validated miRNA interaction, we identified several miRNAs, which target both Jak1 and Stat3. This finding indicates that JAK1 and STAT3 are regulated by the same miRNA. In future, we will alter the miRNA expression levels of these miRNA candidates to study the resulting changes in JAK1 and STAT3 expression in vitro, as well as the resulting effects on the pathway dynamic. So far, our miRNA extended model confirms the measured miRNA expression and identified mRNA and protein turnover rates, which comply with literature data. In the following, we will use this model to study the impact of changing miRNA expression on the phospho-dynamic after IL-6 stimulation in mouse hepatocytes.

6.2.3 JAK1-related microRNAs have a severe impact on signal dynamic and strength

To study the impact of JAK1-related miRNAs on the pathway dynamic, we analyze two scenarios: Case (*i*) JAK1-related miRNAs are expressed as a feedback of active STAT3, which results in a downregulation of JAK1; Case (*ii*) alteration in the expression profile of pre-induced miRNAs results in changes in JAK1 and STAT3 expression (see Figure 6.8). In a recent work Dai and coworkers (45) showed that miR-17 is induced by STAT3, which is among the highly expressed miRNAs in primary hepatocytes. In order to analyze the effect of STAT3 induced miRNAs, we altered the miRNA-extended model by linking the production rate to the pSTAT3 concentration.

Figure 6.7 shows the result of the best fit and indicates that a PSTAT-coupled increase of JAK1-related miRNAs has no effect on the phosphorylation of pJAK1 and pSTAT3. As expected, we see a decrease in expression of mJAK1 based on strong increase of miRNAs during 20 to 40 minutes. An explanation for the stable phosphorylation process is its speed. Even a fast miRNA regulation process is too slow to have a severe impact on the phospho-cascade. Moreover, we observe a phosphorylation peak around 20 minutes, whereas pSTAT3 activated feedbacks like SOCS3 or JAK1

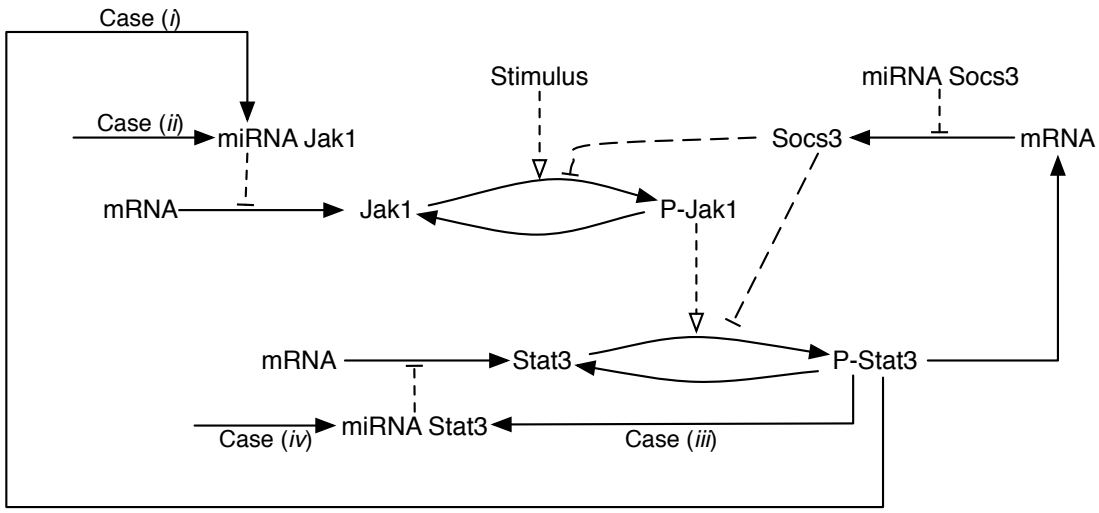

Figure 6.8: **Modification of the gp130-STAT3 pathway**. In order to study various cases of miRNA-mediated regulation, we modify the original miRNA-extended gp130-STAT3 pathway model. Case (*i*) illustrates the linkage of miRNA Jak1 expression to active P-Stat3. For case (*ii*) we increase the miRNA Jak1 expression before IL-6 stimulation. Case (*iii*) illustrates the linkage of miRNA STAT3 expression to active P-Stat3 therefore describing a direct positive feedback loop, whereas we increase the miRNA Stat3 expression before IL-6 stimulation for case (*iv*).

miRNAs have an expression peak around 30 to 40 minutes. This activation pattern is too late to alter the phosphorylation of pJAK1 and pSTAT3. In summary, we were able to show that JAK1-related miRNA activated by the transcription factor STAT3 have no impact on the phosphorylation of pJAK1 and pSTAT3 during the IL-6 stimulation. In a further analysis, we study the effect of pre-induced miRNAs. Therefore, we use the miRNA-extended model and decrease the expression of mJAK1 and JAK1 at time point 0h. This procedure simulates an increase of miRNA expression levels before IL-6 stimulation. We decrease the expression of mJAK1 and JAK1 to 30% of the original initial condition. The resulting simulation of the miRNA-extended model indicates a decrease of pJAK1 and pSTAT3 expression of 37% (see Figure 6.9). In addition, we observe a time delay in the activation peak for pJAK1 and pSTAT3 of 13%. This result indicates that a decrease in mRNA expression, induced by miRNAs, leads to a decrease of pJAK1 expression in a similar amount. Moreover, this illustrates that the saturation level of pJAK1 is already reached in the normal system and alteration of JAK1 has therefore a severe impact of the signal dynamic and strength.

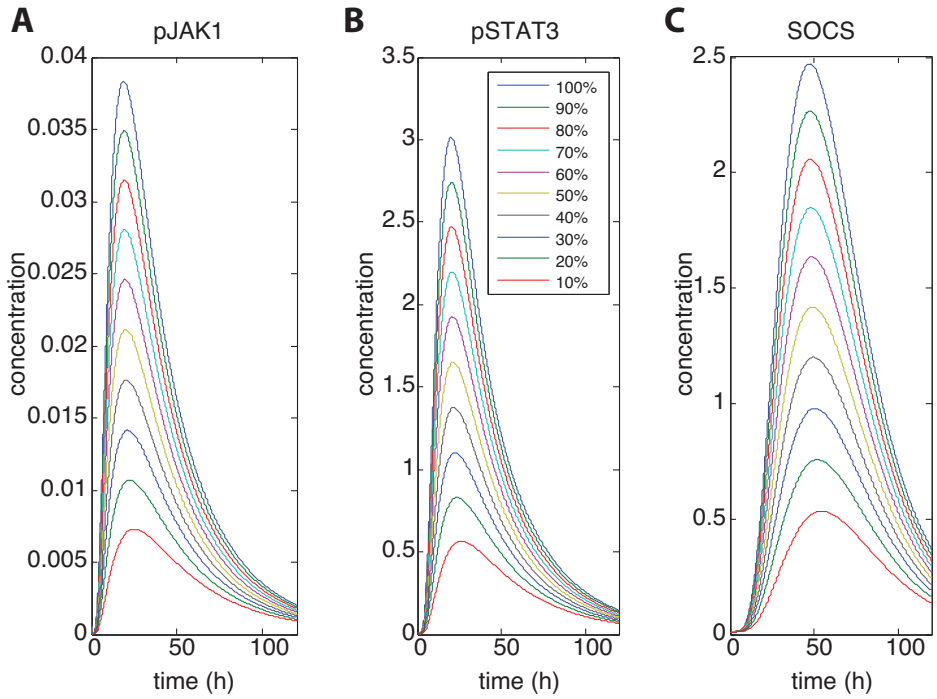

Figure 6.9: **Impact on the signal dynamic by altering JAK1 expression**. We iteratively alter JAK1 protein and mRNA expression before the IL-6 stimulation. (**A**) shows the pJAK1 expression by setting JAK1 mRNA and protein from 100% to 10% of the initial concentration. (**B**) shows the pSTAT3 expression and (**C**) SOCS protein expression predicted by the miRNA-extended model. The x-axis shows the time in hours and the y-axis the concentration of the corresponding protein.

6.2.4 Simulation of STAT3-related microRNAs identify a low level of activated STAT3

To study the impact of STAT3-related miRNA, we again study two scenarios: Case (*iii*), miRNAs regulating mSTAT3 are expressed as a feedback of active STAT3, which results in a downregulation of STAT3. Case (*iv*) alteration in the expression profile of pre-induced miRNAs results in changes in STAT3 expression. To analyze the effect described in case (*iii*) we modified the miRNA-extended model by linking the production rate to the pSTAT3 concentration (see Figure 6.8). This case study confirms the findings for JAK1-related miRNAs. Again, we do not observe any effect on the phosphorylation of pJAK1 and pSTAT3. As expected, we also see a decrease in expression of mSTAT3 based on strong increase of STAT3 miRNAs during 20 to 40 minutes.

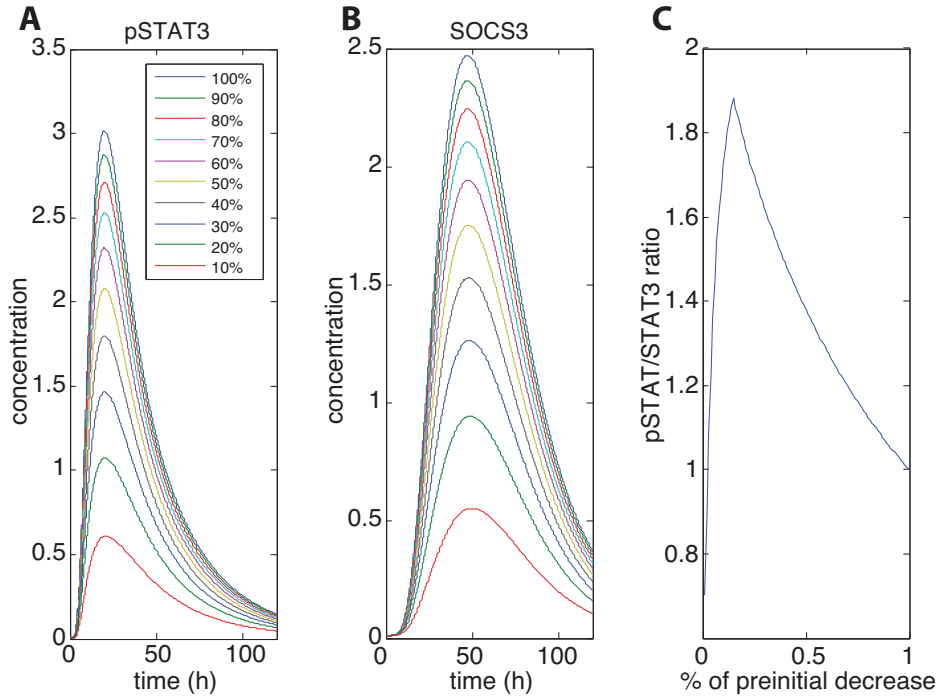

Figure 6.10: **Impact on the signal dynamic by altering STAT3 expression.** We iteratively alter STAT3 protein and mRNA expression before the IL-6 stimulation. (**A**) shows the pSTAT3 expression by setting STAT3 mRNA and protein from 100% to 10% of the initial concentration. (**B**) shows the SOCS3 protein expression. The x-axis shows the time in hours and the y-axis the concentration of the corresponding protein. (**C**) Resulting pSTAT3/STAT3 ratio for setting STAT3 mRNA and protein from 100% to 1% of the initial concentration (x-axis). Y-axis shows the relative ratio compared to the original pSTAT3/STAT3 ratio.

These findings complete the overall result that an increase in miRNA expression induced by pSTAT3 has no effect on the phosphorylation processes in IL-6 dependent activation of gp130-STAT3 in mouse hepatocytes. As we detected several STAT3-related miRNAs in the expression profiles of mouse hepatocytes, we next study the influence of pre-induced miRNAs. Again, we use the miRNA-extended model and decrease the expression of mSTAT3 and STAT3 to 30% of the original initial condition at time point 0h. The simulation of the miRNA-extended model indicates a decrease of expression of 49% for pSTAT3 (see Figure 6.10A). SOCS3 has half of its expression level compared to an unaltered system (see Figure 6.10B). Compared to the JAK1 inference by miRNAs, we obtain no time delay for pSTAT3 and only a small delay of 3% for SOCS3. The results show that the amount of downregulation of STAT3 before stimulation is not complete transferred after the stimuli. A combination of two factors can

be promote this outcome: first, the saturation level of pSTAT3 in the original system is not reached and second, the pJAK1 concentration is not altered by STAT3-related miRNAs. In this case free STAT3 can be phosphorylated by the gp130 receptor and pJAK1. To study the saturation level of pSTAT3, we analyze the ratio between pSTAT3/STAT3. Therefore, we iteratively reduce the concentration of mSTAT3 and STAT3 before IL-6 stimulation up to 1% of the initial value. We identify then the maximal concentration of pSTAT3 and calculate the pSTAT3/STAT3 ratio. Figure 6.10C shows the obtained ratio values for 100% to 1% of the original mSTAT3 and STAT3 concentration. We obtain an increase in relative pSTAT3 concentration for a reduced initial start value of up to 18%, followed by a sharp decrease in relative pSTAT3. This finding indicates that only 20% of STAT3 proteins are activated by phosphorylation. This finding can explain our observation of a reduction of 50% for pSTAT3 while STAT3 was set to 30% of the initial concentration. Moreover, we argue that these results show that STAT3 regulation via miRNAs has no severe impact on the overall pSTAT3 activation but rather miRNA regulation prevent of pSTAT3 and STAT3 expression overshooting.

6.3 Material and Methods

6.3.1 Mathematical model

For the analysis of the miRNA influence on the dynamics of the signaling cascade we set up two different models. Model 6.1 reflects a signaling cascade schematized in Figure 6.1, whereas Model 6.2 represents the gp130 receptor pathway stimulated by IL-6. Within this model, we integrated miRNA influence on central pathway position such as JAK1, STAT3, and SOCS3 to study the impact on the signal dynamic.

Signaling cascade

We set up a mathematical model of non-linear ordinary differential equations (ODEs) in order to analyze the impact of miRNAs on a signaling cascade. Therefore, we define a signaling cascade by taking one protein-kinase, one protein-inhibitor and one readout protein. Upon receptor stimulation $[R]$, the protein kinase (e.g. JAK1), which is translated from its mRNA, is phosphorylated. The readout protein that elicits a cellular response (e.g., activation of a transcription factor STAT3) will then be activated by the protein kinase. In order to suppress or turn out the signal, we added an inhibitor

(e.g. SOCS3) to the system, which inhibits the activation of the readout protein by a non-competitive inhibition. The non-competitive inhibition, phosphorylation as well as the dephosphorylation is modeled by Michaelis-Menten Kinetics. For the remaining model we use mass action kinetics. Equation 6.1 shows the final model and all parameters. In order to study the differences between a miRNA-environment and a system without miRNAs, we used the model described in Equation 6.1. For a system without miRNAs, we turn off the miRNA influence by setting c_6, and c_8 to 0.

$$\begin{aligned}
\frac{\mathrm{d}[P-Kinase]}{\mathrm{d}t} &= \frac{a_1[R][Kinase]}{K_I+[Kinase]} - b_1[P-Kinase] - \frac{v_1[P-Kinase]}{K_{III}+[P-Kinase]} \\
\frac{\mathrm{d}[P-Protein]}{\mathrm{d}t} &= \frac{a_2[Protein][P-Kinase]}{(K_{II}+[Protein])\left(1+\frac{[Inhibitor]}{K_2}\right)} - b_2[P-Protein] - \frac{v_2[P-Protein]}{K_{IV}+[P-Protein]} \\
\frac{\mathrm{d}[Kinase]}{\mathrm{d}t} &= a_3[mKinase] - b_3[Kinase] - \frac{a_1[R][Kinase]}{K_I+[Kinase]} + \frac{v_1[P-Kinase]}{K_{III}+[P-Kinase]} \\
\frac{\mathrm{d}[Protein]}{\mathrm{d}t} &= a_4[mProtein] - b_4[Protein] - \frac{a_2[Protein][P-Kinase]}{(K_{II}+[Protein])\left(1+\frac{[Inhibitor]}{K_2}\right)} + \frac{v_2[P-Protein]}{K_{IV}+[P-Protein]} \\
\frac{\mathrm{d}[Inhibitor]}{\mathrm{d}t} &= a_5[mInhibitor] - b_5[Inhibitor] \\
\frac{\mathrm{d}[mKinase]}{\mathrm{d}t} &= a_6 - (b_6 + c_6[miRKinase])[mKinase] \\
\frac{\mathrm{d}[mProtein]}{\mathrm{d}t} &= a_7 - b_7[mProtein] \\
\frac{\mathrm{d}[mInhibitor]}{\mathrm{d}t} &= a_8 - (b_8 + c_8[miRInhibitor])[mInhibitor] \\
\frac{\mathrm{d}[miRKinase]}{\mathrm{d}t} &= a_9 - b_9[miRKinase] - c_6[miRKinase][mKinase] \\
\frac{\mathrm{d}[miRInhibitor]}{\mathrm{d}t} &= a_{10} - b_{10}[miRInhibitor] - c_8[miRInhibitor][mInhibitor]
\end{aligned} \tag{6.1}$$

gp130-STAT3 model

In addition, to the signaling cascade, we study the impact of miRNAs on the gp130-STAT3 pathways, especially on the IL-6 stimulated gp130 receptor. In this setting, we slightly adapt the model 6.1. Figure 6.5 shows a schematic illustration of the gp130 receptor pathway. The activation of the kinase JAK1 is modeled by an active receptor gp130 and a non-competitive inhibition by SOCS3. The activation of the negative feedback loops of SOCS3 is modeled by the parameter a_8 and the concentration of $[P-STAT3]$, in which a_8 comprises the nuclear import of activated P-STAT3. Again, we study the impact of miRNAs in the gp130-STAT3 pathway, we distinguish between two alternative models ('without miRNAs' and a miRNA-extended model). In case of the 'without miRNA' model, we set c_6, c_7 and c_8 to 0 in order to turn off the miRNA influence on the system.

$$
\begin{aligned}
\frac{\mathrm{d}[P-JAK1]}{\mathrm{d}t} &= \frac{a_1[R][JAK1]}{(K_I+[JAK1])\left(1+\frac{[SOCS3]}{K_1}\right)} - b_1[P-JAK1] - \frac{v_1[P-JAK1]}{K_{III}+[P-JAK1]} \\
\frac{\mathrm{d}[P-STAT3]}{\mathrm{d}t} &= \frac{a_2[STAT3][P-JAK1]}{(K_{II}+[STAT3])\left(1+\frac{[SOCS3]}{K_2}\right)} - b_2[P-STAT3] - \frac{v_2[P-STAT3]}{K_{IV}+[P-STAT3]} \\
\frac{\mathrm{d}[JAK1]}{\mathrm{d}t} &= a_3[mJAK1] - b_3[JAK1] - \frac{a_1[R][JAK1]}{(K_I+[JAK1])\left(1+\frac{[SOCS3]}{K_1}\right)} + \frac{v_1[P-JAK1]}{K_{III}+[P-JAK1]} \\
\frac{\mathrm{d}[STAT3]}{\mathrm{d}t} &= a_4[mSTAT3] - b_4[STAT3] - \frac{a_2[STAT3][P-JAK1]}{(K_{II}+[STAT3])\left(1+\frac{[SOCS3]}{K_2}\right)} + \frac{v_2[P-STAT3]}{K_{IV}+[P-STAT3]} \\
\frac{\mathrm{d}[SOCS3]}{\mathrm{d}t} &= a_5[mSOCS3] - b_5[SOCS3] \\
\frac{\mathrm{d}[mJAK1]}{\mathrm{d}t} &= a_6 - (b_6 + c_6[miRJAK1])[mJAK1] \\
\frac{\mathrm{d}[mSTAT3]}{\mathrm{d}t} &= a_7 - (b_7 + c_7[miRSTAT3])[mSTAT3] \\
\frac{\mathrm{d}[mSOCS3]}{\mathrm{d}t} &= a_8[P-STAT3] - (b_8 + c_8[miRSOCS3])[mSOCS3] \\
\frac{\mathrm{d}[miRJAK1]}{\mathrm{d}t} &= a_9 - b_9[miRJAK1] - c_6[miRJAK1][mJAK1] \\
\frac{\mathrm{d}[miRSTAT3]}{\mathrm{d}t} &= a_{10} - b_{10}[miRSTAT3] - c_7[miRSTAT3][mSTAT3] \\
\frac{\mathrm{d}[miRSOCS3]}{\mathrm{d}t} &= a_{11} - b_{11}[miRSOCS3] - c_8[miRSOCS3][mSOCS3]
\end{aligned}
\tag{6.2}
$$

In order to study the cases of PSTAT3 dependent miRNAs, we modify the model by linking the miRNA production rates to the P-STAT3 concentration:

$$
\begin{aligned}
\frac{\mathrm{d}[miRJAK1]}{\mathrm{d}t} &= a_9[PSTAT3] - b_9[miRJAK1] - c_6[miRJAK1][mJAK1] \\
\frac{\mathrm{d}[miRSTAT3]}{\mathrm{d}t} &= a_{10}[PSTAT3] - b_{10}[miRSTAT3] - c_7[miRSTAT3][mSTAT3]
\end{aligned}
\tag{6.3}
$$

where a_9 is the production rate of miRJAK1, b_9 the turnover rate, and c_6 the mRNA:miRNA binding factor of miRJAK1 and mJAK1. For miRSTAT3, a_{10} defines the production rate, b_{10} the turnover rate, and c_7 the mRNA:miRNA binding factor.

6.3.2 Parameter estimation

For the signaling model, we used a simulated annealing approach (133) to obtain the parameter. Parameter for phosphorylation and dephosphorylation were adapted from (128). Production rate for miRNA was set to 5 times the mRNA production rate, similar to (161). Turnover rates for proteins and mRNAs were set between 1 and 24 hours. miRNA turnover were set between 8 and 24 hours. We sample 10.000 turnover rates for proteins, mRNAs and miRNAs and determine the steady-state level of the system. Simulated annealing is then used to obtain the new production parameter for either mRNAs or miRNAs to shutdown the pathway signal to 50% based on the original steady-state level.

For the gp130-STAT3 model we used a Markov chain Monte Carlo approach (222) to first estimate parameters for three submodules containing either all JAK1, STAT3 or SOCS3 reactions. Based on these first parameter distributions, we define parameters prior based on the MCMC results for all parameters except miRNA related ones. Finally, we combine the three submodules into one whole gp130-STAT3 model and using a simulated annealing approach with normal distributed prior weights for the parameter. We then minimize the error of the sum of weighted residuals between the measurements and the model trajectory. For each measurement we assume a normal-distributed error, whereas deviations of the measured data error were fitted within the model.

6.3.3 Quantification of (phospho)-proteins

Cytosolic extracts of $2x10^6$ hepatocytes were prepared. Lysates were then centrifuged and supernatants were used as cytosolic fraction for immunoprecipitation. For quantification of recombinant proteins, a dilution series of the respective recombinant calibrator protein and a BSA standard series were separated by SDS-PAGE. Randomized quantitative immunoblotting data was processed using GelInspector software (231). The addition of normalizers allowed for quality control and normalization of the raw data. The following normalizers were used: GST-gp130 for gp130 and pgp130, GST-JAK1 for JAK1 and pJAK1, GST-STAT3 for STAT3 and pSTAT3 as well as SBP-SOCS3 for SOCS3. Immunoprecipitation was performed with an equivalent of $2x10^6$ cells by adding target-specific antibody and Protein A sepharose to the lysate. For computational data normalization, recombinant calibrator proteins were added to immunoprecipitations.

6.3.4 Quantification of mRNA transcripts

Quantitative two-step RT-PCR was performed using a LightCycler 480 in combination with the hydrolysis-based Universal ProbeLibrary platform. In general, qRT-PCR amplifications were performed in 96-well format according to the manufacturer's manual. Crossing point values were calculated using the Second Derivative Maximum method of the LightCycler 480 Basic Software. PCR efficiency correction was performed for each PCR setup individually based on a dilution series of template cDNA (170). Relative concentrations were normalized using HPRT as a reference gene (260; 276).

6.3.5 llumina next-generation small RNA sequencing

In the analysis, we stimulate hepatocytes with 1 nM IL-6 for 1 h, 2 h and an unstimulated control (0 h) each performed in duplicates. For each time point total RNA from $2x10^6$ primary hepatocytes was isolated using the RNeasy Mini Kit. RNA from all plates was extracted in parallel according to the manufacturer's instructions for adherent cells. To eliminate traces of DNA, on-column digests using the RNAse-Free DNAse Set were performed. Finally Small RNA Sample Prep Kit was used to extract small RNA. For data processing miRanalyzer (88) was used to extract miRNA time-course data.

6.4 Conclusions and Outlook

Within this chapter, we analyze the post-transcriptional regulation of signaling pathways by miRNAs. We first study a signaling cascade, which is representative of many signaling pathways. We show that alteration of signal transduction proteins and transcripts via gene or miRNA regulation has a severe impact on the signal maintenance and shutdown time. Changes in the expression of signal transduction activators or inhibitors leads either to a fast signal shutdown via the inhibitor or a slow signal repression via the activator. We show that gene or miRNA regulation leads to similar recovery times of the signal. We identify specific turnover rate ranges, which lead to a context specific increase or delay in the recovery time. Extending the pathway model to a whole model of the gp130-STAT3 pathway, we study the impact of miRNAs on the signal dynamic after IL-6 stimulation. We calibrate the model to time-course data from primary mouse hepatocytes and predict a global miRNA influence for JAK1 and STAT3, which is validated by miRNA expression profiles. Moreover, we can show that IL-6 induced miRNAs have no influence on the signal dynamic, whereas pre-induced changes in miRNA expression lead to an alteration of the pSTAT3/STAT3 ratio having a strong impact on the signal dynamic. These analyses show that miRNA regulation as an addition layer of transcriptional control allows the cell to alter the signal transduction in context specific manner. In future work, we will experimentally confirm our prediction made by the gp130-STAT3 model. Therefore, we will use siRNA knockdown experiments to alter JAK1 or STAT3 expression. From a theoretical point of view, we can extend the model in two directions. First, we can refine and extend the model to describe the receptor binding complex and the STAT3 nucleus transport in

more detail. Second, one can use different mathematical approaches such as stochastic differential equation to analyze the stochastic fluctuation of chances in mRNA and protein expression during the active signaling pathway. In the next chapter, we present a novel unsupervised method to analyze multi-scale data. Using this approach, we are able to study the multi-layered and temporal responses of active signaling pathways.

7 Unsupervised method for the analysis of multi-scale data

7.1 Background

With the availability of high-throughput 'omics' data, more and more methods from statistics and signal processing are applied in the field of bioinformatics (262). Direct application of such methods to biological data sets is essentially complicated by three issues, namely *(i)* the large-dimensionality of observed variables (e.g. transcripts or metabolites), *(ii)* the small number of independent experiments and *(iii)* the necessity to take into account prior information in the form of e.g. interaction networks or chemical reactions.

While *(i)* may be tackled by targeted analysis, feature selection or efficient dimension reduction methods, the issue of low number of samples (experiments) may hinder the transfer of methods. For example, with cDNA microarrays, the number of genes (p) is usually much larger than the experiment size n (number of arrays). Quantitative data from experiments are often classified as small-n-large-p problems (142) and algorithms that are currently being developed are tailored for such kind of data. Detailed prior information is in general best handled by Bayesian methods (73), which are however not straight-forward to formulate in small-n-large-p problems.

Here, we focus on the unsupervised extraction of overlapping clusters in data sets exhibiting properties *(i-iii)*. If applied to gene expression profiles acquired by microarrays or metabolic profiles from mass spectrometry, we can interpret these clusters as jointly acting species (cellular processes). While partitioned clustering based on k-means (263) or hierarchical clustering (54) has been successful in some domains and is often the initial tool of choice for data grouping, overlapping clusters are better described by fuzzy techniques (71) or linear models (126). Focusing on

the latter, we can essentially summarize these techniques as matrix factorization algorithms. Constraining the factorization using e.g. decorrelation, statistical independence or non-negativity then leads to algorithms like principal component analysis (PCA), independent component analysis (104) and nonnegative matrix factorization (155), respectively. Although such methods are successfully applied in bioinformatics (167; 229; 266), they partially run into issues *(i-iii)* as described above. In particular, it is not clear how to include prior knowledge, which has been a quite successful strategy in other contexts (258). A first step towards this direction is network component analysis (NCA) (20; 166). It integrates prior knowledge in form of a multiple-input motif to uncover hidden regulatory signals from the outputs of networked systems. Hence, it focuses on the estimation of single gene's expression profiles, not in a linear decomposition of a data set into overlapping clusters. NCA poses strict assumptions on the topology of this predefined network, which makes it hardly applicable to mammalian high-throughput 'omics' data. Moreover, feedbacks from the regulated species back to the regulators are treated only as 'closed-loops', without explicitly modeling the feedback structure.

To overcome these constraints, this contribution provides a novel framework for the linear decomposition of data sets into expression profiles. We present a new matrix factorization method that is computationally efficient *(i)*, able to deal with the low number of experiments *(ii)* and includes as much prior information a possible *(iii)*. In order to achieve computational efficiency coupled with robust estimation, we use delayed correlations instead of higher-order statistics. In signal processing, this strategy has been shown to be advantageous (12; 271) for two reasons: such methods use more information from the data without over-fitting it, and they are second order and therefore computationally efficient. This is crucial for the application to microarray data, since dimensionality tends to be high in this environment.

However, delayed correlations can usually not be computed in the case of biological high-throughput experiments such as in microarray samples. While time-resolved experiments may provide correlations, the number of temporal observations are commonly too small (<10) for the estimation of time-delayed correlations.

Hence, we instead pose factorization conditions along the set of genes or other biological variables. We link these variables using prior knowledge e.g. in the form of a transcription factor or protein-protein interaction (PPI) network, metabolic pathways or *via* explicitly given models. Using this information enables us to define a graph-decorrelation algorithm that combines prior knowledge with source-separation

techniques, for illustration see Figure 7.1. In case of gene expression analysis the input of GraDe are the expression data and an underlying regulatory network. After applying GraDe, we obtain two matrices, a mixing and a source matrix. We interpret the sources as the biological processes and the mixing coefficients as their time-dependent activities. Hereby, the extracted sources group the genes' expression that can be explained by the underlying regulatory network, e.g. different responses of a cell to an external stimulus.

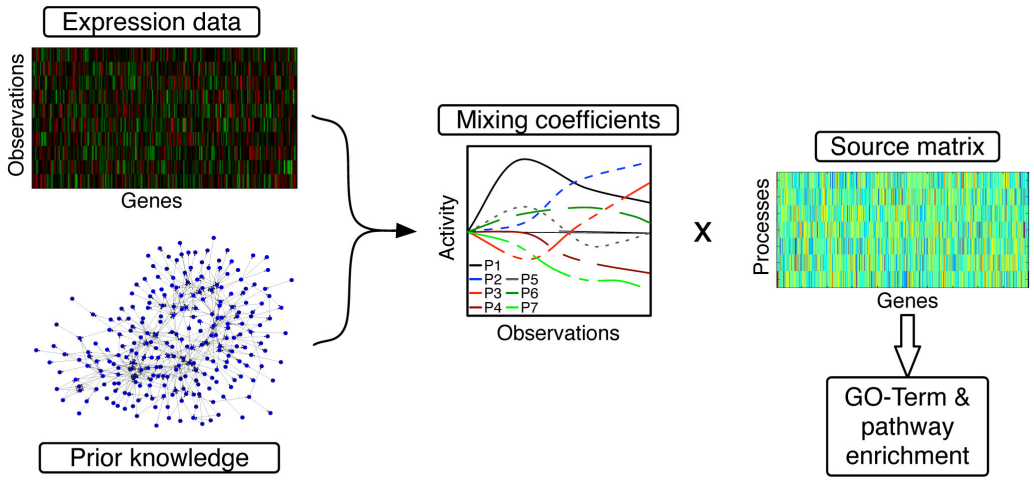

Figure 7.1: **GraDe: Graph-decorrelation algorithm.** In cells, various biological processes are taking place simultaneously. Each of these processes has its own characteristic gene expression pattern, but different processes may overlap. A cell's total gene expression is then the sum of the expression patterns of all active processes, weighted by their current activation level. The GraDe algorithm combines a matrix factorization approach with prior knowledge in form of an underlying regulatory network. The input of GraDe is the transcriptional expression data, where observations can be different conditions or a time points, and the underlying regulatory network (prior knowledge). GraDe decomposes the observed expression data into the underlying sources S and their mixing coefficients A. Analyzing time-course microarray data, we interpret these sources as the biological processes and the mixing coefficients as their time-dependent activities. Observations indicate their expression behavior either in the different conditions or time-points and activity their activation strength. We further filter process-related genes by taking only the genes with the strongest contribution in each process. Finally, we test for enrichment of cellular processes (GO) and biological pathways (KEGG).

We demonstrate the applicable of GraDe on three examples: (*i*): The cytokine interleukin IL-6 mediates the production of acute phase proteins by hepatocytes and pro-

motes liver regeneration (63). In order to unveil the multi-layered temporal signaling pathway responses, we measure gene expression in *IL-6* stimulated mouse hepatocytes by a time-course microarray experiment. Applying GraDe with a literature based gene regulatory network, we are able to infer associated biological processes as well as the dynamic behavior of *IL-6* related gene expression. In addition, we find that the estimated factors are robust against the high number of false positives contained in large-scale biological databases. (*ii*): We apply Grade to microarray data from a stem cell differentiation experiment. In contrast to other factorization techniques, GraDe finds a structured and detailed separation of known biological processes. (*iii*): We apply GraDe to combined mRNA and miRNA data using information about gene regulation and miRNA target genes to link mRNA and miRNA in a regulatory network. Applying GraDe, we are able to identify a core regulatory network of the differentiation process in glutamatergic neuros from high-throughput data.

7.2 Results and Discussion

The activation of gene regulatory processes upon external stimulations induces a rearrangement of cellular gene expression patterns. Matrix factorization techniques are currently explored in the analysis of such multi-layered and overlapping temporal responses. In the following, we propose an algorithm that incorporates prior knowledge as a constraint to the factorization task (see Figure 7.1).

7.2.1 Matrix factorization incorporating prior knowledge

In signal processing, various matrix factorization techniques have been developed that employ intrinsic properties of data to decompose them into underlying sources (12; 271; 270). These methods are based on *delayed correlations* that can be defined for data having a temporal or spatial structure. For instance, the *time-delayed correlation matrix* of a centered, wide-sense stationary multivariate random process $\mathbf{x}(t)$ is defined as

$$\left(\mathbf{C}_{\mathbf{x}}(\tau)\right)_{ij} := E\left(\mathbf{x}_i(t+\tau)\mathbf{x}_j(t)^\top\right), \tag{7.1}$$

where E denotes expectation. Here, off-diagonal elements detect time-shifted correlations between different data dimensions. For $\tau = 0$ this measure reduces to the common cross-correlation. Given l features, e.g. genes, aggregated in a data matrix $\mathbf{X}$,

e.g. mRNA expression data, the cross-correlation matrix can be easily estimated with the unbiased variance estimator:

$$\mathbf{C_X} = \frac{1}{l-1}\,\mathbf{X}\mathbf{X}^\top. \tag{7.2}$$

However, the experimentally generated quantitative data sets we face in bioinformatics rarely imply a natural order like which allows defining a generic kind of delayed correlation. We therefore generalize this concept by introducing prior knowledge that links features (e.g. genes) along a pre-defined underlying network. This network may be large-scale, but can be also an explicitly given small-scale process. Moreover, integrated information may be of qualitative (e.g. interaction) as well as quantitative nature (e.g. interaction strength, reaction rates).

Graph-delayed correlation

We encode prior knowledge in a directed, weighted graph $G := (\mathcal{V}, \mathcal{E}, w)$ defined on vertices $\mathcal{V} \in \{1, \ldots, l\}$ corresponding to our features. The edges $\mathcal{E}$ are weighted with weights $w : \mathcal{E} \to \mathbb{R}$. These are collected in a *weight matrix* $\mathbf{W} \in \mathbb{R}^{l\times l}$, where w_{ij} specifies the weight of edge $i \to j$. Note that our weights may be negative, and G may contain self-loops. For any vertex $i \in \mathcal{V}$, we denote by $S(i) := \{j|(i,j) \in \mathcal{E}\}$ the set of *successors of* i, by $P(i) := \{j|(j,i) \in \mathcal{E}\}$ its *predecessors*.

The graph G introduces a partial ordering on the l features. We use the weight matrix $\mathbf{W}$ as propagator for an activity pattern $\mathbf{x} \in \mathbb{R}^l$ of our features and define the *G-shift* $\mathbf{x}^G$ of $\mathbf{x}$ as the vector with components

$$x_i^G := \sum_{j\in P(i)} \mathbf{W}_{ji}\, x_j\,. \tag{7.3}$$

Recursively, we define any positive shift $\mathbf{x}^G(\tau)$ (see Figure 7.2). For negative shifts we replace predecessors $P(i)$ by successors $S(i)$, which formally corresponds to a transposition of the weight matrix $\mathbf{W}$. Using the convention of trivial weights for non-existing edges of G, we can extend the above sum to all vertices. Gathering available m experiments (rows) into a data matrix $\mathbf{X} \in \mathbb{R}^{m\times l}$, we obtain the simple, convenient formulation of a G-shifted data set

$$\mathbf{X}^G(\tau) = \begin{cases} \mathbf{X}\mathbf{W}^\tau & \tau \geq 0 \\ \mathbf{X}(\mathbf{W}^\top)^\tau & \tau < 0 \end{cases}\,. \tag{7.4}$$

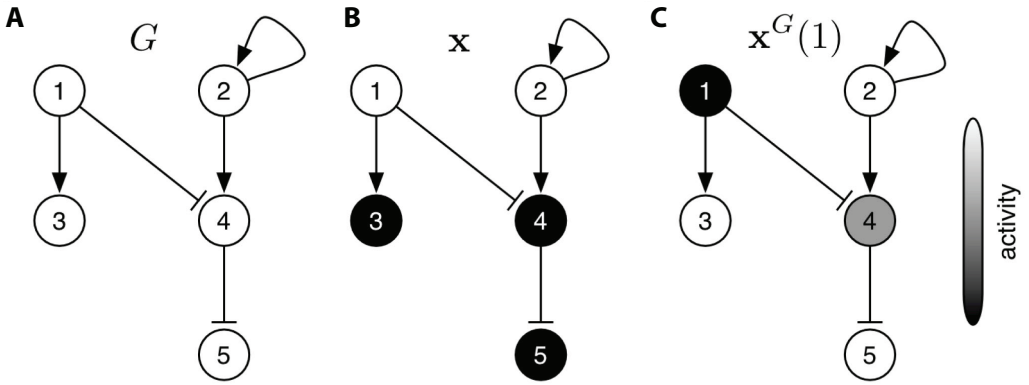

Figure 7.2: **Illustration of the G-shift.** Illustration of the G-shift in the unweighted graph G shown in (**A**). We start with an initial node activity $\mathbf{x}$ depicted in (**B**). We use the graph as propagator for the time evolution of this pattern: after one positive shift we achieve the activity pattern $x^G(1)$ in (**C**).

After mean removal, we may assume that each row of $\mathbf{X}$ is centered. Then, in analogy to the unbiased estimator for cross-correlations in Equation 7.2, we define the *graph-delayed (cross)-correlation*

$$\mathbf{C}^G_{\mathbf{X}}(\tau) := \frac{1}{l-1}\,\mathbf{X}^G(\tau)\mathbf{X}^\top = \frac{1}{l-1}(\mathbf{X}\mathbf{W}^\tau\mathbf{X}^\top)\,. \tag{7.5}$$

Note that our definition includes the standard time-delayed correlation by shifting along the line graph $1 \to 2 \to \ldots \to l-1 \to l$.

The graph-delayed correlation is only symmetric if the used graph shows this feature which is, for instance in regulatory networks, rarely the case. For our following derivations, a symmetric generalized correlation measure however will turn out to be very convenient. In the remainder of this section, we will therefore use the *symmetrized graph-delayed correlation*

$$\overline{\mathbf{C}}^G_{\mathbf{X}}(\tau) = \frac{1}{2}\left(\mathbf{C}^G_{\mathbf{X}}(\tau) + \mathbf{C}^G_{\mathbf{X}}(\tau)^\top\right)\,. \tag{7.6}$$

Enforcing the symmetry property is strategy has been often applied in the case of temporally or spatially delayed correlations. It has also been demonstrated that symmetrization stabilizes the estimation of the cross-correlations from data (104). Moreover, it can be shown that asymptotically using either normal or symmetrized correlations end up giving the same eigenvectors (12).

Factorization model

The linear mixing model for the input data matrix $\mathbf{X} \in \mathbb{R}^{m \times l}$ is given by

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \boldsymbol{\varepsilon}\,. \tag{7.7}$$

Here, the matrix of source contributions $\mathbf{A} \in \mathbb{R}^{m \times n}$ $(m \geq n)$ is assumed to have full column rank. The sources $\mathbf{S} \in \mathbb{R}^{n \times l}$ are uncorrelated, zero-mean stationary processes with nonsingular covariance matrix. We allow for a noise term $\boldsymbol{\varepsilon} \in \mathbb{R}^{m \times l}$, which is modeled by a stationary, white zero-mean process with variance σ^2. We assume white unperturbed data $\tilde{\mathbf{X}} := \mathbf{A}\mathbf{S}$ (possibly after whitening transformation). In other words, we interpret each row of $\mathbf{X}$ as linear mixture of the n sources (rows of $\mathbf{S}$), weighted by mixing coefficients stored in $\mathbf{A}$. Without additional restrictions, this general linear blind source-separation problem is underdetermined.

Here, we assume that the sources have vanishing graph-delayed cross-correlation with respect to some given graph G and all shifts τ. Formally, this means that $\overline{\mathbf{C}}^G_{\mathbf{S}}(\tau)$ is diagonal. We observe

$$\overline{\mathbf{C}}^G_{\mathbf{X}}(\tau) = \begin{cases} \mathbf{A}\overline{\mathbf{C}}^G_{\mathbf{S}}(\tau)\mathbf{A}^\top + \sigma^2\mathbf{I}, & \tau = 0 \\ \mathbf{A}\overline{\mathbf{C}}^G_{\mathbf{S}}(\tau)\mathbf{A}^\top & \tau \neq 0 \end{cases}. \tag{7.8}$$

Clearly, a full identification of $\mathbf{A}$ and $\mathbf{S}$ is not possible, because Equation (7.7) defines them only up to scaling and permutation of columns: Multiplication of a source by a constant scalar can be compensated by dividing the corresponding row of the mixing matrix by the scalar. Similarly, the factorization implies no natural order of the sources. We can take advantage of the scaling indeterminacy by requiring our sources to have unit variance, i.e. $\overline{\mathbf{C}}^G_{\mathbf{S}}(0) = \mathbf{I}$. With this, as we assumed white data $\tilde{\mathbf{X}}$, we see that $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$, i.e. $\mathbf{A}$ is orthogonal. Thus, the factorization in Equation (7.8) represents an eigenvalue decomposition of the symmetric matrix $\overline{\mathbf{C}}^G_{\mathbf{X}}(\tau)$. If additionally we assume that $\overline{\mathbf{C}}^G_{\mathbf{S}}(\tau)$ has pairwise different eigenvalues, the spectral theorem guarantees that $\mathbf{A}$ – and with it $\mathbf{S}$ – is uniquely determined by $\mathbf{X}$ except for permutation. The reason why we focused on the symmetrized instead of the simple graph-delayed correlation matrix was precisely that we wanted to have a symmetric matrix, because then the eigenvalue decomposition is well defined and also simple to compute.

However, we have to be careful, because we cannot expect $\overline{\mathbf{C}}^G_{\mathbf{X}}(\tau)$ to be of full rank. Obviously, we require more features than obtained sources ($l \gg m$), hence in general

$\mathrm{rank}(\mathbf{X}) = m$. If G contains an adequate amount of information, $\mathrm{rank}(\mathbf{W})$ is of order l and since $l \gg m$, $\mathrm{rank}(\overline{\mathbf{C}}^G_{\mathbf{X}}(\tau))$ is essentially determined by (the upper bound) m. Hence, when analyzing high-throughput biological data linked by underlying large-scale networks, we can extract as many sources as observations are available.

The GraDe algorithm

Equation (7.8) also gives an indication of how to solve the matrix factorization task in our setting. The first step consists of whitening the no-noise term $\tilde{\mathbf{X}} = \mathbf{AS}$ of the observed mixtures $\mathbf{X}$. The whitening matrix can be easily estimated from $\mathbf{X}$ by diagonalization of the symmetric matrix $\overline{\mathbf{C}}^G_{\tilde{\mathbf{X}}}(0) = \overline{\mathbf{C}}^G_{\mathbf{X}}(0) - \sigma^2\mathbf{I}$, provided that the noise variance σ^2 is known or reasonably estimated. If more signals than sources are observed, dimension reduction can be performed in this step. Insignificant eigenvalues then allow estimation of the noise variance, compare (12). Now, we may estimate the sources by diagonalization of the single, symmetric graph-delayed correlation matrix $\overline{\mathbf{C}}^G_{\mathbf{X}}(\tau)$. This procedure generalizes AMUSE (271), which employs standard autocorrelation, to the extended definition of graph-correlation.

The performance of AMUSE is relatively sensitive to additive noise. Moreover, the estimation by a finite amount of samples may lead to a badly estimated autocorrelation matrix (270). To alleviate these problems, algorithms like SOBI (12) or TDSEP (301) extend this approach by *joint diagonalization* of *multiple* autocorrelation matrices, calculated with a set of different time-delays. In a similar manner we jointly diagonalize multiple G-autocorrelation matrices obtained from G-shifts with different lags. This approximative joint diagonalization can be achieved by a variety of methods. We use the Jacobi-type algorithm proposed in (32), since we later compare GraDe's performance to the classic SOBI algorithm.

Altogether, we subsume this procedure in the graph-decorrelation algorithm (GraDe). At `http://cmb.helmholtz-muenchen.de/grade` an implementation is freely available. When shifting along the line graph, GraDe with a single lag reduces to AMUSE, and GraDe with multiple shifts corresponds to SOBI. In summary, the input of GraDe is (*i*) a expression matrix $\mathbf{X} \in \mathbb{R}^{m \times l}$ containing m experiments and l genes and (*ii*) a *weight matrix* $\mathbf{W} \in \mathbb{R}^{l \times l}$ containing the prior knowledge. We obtain a mixing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}\,(m \geq n)$ and a source matrix $\mathbf{S} \in \mathbb{R}^{n \times l}$. In the case of gene expression data analysis the sources can be interpreted as biological processes and the mixing coefficients as their time-dependent activities.

Including prior knowledge into the source-separation task may introduce bias in the patterns that are pre-defined and, in turn, the analysis and results obtained. It is important to note that annotation of biological knowledge is biased and under permanent change. Therefore, when using gene regulatory networks as prior knowledge one has to keep in mind that this information is subject to annotation bias. Thus the density of connections in certain regions of the network might be higher due to the fact that these parts are better explored. In the case of classification problem, recent studies have shown that methods can be improved in terms of classification accuracy by including prior knowledge into the classification process (113). These methods benefit from the fact that genes are not treated as independent. Hence, most of these methods are based on the hypothesis that genes in close proximity, which are connected to each other, should have similar expression profiles. The same assumption may be transferred to source-separation methods. Applying standard methods like ICA or PCA, implies the assumption that all data points, i.e. in our setting the expression levels of different genes are sampled i.i.d. from an underlying probability density. This assumption is obviously not fulfilled, since the genes' expression values are the read-outs of different states of a complex dynamical system: Genes obey dynamics along a transcription factor network. Instead of ignoring the genes' dependencies, we here proposed to explicitly model them using prior knowledge given within a gene-regulatory network. Therefore, one of the key advantages of GraDe is to overcome the assumption of the independencies. Applying GraDe to time-course expression data (see section *Validation of the time-dependent signals*), we will show that including prior knowledge into the source separation task leads to an improvement compared to fully-blind methods like PCA. Finally, we believe that with increasing quality and amount of biological knowledge, methods incorporating prior knowledge will increase in performance as well.

7.2.2 Illustration of GraDe

In order to illustrate GraDe, we analyze two toy examples. We first focus on a bifan structure shown in Figure 7.3A and assume to have six genes from the time-courses of expression levels depicted in Figure 7.3B. For data generation, the system is simulated

by ordinary differential equations:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = -\gamma_i x_i(t) + \sum_{j \in P(i)} f_{ji}\left(x_j(t)\right). \tag{7.9}$$

where we model interactions by sigmoidal Hill functions. In this case, one input x_1 is active until time-point 10, when it is turned off and instead production of x_2 is switched on. Consequently, x_3 peaks at time 10, but also x_4 shows an early activation due to low expression of its inhibitor. Applying GraDe (with the known bifan topology, but without access to the underlying ODE system), we find that three sources are sufficient to explain the data (Figure 7.3B). From the extracted sources and their time-courses (shown in Figure 7.3E and 7.3D) we see that the strongest source s_1 represents the externally controlled inputs and the network topology: the source couples x_1 and x_3, and in opposite direction x_2 and x_4. Therefore, GraDe is able to recover the two processes. Source s_2 has the lowest contribution to the total expression values and is needed for fine-tuning the combined dynamics, as we obtain an early activation of x_4 due to low expression of its inhibitor. Consequently, the source s_2 is active at time-points 2 and 4, i.e. immediately after the switching operations. Source s_3 again reflects the crossover inhibitions, accordingly its time-course is flat. This source groups the input of the network, which could be linked e.g. to pathway stimulation. For our second example we use the funnel structure in Figure 7.3F, where we defined the expression values for three different input conditions (Figure 7.3G). Eigenvalues and the factorization obtained by GraDe are visualized in Figure 7.3H-J. Source s_1 again reflects the network topology, by grouping the cascade genes, while s_2 allows the reconstruction of the last condition. As we expect, GraDe are able to recover the two independent inputs. Applying GraDe to two different toy examples, we are able to show that GraDe is applicable both time-course as well as conditional experiments. In both cases, GraDe identifies the different responses and inputs of the system.

7.2.3 Evaluation on artificial data

In order to evaluate GraDe, we generated random mixtures of artificial G-decorrelated signals. A common way to create standard-autocorrelated signals are *moving average* (MA) models (104): For a white noise process $\boldsymbol{\varepsilon}$ and real coefficients $\theta_1 \ldots \theta_q$, a MA(q) model $\mathbf{x}$ is defined by $\mathbf{x}_t = \boldsymbol{\varepsilon}_t + \theta_1 \boldsymbol{\varepsilon}_{t-1} + \cdots + \theta_q \boldsymbol{\varepsilon}_{t-q}$. In our notation, we interpret this MA signal $\mathbf{x}$ as a weighted sum of G-shifted versions of $\boldsymbol{\varepsilon}$, shifted q times along

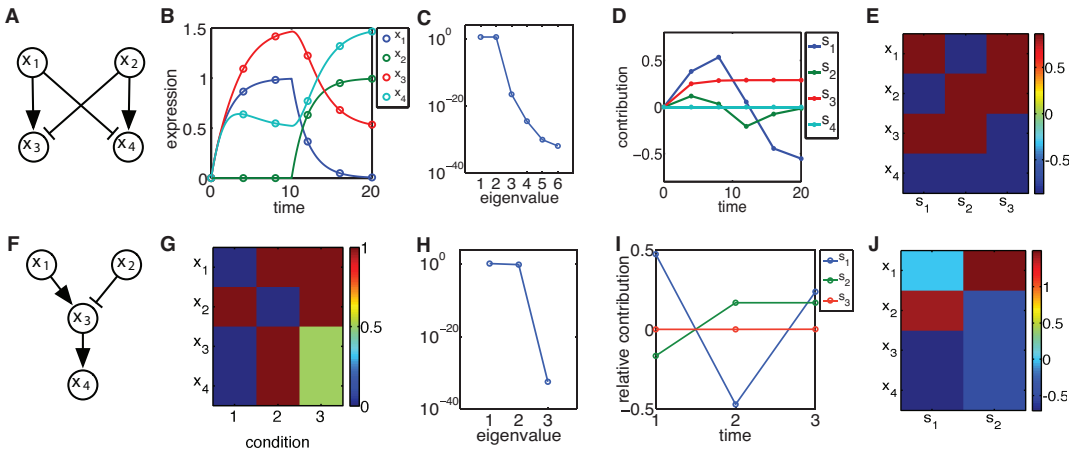

Figure 7.3: **Illustration of GraDe.** For the bifan motif in **A** we take 6 genes (dots) from the simulated time-courses in **B** and apply GraDe: **C** shows the eigenvalues of the decomposition in GraDe. In **D** we plot the time-courses of the extracted sources $s_1 \ldots s_6$, hence the curves are the columns of the mixing matrix. From **C** we see that only the first three sources are relevant, which are visualized as heat-map **E**. For our second example **F** we assume to know expressions in different conditions as shown in **G**. The factorization by GraDe is visualized in subfigures **H** to **J**.

the line graph G. Therefore, for an arbitrary graph G we define a q-th order *G–MA(q) model* as

$$\mathbf{x} = \boldsymbol{\varepsilon} + \theta_1 \boldsymbol{\varepsilon}^G(1) + \theta_2 \boldsymbol{\varepsilon}^G(2) + \ldots + \theta_q \boldsymbol{\varepsilon}^G(q) \tag{7.10}$$

Any G-MA(q) process is equivalent to a G-MA(1) process with a modified graph.

In a first simulation, we used directed Erdös-Rényi random graphs (57) with mean connectivity 17.5 and random weights in $(-1, 1)$ to generate $m = 2$ G–decorrelated G-MA(1) signals with $l = 5000$ samples. Data were normalized to unit variance and mixed with a random mixing matrix. We added Gaussian uncorrelated noise of variable strength σ and applied GraDe (without noise estimation) with one and 30 shifts, respectively. Reconstruction quality was estimated using the Amari-index that quantifies the deviation between the correct and the estimated mixing matrix (41). From Figure 7.4A we see that for G-MA(1) processes GraDe with a single-shift performs well in the low-noise setting, in contrast to multiple shifts. This is a consequence of the complex short-distance, but vanishing long-distance autocorrelation structures. When performing multiple shifts, each lag is weighted equally, which deteriorates the algorithm's performance.

Accordingly, as shown in Figure 7.4B, GraDe with multiple shifts outperformed

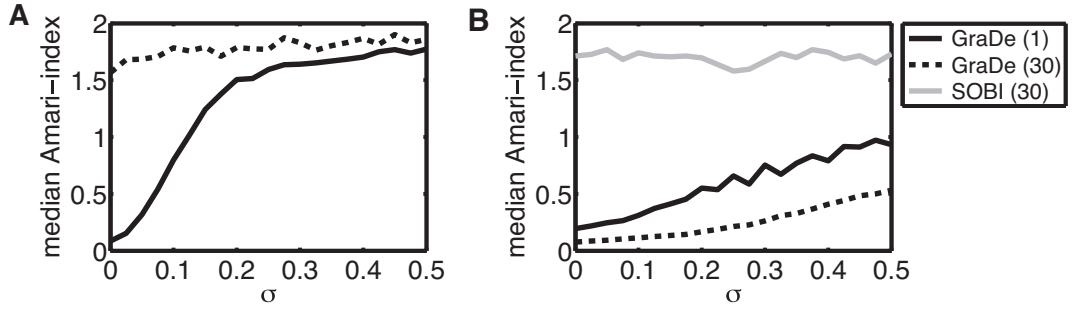

Figure 7.4: **Performance on artificial data:** Mixtures of (**A**) two G-MA(1) processes with random graphs G, (**B**) mixtures of two G-MA(20) processes with signed line graphs. The plots show the dependence of median Amari–indices on the noise level σ over 1000 runs. We compare GraDe with one and 30 shifts, in (**b**) in addition SOBI with 30 shifts.

single–shift GraDe when applied to higher order G–MA processes. We generated G–MA(20) processes of sample size $l = 1500$ with a signed line graph, where the edges had weights ± 1 with equal probability. The unsigned line graph used by SOBI was not sufficient to reconstruct these signals in a proper way, whereas GraDe with the true graph separated them even using a single shift only. However, similar to the behavior in the standard-autocorrelation case, here multiple shifts dramatically enhance GraDe's robustness against additive noise.

7.2.4 *IL-6* mediated responses in primary hepatocytes

In liver, the cytokine interleukin *IL-6* mediates two major responses. First, it induces hepatocytes to produce acute phase proteins upon infection-associated inflammation. These proteins include complement factors to destroy or inhibit growth of microbes. In addition, *IL-6* promotes liver regeneration and protects against liver injury (63). *IL-6* regulates several cellular processes such as proliferation, differentiation and the synthesis of acute phase proteins (72). Upon binding to its cell surface receptor, *IL-6* activates the receptor associated Janus tyrosine kinase (JAK) 1 - signal transducer and activator of transcription (STAT) 3 - signal transduction pathway. The latent transcription factor STAT3 is translocated to the nucleus after activation and subsequently alters gene expression.

To identify the biological responses to *IL-6* in a time-resolved manner, we stimulated primary mouse hepatocytes with 1 nM *IL-6* up to 4 hours and analyzed the changes

in gene expression by microarray analysis. In a first approach, we extracted all genes that were significantly regulated compared to time point 0h. In total, we obtained 121 genes and applied k-means clustering to detect groups within this set. Based on this approach, we could not identify any time-resolved responses upon *IL-6* stimulation. Due to the small number of significantly regulated genes, we decided to employ a genome-wide approach using GraDe to resolve the cellular responses upon *IL-6* in more detail.

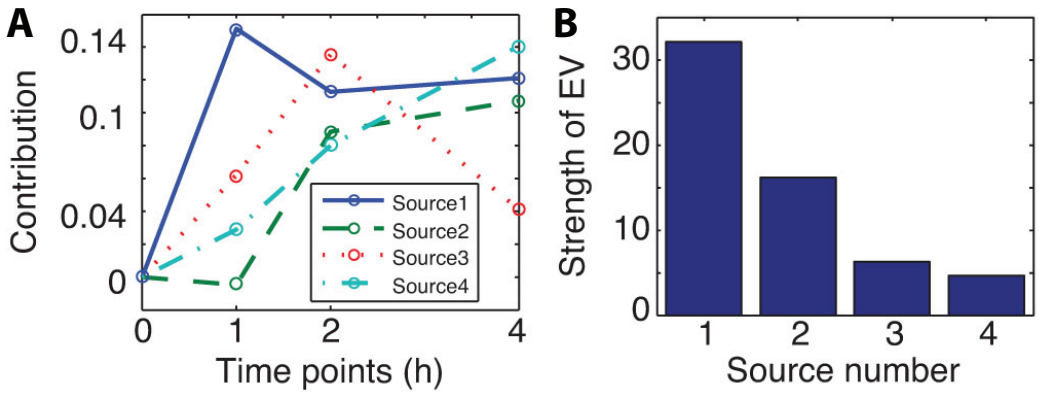

Figure 7.5: **Result of GraDe.** This figure illustrates the decomposition of the time-course microarray experiment on *IL-6* stimulated hepatocytes with GraDe. As underlying network we used interactions from the TRANSPATH database (see Methods). (**A**) shows the time-courses of the four extracted sources, centered to time point 0 h. The x-axis shows the measured time-points and the y-axis the contribution of the mixing matrix. In (**B**), we plot the strength of the eigenvalues (EV) of the resulting sources. All four extracted sources have significant contributions.

GraDe discovers time-dependent biological processes upon *IL-6* stimulation

Using GraDe, we linked all 5709 expressed genes along a gene regulatory network derived from the TRANSPATH database (see Methods). We obtained four graph-decorrelated gene expression sources (GES), which we labeled from 1 to 4 according to their decreasing eigenvalues (Figure 7.5B). We see that dimension reduction and with it noise level estimation were not possible in our case. The estimated mixing matrix is shown in Figure 7.5A. The matrix of source contributions contains positive and negative components. We partitioned a source into submodes that contain either the negative signals or the positive signals, respectively. We selected all genes in the positive submodes by choosing a threshold ≥ 2 as well as all genes in the negative

submodes with a threshold $\leq$ -2, respectively. These sets were subsequently used for GO enrichment analysis using a p-value < 0.05 after correction by False Discovery Rate.

Differentially expressed genes within GES 1 display an immediate strong increase in expression following *IL-6* stimulation. After peaking at one hour, expression decreases to elevated levels compared to untreated samples. Significantly enriched GO-Terms within this GES correspond to responses triggered by external stimuli and inflammation (see Table 7.1). In liver, upon infection- or injury-associated inflammation *IL-6* mediates production of acute phase proteins (APP) by hepatocytes as represented by the GO-Term "(acute) inflammatory response" (e.g. *Saa4*, *Fgg*, *Pai1*). *Angptl4* is a positive acute phase protein (171) showing a strong increase in expression during the first hour after stimulation followed by a decrease after two hours. GraDe reconstruct the expression pattern by the mixing of the four different source time-patterns GES 1 to 4. We identify *Angptl4* in GES 1 and 3 having a source contribution ≥ 2. The combination of both GESs showing perfectly the strong increase after *IL-6* (GES 1) and the induced decreased after 2 hours (GES 3). The GO-Term "(external) stimulus" includes genes of the JAK-STAT signaling pathway like *STAT3* as well as several genes encoding for signaling components such as *Hamp*, *Cepbd* and *Osmr*. These entities represent regulatory processes like negative feedbacks as well as secondary signaling events. Genes with negative contribution in GES 1 were associated with metabolic processes like "L-serine biosynthesis" or "fructose metabolic processes". This is in line with the function of *IL-6* as a priming factor, mediating the conversion of quiescent hepatocytes from G0 to G1 phase of the cell cycle during liver regeneration (63). It can be argued that down-regulation of genes associated with metabolic processes is due to the transformation of differentiated metabolically active hepatocytes into proliferative cells. The down-regulated metabolic functions at least partially take place in mitochondria. Accordingly, parts of the glycolysis pathway were down-regulated in primary hepatocytes.

GES 2 shows a slight decrease after stimulation followed by a late-phase increase in expression. We identify several biological processes associated with "cell cycle and division" within this GES. A representative gene of GES 2 is the cell cycle inhibitor *Cdkn1b*. Its reduction of expression corresponds to the induction of cell cycle progression and in particular to the transfer from G0 to G1. These characteristics are further supported by the negative contribution of *Cdkn1b* in GES 3. Analyzing genes with a positive contribution in GES 2 only, we found, in addition to involvement in

early cell cycle events, genes showing an association with (programmed) cell death and apoptosis. It was already indicated that *IL-6* promotes liver regeneration and protects against liver injury by inducing anti-apoptotic and survival genes (63; 254). GO-Terms corresponding to genes found in GES 2 having a negative contribution are more heterogeneous. Within the top GO-Terms we identified several biological functions associated with the *IL-6* stimulus. Based on the induction of the acute inflammatory response, coagulation factors were activated. Moreover, several genes associated with gene translation were found. In addition, genes associated with metabolic processes are represented by this GES.

The time course behavior of GES 3 shows a delayed activation subsequent to stimulation with *IL-6*. We identified several GO-Terms associated with "cell cycle" and "cell division" similar to GES 2. However, GES 3 includes mainly genes related to late events in the cell cycle, i.e. during G2 and M phase (e.g. *Gmnn*, *Mcm2*, *Plk2*). *Wee1* as a main regulator of *Cdc2* displays a negative contribution to GES 3, hence indicating *Wee1* down-regulation and subsequent progression through the G2-M check point. In addition, we identify *Ccnb2* a late cell cycle genes, which repression leads to cell cycle arrest in the G2 phase. The time-course expression pattern, shows a strong increase after *IL-6* stimulation followed by a decrease after two hours. We identify *Ccnb2* in GES 1 and GES 3 perfectly reconstruct the strong increase after the stimulation and the inactivation after two hours. The *IL-6*-induced priming phase is characterized by the activation of the latent transcription factor *STAT3*. This immediate response induces the expression of early responsive genes like the transcription factor *AP-1* (282) subsequently inducing a secondary gene response leading to transcription of cyclins *A*-*E*, *p53*, and the cyclin dependent kinase *P34-cdc2* (2).

Applying KEGG pathway enrichment, we found the cell cycle, with DNA replication in particular, and p53 pathway enriched within this GES. Interestingly, *IL-6* stimulation alone is not sufficient to efficiently induce proliferation of primary mouse hepatocytes *in vitro*. Hence, despite the persistent re-organization of the induced gene expression profile and the induction of early cell cycle players such as cyclin *A*, additional stimuli may be necessary to initiate a strong proliferative response of primary mouse hepatocytes.

GES 4 shows the lowest eigenvalue. It has a strong increase in expression following the *IL-6* stimulus. GO-Term enrichment reveals several biological processes found in GES $1-3$ like coagulation, translation, acute phase, and response of the stimulus. Genes having a negative contribution in GES 4, indicating a decrease in expression

after the stimulus, are again associated with metabolic processes. Both, GES 3 and 4 imply that hepatocytes stimulated with *IL-6* show affection for division causing a down-regulation of genes associated with the metabolic processes.

Source	Mode	Biological process
1	positive	(external) stimulus, inflammatory response
	negative	(fructose) metabolic process
2	positive	early cell cycle and division
	negative	metabolic process, apoptosis
3	positive	late cell cycle and division
	negative	-
4	positive	translation, coagulation
	negative	(protein) metabolic process

Table 7.1: **Main biological processes in response to *IL-6*.** Summary of the main biological processes in hepatocytes regulated as response to *IL-6*. Mode indicates genes with significant positive (≥ 2) or negative (≤ -2) contribution to the source. The main biological processes found for the corresponding group of genes are given in the last column.

Validation of the time-dependent signals

In order to evaluate our findings, we compared the outcome of GraDe with standard methods. As there is no established matrix factorization technique that incorporates prior knowledge, we employed PCA (221), k-means clustering (123) and FunCluster (95), a clustering method that incorporates Gene Ontology information into the clustering task.

To test the biological findings obtained by GraDe, we applied a similar approach as proposed by Teschendorff *et al.* (266). We first asked how well biological pathways can be mapped to the inferred submodes or clusters. For GraDe and PCA, we selected in each submode the genes having an absolute source contribution above 2 standard-deviations. The average number of selected genes in each submode ranges from 75 to 280. For k-means clustering, we infer 8 clusters on a subset of the top 15% most variable genes to ensure that the average number of selected genes is comparable to GraDe and PCA.

To evaluate the mapping of pathways to submodes or clusters we applied the pathway enrichment index (PEI). For each submode or cluster we evaluated significantly enriched pathways by using a hypergeometric test (see Methods). The PEI is then defined as the fraction of significant pathways mapped to at least one submode or cluster. The PEI for each method is shown in Figure 7.6. We find that the PEI is higher for GraDe compared to PCA, k-means clustering or FunCluster indicating that GraDe maps submodes closer to biological pathways.

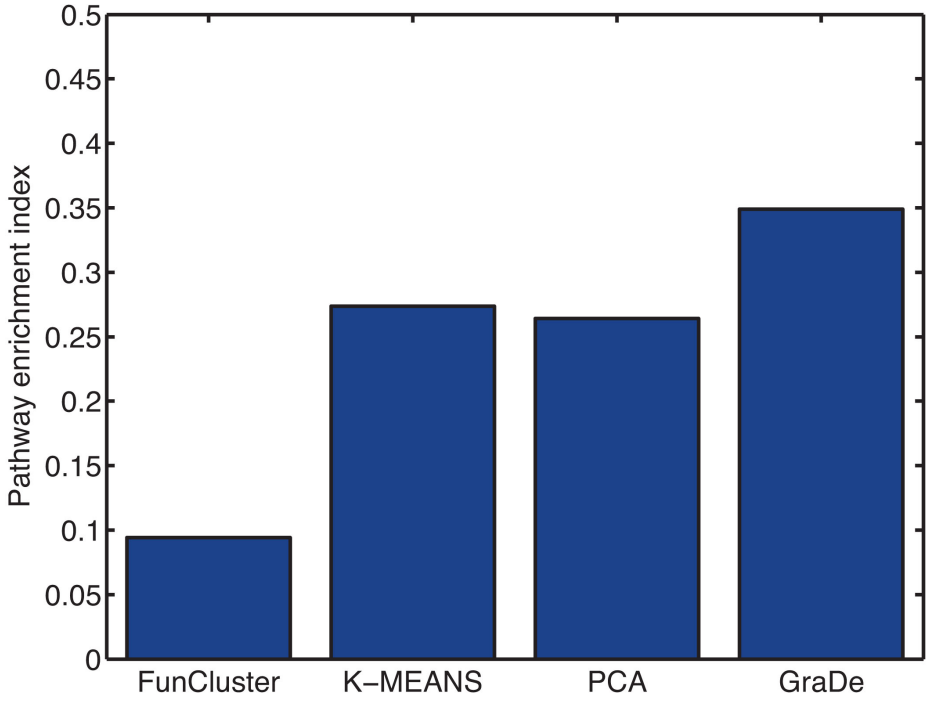

Figure 7.6: **Pathway enrichment.** Result of the pathway enrichment analysis. For each method applied to our data set, we plotted the pathway enrichment index (PEI). This index gives the fraction of KEGG pathways found enriched in at least one submode or cluster (see Methods). GraDe obtained a much higher PEI than PCA or k-means clustering. This indicates that sources obtained by GraDe map much closer to biological pathways.

In addition, we validated the time-dependent responses upon *IL-6* stimulation in more detail by searching for enriched GO-Terms. Applying PCA, we found that the first principle component (PC) contains 99% of data variance. GO-Term enrichment analysis revealed that PC 1 contains genes linked to (blood) coagulation and hemostasis. A second major response after *IL-6* is the activation of cell cycle or cell division. We found an enrichment of these biological processes in PC 2 and PC 4. PC 2 shows

a decreased time-course behavior after the stimulation. Genes linked to cell cycle and corresponding pathways have a negative contribution in PC 2 indicating an increased time-course expression pattern after IL-6 stimulation. This finding is analogous to GraDe, where we find cell-cycle pathways in GES 1 and 3 showing also an increasing expression pattern after the stimulation. With GraDe we identified several genes that are associated with metabolic processes showing a down-regulation after stimulus. PCA covers these biological processes by two components PC 2 and PC 3, where PC 3 shows a strong increase and PC 2 a decrease of expression after the stimulus (see Figure 7.7A). The direct response of *IL-6* was found in PC 4, but we identified only acute inflammatory response. Moreover, PCA grouped cell cycle (negative mode) and the direct response (positive mode) into PC 4 and was not able to separate the cell cycle processes into the early (e.g. *Cdkn1b*) and late (e.g. *Mcm2*) responses after *IL-6* stimulation.

Focusing on the results of the k-means clustering, we obtained an enrichment of cell cycle processes in cluster 3 (see Figure 7.7B). This cluster shows only a marginal increase in expression after the stimulus and therefore does not reflect the strong activation of cell cycle found by GraDe and PCA. Genes associated with metabolic processes are grouped in cluster 5, which has a constant expression level after *IL-6* stimulus. Hence, k-means clustering failed to infer a cluster associated to the downregulation of metabolic processes upon *IL-6*. Cluster 4 shows a characteristic time-course pattern after *IL-6* stimulation, but we were not able to reveal any significant biological processes associated to *IL-6*. Altogether, k-means clustering neither identifies the direct response upon *IL-6* nor the separation between early and late cell cycle genes.

Comparing the result of FunCluster, we also identify a set of co-regulated genes associated with cell cycle (Cluster 3; see Figure 7.7C). Genes grouped in this cluster show an increase in expression after the stimulus. However, FunCluster was also not able to separate the early and late cell cycle processes, observed by GraDe. Genes associated with metabolic processes are grouped in cluster 5, showing a decreasing expression pattern after one hour of stimulation. Therefore, FunCluster also identifies the downregulation of metabolic processes indicating that IL-6 reduces expenditures for the energy metabolism. However, FunCluster was not able to identify the primary response of IL-6 mediating the production of acute phase proteins (APP) by hepatocytes. Moreover, FunCluster also did not find any significant processes related to the JAK-STAT related genes, such as *Stat3*, *Hamp*, *Cepbd* and *Osmr*, showing an increased expression pattern.

These results show that the decomposition obtained by GraDe provided much more detailed biological insights than PCA, k-means clustering or FunCluster. PCA was able to identify three main biological processes upon *IL-6* stimulus. However, it failed to give a correct time-resolved pattern of these biological processes, whereas sources from GraDe reproduce the characteristic time-course behavior of the *IL-6* response. Moreover, GraDe reveals a much more structured and time-resolved result, which allows assigning each source to a different main process.

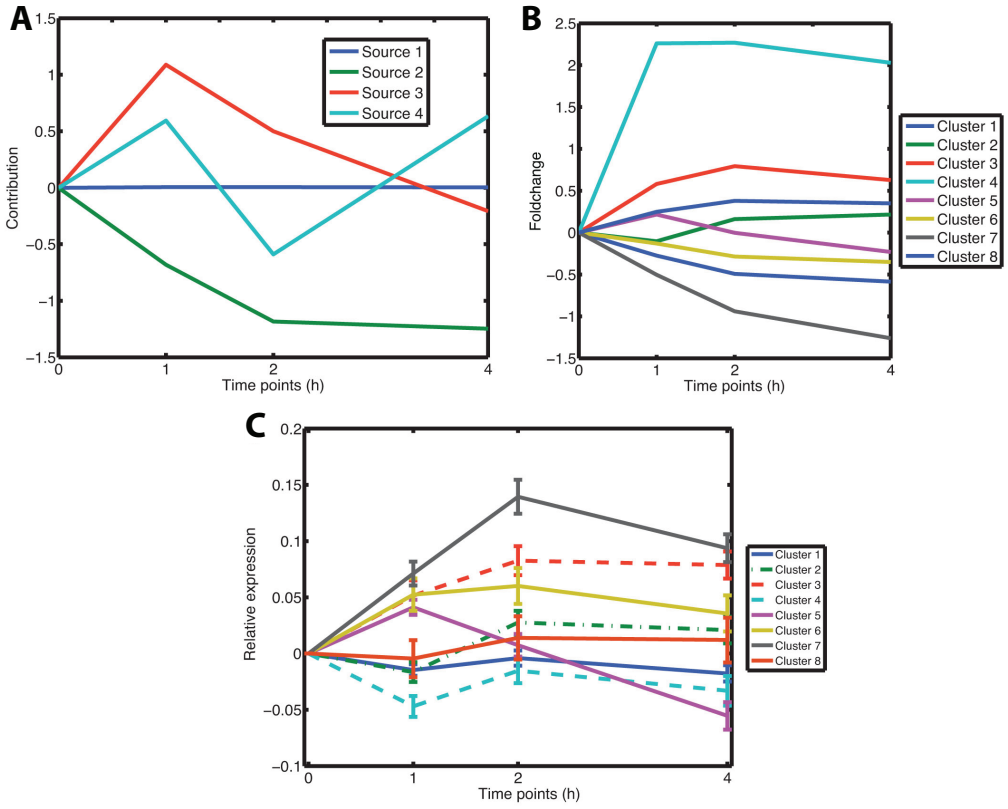

Figure 7.7: **Result of PCA, k-means clustering and FunCluster.** (**A**) illustrates the result of PCA for the time-course data of *IL-6* stimulated hepatocytes. The x-axis corresponds to the measured time-points and the y-axis gives the centered (to time point 0h) contributions of the mixing matrix. The result of the k-means clustering is shown in (**B**). The x-axis shows the measured time-points and the y-axis shows the fold-change values of the centroids at that time-points. (**C**) shows the result of FunCluster. The plot shows the mean expression of the different cluster and the bars indicates the standard deviation at a particular time-point. The x-axis shows the measured time-points and the y-axis shows the relative expression at that time-points.

Robustness analysis

Detailed knowledge about gene regulation is often not available and far from complete. Therefore, the quality of a large-scale gene regulatory network is not perfect. In order to test the effect of network errors on the output of GraDe, we performed two robustness analyses. Starting with our TRANSPATH network, we generated randomized versions by either shuffling the network content or adding random information (see Methods). By shuffling edge information of the gene regulatory network between 0.1 and 100% of all original edges, we simulated a loss of information. To quantify robustness, we employed the Amari-index, which measures the deviation between two mixing matrices. We obtained significantly low Amari-indices for up to 3% reshuffled edges within the gene regulatory networks (mean Amari-index = 3.83, $p = 0.034$), whereas a complete randomization of the network results in an Amari-index of 9.63 (see Figure 7.8B). This shows that the quality of the regulatory network has of course a strong influence on the output of the GraDe algorithm. It is obvious that GraDe depends on the regulatory network, and replacing gene interaction through random information will lead to loss of the signals.

We ran a second robustness analysis by adding random information to the existing gene regulatory network. This is important because we expect large-scale networks extracted from literature to contain many false-positives. Significantly low Amari-indices were obtained by adding up to 13% random information (mean Amari-index = 3.94, $p = 0.046$) to the network (see Figure 7.8B). This result shows that GraDe is able to detect the signals even after adding a large amount of probably wrong information to the network. The tolerance of the algorithm to the second randomization strategy is much higher, as here no correct information is destroyed. Overall, with both randomization procedures we were able to prove that GraDe is robust against a reasonable amount of both, false positives and missing information.

In addition, we analyzed the noise effect of gene expression data by randomly choosing between one and three replicates for each time point. We found significantly low Amari-indices (mean Amari-index = 4.16 $p = 0.026$) by comparing the 95% quantile of the resulting Amari-index with a random sampling. Thus, GraDe is also robust against biological noise.

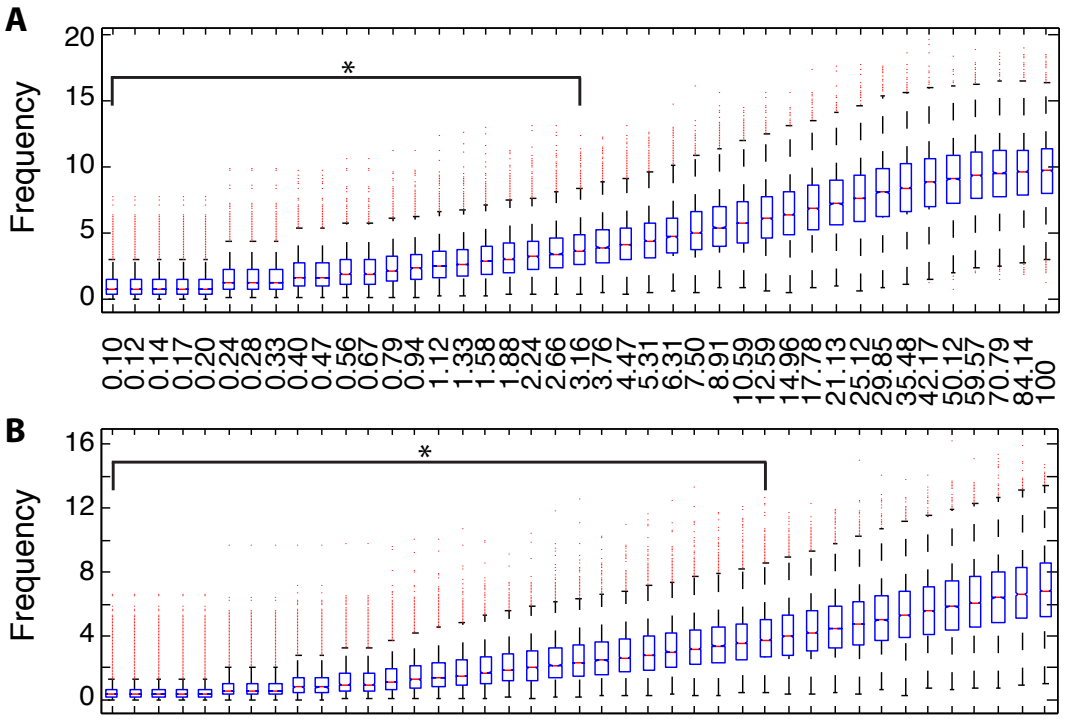

Figure 7.8: **Robustness analysis.** Robustness analysis: We evaluated the robustness of GraDe against errors in the underlying graph. To this end, we compared the mixing matrix that we extracted with the TRANSPATH network with those obtained based on perturbed versions. For this comparison we use the Amari index (see Methods). The boxplots show Amari-indices obtained with (**A**) a network rewiring approach and (**B**) when adding random information to the network. The x-axis shows the amount of information randomized (in %), the y-axis gives the obtained Amari-index. $*$ indicates significant 95% quantiles compared to a random sampling (p-value ≤ 0.05). We see that GraDe is robust against a reasonable amount of wrong information.

7.2.5 A microarray experiment on stem cell differentiation

The regulation of gene expression is essential to proper cell functioning. The first step of gene expression is mRNA transcription. Here, a copy of a gene from the DNA to messenger RNA (mRNA) is made, encoding a chemical "blueprint" for a protein product. Microarrays are the state-of-the-art technology for the genome-wide measurement of these transcript levels. They are known to be quite noisy, and the still high costs keep the number of replicates small. This makes gene expression analysis a particular challenge for machine learning. Matrix factorization techniques are currently explored as unsupervised approaches to such data (229). Here, the extracted gene expression

sources (GESs) can be interpreted as distinct biological processes, which are active on a level quantified in the mixing matrix. Applying GraDe, we require that biological processes that can be explained by the underlying network are not split up between different GES.

In the following, we interpret a microarray experiment investigating the crucial role of the transcription factor STAT5 during hematopoietic stem cell differentiation. STAT5 is strongly activated in 30% of patients with acute myelogenous leukemia (AML) (132). The cytokine GM-CSF activates STAT5 and controls the differentiation of progenitor cells (GMPs) into granulocytes and macrophages, two types of white blood cells. AML cells are mostly from the granulocyte lineage, hence it is important to elucidate the role of STAT5 upon GM-CSF stimulation.

In (132), both normal and STAT5-knockout GMPs were stimulated with GM-CSF or left unstimulated. RNA from these four samples was measured with microarrays. Data are available at GEO under accession number GSE14698. We used 1601 differentially expressed genes for further analysis (t-test on stimulated vs. unstimulated gene expression significant with a p-value< 0.05).

To interconnect these genes we used the known gene–gene interactions that are collected in the database TRANSPATH (150). Applying GraDe, we preferred a single shift, as multiple shifts may lead to an accumulation of errors in the interaction network. We obtained four GES (Figure 7.9A). We selected genes in the GES that were expressed above the threshold ± 2 and mapped these sets onto biological processes by performing Gene-Ontology (GO) enrichment analysis.

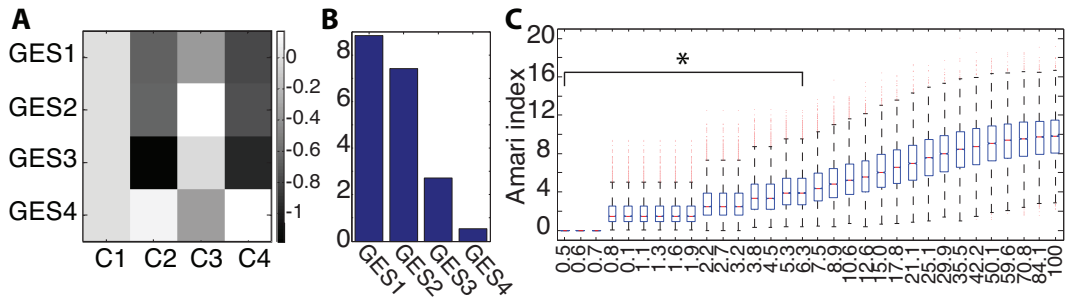

Figure 7.9: **GraDe result and robustness analysis:** The heatmap (**A**) shows the mixing matrix (centered to C1). Conditions C1–C4 correspond to stimulated/unstimulated GMPs (C1-C2) and STAT5–ko cells (C3-C4). (**B**): Eigenvalues of the four GES. (**C**) Amari-indices for the randomized networks against the fraction of randomized edges (in %). The $*$ indicates significant Amari-indices.

GraDe separates GM-CSF and STAT5 dependent processes

The sources extracted by GraDe separate GM-CSF and STAT5 dependent biological processes. Figure 7.9A shows that the contribution of GES 1 differs between stimulated and unstimulated GMPs, but also between wild-type and STAT5–ko. Enriched GO–Terms correspond to responses triggered by external stimuli, activation or regulation of signal transduction as well as responses to STAT5 such as the MAPK or JNK signaling cascades. This is line with previous work (242). In addition, we found processes linked to cell differentiation, the primary response to GM-CSF stimulation (132). GES 2 separates stimulated and unstimulated cells, independently of STAT5 condition. Consequently, we identified biological processes linked to immune response. GM-CSF is an important hematopoietic growth factor for enhancing immune responses and is known to recruit and activate antigen-presenting cells (148). It also has profound effects on the functional activities of various circulating leukocytes, which are involved in defense processes (242).

Genes in GES 3 can be linked to telomere organization and maintenance. Telomerase activity is reported to be nearly 3-fold higher in GM-CSF stimulated cells (67). Telomere length is a critical factor in determining the replicative potential of mitotic cells. The accelerated telomere shortening due to excessive replication may hint at hematopoietic stem cell premature aging.

GES 4 has a different contribution in stimulated and unstimulated STAT5–knockouts. Similarly to GES 2, we found several processes linked to immune responses. Additionally, GES 4 contains genes associated with the activation of macrophages or myeloid leukocytes. GM-CSF is involved in the generation of granulocytes and macrophages, responsible for non-specific defense processes.

Applying PCA or ICA (104) to the data, we found only in two sources significantly enriched GO–Terms. Note that we cannot employ SOBI since a gene's position on the microarray chip is completely arbitrary. PCA extracted a source linked to immune response and aggregates various biological processes in a second one. We obtained a similar result performing ICA with JADE (31), which also separates the immune response from other biological processes. Thus, GraDe finds a much more structured, detailed response than fully blind approaches.

Robustness to graph errors

The quality of the employed regulatory network is not perfect. It may contain false interactions and is far from complete. To analyze the robustness of GraDe, we randomized between 0.5 and 100% of the edges in the TRANSPATH graph. In each step we reshuffled 10.000 times the corresponding percentage of edges using degree-preserving rewiring (284). Applying GraDe with these graphs we obtained new factorizations. We used the Amari index to quantify changes and determined a p-value to detect significantly low changes. This p-value was calculated by comparing the 95% quantile of Amari-indices for each randomization step with Amari-indices for normally distributed random separating matrices. We obtained significantly low ($p < 0.05$) Amari-indices for up to 6% of rewired edges (Figure 7.9C). Thus, the underlying graph has obviously a strong influence on, but we showed that GraDe is robust against a reasonable amount of graph errors.

7.2.6 Differentiation of glutamatergic neuros: Combined analysis of mRNA and microRNA data

Stem cell differentiation is the development of a stem cell to a fully differentiated cell. The definition of a stem cell includes two properties: (*i*) Self-renewal, denote the ability to maintaining the undifferentiated state while go through cell division. (*ii*) Potency, denote the capacity to differentiate into specialized cell types. Embryonic stem cells are derived from the epiblast tissue of the inner cell mass. Embryonic stem cells are pluripotent and differentiate during development to all derivatives of the three primary germ layers: ectoderm, endoderm and mesoderm. In this application, mouse embryonic stem cells were stimulated with retinoic acid after day 4 (CA4d). After day 6 (CA6d), cells were restimulated with retinoic acid, which leads after day 9 to radial glia cells. After day 15, stem cells derived neuronal progenitor cells differentiated into glutamatergic neurons The whole procedure is summarized in Figure 7.10.

During the differentiation process both mRNA and miRNA expression was measured at five different developmental stages. The goal of the analysis is to define a small regulatory network around Pax6, a transcription factor playing a important role in the differentiation process. We apply GraDe with the underlying regulatory network to the combined mRNA and miRNA data. To interconnect genes we used known gene-gene interactions that are provided by the TRANSPATH database (150). For the

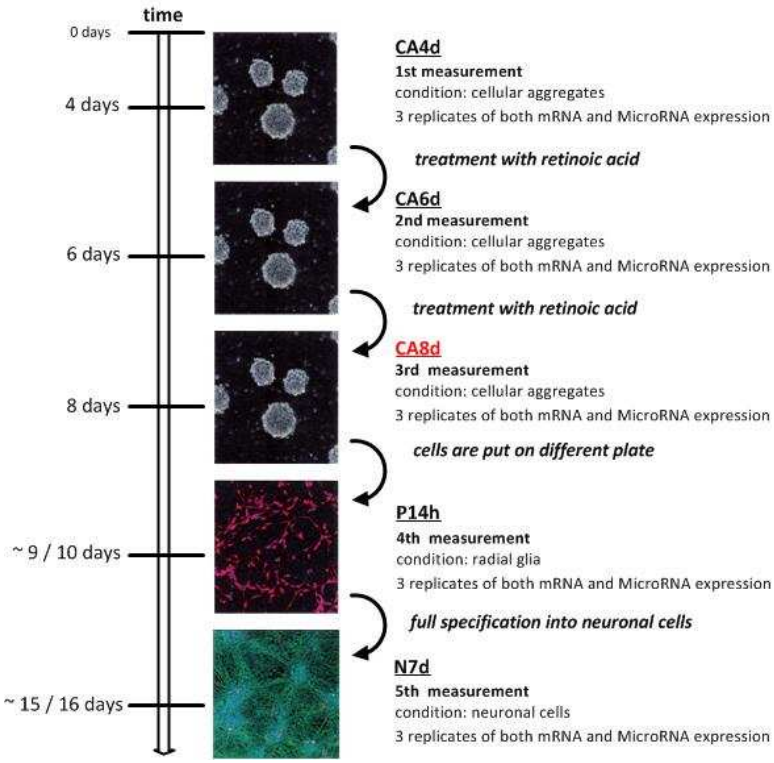

Figure 7.10: **Workflow of the measured mRNA and miRNA data:** The cellular aggregates (CA) were treated with retinoic acid twice between measurements 1 & 2 and 2 & 3 to induce neuronal differentiation. After measurement 3 on day 8, cells were put on a different plate where they differentiated into radial glia cells. The last measurement was performed more than 6 days later at the fully differentiated neuronal state. CA8d indicates in red is the crucial differentiation process since it is after full induction of neuronal fate by retinoic acid and $\sim$1-2 days before glia stage, which are already neuronal progenitors. Figure provided by (174).

interaction of miRNAs and genes, we used TargetScanS and PicTar prediction tools. We obtain five GES and the resulting strength of the Eigenvalues (see Figure 7.11B). Based on the Eigenvalues we choose the top three GESs to analyze Pax6 related genes and miRNAs. Figure 7.11A shows the time-course behavior of the resulting Sources.

Pax6 has the strongest eigenvalue in GES3, but will be also represented by a negative contribution of GES1, which results in total to the obtained time-course illustrated in Figure 7.12A. To set up a small regulatory model centered by Pax6, we first extracted all miRNAs targeting Pax6. Using TargetScanS and PicTar we identify six miRNAs, which are expressed at least one time point. Next, we maintain only those miRNAs having an absolute source contribution ≥ 1.5. This filter steps results in miR-129-5p

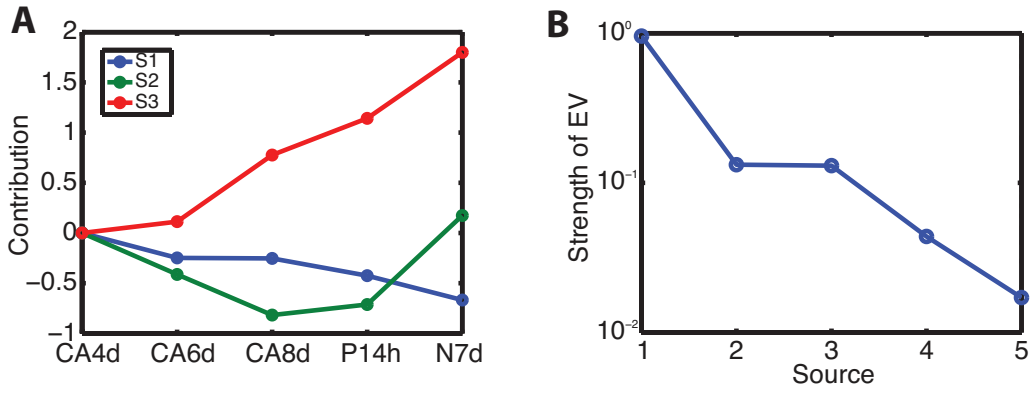

Figure 7.11: **This figure illustrates the decomposition of the mouse embryonic stem cell differentiation experiment.** As underlying network we used interactions from the TRANSPATH database and miRNA target prediction using TargetScanS. (**A**) shows the time-courses of the top three extracted sources, centered to time point CA4d. The x-axis shows the measured time-points and the y-axis the contribution of the mixing matrix. In (**B**), we plot the strength of the eigenvalues (EV) of the resulting sources.

and miR-300. To extend this model, we then identify transcription factors (TFs), which either regulate Pax6 or the identified miRNAs. One important transcription factor is Sp1, which has an high impact on the regulation of Pax6 (299). Based on miRNA target prediction, we identify miR-495, which is (*i*) located in the miR-300 genomic cluster and (*ii*) target Sp1 and (*iii*) has a source contribution ≥ 1.5 for GES 3. We therefore include miR-495 into our set of miRNAs to include miRNA regulation on PAx6 and Sp1. In order to further extend the model by transcription factors we applying a transcription factor binding sites prediction approach using the MatInspector software (33). MatInspector returns a results a large database of transcription factors for different locations within the transcription factor binding sites. For further analysis, we used only those transcription factors having a score of 1, which can be translated as a perfect hit of the resulting transcription factor matrix to the binding site. We filtered for predicted transcription factor for Pax6, miR-129-5p, miR-300 and miR-495 and having a source contribution of ≥ 1.5. Reducing the set of transcription factors to those, results in regulatory genes showing a strong (graph)-correlated expression pattern with either Pax6 or identified miRNAs. Figure 7.12A shows all interaction and regulation, which are either literature based (known edges) or predicted (putative edge). The minimal model containing all target genes and regulation transcription factors is shown in Figure 7.12B. The resulting model can then used for network inference

methods using prior knowledge. Lutter et al. (175) developed a novel approach based on Bayesian inference of boolean networks to construct gene regulatory models from combined mRNA and microRNA (miRNA) expression data. In order to determine an 'a priori model' space, the resulting model obtained by GraDe, TF and miRNA target prediction is used. Sample from this boolean model space and the likelihood for each model is estimated to predict the measured gene expression profiles. The discrete models comprised multiple instances of weighted, directed gene-interactions, where nodes represent either genes or miRNAs and the edges the various regulatory interactions. Using a parallel optimization method based on evolutionary computing, we obtain an approximation to a maximum posterior of our model space.

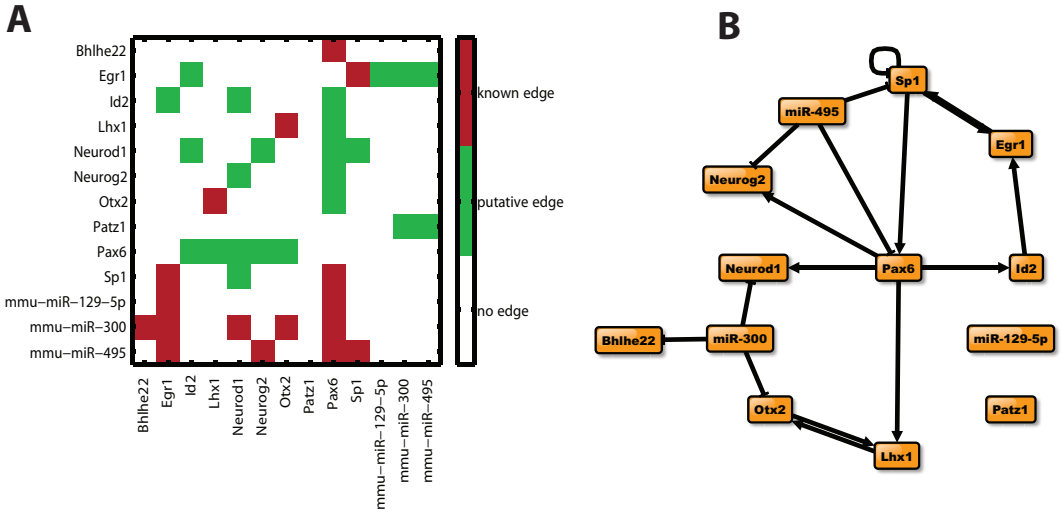

Figure 7.12: **The resulting regulatory model of the mouse embryonic stem cell differentiation process.** (**A**) Connectivity matrix of all interaction and regulation, which are either literature based (known edges) or predicted (putative edge). (**B**) The regulatory model based on the Bayesian inference of boolean networks approach.

We used GraDe to analyze combined data set of mRNAs and miRNAs from a high-throughput experiment. Applying GraDe on the mouse embryonic stem differentiation data, we identify a set of genes and miRNAs, that have strong (graph)-correlated time-course pattern compared to Pax6, a central player in the differentiation process. In a first step, we filtered genes and miRNAs based on their graph-correlation and second we used miRNA target and transcription factor binding site prediction to finally connect the selected set of differentiation player in a small regulatory model. The result of this application shows that GraDe can successfully applied to multi-scaled data and

infer beside large-scale relations also small scale networks.

7.3 Material and Methods

7.3.1 *IL-6* stimulated mouse hepatocytes

RNA probes from primary mouse hepatocytes were assessed with the Bioanalyzer 2100 (Agilent) to ensure that 28S/18S rRNA ratios were in the range of 1.5 to 2.0 and concentrations were comparable between probes. For each time point, 4 μg of total RNA were used for the hybridization procedure using the One-Cycle Target Labeling Kit (Affymetrix). Fluorescence intensities were acquired with the GeneChip Scanner 3000 and the GCOS software (Affymetrix). GeneChip Mouse Genome 430 2.0 Arrays (Affymetrix) were used in the analysis comprising stimulations with 1 nM *IL-6* for 1 h, 2 h, 4 h and an unstimulated control (0 h) each performed in triplicates. As a probe level model (PLM) for microarray data an additive-multiplicative error model was used. Data processing was performed using the Limma toolbox (251) provided by Bioconductor (74). The RMA approach was used for normalization and background correction. Probe sets were filtered out by the genefilter package. A gene was considered as expressed if the signal was above 100 (unlogged data) for at least one time point. Finally, we obtained a data set of 5709 genes. Significantly regulated genes compared to time point 0 h were determined by using the LIMMA (Linear Models for Microarray Data) method (250). The Limma toolbox uses the moderated t-statistics to identify significant regulated genes. Moreover the moderated t-statistics is advisable for a small number of arrays (273; 250). A gene was determined as significant regulated if the p-value was < 0.05 after multiple testing correction by the Benjamini-Hochberg procedure (13). Raw data are available at GEO with accession number GSE21031.

7.3.2 Gene Regulatory network

In order to link genes along an underlying network we used the TRANSPATH database (150) that provides detailed knowledge of intracellular signaling information based on changes in transcription factor activity. We searched for direct gene or protein interactions within the TRANSPATH database using the terms: transactivation, increase of abundance, expression, activation, DNA binding, increase of DNA binding, transre-

pression, decrease of abundance, decrease of DNA binding, and inhibition.

7.3.3 Principle component analysis

For principle component analysis (PCA) we performed an eigenvalue decomposition of the covariance matrix of the data set X. Thereby we obtained a decomposition into an orthonormal source matrix S and an orthogonal mixing matrix A. We applied PCA to the same set of expressed genes as GraDe and also inferred four sources. We defined for each component two submodes by grouping genes with a threshold $\geq +2$ standard-deviations and a second set of genes having a source weight of ≤ -2 standard-deviations.

7.3.4 k-Means clustering

In order to ensure a fair comparison of k-means clustering with GraDe and PCA, we first applied a gene selection step to provide that all methods selected an approximately equal number of genes, as proposed in (266). We ranked all expressed genes according to their expression variance across the time-course and then selected the top 15% variable genes. Having the selected genes, clustering was then performed using k-means clustering (123), where k was set to 8 in order to match the same number of submodes inferred by GraDe and PCA.

7.3.5 FunCluster

In addition to k-Means clustering, we also include a clustering method which incorporates Gene Ontology information into the clustering task. We use the FunCluster method, which performs functional analysis of gene expression data (29; 95). FunCluster detects co-regulated biological processes through a specially designed clustering procedure involving biological annotations (GO and KEGG) and gene expression data. We apply the FunCluster implementation provided within the R environment (264) and using standard parameters.

7.3.6 Enrichment analysis

For gene sets grouped in sources obtained by GraDe and PCA or k-means clusters we performed an enrichment analysis to determine significantly enriched biological pro-

cesses and pathways. For biological processes we performed a Gene Ontology (28) term enrichment analysis, which was carried out with the R package GOstats (74). For pathway enrichment analysis we used non-metabolic pathways that are manually curated by KEGG (121). Pathway enrichment was also evaluated with the GOstats package. To correct for multiple testing, we used the Benjamini-Hochberg procedure (13) and called an association significant if the p-value was less than 0.05. To evaluate the mapping of pathways to submodes or clusters we applied the pathway enrichment index (PEI), as proposed by (266). For each submode or cluster we evaluated the significance of enrichment of a set of genes in a particular pathway by using a hypergeometric test. A pathway association was considered as significant if the p-value was below 0.05 after multiple testing correction using the Benjamini-Hochberg procedure. The PEI was then defined as the fraction of significant pathway mapped to at least one submode or cluster.

7.3.7 Robustness analysis

Robustness analysis was performed by two network randomizations. The gene regulatory network is interpreted as a weighted bipartite graph, i.e. a graph with two sets of nodes (regulators and regulated genes). Weighted edges indicate interactions either activating or inhibiting. First, we randomized existing edge information within the network between 0.1 and 100%. In each step we shuffled 10.000 times the corresponding amount of edges using a degree-preserving rewiring (184; 284). Applying GraDe with the resulting networks we obtained new factorizations. To compare the original and new results in a quantitative way, we used the Amari-index (41). For each step we took the 95% quantile of the random sampling and calculated a p-value by comparing this quantile to Amari-indices obtained comparing normally distributed random separating matrices and the original mixing matrix. In a second randomization approach, we added 10.000 times new information (edges) between 0.1 and 100% of the original network content and calculated the 95% quantile of the resulting Amari-indices. Again, the p-value was calculated by comparing each quantile with a random sampling.

For analysis of robustness against noise we randomly chose between one and three replicates for each time point and ran GraDe. For each run, we calculated the Amari-index. Again, we compared the 95% quantile of the resulting distribution with a random sampling to obtain the corresponding p-value.

7.4 Conclusions and outlook

Matrix factorization techniques provide efficient tools for the detailed analysis of large-scale biological and biomedical data. While underlying algorithms usually work fully blindly, we propose to incorporate prior knowledge about gene, miRNA or even protein regulation encoded in a graph model. This graph introduces a partial ordering in data without intrinsic (e.g. temporal or spatial) structure, which allows the definition of a graph-autocorrelation function. Using this framework as constraint to the matrix factorization task, we develop a second-order source separation algorithm called graph-decorrelation algorithm (GraDe). First, we demonstrate applicability of GraDe by two different toy examples (time-series and condition experiments), which reflects the common experimental designs. Furthermore, we introduce G-MA processes, which we used to evaluate the performance of GraDe. These processes are theoretically very interesting and need further investigations in upcoming projects. In order to show the robustness and applicability of GraDe for biological data, we discuss three different applications in this thesis. We demonstrate the robustness against noise in the underlying prior network. Moreover, we show within these applications that GraDe obtains many more reliable results compared to clustering and even standard matrix factorization approaches. These results indicate that the integration of prior knowledge into the matrix factorization task improves the inference of biological findings in large-scale data. In future work, we will investigate whether GraDe can be used for model selection when given different alternative underlying graphs of small-scale models. Furthermore, we are interested in using the concept of GraDe to analyze single signaling pathway responses in large-scale biological data. Within the European Research Council grant latent causes in molecular networks, we will work on a Bayesian formulation of the non-linear source separation problems, which is used in GraDe.

8 Conclusions and Outlook

Since the discovery of miRNAs around ten years ago, enormous efforts have been made to study and unveil the regulatory role of these molecules. One important research field was the identification of miRNA target genes. Various studies proposed several features, which seem to be important in characterizing miRNA target genes. With the resulting identification of thousands of potential miRNA targets genes, the challenge is now to further unveil their biological functions and the corresponding regulatory role of the miRNAs. In the last years, new large-scale approaches, such as Hits-Clip or Par-Clip have been developed, which are able to identify the transcriptome-wide miRNA-binding sites on mRNAs. Decreasing the number of false positive target genes will help to clear the picture and will lead to further findings. A future challenge will be the systematical identification of all miRNAs affecting and regulated by cell signaling. Although we are far from this goal, the experimental tools are definitively in place, including the capacity to screen for miRNAs that contribute to discrete signaling events using unambiguous and pathway-specific readouts in cultured cells or other model systems (106). Many disorders, such as heart disease, autoimmunity and cancer arise from defects in signal transduction networks. Recent studies identified miRNAs as important members of signaling networks acting either as backups of post transcriptional control or as important control instances of feed-forward and feedback loops. Located at these central and important switching points they confer the robustness of the networks. Therefore an intensive study of the regulatory role of miRNAs on these networks may lead to novel drug targets in the future. The goal of this thesis was to study the regulatory motifs of miRNA-mediated regulation in signaling pathways. We presented novel findings first from a general perspective in Chapter 3 and 4. To improve the inference of functional miRNA-pathway associations, we presented a novel approach in Chapter 5. In Chapter 6, we showed the results of a detailed mathematical model to unveil the regulatory role of miRNAs on the pathway dynamic. Finally

in Chapter 7, we presented a novel approach to study the responses of active signal transduction network in large-scale biological data.

8.1 Disease-associated microRNAs and their role in signaling pathways

Our analysis of deregulated miRNAs linked them to various diseases starting from neurodegenerative disorders to cancer. For this study, we used the novel manually curated database PhenomiR. First, we study the difference of disease-related miRNAs and their corresponding expression patterns between cell cultures and living organisms. Our analysis reveals that especially in cancer expression profiles between in vitro and in vivo systems do not correlate. The result indicates that depending on disease type, integration of independent information from cell culture studies are in conflict to conclusions drawn from patient studies. These observations and the results of our study show that the potential of cell cultures to investigate miRNA expression in diseases is limited. As a consequence, the suitability of cell cultures has to be verified for each disease and cell line before using such data as tool for the prognosis of diseases in human beings. Analyzing the PhenomiR database, we identified several studies, which did not investigate the impact of single deregulated miRNAs but rather groups of miRNAs organized in genomic clusters. Linking these genomic clusters to phenotypes, we identify for the first time that deregulated miRNA genomic clusters are significantly overrepresented in the majority of investigated diseases compared to single miRNA genes. The pivotal role of miRNA clusters in human diseases suggests that effective treatment of various diseases may require a combinatorial approach to target not singular miRNAs but rather miRNA clusters. In order to further study the role of disease-related miRNAs, we analyze their role in regulating signal transduction pathways. Therefore, we set up a multipartite graph consisting of five sets of nodes and links, established by different data resources. Using a thorough statistical analysis of a multipartite graph consisting of miRNAs, proteins, diseases, and signaling pathways in a tissue-specific manner, we link diseases via miRNAs on signaling pathways and uncover the impact of disease-associated miRNAs on human signaling pathways. Further improvements in the accuracy of miRNA target prediction will lead to more detailed findings. In addition, novel techniques such as our proximity measurement will help to better understand the functional relation between signaling pathways and

their regulation through miRNAs.

8.2 Inferring functional microRNA-pathway associations

In this thesis, we present a novel approach to link miRNAs and signaling pathways using our proximity measure. This technique goes beyond the common enrichment approach by incorporating the topology of the biological networks. Applying the proximity score to a global set of experimentally validated miRNA targets, we identify functional miRNA-pathway associations that significantly differ from those inferred with the conventionally used enrichment score. We are able to identify additional subclasses of miRNA pathway associations in addition to the enrichment of miRNA target patterns. Using gene ontology annotation, we show that proximal target patterns correspond to a specific function in cell signaling. We were able to show that the application of concepts from graph theory to signal transduction allows the identification of novel miRNA-pathway associations. In addition, we presented our novel web server miTALOS that provides novel features for the functional analysis of miRNA-mediated regulation in biological pathways. MiTALOS offers two different methods to identify functional miRNA-pathway associations. The two measures provide significant miRNA-pathway associations for two alternative forms of miRNA control. As miRNAs and their target genes show highly tissue-specific expression signatures, miTALOS provides a tissue filter. This is a novel feature in contrast to already existing resources, where the functional analysis is corrupted by targets that are not expressed in the tissue under consideration. In a functional analysis of prostate cancer related miRNAs, we showed the benefit of the proximity score and novel features provided by miTALOS to identify biological meaningful miRNA-pathway associations. We think that the concept of proximity can serve as a powerful tool to identify patterns in networks beyond miRNA regulation in signal transduction. For drug targets in metabolic networks, disease genes in signaling pathways, or other network medicine approaches, our novel measurement might generate useful hypotheses beyond the commonly used enrichment method. One possible extension is the application for user specific gene or protein lists. In case of a large-scale analysis, one is interested in linking functional annotation to a (large) list of genes or proteins of interests. So far, functional annotation for e.g. highly expressed or process related genes is commonly obtained using the

enrichment method. Applying our novel proximity measure to these data, we suggest to infer novel functional relations, which are beyond the simple enrichment feature.

8.3 Modeling microRNA-mediated regulation of signaling pathways

The research area of post-transcriptional regulation obtained new impulses after the discovery of miRNA molecules. Within this thesis, we study the regulatory role of these small molecules from a dynamical point of view. We first analyze a signaling cascade, which is representative of many signaling pathways. We show that alteration of signal transduction proteins and transcripts via gene or miRNA regulation has a severe impact on signal maintenance and shutdown. Changes in the expression of signal activators or inhibitors lead either to a fast signal shutdown via the inhibitor or to a slow signal repression via the activator. In addition, our results indicate that gene or miRNA regulation induces similar recovery times of the signal. In summary, our analysis shows that miRNA regulation as an addition layer of transcriptional control allows the cell to alter the signal transduction in context specific manner. In future work, other highly representative signaling patterns, such as scaffold proteins or even pathway crosstalks, can be analyzed to further unveil the role of miRNAs on the signal transduction dynamic.

Our miRNA-model of the gp130-STAT3 pathway can be extended into two directions. From a theoretical point, we can use a more complex model, which may describe the receptor binding complex in more detail, the nucleus transport or the negative feedback loop. Moreover, one could use a stochastic modeling approach instead of ODEs to capture the fluctuations of mRNA and protein expression during signal transduction. The most critical point is the modeling of miRNA regulation. First of all, one could integrate the mRNA:miRNA complex and its release. Here, further biological knowledge is needed to understand the biological principles of this complex and its impact on mRNA cleavage and inhibition of protein translation. From the biological point of view, miRNA regulation can be extended in different ways. Further investigation of the miRNA biogenesis, especially of the regulatory mechanism of miRNAs, is necessary. So far, we model miRNA impact on mRNA expression by combining miRNAs to a single miRNA regulation process. In case of explicit knowledge of single miRNAs we could extend the pathway model by integrating the corresponding single miRNAs. Fi-

nally, one important perspective is the experimental validation of our prediction made by the gp130-STAT3 model. To confirm these results, we will use siRNA knockdown experiments to alter JAK1 and STAT3 expression. Using these experiments, we are then able to study the importance of the PSTAT3/STAT3 ratio for the efficiency of signal transduction within gp130-STAT3.

8.4 Using biological knowledge to infer functional relationships

Matrix factorization techniques provide efficient tools for the detailed analysis of large-scale biological and biomedical data. In Chapter 7, we present a matrix factorization algorithm that incorporates prior knowledge about gene, miRNA or even protein regulation encoded in a graph model. This graph introduces a partial ordering in data without intrinsic (e.g. temporal or spatial) structure, which allows the definition of a graph-autocorrelation function. Using this framework as constraint to the matrix factorization task we developed the second-order source separation algorithm GraDe. We proved identifiability in our factorization model, where we posed constraints that are based on the novel concept of graph-delayed correlations. Starting with proof for artificial data, we also demonstrated the applicability as well as robustness of the proposed approach within three studies. The methodological strength of the proposed approach is two-fold: first, it naturally arises from a network approximation of the general ODE model of gene regulation. Second, instead of ignoring the sample dependencies in biological high-throughput data by assuming i.i.d. samples, we explicitly model them. For further work, various extensions of our GraDe algorithm can be considered. So far, we used large-scale biological networks to obtain prior knowledge and link mRNAs and miRNAs. Beside the large-scale applications, GraDe can be used for e.g. model selection. Using different (ODE) models, the concept of graph-autocorrelation could be used to select the 'best' model that is able to infer the data. Therefore, new methods for model selection based on the concept of graph-decorrelation have to be developed. Furthermore, GraDe could be used to identify pathway specific responses in the data. The idea of Principle Pathway Analysis is to distinguish between active and inactive pathways in biological data, e.g. after a cell stimulus. New concepts to evaluate the outcome of the resulting pathway autocorrelation based on a non-linear source separation have to be developed. Finally, GraDe can be used to derive gene-regulatory

or signaling models based on multivariate biological data. However any model only approximates reality. In case of a failing model, the question is how we are able to refine the model using this experimental knowledge. In signal processing, namely the blind identification of hidden (latent) variables in a mixing model technique, which is able to solve this problem. The concept of graph-autocorrelation can here be used to develop new approaches. Finally, a fully Bayesian formulation of the non-linear source separation problems, which is used in GraDe, may allow an alternative way of the efficient but crucial inclusion of prior biological information. Within the European Research Council grant latent causes in molecular networks, new approaches dealing with these topics will be developed over the next years.

References

[1] Adams, G. P. and Weiner, L. M., 2005: Monoclonal antibody therapy of cancer. *Nat Biotechnol* **23**:1147–1157.

[2] Albrecht, J. H. and Hansen, L. K., 1999: Cyclin D1 promotes mitogen-independent cell cycle progression in hepatocytes. *Cell Growth Differ* **10**(6):397–404.

[3] Alon, U., 2007: Network motifs: theory and experimental approaches. *Nature reviews Genetics* **8**(6):450–61.

[4] Amberger, J., Bocchini, C. A., Scott, A. F. and Hamosh, A., 2009: McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research* **37**(Database issue):D793–6.

[5] Ambs, S., Prueitt, R. L., Yi, M., Hudson, R. S., Howe, T. M., Petrocca, F., Wallace, T. A., Liu, C.-g., Volinia, S., Calin, G. A., Yfantis, H. G., Stephens, R. M. and Croce, C. M., 2008: Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer. *Cancer Res* **68**(15):6162–70.

[6] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L.-S. L., 2004: UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**(Database issue):D115–D119.

[7] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G., 2000: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**(1):25–9.

[8] Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P., 2008: The impact of microRNAs on protein output. *Nature* **455**(7209):64–71.

References

[9] Balázsi, G., Barabási, A.-L. and Oltvai, Z. N., 2005: Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **102**(22):7841–6.

[10] Bartel, D. P., 2004: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**:281–297.

[11] Baskerville, S. and Bartel, D. P., 2005: Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**:241–247.

[12] Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F. and Moulines, E., 1997: A blind source separation technique using second-order statistics. *IEEE Trans Signal Proces* **45**(2):434–444.

[13] Benjamini, Y. and Hochberg, Y., 1995: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol* **57**(1):289–300.

[14] Bentwich, I., 2005: Prediction and validation of microRNAs and their targets. *FEBS letters* **579**(26):5904–10.

[15] Berger, S. M., Pesold, B., Reber, S., Schönig, K., Berger, A. J., Weidenfeld, I., Miao, J., Berger, M. R., Gruss, O. J. and Bartsch, D., 2010: Quantitative analysis of conditional gene inactivation using rationally designed, tetracycline-controlled miRNAs. *Nucleic acids research* **38**(17):e168.

[16] Blöchl, F., Kowarsch, A. and Theis, F. J., 2010: Second-Order Source Separation Based on Prior Knowledge Realized in a Graph Model. *Lecture Notes in Computer Science* **6365**(LATENT VARIABLE ANALYSIS AND SIGNAL SEPARATION):434–441.

[17] Blöchl, F., Rascle, A., Kastner, J., Witzgal, R., Lang, E. and Theis, F., 2010: *Are we to integrate previous information into microarray analyses? Interpretation of a Lmx1b-knockout experiment*. Bentham Science Publishers.

[18] Bolognani, F. and Perrone-Bizzozero, N. I., 2008: RNA-protein interactions and control of mRNA stability in neurons. *Journal of neuroscience research* **86**(3):481–9.

[19] Bonizzi, G. and Karin, M., 2004: The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends Immunol* **25**:280–288.

[20] Boscolo, R., Sabatti, C., Liao, J. C. and Roychowdhury, V. P., 2005: A generalized framework for network component analysis. *IEEE/ACM Trans Comput Biol Bioinformatics* **2**(4):289–301.

[21] Bossis, G., Malnou, C. E., Farras, R., Andermarcher, E., Hipskind, R., Rodriguez, M., Schmidt, D., Muller, S., Jariel-Encontre, I. and Piechaczyk, M., 2005: Down-regulation of c-Fos/c-Jun AP-1 dimer activity by sumoylation. *Mol Cell Biol* **25**:6964–6979.

[22] Box, G. E. P. and Norman, R. D., 1987: *Empirical Model-Building and Response Surfaces.* Wiley.

[23] Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. and Cohen, S. M., 2003: bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113**(1):25–36.

[24] Bromberg, J. and Darnell, J. E., 2000: The role of STATs in transcriptional control and their impact on cellular function. *Oncogene* **19**(21):2468–73.

[25] Bruning, J. C., Gillette, J. A., Zhao, Y., Bjorbaeck, C., Kotzka, J., Knebel, B., Avci, H., Hanstein, B., Lingohr, P., Moller, D. E., Krone, W., Kahn, C. R. and Muller-Wieland, D., 2000: Ribosomal subunit kinase-2 is required for growth factor-stimulated transcription of the c-Fos gene. *Proc Natl Acad Sci U S A* **97**:2462–2467.

[26] Bushati, N. and Cohen, S. M., 2007: microRNA functions. *Annual review of cell and developmental biology* **23**:175–205.

[27] Calin, G. A. and Croce, C. M., 2006: MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**:857–866.

[28] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R., 2004: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32**(Database issue):D262–D266.

[29] Cancello, R., Henegar, C., Viguerie, N., Taleb, S., Poitou, C., Rouault, C., Coupaye, M., Pelloux, V., Hugol, D., Bouillot, J.-L., Bouloumié, A., Barbatelli, G., Cinti, S., Svensson, P.-A., Barsh, G. S., Zucker, J.-D., Basdevant, A., Langin, D. and Clément, K., 2005: Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss. *Diabetes* **54**(8):2277–86.

References

[30] Cano, C. E., Gommeaux, J., Pietri, S., Culcasi, M., Garcia, S., Seux, M., Barelier, S., Vasseur, S., Spoto, R. P., Pébusque, M.-J., Dusetti, N. J., Iovanna, J. L. and Carrier, A., 2009: Tumor protein 53-induced nuclear protein 1 is a major mediator of p53 antioxidant function. *Cancer Res* **69**(1):219–26.

[31] Cardoso, J. F., 1999: High-order contrasts for independent component analysis. *Neural computation* **11**(1):157–92.

[32] Cardoso, J. F. and Souloumiac, A., 1995: Jacobi angles for simultaneous diagonalization. *{SIAM} J Mat Anal Appl* **17**(1):161–164.

[33] Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T., 2005: MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics (Oxford, England)* **21**(13):2933–42.

[34] Castoldi, M., Schmidt, S., Benes, V., Noerholm, M., Kulozik, A. E., Hentze, M. W. and Muckenthaler, M. U., 2006: A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). *RNA (New York, NY)* **12**(5):913–20.

[35] Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D., 2009: BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic acids research* **37**(Database issue):D588–92.

[36] Chen, C.-Z., Li, L., Lodish, H. F. and Bartel, D. P., 2004: MicroRNAs modulate hematopoietic lineage differentiation. *Science (New York, NY)* **303**(5654):83–6.

[37] Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B., 2009: Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**(7254):479–86.

[38] Chiao, P. J., Na, R., Niu, J., Sclabas, G. M., Dong, Q. and Curley, S. A., 2002: Role of Rel/NF-kappaB transcription factors in apoptosis of human hepatocellular carcinoma cells. *Cancer* **95**:1696–1705.

[39] Choi, P. S., Zakhary, L., Choi, W.-Y., Caron, S., Alvarez-Saavedra, E., Miska, E. A., McManus, M., Harfe, B., Giraldez, A. J., Horvitz, H. R., Schier, A. F. and Dulac, C., 2008: Members of the miRNA-200 family regulate olfactory neurogenesis. *Neuron* **57**(1):41–55.

[40] Choi, W.-Y., Giraldez, A. J. and Schier, A. F., 2007: Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. *Science (New York, NY)* **318**(5848):271–4.

[41] Cichocki, A. and Amari, S., 2002: *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications.* Wiley, New York.

[42] Couzin, J., 2008: MicroRNAs make big impression in disease after disease. *Science* **319**:1782–1784.

[43] Cui, Q., Ma, Y., Jaramillo, M., Bari, H., Awan, A., Yang, S., Zhang, S., Liu, L., Lu, M., O'Connor-McCourt, M., Purisima, E. O. and Wang, E., 2007: A map of human cancer signaling. *Mol Syst Biol* **3**:152.

[44] Cui, Q., Yu, Z., Purisima, E. O. and Wang, E., 2006: Principles of microRNA regulation of a human cellular signaling network. *Mol Syst Biol* **2**:46.

[45] Dai, B., Meng, J., Peyton, M., Girard, L., Bornmann, W. G., Ji, L., Minna, J. D., Fang, B. and Roth, J. A., 2011: STAT3 mediates resistance to MEK inhibitor through microRNA miR-17. *Cancer research* .

[46] Darnell, J. E., 1997: STATs and gene regulation. *Science (New York, NY)* **277**(5332):1630–5.

[47] Davis, B. N. and Hata, A., 2009: Regulation of MicroRNA Biogenesis: A miRiad of mechanisms. *Cell communication and signaling : CCS* **7**:18.

[48] Davis, B. N., Hilyard, A. C., Lagna, G. and Hata, A., 2008: SMAD proteins control DROSHA-mediated microRNA maturation. *Nature* **454**:56–61.

[49] DeMaria, S. and Ngai, J., 2010: The cell biology of smell. *The Journal of cell biology* **191**(3):443–52.

[50] Donald, C. D., Cooper, C. R., Harris-Hooker, S., Emmett, N., Scanlon, M. and Cooke, D. B., 2001: Cytoskeletal organization and cell motility correlates with metastatic potential and state of differentiation in prostate cancer. *Cell Mol Biol* **47**(6):1033–8.

[51] Dong, H., Siu, H., Luo, L., Fang, X., Jin, L. and Xiong, M., 2010: Investigation gene and microRNA expression in glioblastoma. *BMC genomics* **11 Suppl 3**:S16.

[52] Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. and Vogelstein, B., 2003: Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**(15):8817–22.

References

[53] Du, T., Li, B., Liu, S., Zang, P., Prevot, V., Hertz, L. and Peng, L., 2009: ERK phosphorylation in intact, adult brain by alpha(2)-adrenergic transactivation of EGF receptors. *Neurochemistry international* **55**(7):593–600.

[54] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D., 1998: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**(25):14863–8.

[55] Eisenberg, I., Eran, A., Nishino, I., Moggio, M., Lamperti, C., Amato, A. A., Lidov, H. G., Kang, P. B., North, K. N., Mitrani-Rosenbaum, S., Flanigan, K. M., Neely, L. A., Whitney, D., Beggs, A. H., Kohane, I. S. and Kunkel, L. M., 2007: Distinctive patterns of microRNA expression in primary muscular disorders. *Proceedings of the National Academy of Sciences of the United States of America* **104**(43):17016–21.

[56] Ellwanger, D. C., Büttner, F. A., Mewes, H.-W. and Stümpflen, V., 2011: The sufficient minimal set of miRNA seed types. *Bioinformatics (Oxford, England)* .

[57] Erdös, P. and Renyi, A., 1959: On Random Graphs. I. *Publicationes Mathematicae* **6**:290–297.

[58] Erlich, Y., Mitra, P. P., DelaBastide, M., McCombie, W. R. and Hannon, G. J., 2008: Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature methods* **5**(8):679–82.

[59] Evangelisti, C., Florian, M. C., Massimi, I., Dominici, C., Giannini, G., Galardi, S., Buè, M. C., Massalini, S., McDowell, H. P., Messi, E., Gulino, A., Farace, M. G. and Ciafrè, S. A., 2009: MiR-128 up-regulation inhibits Reelin and DCX expression and reduces neuroblastoma cell motility and invasiveness. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **23**(12):4276–87.

[60] Fang, J., Ding, M., Yang, L., Liu, L.-Z. and Jiang, B.-H., 2007: PI3K/PTEN/AKT signaling regulates prostate tumor angiogenesis. *Cell Signalling* **19**(12):2487–97.

[61] Farh, K. K.-H., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B. and Bartel, D. P., 2005: The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**(5755):1817–21.

[62] Farooq, A. and Zhou, M.-M., 2004: Structure and regulation of MAPK phosphatases. *Cell Signalling* **16**(7):769–79.

[63] Fausto, N., 2000: Liver regeneration. *Journal of hepatology* **32**(1 Suppl):19–31.

[64] Fedurco, M., Romieu, A., Williams, S., Lawrence, I. and Turcatti, G., 2006: BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research* **34**(3):e22.

[65] Fevotte, C. and Doncarli, C., 2004: Two Contributions to Blind Source Separation Using Time-Frequency Distributions. *IEEE Signal Processing Letters* **11**(3):386–389.

[66] Fisher, R. A., 1922: On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *J Roy Statistical Society* **85**(1):87.

[67] Flanary, B. E. and Streit, W. J., 2004: Progressive telomere shortening occurs in cultured rat microglia, but not astrocytes. *Glia* **45**(1):75–88.

[68] Friedman, R. C., Farh, K. K.-H., Burge, C. B. and Bartel, D. P., 2009: Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**:92–105.

[69] Gandellini, P., Folini, M., Longoni, N., Pennati, M., Binda, M., Colecchia, M., Salvioni, R., Supino, R., Moretti, R., Limonta, P., Valdagni, R., Daidone, M. G. and Zaffaroni, N., 2009: miR-205 Exerts tumor-suppressive functions in human prostate through down-regulation of protein kinase Cepsilon. *Cancer Res* **69**(6):2287–95.

[70] Gangaraju, V. K. and Lin, H., 2009: MicroRNAs: key regulators of stem cells. *Nature reviews Molecular cell biology* **10**(2):116–25.

[71] Gasch, A. P. and Eisen, M. B., 2002: Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol* **3**(11):research0059.1–research0059.22.

[72] Gauldie, J., Richards, C., Harnish, D., Lansdorp, P. and Baumann, H., 1987: Interferon beta 2/B-cell stimulatory factor type 2 shares identity with monocyte-derived hepatocyte-stimulating factor and regulates the major acute phase protein response in liver cells. *Proc Natl Acad Sci U S A* **84**(20):7251–5.

[73] Gelman, A., Carlin, J., Stern, H. and Rubin, D., 2004: *Bayesian Data Analysis.* Chapman and Hall, New York.

[74] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y. and Gentry, J., 2004: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10):R80.

[75] George, D., 2003: Targeting PDGF receptors in cancer–rationales and proof of concept clinical trials. *Advances in experimental medicine and biology* **532**:141–51.

References

[76] Gerondakis, S., Grossmann, M., Nakamura, Y., Pohl, T. and Grumont, R., 1999: Genetic approaches in mice to understand Rel/NF-kappaB and IkappaB function: transgenics and knockouts. *Oncogene* **18**:6888–6895.

[77] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L., 2007: The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**(21):8685–90.

[78] Gottardo, F., Liu, C. G., Ferracin, M., Calin, G. A., Fassan, M., Bassi, P., Sevignani, C., Byrne, D., Negrini, M., Pagano, F., Gomella, L. G., Croce, C. M. and Baffa, R., 2007: Micro-RNA profiling in kidney and bladder cancers. *Urologic oncology* **25**(5):387–92.

[79] Grasso, a. W., Wen, D., Miller, C. M., Rhim, J. S., Pretlow, T. G. and Kung, H. J., 1997: ErbB kinases and NDF signaling in human prostate cancer cells. *Oncogene* **15**(22):2705–16.

[80] Greshock, J., Nathanson, K., Martin, A.-M., Zhang, L., Coukos, G., Weber, B. L. and Zaks, T. Z., 2007: Cancer cell lines as genetic models of their parent histology: analyses based on array comparative genomic hybridization. *Cancer research* **67**(8):3594–600.

[81] Griffiths-Jones, S., 2004: The microRNA Registry. *Nucleic Acids Res* **32**:D109–D111.

[82] Griffiths-Jones, S., van Dongen, S., Saini, H. K. and Enright, A. J., 2008: miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**(Database issue):D154–D158.

[83] Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S. and Bartel, D. P., 2008: Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**(7217):1193–7.

[84] Guldberg, C. and Waage, P., 1879: Studies Concerning Affinity" C. M. Forhandlinger: Videnskabs-Selskabet i Christiana (1864), 35. *Erdmann's Journal für Practische Chemie* **127**:64–114.

[85] Gunaratne, P. H., 2009: Embryonic stem cell microRNAs: defining factors in induced pluripotent (iPS) and cancer (CSC) stem cells? *Current stem cell research & therapy* **4**(3):168–77.

[86] Guo, H., Ingolia, N. T., Weissman, J. S. and Bartel, D. P., 2010: Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**(7308):835–840.

[87] Haan, S., Ferguson, P., Sommer, U., Hiremath, M., McVicar, D. W., Heinrich, P. C., Johnston, J. A. and Cacalano, N. A., 2003: Tyrosine phosphorylation disrupts elongin

interaction and accelerates SOCS3 degradation. *The Journal of biological chemistry* **278**(34):31972–9.

[88] Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J. M. and Aransay, A. M., 2009: miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic acids research* **37**(Web Server issue):W68–76.

[89] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T., 2010: Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**(1):129–41.

[90] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and Mckusick, V. A., 2005: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res 33, Database issue, D514-D517* .

[91] Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D. and Zavolan, M., 2009: Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res* **19**(11):2009–20.

[92] He, L. and Hannon, G. J., 2004: MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews Genetics* **5**(7):522–31.

[93] Hébert, S. S. and De Strooper, B., 2009: Alterations of the microRNA network cause neurodegenerative disease. *Trends in neurosciences* **32**(4):199–206.

[94] Heldin, C. H., Johnsson, A., Wennergren, S., Wernstedt, C., Betsholtz, C. and Westermark, B., 1986: A human osteosarcoma cell line secretes a growth factor structurally related to a homodimer of PDGF A-chains. *Nature* **319**:511–514.

[95] Henegar, C., Cancello, R., Rome, S., Vidal, H., Clément, K. and Zucker, J.-D., 2006: Clustering biological annotations and gene expression data to identify putatively coregulated biological processes. *J Bioinform Comput Biol* **4**(4):833–52.

[96] Herranz, H. and Cohen, S. M., 2010: MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. *Genes & development* **24**(13):1339–44.

[97] Hill, A., 1910: The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J Physiol* **40**.

References

[98] Hollams, E. M., Giles, K. M., Thomson, A. M. and Leedman, P. J., 2002: MRNA stability and the control of gene expression: implications for human disease. *Neurochem Res* **27**:957–980.

[99] Hornstein, E. and Shomron, N., 2006: Canalization of development by microRNAs. *Nature genetics* **38 Suppl**:S20–4.

[100] Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., Chien, C.-H., Wu, M.-C., Huang, C.-Y., Tsou, A.-P. and Huang, H.-D., 2010: miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res* **39**(Database issue):D163–D169.

[101] Huang, D. W., Sherman, B. T. and Lempicki, R. A., 2009: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**(1):44–57.

[102] Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C. and Lempicki, R. A., 2007: The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology* **8**(9):R183.

[103] Hyvärinen, A., 1999: Fast and robust fixed-point algorithms for independent component analysis. *{IEEE} Transactions on Neural Networks* **10**(3):626–634.

[104] Hyvärinen, A., Karhunen, J. and Oja, E., 2001: *Independent Component Analysis*. John Wiley & Sons.

[105] Ikeda, S., Kong, S. W., Lu, J., Bisping, E., Zhang, H., Allen, P. D., Golub, T. R., Pieske, B. and Pu, W. T., 2007: Altered microRNA expression in human heart disease. *Physiological genomics* **31**(3):367–73.

[106] Inui, M., Martello, G. and Piccolo, S., 2010: MicroRNA control of signal transduction. *Nat Rev Mol Cell Biol* **11**(4):252–63.

[107] Irminger-Finger, I. and Jefford, C. E., 2006: Is there more to BARD1 than BRCA1? *Nature reviews Cancer* **6**(5):382–91.

[108] Ivanovska, I., Ball, A. S., Diaz, R. L., Magnus, J. F., Kibukawa, M., Schelter, J. M., Kobayashi, S. V., Lim, L., Burchard, J., Jackson, A. L., Linsley, P. S. and Cleary, M. A., 2008: MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression. *Mol Cell Biol* **28**(7):2167–74.

[109] Ivanovska, I. and Cleary, M. A., 2008: Combinatorial microRNAs: working together to make a difference. *Cell cycle* **7**(20):3137–42.

[110] Jansen, R. P., 2001: mRNA localization: message on the move. *Nature reviews Molecular cell biology* **2**(4):247–56.

[111] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G. and Liu, Y., 2009: miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic acids research* **37**(Database issue):D98–104.

[112] Jin, Y., Zhang, H., Tsao, S. W., Jin, C., Lv, M., Strömbeck, B., Wiegant, J., Wan, T. S. K., Yuen, P. W. and Kwong, Y.-L., 2004: Cytogenetic and molecular genetic characterization of immortalized human ovarian surface epithelial cell lines: consistent loss of chromosome 13 and amplification of chromosome 20. *Gynecologic oncology* **92**(1):183–91.

[113] Johannes, M., Brase, J. C., Fröhlich, H., Gade, S., Gehrmann, M., Fälth, M., Sültmann, H. and Beissbarth, T., 2010: Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics (Oxford, England)* **26**(17):2136–44.

[114] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S., 2004: Human microRNA targets. *PLoS Biol* **2e363**.

[115] Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. and Kinzler, K. W., 2008: Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**:1801–1806.

[116] Jovanovic, M. and Hengartner, M. O., 2006: miRNAs and apoptosis: RNAs to die for. *Oncogene* **25**(46):6176–87.

[117] Jurica, M. S. and Moore, M. J., 2003: Pre-mRNA splicing: awash in a sea of proteins. *Molecular cell* **12**(1):5–14.

[118] Kabnick, K. S. and Housman, D. E., 1988: Determinants that contribute to cytoplasmic stability of human c-fos and beta-globin mRNAs are located at several sites in each mRNA. *Mol Cell Biol* **8**:3244–3250.

References

[119] Kai, Z. S. and Pasquinelli, A. E., 2010: MicroRNA assassins: factors that regulate the disappearance of miRNAs. *Nature structural & molecular biology* **17**(1):5–10.

[120] Kan, T., Sato, F., Ito, T., Matsumura, N., David, S., Cheng, Y., Agarwal, R., Paun, B. C., Jin, Z., Olaru, A. V., Selaru, F. M., Hamilton, J. P., Yang, J., Abraham, J. M., Mori, Y. and Meltzer, S. J., 2009: The miR-106b-25 polycistron, activated by genomic amplification, functions as an oncogene by suppressing p21 and Bim. *Gastroenterology* **136**(5):1689–700.

[121] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y., 2008: KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**(Database issue):D480–D484.

[122] Karginov, F. V., Conaco, C., Xuan, Z., Schmidt, B. H., Parker, J. S., Mandel, G. and Hannon, G. J., 2007: A biochemical approach to identifying microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America* **104**(49):19291–6.

[123] Kaufman, L. and Rousseew, P. J., 2005: *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probaility and Statistics. John Wiley & Sons.

[124] Kees, U. R., Rudduck, C., Ford, J., Spagnolo, D., Papadimitriou, J., Willoughby, M. L. and Garson, O. M., 1992: Two malignant peripheral primitive neuroepithelial tumor cell lines established from consecutive samples of one patient: characterization and cytogenetic analysis. *Genes, chromosomes & cancer* **4**(3):195–204.

[125] Kennell, J. A., Gerin, I., MacDougald, O. A. and Cadigan, K. M., 2008: The microRNA miR-8 is a conserved negative regulator of Wnt signaling. *Proc Natl Acad Sci U S A* **105**:15417–15422.

[126] Kerr, M. K. and Churchill, G. A., 2001: Experimental design for gene expression microarrays. *Biostatistics* **2**(2):183–201.

[127] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E., 2007: The role of site accessibility in microRNA target recognition. *Nat Genet* **39**:1278–1284.

[128] Kholodenko, B. N., 2000: Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *European journal of biochemistry / FEBS* **267**(6):1583–8.

[129] Kim, H., Hawley, T. S., Hawley, R. G. and Baumann, H., 1998: Protein tyrosine phosphatase 2 (SHP-2) moderates signaling by gp130 but is not required for the induction of acute-phase plasma protein genes in hepatic cells. *Molecular and cellular biology* **18**(3):1525–33.

[130] Kim, J. B., Porreca, G. J., Song, L., Greenway, S. C., Gorham, J. M., Church, G. M., Seidman, C. E. and Seidman, J. G., 2007: Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science (New York, NY)* **316**(5830):1481–4.

[131] Kim, Y.-K., Yu, J., Han, T. S., Park, S.-Y., Namkoong, B., Kim, D. H., Hur, K., Yoo, M.-W., Lee, H.-J., Yang, H.-K. and Kim, V. N., 2009: Functional links between clustered microRNAs: suppression of cell-cycle inhibitors by microRNA clusters in gastric cancer. *Nucleic Acids Res* **37**:1672–1681.

[132] Kimura, A., Rieger, M. A., Simone, J. M., Chen, W., Wickre, M. C., Zhu, B.-M., Hoppe, P. S., O'Shea, J. J., Schroeder, T. and Hennighausen, L., 2009: The transcription factors STAT5A/B regulate GM-CSF-mediated granulopoiesis. *Blood* **114**(21):4721–8.

[133] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P., 1983: Optimization by simulated annealing. *Science (New York, NY)* **220**(4598):671–80.

[134] Kisseleva, T., Bhattacharya, S., Braunstein, J. and Schindler, C. W., 2002: Signaling through the JAK/STAT pathway, recent advances and future challenges. *Gene* **285**(1-2):1–24.

[135] Kitano, H., 2002: Systems biology: a brief overview. *Science (New York, NY)* **295**(5560):1662–4.

[136] Klamt, S., Haus, U.-U. and Theis, F., 2009: Hypergraphs and cellular networks. *PLoS Comput Biol* **5**:e1000385.

[137] Kloc, M., Zearfoss, N. R. and Etkin, L. D., 2002: Mechanisms of subcellular mRNA localization. *Cell* **108**(4):533–44.

[138] Kloosterman, W. P. and Plasterk, R. H. A., 2006: The diverse functions of microRNAs in animal development and disease. *Developmental cell* **11**(4):441–50.

[139] Kohlhuber, F., Rogers, N. C., Watling, D., Feng, J., Guschin, D., Briscoe, J., Witthuhn, B. A., Kotenko, S. V., Pestka, S., Stark, G. R., Ihle, J. N. and Kerr, I. M., 1997: A JAK1/JAK2 chimera can sustain alpha and gamma interferon responses. *Molecular and cellular biology* **17**(2):695–706.

References

[140] Kong, W., Zhao, J.-J., He, L. and Cheng, J. Q., 2009: Strategies for profiling microRNA expression. *Journal of cellular physiology* **218**(1):22–5.

[141] Korpal, M., Lee, E. S., Hu, G. and Kang, Y., 2008: The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. *The Journal of biological chemistry* **283**(22):14910–4.

[142] Kosorok, M. R. and Ma, S., 2007: Marginal asymptotics for the large p, small n paradigm: With applications to microarray data. *The Annals of Statistics* **35**(4):1456–1486.

[143] Kouzarides, T., 2007: Chromatin modifications and their function. *Cell* **128**(4):693–705.

[144] Kowarsch, A., Blochl, F., Bohl, S., Saile, M., Gretz, N., Klingmuller, U. and Theis, F. J., 2010: Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation. *BMC Bioinformatics* **11**(1):585.

[145] Kowarsch, A., Marr, C., Schmidl, D., Ruepp, A. and Theis, F. J., 2010: Tissue-Specific Target Analysis of Disease-Associated MicroRNAs in Human Signaling Pathways. *PLoS ONE* **5**(6):e11154.

[146] Kowarsch, A., Preusse, M., Marr, C. and Theis, F. J., 2011: miTALOS: Analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA (New York, NY)* **17**(5).

[147] Kowarsch, A., Schmidl, D., Braun, S., Bohl, S., Merkle, R., Klingmüller, U. and Theis, F., 2011: MicroRNA-mediated regulation has an impact on the dynamic behaviour of the JAK-STAT pathway. *Manuscript in preparation* .

[148] Krakowski, M., Abdelmalik, R., Mocnik, L., Krahl, T. and Sarvetnick, N., 2002: Granulocyte macrophage-colony stimulating factor (GM-CSF) recruits immune cells to the pancreas and delays STZ-induced diabetes. *The Journal of pathology* **196**(1):103–12.

[149] Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. and Rajewsky, N., 2005: Combinatorial microRNA target predictions. *Nat Genet* **37**(5):495–500.

[150] Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O. and Wingender, E., 2006: TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res* **34**(Database issue):D546–D451.

[151] Kullback, S. and Leibler, R. A., 1951: On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**(1):79–86.

[152] Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Socci, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foà, R., Schliwka, J., Fuchs, U., Novosel, A., Müller, R.-U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H.-I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt, A., Russo, J. J., Sander, C., Zavolan, M. and Tuschl, T., 2007: A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**(7):1401–14.

[153] Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P., 2001: An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**(5543):858–62.

[154] Leamon, J. H., Lee, W. L., Tartaro, K. R., Lanza, J. R., Sarkis, G. J., DeWinter, A. D., Berka, J., Weiner, M., Rothberg, J. M. and Lohman, K. L., 2003: A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**(21):3769–77.

[155] Lee, D. D. and Seung, H. S., 1999: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755):788–91.

[156] Lee, R., Feinbaum, R. and Ambros, V., 2004: A short history of a short RNA. *Cell* **116**(2 Suppl):S89–92, 1 p following S96.

[157] Lee, R. C., Feinbaum, R. L. and Ambros, V., 1993: The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**(5):843–54.

[158] Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H. and Kim, V. N., 2004: MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal* **23**(20):4051–60.

[159] Legewie, S., Herzel, H., Westerhoff, H. V. and Blüthgen, N., 2008: Recurrent design patterns in the feedback regulation of the mammalian signalling network. *Molecular systems biology* **4**(190):190.

[160] Leucht, C., Stigloher, C., Wizenmann, A., Klafke, R., Folchert, A. and Bally-Cuif, L., 2008: MicroRNA-9 directs late organizer activity of the midbrain-hindbrain boundary. *Nature Neurosci* **11**(6):641–8.

References

[161] Levine, E., Zhang, Z., Kuhlman, T. and Hwa, T., 2007: Quantitative characteristics of gene regulation by small RNA. *PLoS biology* **5**(9):e229.

[162] Lewis, B. P., Burge, C. B. and Bartel, D. P., 2005: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**(1):15–20.

[163] Lewis, B. P., Shih, I.-H., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B., 2003: Prediction of mammalian microRNA targets. *Cell* **115**(7):787–98.

[164] Li, X. and Carthew, R. W., 2005: A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye. *Cell* **123**(7):1267–77.

[165] Li, X., Ponten, A., Aase, K., Karlsson, L., Abramsson, A., Uutela, M., Backstrom, G., Hellstrom, M., Bostrom, H., Li, H., Soriano, P., Betsholtz, C., Heldin, C. H., Alitalo, K., Ostman, A. and Eriksson, U., 2000: PDGF-C is a new protease-activated ligand for the PDGF alpha-receptor. *Nat Cell Biol* **2**:302–309.

[166] Liao, J. C., Boscolo, R., Yang, Y.-L., Tran, L. M., Sabatti, C. and Roychowdhury, V. P., 2003: Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A* **100**(26):15522–7.

[167] Liebermeister, W., 2002: Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **18**(1):51–60.

[168] Liu, G., Ding, M., Chen, J., Huang, J., Wang, H., Jing, Q. and Shen, B., 2010: Computational analysis of microRNA function in heart development. *Acta Biochim Biophy Sin* **42**(9):662–70.

[169] Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J.-J., Hammond, S. M., Joshua-Tor, L. and Hannon, G. J., 2004: Argonaute2 is the catalytic engine of mammalian RNAi. *Science (New York, NY)* **305**(5689):1437–41.

[170] Livak, K. J. and Schmittgen, T. D., 2001: Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif)* **25**(4):402–8.

[171] Lu, B., Moser, A., Shigenaga, J. K., Grunfeld, C. and Feingold, K. R., 2010: The acute phase response stimulates the expression of angiopoietin like protein 4. *Biochem Biophys Res Commun* **391**(4):1737–41.

[172] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R. and Golub, T. R., 2005: MicroRNA expression profiles classify human cancers. *Nature* **435**(7043):834–8.

[173] Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W. and Cui, Q., 2008: An analysis of human microRNA and disease associations. *PloS one* **3**(10):e3420.

[174] Lutter, D., Marr, C., Krumsiek, J., Lang, E. W. and Theis, F. J., 2010: Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. *BMC genomics* **11**:224.

[175] Lutter, D., Walcher, T., Lerch, M., Röh, S., Kowarsch, A., Götz, M., Ninkovic, J. and Theis, F., 2011: A Bayesian approach to infer boolean models for neuronal progenitor cell differentiation. *Manuscript in preparation* .

[176] Maiwald, T., Kreutz, C., Pfeifer, A. C., Bohl, S., Klingmüller, U. and Timmer, J., 2007: Dynamic pathway modeling: feasibility analysis and optimal experimental design. *Annals of the New York Academy of Sciences* **1115**:212–20.

[177] Makeyev, E. V., Zhang, J., Carrasco, M. A. and Maniatis, T., 2007: The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Molecular cell* **27**(3):435–48.

[178] Malliri, A. and Collard, J. G., 2003: Role of Rho-family proteins in cell adhesion and cancer. *Curr Opin Cell Biol* **15**(5):583–9.

[179] Maquat, L. E. and Carmichael, G. G., 2001: Quality control of mRNA function. *Cell* **104**(2):173–6.

[180] Marc, P., Margeot, A., Devaux, F., Blugeon, C., Corral-Debrinski, M. and Jacq, C., 2002: Genome-wide analysis of mRNAs targeted to yeast mitochondria. *EMBO reports* **3**(2):159–64.

[181] Marks, F. and Klingmüller, U., 2008: *Cellular Signal Processing.* Garland Science, New York.

[182] Marr, C., Kowarsch, A., Preusse, M., Backofen, R. and Theis, F., 2011: Beyond enrichment: Measuring microRNA-pathway associations in signaling networks. *submitted* .

[183] Martinez, J. and Tuschl, T., 2004: RISC is a 5' phosphomonoester-producing RNA endonuclease. *Genes & development* **18**(9):975–80.

References

[184] Maslov, S. and Sneppen, K., 2002: Specificity and stability in topology of protein networks. *Science* **296**(5569):910–3.

[185] Mattick, J. S. and Makunin, I. V., 2006: Non-coding RNA. *Human molecular genetics* **15 Spec No**:R17–29.

[186] McClintick, J. N. and Edenberg, H. J., 2006: Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics* **7**:49.

[187] McDonald, D., 2008: Understanding miRNA Turnover: A Study of miRNA Half-Life. *Technical report*, Massachusetts Institute of Technology, Broad Institute of MIT and Harvard.

[188] McKenzie, B. S., Kastelein, R. A. and Cua, D. J., 2006: Understanding the IL-23-IL-17 immune pathway. *Trends Immunol* **27**:17–23.

[189] Medina, P. P., Nolde, M. and Slack, F. J., 2010: OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. *Nature* **467**(7311):86–90.

[190] Meisner, L. F., Wu, S. Q., Christian, B. J. and Reznikoff, C. A., 1988: Cytogenetic instability with balanced chromosome changes in an SV40 transformed human uroepithelial cell line. *Cancer research* **48**(11):3215–20.

[191] Mendell, J. T., 2008: miRiad roles for the miR-17-92 cluster in development and disease. *Cell* **133**:217–222.

[192] Metzker, M. L., 2009: Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**(1):31–46.

[193] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U., 2002: Network motifs: simple building blocks of complex networks. *Science* **298**(5594):824–827.

[194] Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B. and Rigoutsos, I., 2006: A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**:1203–1217.

[195] Molgedey, L. and Schuster, H. G., 1994: Separation of a mixture of independent signals using time-delayed correlations. *Physical Review Letters* **72**(23):3634–3637.

[196] Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I. and Proudfoot, N. J., 2008: Primary microRNA transcripts are processed co-transcriptionally. *Nature structural & molecular biology* **15**(9):902–9.

[197] Mouillet, J.-F., Chu, T., Nelson, D. M., Mishima, T. and Sadovsky, Y., 2010: MiR-205 silences MED1 in hypoxic primary human trophoblasts. *FASEB J* **24**(6):2030–9.

[198] Naka, T., Narazaki, M., Hirata, M., Matsumoto, T., Minamoto, S., Aono, A., Nishimoto, N., Kajita, T., Taga, T., Yoshizaki, K., Akira, S. and Kishimoto, T., 1997: Structure and function of a new STAT-induced STAT inhibitor. *Nature* **387**(6636):924–9.

[199] Nam, S., Kim, B., Shin, S. and Lee, S., 2008: miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res* **36**(Database issue):D159–D164.

[200] NIH, 2002: Know Your Options: Understanding Treatment Choices for Prostate Cancer. *Technical report*, NIH, Bethesda.

[201] Nikiforova, M. N., Tseng, G. C., Steward, D., Diorio, D. and Nikiforov, Y. E., 2008: MicroRNA expression profiling of thyroid tumors: biological significance and diagnostic utility. *The Journal of clinical endocrinology and metabolism* **93**(5):1600–8.

[202] O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. and Baltimore, D., 2007: MicroRNA-155 is induced during the macrophage inflammatory response. *Proc Natl Acad Sci U S A* **104**:1604–1609.

[203] Okazaki, K. and Sagata, N., 1995: The Mos/MAP kinase pathway stabilizes c-Fos by phosphorylation and augments its transforming activity in NIH 3T3 cells. *EMBO J* **14**:5048–5059.

[204] Osella, M., Bosia, C., Corá, D. and Caselle, M., 2011: The Role of Incoherent MicroRNA-Mediated Feedforward Loops in Noise Buffering. *PLoS Computational Biology* **7**(3):e1001101.

[205] Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., Zhang, X., Song, J. S. and Fisher, D. E., 2008: Chromatin structure analyses identify miRNA promoters. *Genes & development* **22**(22):3172–83.

[206] Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J., 2008: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* **40**(12):1413–5.

[207] Papadopoulos, G. L., Alexiou, P., Maragkakis, M., Reczko, M. and Hatzigeorgiou, A. G., 2009: DIANA-mirPath: Integrating human and mouse microRNAs in pathways. *Bioinformatics* **25**(15):1991–3.

References

[208] Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P. and Hatzigeorgiou, A. G., 2009: The database of experimentally supported targets: a functional update of TarBase. *Nucleic acids research* **37**(Database issue):D155–8.

[209] Parker, R. and Song, H., 2004: The enzymes and control of eukaryotic mRNA turnover. *Nature structural & molecular biology* **11**(2):121–7.

[210] Patterson, S. L., Pittenger, C., Morozov, A., Martin, K. C., Scanlin, H., Drake, C. and Kandel, E. R., 2001: Some forms of cAMP-mediated long-lasting potentiation are associated with release of BDNF and nuclear translocation of phospho-MAP kinase. *Neuron* **32**(1):123–40.

[211] Pawlicki, J. M. and Steitz, J. A., 2008: Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. *The Journal of cell biology* **182**(1):61–76.

[212] Pearson, K., 1901: On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559–572.

[213] Perkins, N. D., 2007: Integrating cell-signalling pathways with NF-kappaB and IKK function. *Nat Rev Mol Cell Biol* **8**:49–62.

[214] Perrone, G., Vincenzi, B., Zagami, M., Santini, D., Panteri, R., Flammia, G., Verzì, A., Lepanto, D., Morini, S., Russo, A., Bazan, V., Tomasino, R. M., Morello, V., Tonini, G. and Rabitti, C., 2007: Reelin expression in human prostate cancer: a marker of tumor aggressiveness based on correlation with grade. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **20**(3):344–51.

[215] Poliseno, L., Salmena, L., Riccardi, L., Fornari, A., Song, M. S., Hobbs, R. M., Sportoletti, P., Varmeh, S., Egia, A., Fedele, G., Rameh, L., Loda, M. and Pandolfi, P. P., 2010: Identification of the miR-106b˜25 microRNA cluster as a proto-oncogenic PTEN-targeting intron that cooperates with its host gene MCM7 in transformation. *Science signaling* **3**(117):ra29.

[216] Raghavan, A., 2002: Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Research* **30**(24):5529–5538.

[217] Ramachandran, V. and Chen, X., 2008: Degradation of microRNAs by a family of exoribonucleases in Arabidopsis. *Science (New York, NY)* **321**(5895):1490–2.

[218] Rane, S. G. and Reddy, E. P., 2000: Janus kinases: components of multiple signaling pathways. *Oncogene* **19**(49):5662–79.

[219] Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. and Ruvkun, G., 2000: The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**(6772):901–6.

[220] Ricarte-Filho, J. C. M., Fuziwara, C. S., Yamashita, A. S., Rezende, E., Da-Silva, M. J. and Kimura, E. T., 2009: Effects of let-7 microRNA on Cell Growth and Differentiation of Papillary Thyroid Cancer. *Transl Oncol* **2**(4):236–41.

[221] Ringnér, M., 2008: What is principal component analysis? *Nat Biotechnol* **26**(3):303–4.

[222] Robert, C. and Casella, G., 2004: *Monte Carlo statistical methods*. Springer Verlag.

[223] Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., Benjamin, H., Shabes, N., Tabak, S., Levy, A., Lebanony, D., Goren, Y., Silberschein, E., Targan, N., Ben-Ari, A., Gilad, S., Sion-Vardy, N., Tobar, A., Feinmesser, M., Kharenko, O., Nativ, O., Nass, D., Perelman, M., Yosepovich, A., Shalmon, B., Polak-Charcon, S., Fridman, E., Avniel, A., Bentwich, I., Bentwich, Z., Cohen, D., Chajut, A. and Barshack, I., 2008: MicroRNAs accurately identify cancer tissue origin. *Nature biotechnology* **26**(4):462–9.

[224] Ross, R., Glomset, J., Kariya, B. and Harker, L., 1974: A platelet-dependent serum factor that stimulates the proliferation of arterial smooth muscle cells in vitro. *Proc Natl Acad Sci U S A* **71**:1207–1210.

[225] Ruby, J. G., Jan, C. H. and Bartel, D. P., 2007: Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**(7149):83–6.

[226] Ruepp, A., Kowarsch, A., Schmidl, D., Bruggenthin, F., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C. and Theis, F. J., 2010: PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol* **11**(1):R6.

[227] Ruvkun, G. and Giusto, J., 1989: The Caenorhabditis elegans heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. *Nature* **338**(6213):313–9.

[228] Saumet, A., Vetter, G., Bouttier, M., Portales-Casamar, E., Wasserman, W. W., Maurin, T., Mari, B., Barbry, P., Vallar, L., Friederich, E., Arar, K., Cassinat, B., Chomienne, C. and Lecellier, C.-H., 2009: Transcriptional repression of microRNA genes by PML-RARA increases expression of key cancer proteins in acute promyelocytic leukemia. *Blood* **113**(2):412–21.

References

[229] Schachtner, R., Lutter, D., Knollmüller, P., Tomé, A. M., Theis, F. J., Schmitz, G., Stetter, M., Vilda, P. G. and Lang, E. W., 2008: Knowledge-based Gene Expression Classification via Matrix Factorization. *Bioinformatics* **24**(15):1688–1697.

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

[231] Schilling, M., Maiwald, T., Bohl, S., Kollmann, M., Kreutz, C., Timmer, J. and Klingmüller, U., 2005: Computational processing and error reduction strategies for standardized quantitative data in biological networks. *The FEBS journal* **272**(24):6400–11.

[232] Schindler, C. and Strehlow, I., 2000: Cytokines and STAT signaling. *Advances in pharmacology (San Diego, Calif)* **47**:113–74.

[233] Schotte, D., Chau, J. C. K., Sylvester, G., Liu, G., Chen, C., van der Velden, V. H. J., Broekhuis, M. J. C., Peters, T. C. J. M., Pieters, R. and den Boer, M. L., 2009: Identification of new microRNA genes and aberrant microRNA profiles in childhood acute lymphoblastic leukemia. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK* **23**(2):313–22.

[234] Schulze, a. and Downward, J., 2001: Navigating gene expression using microarrays–a technology review. *Nature cell biology* **3**(8):E190–5.

[235] Scott, S. L., Earle, J. D. and Gumerlock, P. H., 2003: Functional p53 Increases Prostate Cancer Cell Survival After Exposure to Fractionated Doses of Ionizing Radiation 1. *Cancer* **169**(36):7190–7196.

[236] Seigel, G. M., Hackam, A. S., Ganguly, A., Mandell, L. M. and Gonzalez-Fernandez, F., 2007: Human embryonic and neuronal stem cell markers in retinoblastoma. *Molecular vision* **13**:823–32.

[237] Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N., 2008: Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**(7209):58–63.

[238] Sethupathy, P., Megraw, M. and Hatzigeorgiou, A. G., 2006: A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods* **3**:881–886.

[239] Sharova, L. V., Sharov, A. A., Nedorezov, T., Piao, Y., Shaik, N. and Ko, M. S. H., 2009: Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA research : an international journal for rapid publication of reports on genes and genomes* **16**(1):45–58.

[240] Shendure, J. and Ji, H., 2008: Next-generation DNA sequencing. *Nature biotechnology* **26**(10):1135–45.

[241] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M., 2005: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, NY)* **309**(5741):1728–32.

[242] Shi, Y., Liu, C. H., Roberts, A. I., Das, J., Xu, G., Ren, G., Zhang, Y., Zhang, L., Yuan, Z. R., Tan, H. S. W., Das, G. and Devadas, S., 2006: Granulocyte-macrophage colony-stimulating factor (GM-CSF) and T-cell responses: what we do and don't know. *Cell research* **16**(2):126–33.

[243] Shida, Y., Igawa, T., Hakariya, T., Sakai, H. and Kanetake, H., 2007: p38MAPK activation is involved in androgen-independent proliferation of human prostate cancer cells by regulating IL-6 secretion. *Biochem Biophys Res Commun* **353**(3):744–9.

[244] Shimada, K., Nakamura, M., Ishida, E., Higuchi, T., Tanaka, M., Ota, I. and Konishi, N., 2007: c-Jun NH2 terminal kinase activation and decreased expression of mitogen-activated protein kinase phosphatase-1 play important roles in invasion and angiogenesis of urothelial carcinomas. *Am J Pathol* **171**(3):1003–12.

[245] Shimodaira, H., 2002: An approximately unbiased test of phylogenetic tree selection. *Systematic biology* **51**(3):492–508.

[246] Shingara, J., Keiger, K., Shelton, J., Laosinchai-Wolf, W., Powers, P., Conrad, R., Brown, D. and Labourier, E., 2005: An optimized isolation and labeling platform for accurate microRNA expression profiling. *RNA (New York, NY)* **11**(9):1461–70.

[247] Shukla, S., Maclennan, G. T., Hartman, D. J., Fu, P., Resnick, M. I. and Gupta, S., 2007: Activation of PI3K-Akt signaling pathway promotes prostate cancer cell invasion. *Int J Cancer* **121**(7):1424–32.

[248] Siewert, E., Müller-Esterl, W., Starr, R., Heinrich, P. C. and Schaper, F., 1999: Different protein turnover of interleukin-6-type cytokine signalling components. *European journal of biochemistry / FEBS* **265**(1):251–7.

[249] Silber, J., Lim, D. A., Petritsch, C., Persson, A. I., Maunakea, A. K., Yu, M., Vandenberg, S. R., Ginzinger, D. G., James, C. D., Costello, J. F., Bergers, G., Weiss, W. A., Alvarez-Buylla, A. and Hodgson, J. G., 2008: miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. *BMC medicine* **6**:14.

[250] Smyth, G. K., 2004: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Molec Biol* **3**:Article3.

[251] Smyth, G. K., Ritchie, M., Thorne, N. and Wettenhall, J., 2005: Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, (pages 397 – 420). Springer.

[252] Society, A. C., 2002: ACS Cancer Facts and Figures. *Technical report*, GA: American Cancer Society, Atlanta.

[253] Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M. and Cooper, D. N., 2003: Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**:577–581.

[254] Streetx, K. L., Luedde, T., Manns, M. and Trautwein, C., 2000: Interleukin 6 and liver regeneration. *Gut* **47**(2):309–312.

[255] Sturm, M., Hackenberg, M., Langenberger, D. and Frishman, D., 2010: TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics* **11**:292.

[256] Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G. and Hogenesch, J. B., 2002: Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**:4465–4470.

[257] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R. and Hogenesch, J. B., 2004: A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**(16):6062–7.

[258] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P., 2005: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**(43):15545–50.

[259] Suzuki, H. I., Yamagata, K., Sugimoto, K., Iwamoto, T., Kato, S. and Miyazono, K., 2009: Modulation of microRNA processing by p53. *Nature* **460**:529–533.

[260] Szabo, A., Perou, C. M., Karaca, M., Perreard, L., Palais, R., Quackenbush, J. F. and Bernard, P. S., 2004: Statistical modeling for selecting housekeeper genes. *Genome biology* **5**(8):R59.

[261] Takamizawa, J., Konishi, H., Yanagisawa, K., Tomida, S., Osada, H., Endoh, H., Harano, T., Yatabe, Y., Nagino, M., Nimura, Y., Mitsudomi, T. and Takahashi, T., 2004: Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer research* **64**(11):3753–6.

[262] Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R. and Draghici, S., 2007: Machine learning and its applications to biology. *PLoS Comput Biol* **3**(6):e116.

[263] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M., 1999: Systematic determination of genetic network architecture. *Nat Genet* **22**(3):281–5.

[264] Team, R. D. C., 2008: R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria ISBN* **3**:–900051–07–0.

[265] Tekotte, H. and Davis, I., 2002: Intracellular mRNA localization: motors move messages. *Trends in genetics : TIG* **18**(12):636–42.

[266] Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R. and Caldas, C., 2007: Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput Biol* **3**(8):e161.

[267] Theis, F. and Meyer-Baese, A., 2010: *Biomedical Signal Analysis - Contemporary Methods and Applications.* MIT Press.

[268] Theis, F. J., 2004: A new concept for separability problems in blind source separation. *Neural Computation* **16**:1827–1850.

[269] Theis, F. J. and Gruber, P., 2004: Separability of analytic postnonlinear blind source separation with bounded sources. In *Proc. {ESANN} 2004*, (pages 217–222). d-side, Evere, Belgium, Bruges, Belgium.

[270] Theis, F. J., Meyer-Bäse, A. and Lang, E. W., 2004: Second-order blind source separation based on multi-dimensional autocovariances. In *Proc. {ICA} 2004*, volume 3195 of *LNCS*, (pages 726–733). Springer, Granada, Spain.

[271] Tong, L., Inouye, Y., Soon, V. C. and Huang, Y.-F., 1991: Indeterminacy and identifiability of blind identification. *IEEE Trans Circuits Sys* **38**:499–509.

[272] Tsang, J., Zhu, J. and van Oudenaarden, A., 2007: MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell* **26**:753–767.

[273] Tusher, V. G., Tibshirani, R. and Chu, G., 2001: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**(9):5116–21.

[274] Ueda, T., Bruchovsky, N. and Sadar, M. D., 2002: Activation of the androgen receptor N-terminal domain by interleukin-6 via MAPK and STAT3 signal transduction pathways. *J Biol Chem* **277**(9):7076–85.

[275] Valentino, L. and Pierre, J., 2006: JAK/STAT signal transduction: regulators and implication in hematological malignancies. *Biochemical pharmacology* **71**(6):713–21.

[276] Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F., 2002: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology* **3**(7):RESEARCH0034.

[277] Volinia, S., Calin, G. a., Liu, C.-G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., Prueitt, R. L., Yanaihara, N., Lanza, G., Scarpa, A., Vecchione, A., Negrini, M., Harris, C. C. and Croce, C. M., 2006: A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America* **103**(7):2257–61.

[278] Voorhoeve, P. M., le Sage, C., Schrier, M., Gillis, A. J. M., Stoop, H., Nagel, R., Liu, Y.-P., van Duijse, J., Drost, J., Griekspoor, A., Zlotorynski, E., Yabuta, N., De Vita, G., Nojima, H., Looijenga, L. H. J. and Agami, R., 2006: A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell* **124**(6):1169–81.

[279] Wang, W., Chen, J. X., Liao, R., Deng, Q., Zhou, J. J., Huang, S. and Sun, P., 2002: Sequential activation of the MEK-extracellular signal-regulated kinase and MKK3/6-p38 mitogen-activated protein kinase pathways mediates oncogenic ras-induced premature senescence. *Mol Cell Biol* **22**:3389–3403.

[280] Wang, X., 2008: miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**(6):1012–7.

[281] Wang, X. and Wang, X., 2006: Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic acids research* **34**(5):1646–52.

[282] Westwick, J., Weitzel, C., Minden, A., Karin, M. and Brenner, D., 1994: Tumor necrosis factor alpha stimulates AP-1 activity through prolonged activation of the c-Jun kinase. *J Biol Chem* **269**(42):26396–26401.

[283] Wilson, D., Charoensawan, V., Kummerfeld, S. K. and Teichmann, S. A., 2008: DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic acids research* **36**(Database issue):D88–92.

[284] Wong, P., Althammer, S., Hildebrand, A., Kirschner, A., Pagel, P., Geissler, B., Smialowski, P., Blöchl, F., Oesterheld, M., Schmidt, T., Strack, N., Theis, F. J., Ruepp, A. and Frishman, D., 2008: An evolutionary and structural characterization of mammalian protein complex organization. *BMC Genomics* **9**(1):629.

[285] Wormald, S., Zhang, J.-G., Krebs, D. L., Mielke, L. A., Silver, J., Alexander, W. S., Speed, T. P., Nicola, N. A. and Hilton, D. J., 2006: The comparative roles of suppressor of cytokine signaling-1 and -3 in the inhibition and desensitization of cytokine signaling. *The Journal of biological chemistry* **281**(16):11135–43.

[286] Wu, X., Jiang, R., Zhang, M. Q. and Li, S., 2008: Network-based global inference of human disease genes. *Mol Syst Biol* **4**:189.

[287] Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. and Li, T., 2009: miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research* **37**(Database issue):D105–10.

[288] Xie, Z.-R., Yang, H.-T., Liu, W.-C. and Hwang, M.-J., 2007: The role of microRNA in the delayed negative feedback regulation of gene expression. *Biochemical and biophysical research communications* **358**(3):722–6.

[289] Xiong, J., Yu, D., Wei, N., Fu, H., Cai, T., Huang, Y., Wu, C., Zheng, X., Du, Q., Lin, D. and Liang, Z., 2010: An estrogen receptor alpha suppressor, microRNA-22, is downregulated in estrogen receptor alpha-positive human breast cancer cell lines and clinical samples. *The FEBS journal* **277**(7):1684–94.

[290] Xu, J. and Wong, C., 2008: A computational screen for mouse signaling pathways targeted by microRNA clusters. *RNA* **14**(7):1276–83.

[291] Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M. and Darnell, J. E., 2003: Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome research* **13**(8):1863–72.

[292] Yardy, G. W. and Brewster, S. F., 2005: Wnt signalling and prostate cancer. *Prostate Cancer Prostatic Dis* **8**(2):119–26.

[293] Yeilding, N. M., Rehman, M. T. and Lee, W. M., 1996: Identification of sequences in c-myc mRNA that regulate its steady-state levels. *Mol Cell Biol* **16**:3511–3522.

[294] Yin, J. Q., Zhao, R. C. and Morris, K. V., 2008: Profiling microRNA expression with microarrays. *Trends in biotechnology* **26**(2):70–6.

[295] Ying, S.-W., Futter, M., Rosenblum, K., Webber, M. J., Hunt, S. P., Bliss, T. V. P. and Bramham, C. R., 2002: Brain-derived neurotrophic factor induces long-term potentiation in intact adult hippocampus: requirement for ERK activation coupled to CREB and upregulation of Arc synthesis. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **22**(5):1532–40.

[296] Yoshimura, A., Ohkubo, T., Kiguchi, T., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., Hara, T. and Miyajima, A., 1995: A novel cytokine-inducible gene CIS encodes an SH2-containing protein that binds to tyrosine-phosphorylated interleukin 3 and erythropoietin receptors. *The EMBO journal* **14**(12):2816–26.

[297] Zhang, B., Pan, X., Cobb, G. P. and Anderson, T. A., 2007: microRNAs as oncogenes and tumor suppressors. *Dev Biol* **302**(1):1–12.

[298] Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M. S., Giannakakis, A., Liang, S., Naylor, T. L., Barchetti, A., Ward, M. R., Yao, G., Medina, A., O'brien-Jenkins, A., Katsaros, D., Hatzigeorgiou, A., Gimotty, P. A., Weber, B. L. and Coukos, G., 2006: microRNAs exhibit high frequency genomic alterations in human cancer. *Proceedings of the National Academy of Sciences of the United States of America* **103**(24):9136–41.

[299] Zheng, J. B., Zhou, Y. H., Maity, T., Liao, W. S. and Saunders, G. F., 2001: Activation of the human PAX6 gene through the exon 1 enhancer by transcription factors SEF and Sp1. *Nucleic acids research* **29**(19):4070–8.

[300] Zhou, X., Ren, Y., Moore, L., Mei, M., You, Y., Xu, P., Wang, B., Wang, G., Jia, Z., Pu, P., Zhang, W. and Kang, C., 2010: Downregulation of miR-21 inhibits EGFR pathway and suppresses the growth of human glioblastoma cells independent of PTEN status. *Laboratory investigation; a journal of technical methods and pathology* **90**(2):144–55.

[301] Ziehe, A. and Mueller, K.-R., 1998: TDSEP - an efficient algorithm for blind separation using time structure. In Niklasson, L., Bodén, M. and Ziemke, T. (editors), *Proc. of ICANN 1998*, (pages 675–680). Springer Verlag, Berlin, Skövde, Sweden.

[302] Zohrabian, V. M., Forzani, B., Chau, Z., Murali, R. and Jhanwar-Uniyal, M., 2009: Rho/ROCK and MAPK signaling pathways are involved in glioblastoma cell migration and proliferation. *Anticancer Res* **29**(1):119–23.