INAUGURAL - DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich-Mathematischen Gesamtfakultät

der

Ruprecht-Karls-Universität

Heidelberg

vorgelegt von

Diplom-Informatikerin (Bioinformatik) Ramona Schmid

aus Ellwangen/Jagst

Tag der mündlichen Prüfung: 26.09.2012

Analyzing Compounds' Mode of Action

A Use Case for New Approaches Utilizing Protein Interaction Networks and Prior Knowledge to Complement State-of-the-Art Gene Expression Analyses.

1. Gutachter: Prof. Dr. Roland Eils
2. Gutachter: Prof. Dr. Ulrike Müller

Zusammenfassung

Hintergrund: Wissenschaftler der pharmazeutischen Industrie und der akademischen Forschung arbeiten gemeinsam an der Erforschung der grundlegenden Ursache einer Erkrankung auf zellulärer Ebene bis hin zum zugelassenen neuen Medikament. Analysen der Wirkungsweise (*mode of action*) neuer Substanzen sind unter zunehmenden Sicherheits- und Nutzenanforderungen ein immer wichtiger werdender Beitrag in der Entwicklung eines neuen Wirkstoffs. Dabei wird zwischen den Effekten am gewünschten Zielprotein (*on-target*) und den Effekten an möglicherweise unbekannten Zielproteinen (*off-targets*) unterschieden. Häufig ist das Wissen über diese Effekte sehr begrenzt. Da die Wirkung hauptsächlich durch Wechselwirkungen von Proteinen oder Signalkaskaden vermittelt wird, ist ihre Untersuchung auf der Basis von Proteininteraktionsnetzwerken (PI-Netzwerke) ein vielversprechender Ansatz. Die Menge an verfügbaren biologische Daten aus verschiedensten Quellen steigt stetig an. Die Integration dieses Wissens ist wichtig, um ein tieferes Verständnis der zugrundeliegenden Biologie zu erlangen. Häufig werden Genexpressionsstudien von erkranktem Gewebe und/oder wirkstoffbehandelten biologischen Proben durchgeführt, um die Wirkungsweise neuer Wirkstoffe unter Berücksichtigung von transkriptionellen Änderungen verstehen zu können.

***Status quo*:** Die Wirkungsweise von Substanzen kann analysiert werden, indem die Teile von Proteininteraktionsnetzwerken untersucht werden, in denen aufgrund von Wirkstoffbehandlung Änderungen zu beobachten sind. Mathematische oder graphtheoretische *in silico* Methoden, die interessante Teile von Netzwerken identifizieren, finden weit verbreiteten Einsatz. Fragestellungen reichen dabei von der Ermittlung stark vernetzter Subgraphen über die Identifizierung kürzester Wege bis hin zur Berechnung von Subgraphen oder Modulen, die bestimmte Zielfunktionen optimieren. Entsprechende Algorithmen können auf biologische Fragestellungen angewandt werden, um beispielsweise konditionsresponsive Subnetzwerke in verschiedensten Arten von biologischen Netzwerken zu identifizieren. Gegenwärtige Methoden, die zur Analyse der Wirkungsweise eingesetzt werden können, befassen sich hauptsächlich mit der Detektion von Subnetzwerken, in denen Informationen aus dem Bereich der funktionellen Genomik angereichert sind, z.B. Anreicherung an deregulierten Genen. Diese Methoden vernachlässigen die Existenz von regulatorischen Mechanismen auf post-transkriptionaler und auch post-translationaler Ebene wie miRNA-Interferenz

oder Protein-Phosphorylierung. Außerdem detektieren gängige Methoden häufig relativ große Module. Dabei deckt ein solches Modul möglicherweise mehrere Prozesse gleichzeitig ab, beispielsweise sowohl *on-* als auch *off-target* Effekte. Zur Detektion und Interpretation von Einzeleffekten innerhalb eines biologischen Systems wäre es hilfreich, kleinere Module identifizieren zu können.
Einige frühere Arbeiten fokussieren sich auf die Vorhersage und Gewichtung von Proteininteraktionen unter Berücksichtigung von vorhandenem Wissen. In den meisten Fällen werden die Interaktionen dabei bezüglich eines als Gold-Standard betrachteten Datensatzes gewichtet. Da unser Wissen über die Rolle von Genen und Proteinen nach wie vor sehr unvollständig ist, gibt es keinen Gold-Standard, der die Realität exakt widerspiegelt. Des Weiteren sind gegenwärtig verwendete Gewichtungsmethoden häufig weder einfach zu verwenden noch einfach zu interpretieren. Eine ideale Gewichtung sollte die einfache Integration zusätzlicher Datenquellen und deren individuelle Gewichtung basierend auf Expertenwissen ermöglichen.

Vorgehen & Ergebnisse: In der vorliegenden Arbeit werden Genexpressionsdaten analysiert, die der Aufdeckung von *on-* und *off-target* Effekten verschiedener Wirkstoffe zur Inhibition von TGF-βR1 dienen sollen. Um eine verlässliche Basis für die Datenanalyse zu schaffen, werden im ersten Teil der Arbeit verschiedene Aspekte zur Auswahl einer geeigneten Normalisierungsmethode vorgestellt. Unter deren Berücksichtigung wird schließlich eine optimale Normalisierungsstrategie gewählt.
Um die Wirkmechanismen der verschiedenen Substanzen zu analysieren, wird ein Verfahren vorgeschlagen, das die Interaktionen zwischen Proteinen mittels verschiedener Evidenzien gewichtet. Die Relevanz der Proteine wird dabei nicht nur über die Expression ihrer kodierenden Gene sondern auch durch ihre Beziehung zu anderen Proteinen bewertet. Dadurch werden Analysen über die Genexpressionsebene hinaus erweitert. Die Bewertung dieser Beziehungen erfolgt über die Gewichtung der Proteininteraktionen. Dazu werden Informationen über molekulare Funktionen, biologische Prozesse, zelluläre Kompartimente, Transkriptionsfaktorbindestellen und literaturbasierte Konfidenzwerte integriert, um die entsprechenden Kanten im Netzwerk zu gewichten. Expressionsdaten dienen als Ankerpunkt der Analysen, um das Netzwerk schließlich in den biologischen Kontext zu transferieren.
Des Weiteren wird in dieser Arbeit eine neue Methode zur Extraktion von Modulen aus gewichteten PI-Netzwerken entwickelt, modEx. Mittels der durch modEx extrahierten Module ist es möglich, Einzeleffekte innerhalb des biologischen Systems

abzugreifen.
Für den vorliegenden Expressiondatensatz kann gezeigt werden, dass die vorgeschlagene Kantengewichtung der weit akzeptierten STRING-Gewichtung überlegen ist. Darüber hinaus können unter Verwendung von modEx Module extrahiert werden, die den zugrundeliegenden biologischen Mechanismus besser repräsentieren als Module, die durch das gängige jActiveModule identifiziert werden.
Die vorgestellten Methoden werden verwendet, um den Wirkungmechanismus, d. h. sowohl die *on-* als auch *off-target* Effekte verschiedener Wirkstoffe zu analysieren. Es kann gezeigt werden, dass dadurch ein fokussierterer Blick auf die Effekte der Wirkstoffe möglich ist als durch gegenwärtige *state-of-the-art* Analysen eines Genexpressionsdatensatzes.

Abstract

Background: Scientists in pharmaceutical as well as academic research work together to solve the challenging puzzle from the basic causes of disease at the level of genes, proteins and cells up to a marketed new drug. Analyses of mode of action (MoA) of new chemical entities (NCEs) are a very important step in the development of new drugs. One distinguishes between effects induced by modulating the compounds' actual target protein (on-target effects) and effects induced by additional, possibly unknown targets (off-target effects). Quite often knowledge about either of these effects is limited. Since MoA is mainly triggered by the interplay of proteins or signaling cascades, investigating the change and subsequent influence of the changed molecules in a protein interaction (PI) network is a promising initial step to further analyses. As more and more data from diverse sources becomes available, the integration of this knowledge is important for generating a deeper insight into biology. In addition, expression experiments based on disease tissue and/or compound treatment are frequently conducted to get insight into transcriptional changes that could explain compounds' MoA.

***Status quo*:** MoA could be analyzed by investigating those parts of a PI network that show changes based on compound treatment. Mathematical or graph theoretical *in silico* methods to identify interesting parts of a network based on different criteria are widely used. Criteria range from detection of highly connected subgraphs to subgraphs maximizing weights assigned to parts of the network under investigation. These methods can be transferred to biology and can be used to, e. g. identify condition responsive subnetworks on various types of molecular networks. Present questions addressed mainly focus on the detection of subnetworks enriched in information from functional genomics, e.g. differentially expressed genes. They neglect the existence of distance regulatory functions on the post-transcriptional as well as post-translational level like miRNA interference or protein phosphorylation. Further, available methods usually detect relatively large modules. It is easily possible that more processes, i. e. the on- and several off-target effects, are covered by one larger module. Thus, the individual effects are difficult to detect and interpret. To be able to derive individual effects, it is necessary to reveal small modules that are related to the individual effects present in the biological system under investigation. Previous works focus on predicting and weighting interactions between proteins

based on prior knowledge. In most cases interactions are weighted according to a gold standard which always depends on the current knowledge. Our knowledge about the role of genes/proteins is far from complete and still accumulating and evolving, thus, a gold standard does not reflect reality. Present scoring methods lack ease of use as well as ease of interpretation of scoring function to describe the pairwise relatedness of proteins. Further, an ideal score should be highly flexible by allowing easy integration of newly gained knowledge and it should offer the possibility to differentially weight individual evidence based on expert knowledge.

Methods & Results: In this work, I made use of a gene expression data set investigating the inhibition of the TGF-β signaling pathway by different compounds targeting TGF-βR1. To gain a sound basis for follow-up analyses, different aspects of how to select the best suited normalization procedure for the underlying expression data are proposed in the first part of this thesis.
To analyze compounds' MoA, I propose a method that weights interactions between proteins based on different kinds of evidence. In this method, the relevance of the proteins is based on the biological relatedness to other possibly not deregulated protein coding genes. Thereby, analyses are expanded beyond transcriptional deregulation. To elucidate the biological relatedness, information on molecular function, biological processes and cellular compartment, information on transcription factor binding sites and literature-based confidence scores are integrated for weighting the edges between proteins. To transfer the network into the biological context of interest, expression experiments are used as anchoring points for the analyses.
Further, I introduce modEx, a method to extract small modules out of a weighted protein interaction network. Modules extracted using modEx reflect the individual effects present in the biological system under investigation.
For the expression data set used, the proposed edge scoring is shown to be superior to the widely accepted STRING scoring. Furthermore, modEx extracts modules that represent the underlying mechanism better than jActiveModule, a commonly used subgraph extraction method. These newly proposed approaches are applied to elucidate the MoA, i. e. the on- as well as off-target effects, of compounds. They are shown to grant a more focused view on the effects of compounds than current state-of-the-art methods applied for the analysis of gene expression data.

Acknowledgements

I am truly indebted to my supervisors **Prof. Roland Eils**, **Dr. Benedikt Brors** and **Dr. Karsten Quast** for giving me the opportunity to work in such an interesting field, for their continuous support as well as for the very helpful discussions and advice. For funding my PhD thesis, I would like to thank **Boehringer Ingelheim**.

For statistical advice and for her amicable encouragement I owe sincere thankfulness to **Dr. Carina Ittrich**. I highly appreciate your reliable, tireless, and unhesitant support.

This dissertation would not have been possible without the support of **Dr. Patrick Baum**. I would like to deeply thank you for the hours spent in the lab working in a very conscious way to supply me with excellent data - the basis for my work. I am really thankful for your patience with all my biological and especially TGF-β related questions. It was an honor and a pleasure working with you.

For his reliable, pushing and well-structured support I am very grateful to **Dr. Detlev (Menne) Mennerich**.

For great help with the theoretical part of my work I would like to truly thank **Dr. Nadja Betzler** and **Dr. Johannes Uhlmann** who advised me from somewhere in the world.

To many people at BI: I am really grateful to the whole **TDR-team** for the pleasant working atmosphere. I am much obliged to **Dr. Katrin Fundel-Clemens**, **Bärbel Lämmle**, **Stephen-James Gelling**, **Dr. Fabian Birzele**, **Dr. Eric Simon** and **Dr. Gerald Birk**. All of you supported me in many different ways. For their patience and for trusting in me, I am grateful to **Dr. Tobias Hildebrandt**, **Dr. Udo Maier**, and **Dr. Andreas Weith**. I also would like to thank **Susanne Acker**, **Dagmar Knebel**, **Catherine Nicolo**, **Sandra Vogel**, and **Werner Rust** for their encouragement. Many thanks also to **Dr. Katja Kroker**, **Dr. Heiko Stahl**, and all other PhD students for the companionship on an essential part of this tough route.

Dr. Dilafruz Juraeva, **Dr. Nathalie Harder**, **Dr. Rainer König**, **Nora Rieber**, **Dr. Gunnar Schramm**, **Thomas Wolf**, and **Dr. Marc Zapatka**: Thanks for any

advice and the discussions and nice time during the iBios retreats.

I am very grateful to all the people I got to know at the European Bioinformatics Institute. **Dr. Wolfgang Huber**, thank you very much for giving me this great opportunity and for all the support during that time and beyond. To **Dr. Simon Anders**, **PhD Richard Bourgan**, **PhD Jörn Tödling**, **PhD Elin Axelsson**, **Dr. Bernd Fischer**, **PhD Audrey Kauffmann**, **PhD Kristen Feher**, and **PhD Julia Fisher**: Due to you, I had an inspiring, fulfilling and unforgettable time in Cambridge. Many thanks to **EMBL/EBI** for funding my half year stay as a visiting PhD student.

This dissertation would not have been possible without technical and administrative support. Hence, I would like to thank the IT-people **Karlheinz Groß**, **Wolfgang Marquardt**, **Jürgen Schindler** and **Dietmar Haag**. For administrative support, I owe thanks to **Mareike Gaus**, **Silke Heusel-Stütz**, and **Corinna Sprengart**. Thanks for supporting me with your organizational skills.

Last but not least, I thank my family, especially my parents **Peter** and **Angelika**, my siblings **Daniel** and **Anja**, and **Jan** for supporting me in any possible way, for giving me good advice whenever needed, for always encouraging me, and for their infinite understanding.

Sincere thanks to all of you!

Contents

List of Figures

List of Tables

List of Abbreviations

AAV	Adeno-associated virus
APPROVe	Adenomatous Polyp PRevention On Vioxx
ATP	Adenosine triphosphate
AUC	Area under the curve
BI1	Boehringer Ingelheim compound #46 [1]
BI2	Boehringer Ingelheim compound #5/BIBF0775 [1]
BI3	Boehringer Ingelheim compound #47i/BI34659 [1]
BI4	Boehringer Ingelheim compound #47l [1]
BI5	Boehringer Ingelheim compound #47p [1]
BIND	Biomolecular Interaction Network Database
BioGrid	Biological General Repository for Interaction Datasets
BP	GO: Biological process
CC	GO: Cellular component
CDF	Cumulative distribution function
CETP	Cholesteryl ester transfer protein
CEZANNE	Co-Expression Zone ANalysis using NEtworks
COPD	Chronic obstructive pulmonary disease
DAG	Directed acyclic graph
DASMI	Distributed Annotation System for Molecular Interactions
DIP	Database of Interacting Proteins
DKFZ	Deutsches Krebsforschungszentrum
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
ECM	Extracellular matrix
EMA	European Medicines Agency
EMT	Epithelial-mesenchymal transition

Ex1	External compound PD166285, PubChem CID 5311382 [2]
Ex2	External compound PD166285, PubChem CID 5327885 [2]
FCS	Fetal calf serum
FDA	Food and Drug Administration
FDR	False discovery rate
FP	False positive
GGA	Genomatix Genome Analyzer
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GXNA	Gene eXpression Network Analysis
HPRD	Human Protein Reference Database
HTS	High-throughput screening
IC	Information content
ILLUMINATE	Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events
IND	Investigational New Drug
IPA	Ingenuity Pathway Analyses
IPF	Idiopathic pulmonary fibrosis
iRefIndex	Interaction Reference Index
KEGG	Kyoto Encyclopedia of Genes and Genomes
LADME/Tox	Liberation, absorption, distribution, metabolism, elimination, and toxicity
LCA	Lowest common ancestor
LLS	Log likelihood score
MATISSE	Module Analysis via Topology of INteractions and Similarity SEts
MF	GO: Molecular function
MHLW	Japan's Ministry of Health, Labour and Welfare
MiMI	Michigan Molecular Interactions
MINT	Molecular INTeraction database
MIPS	Munich Information Center for Protein Sequences
MoA	Mode of action
MPact	MIPS protein interaction resource on yeast
MPPI	MIPS Mammalian Protein-Protein Interaction Database

MSQ	Mean sum of squares
NBE	New biological entity
NCE	New chemical entity
NDA	New Drug Application
NLP	Natural language processing
OPHID	Online Predicted Human Interaction Database
PI	Protein interaction
PINA	Protein Interaction Network Analysis
PPI	Protein-protein interactions
qRT-PCR	Quantitative real time polymerase chain reaction
QSAR	Quantified structure activity relationship
R&D	Research and development
ROC	Receiver Operator Characteristic
rsn	Robust spline normalization
SA	Simulated annealing
SAR	Structure activity relationship
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TGF-β	The transforming growth factor β
TGF-βR1	Tranforming growth factor β receptor 1
TGF-βR2	Transforming growth factor β receptor 2
TN	True negative
TP	True positive
vsn	Variance stabilization and normalization
vst	Variance-stabilizing transformation
1-vCMECG	1-vertex Constrained Maximum Edge-Weight Connected Graph
k-vCMECG	k-vertex Constrained Maximum Edge-Weight Connected Graph
k-veCMECG	k-vertex/edge Constrained Maximum Edge-Weight Connected Graph
MWCS	Maximum-Weight Connected Subgraph Problem
PCST	Price Collecting Steiner Tree Problem
vCMECG	vertex Constrained Maximum Edge weight Connected Graph

Chapter 1

Background

Scientists in pharmaceutical as well as academic research work together to solve the challenging puzzle of basic causes of disease at the level of genes, proteins and cells up to a marketed new drug. This drug in the ideal case inhibits or reverses the disease progression or at least treats the symptoms of the disease to relieve patients of their suffering. Researchers work to identify and validate disease related target molecules, discover and optimize the right new chemical or biological entity (NCE or NBE, respectively) to interact with that molecule, test for safety and efficacy and gain approval to get the new drug into clinical practice. This whole process takes 10 to 15 years with an estimated average cost for research and development (R&D) of a successful drug in the range of $800 million to $1 billion. This number includes the cost of the thousands of failures: For every 5,000 - 10,000 compounds that enter the R&D pipeline, ultimately only one receives approval. [3, 4]

In this chapter, I first give an overview over the drug development process in Section 1.1, reveal the importance of analyses of compounds' mode of action in Section 1.2, summarize the objective of this thesis in Section 1.3, and describe the structure of this thesis in Section 1.4.

1.1 Drug Development Process

The development of a new drug can be split into two phases, namely drug discovery and clinical trials. This section gives a short introduction into both. A schematic overview of the complete process together with expected times needed for the indi-

vidual steps is given in Figure 1.1.

1.1.1 Drug Discovery Process

Pre-discovery

The basis of every drug development process is the sound understanding of the disease mechanisms to treat. This could for example be achieved by linking an induced disease state to altered gene expression in diseased versus healthy samples. Next, the role of the respective proteins, how they interact with each other in living cells, and how this ultimately leads to the disease is investigated. This knowledge provides the basis for revealing the relevant pathomechanisms. However, even with new tools and insights, this research takes many years of work and quite often leads to abrupt ends.

Target Identification

Once researchers gain enough understanding of the underlying disease, a target is selected. A target is commonly a single molecule, in most cases a protein, which is involved in the mechanism of a particular disease. To achieve a desired effect, it is critical that researchers pick a druggable target. A biological entity is druggable if its behavior can be modulated by a drug molecule thereby achieving a desired effect like for example inhibition of enzymatic activity.

New approaches that are applied for identifying this type of target are text- and data mining [5–8]. In text mining, literature is screened by computational approaches. Applied methods range from investigating co-occurrences of words, for example a gene mentioned together with the disease of interest, to more sophisticated methods like natural language processing (NLP) which are capable of "reading" the text in a human like fashion. By data mining different kind of information sources are integrated. Applying text mining as well as data mining, the list of possible targets could be narrowed down, ultimately leading to a handful of targets. In the following, these targets need to be validated.

Target Validation

After a target has been identified on a more or less theoretical basis, researchers need to prove the positive effect of modulating the target towards the healthy state. By conscientious target validation dead ends can be identified early in the pipeline. Thereby costly failures of projects in later phases are prevented. State of the art methods for target validation range from gene expression analysis, siRNA screening, use of tool compounds, image analysis like high content screening [9] or tissue array investigations [10] to studies using viral vectors [11] or genetically modified mice [12–15]. siRNAs, for instance, could be used to mimic the effects of compounds inhibiting the protein encoded by the siRNA's target transcript. A further example in this regard is target validation using Adeno-associated virus (AAV) [16]. AAV infects humans and some other primate species. Only inducing a very mild immune response, is currently not known to cause disease. The virus specifically integrates its genome into that of the host cell. Thus, it is a very attractive approach to create viral vectors for the over-expression of genes, thereby elevating the level of the respective proteins. Targets like E3-ligases, leading to the ubiquitination and subsequent degradation of target proteins, could be validated using such an approach. If the enrichment of proteins leads to the desired effect, the target would be validated. After successful validation of a target, a lead has to be identified that modulates the respective target in the desired way.

Lead Identification

The main goal of lead identification is to find a chemical, a so called "lead compound" or simply "lead", that may act on the target to alter the disease. This molecule constitutes the initial scaffold for the new drug.

High-throughput screening (HTS) is the most common way that leads are found. Advances in robotics and computational power allow researchers to test hundreds of thousands of compounds against a target. The compounds present in screening pools cover a wide range of the chemical space. Many successful drugs are derived from naturally occurring molecules. Penicillin which is produced by the fungus *Penicillium* when its growth is inhibited by stress [17] is possibly the most prominent example. But also butylscopolamine and acetylsalicylic acid which are the active

ingredients for Buscopan® [18, 19] and Aspirin® [20], respectively, constitute very successful drugs of natural origin. Thus, parts of compounds present in libraries for HTS are "nature product-like" [21, 22]. Based on the results of HTS, several lead compounds are usually selected for further study.

Besides HTS, in "rational design" detailed structural knowledge of the targets' active site can be used by chemists to design molecules that interact with this site. Additionally, in *de novo* synthesis by computational chemistry sophisticated computer modeling is used to predict what type of molecule could be used as a lead. Typically the 3D-structure of at least the target protein's active site has to be resolved in advance, again, either by computational modeling or by X-ray crystallography [23–25].

After a lead has been successfully identified, this structure is subsequently optimized towards different parameters during lead optimization.

Lead Optimization

Compounds that were selected in the initial screening are optimized or altered in order to maximize efficacy, specificity and period of impact while at the same time minimizing side effects and toxicity. Based on their chemical and physical properties and the resulting biological effects researchers try to infer the behavior of leads to minimize the failure rate in a later stage of the drug development process.

The structure activity relationship (SAR) describes the relationship of the lead's structure to its pharmacodynamic as well as pharmacokinetic properties. Physical and chemical properties are used to quantify the structure activity relationships (QSAR). Pharmacodynamic refers to mechanism of the drug at the target, pharmacokinetic to its distribution in the organism and in which concentration the drug and its metabolites are present. Pharmacokinetic is described in terms of LADME/Tox parameters which characterize the compound with regard to liberation, absorption, distribution, metabolism, elimination, and toxicity: Liberation refers to the release of the active ingredient from the drug, absorption to the entering into the blood stream, distribution to the circulation to the proper site of action, metabolism to the decomposition, elimination to the excretion of the drug and toxicity to potential or real toxicity. Resorption and bioavailability are closely linked to LADME/Tox

parameters. They describe the uptake of substances into the blood and the fraction of active ingredients available to the organism, respectively. For an optimal resorption and bioavailability and thereby for a better efficacy and duration of effect the lipophilicity, the size of the molecule and the related metabolic stability of a molecule are decisive. Quite often leads fail in early stages due to low bioavailabilty and due to toxic metabolites. Thus, LADME parameters are taken into account in early stages. The most famous parameters which are considered with respect to LADME are probably Lipinski's rule of five. Analyzing the properties of about 2,250 compounds of the Derwent World Drug Index (WDI) Lipinski *et al.* were able to summarize properties that were common to 90% of the compounds [26]. Lipinski assumes that as soon as more than one property is not met by a compound the success rate gets relatively low since absorption and permeability are decreased. However, Lipinski's rule of five should be considered as guidelines, not as hard rules.

LADME/Tox studies are performed in living cells, in animals and via computational models. They help researchers prioritize lead compounds early in the discovery process. Technologies such as magnetic resonance imaging and X-ray crystallography, along with powerful computer modeling capabilities support chemists to design chemical structures to optimize LADME/Tox properties. Additionally, the molecules are changed to minimize possible interactions with other molecules, thus reducing the potential for side effects.

Different variations or analogues of the initial leads are designed and tested. The resulting compounds represent the candidate drugs that enter pre-clinical testing.

Pre-clinical Testing

Regulatory institutions require extremely thorough testing before the candidate drug can be studied in humans. To meet these high standards, scientists carry out in-depth *in vitro* and *in vivo* tests to understand how the drug works and what its safety profile looks like in pre-clinical tests.

After starting with 5,000 up to 10,000 compounds (Figure 1.1), between one and five molecules will be studied in clinical trials. Different authorities around the world are in charge of approving drugs for clinical testing and marketing. The most

influential ones are probably the U.S. Food and Drug Administration (FDA) [27], the European Medicines Agency (EMA) [28] and Japan's Ministry of Health, Labour and Welfare (MHLW) [29]. Great efforts are made by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) to harmonize the process of registration. Thus, in the following I focus on the guidelines as set by the FDA, which can be found on the respective web page [27]. Guidelines set by other institutions are, by and large, comparable.

Investigational New Drug (IND) Application and Safety

The goals of the IND are to provide enough information to permit FDA reviewers to determine whether the drug is safe enough to start first human trials. Basis for this application are all results gathered during the pre-clinical work in the drug discovery process: the candidate drug's chemical structure, how it is thought to work in the body, a listing of any side effects and manufacturing information. Additionally, the IND also provides a detailed clinical trial plan.

1.1.2 Clinical Trials

To provide confident results, drug trials are generally placebo-controlled, randomized and double-blinded.

- Placebo-controlled: Some subjects will receive the new drug candidate and others will receive a placebo. In some instances, the drug candidate may be tested against another treatment rather than a placebo.
- Randomized: Each of the study subjects in the trial is assigned randomly to one of the treatments.
- Double-blinded: Neither the researchers nor the subjects know which treatment is being delivered until the study is over.

Such a study design provides the best evidence of a direct relationship between the drug and its effect on the disease.

The number of subjects participating in a trial has to be carefully considered: On the one hand, the more subjects take part in the study, the more likely real effects are detected. On the other hand, the more subjects are investigated, the

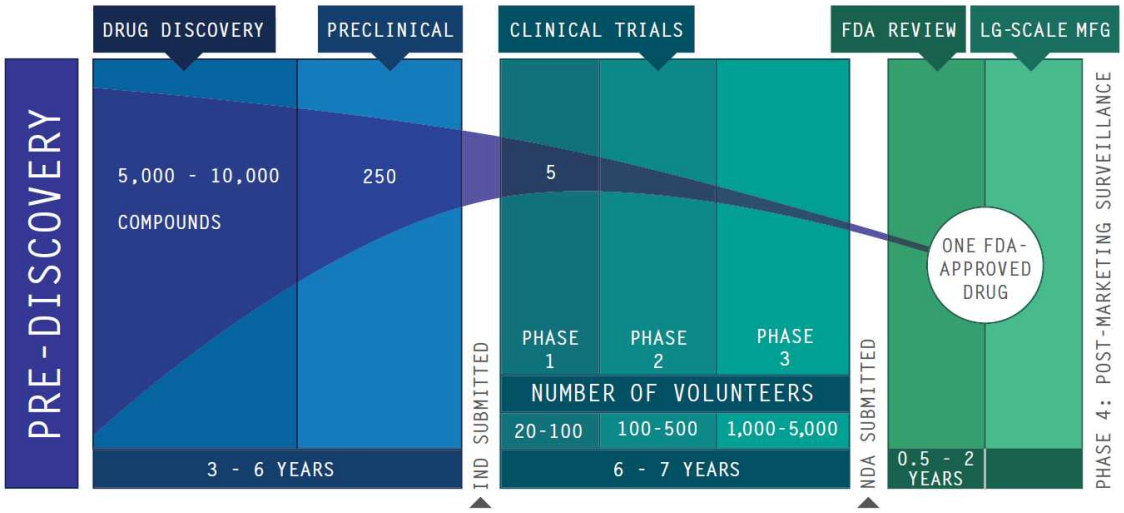


Figure 1.1: **Drug Development Process.** The different phases as well as the compounds left for consideration and the approximate time necessary for each phase are displayed. Additionally, the number of participants in the clinical trials are listed. This figure was taken from [3, 4].

more expensive and difficult is the trial. Additionally futility has to be taken into account. It is unethical to expose unnecessary many people to the unavoidable risk of clinical trials as long as the effects of the drug candidate is not ensured.

The Phases of Clinical Trials

Recently, the FDA has established the Phase 0 trial, which allows researchers to test a very low drug dose in a small cohort of volunteers to quickly identify drug candidates that are ineffective. Thereby, costs and lengths of clinical trials can be efficiently reduced.

In Phase 1 trials the candidate drug is tested in a bigger cohort of volunteers for the first time. These studies are usually conducted with about 20 to 100 healthy volunteers. The main goal of a Phase 1 trial is to discover if the drug is safe in humans. Researchers look at the pharmacokinetics and the pharmacodynamics of a drug. People participating in these trials are at any time monitored thoughtfully. Based on these trials, the safe dosing range is determined and it is decided whether the drug should be pushed forward to the next phase.

In Phase 2 trials researchers evaluate the candidate drug in about 100 to 500 patients suffering from the disease, and examine possible short-term side effects. This phase could also be split up into Phase 2a and 2b. Phase 2a would investigate efficacy and dosage in a small group of patients. Based on these results, Phase 2b could be optimally designed for a larger group of patients. During Phase 2 mechanism and efficacy as well as optimal dose strength and schedules are investigated to optimally design a Phase 3 trial.

In Phase 3 trials researchers study the drug candidate in about 1,000-5,000 patients to generate statistically significant data. This phase is key in determining whether the drug is effective and safe. Phase 3 trials are both the most expensive and the most time-consuming trials.

During all clinical phases researchers are conducting many other critical studies like investigations for large scale production. Further, the complex application required by the authorities for approval of the new drug has to be prepared.

New Drug Application (NDA)

The goals of the NDA are to provide enough information to permit FDA reviewers to reach the following key decisions:

- Whether the drug is safe and effective and whether the benefits of the drug outweigh the risks.
- Whether the package insert is appropriate and what it should contain.
- Whether the methods used in manufacturing the drug and the controls used to maintain the drug's quality are sufficient.

If the review is positive, the new drug gets approved.

Post-marketing phase

Even if a new drug has been approved and is already marketed, it is still monitored with respect to detection, assessment, understanding and prevention of adverse effects as well as the patients overall satisfaction related to the treatment. This process is called pharmacovigilance. Surveys are conducted and information is collected to

further evaluate the drug. The most important aim is to identify possible hazards associated with drugs as soon as possible to prevent unnecessary harm to patients.

1.2 Mode of Action

Analyzing mode of action (MoA) of compounds or new drugs is one of the most important but also probably the most difficult part during the drug development process. Negligent analyses of mode of action can easily lead to costly late failures of compounds, in the worst case constituted by severe effects observed in patients as it was the case for torcetrapib, a cholesteryl ester transfer protein (CETP) inhibitor.

At the time Pfizer discontinued its so-called ILLUMINATE (Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events) trial, the company already had spent $800 millions on R&D. Analyses of the clinical trial led to the suggestion of adverse events that may have been responsible for an observed increase in mortality rate of patients treated with torcetrapib in combination with atorvastatin (82) compared to patients only treated with atorvastatin (51) [30]. Since research on compounds targeting CETP was and is still going on at other companies, it has indeed been confirmed in animal experiments that an off-target mechanism of torcetrapib increases blood pressure [31]. Further, in rats, torcetrapib was shown to be associated with an increase in plasma levels of aldosterone and corticosterone and, *in vitro*, with an release of aldosterone from adrenocortical cells. The increase in blood pressure was not mediated by the increased levels of steroids but was shown to be dependent on intact adrenal glands. In later studies the off-target effect of torcetrapib could be further narrowed down to be related to an increase in the expression of the alpha subunit 1C from the voltage-gated L -type Ca^{2+} channel [32]. By siRNA mediated knockdown of L-type Ca^{2+} channel subunits alpha 1C and alpha 1D, Clerc *et al.* [32] could show the decrease of aldosterone and aldosterone synthase (CYP11B2). Thereby, they provided a mechanistic link between torcetrapib and aldosterone that is related to activation of the L -type Ca^{2+} channel. In rats, torcetrapib has been shown to induce a potent hypertensive effect mediated by the L-type Ca^{2+} channel. Clerc *et al.* conclude that steroidogenic and hypertensive side effects of torcetrapib may be linked and involve voltage-gated L-type Ca^{2+} channels.

Additional to this prominent example, analyses of Thomson Reuters Life Sciences Consulting reveals 83 failures during phase 3 trials and submissions for the period from 2007 to 2010 [33]. Main reasons for these failures were lack of efficacy (66%) and safety issues (21%). Large proportion of failures are observed for drugs with novel machanism of action that at the same time are located in areas of high unmet medical needs. Such indication areas face reasearchers with challenging science as well as put high competitive pressure on companies to fill their pipelines with compounds that turn to account. To achieve fast success, companies are willing to take higher risk and move forward their compounds although they only display marginal statistical significance. Additionally, indication areas like cancer tempt researchers to prematurely try to reposition drugs/targets from one cancer type to another without sufficiently testing the relevance of the mechanism of action. Examples are sunitinib (Pfizer) in hepatic cancer and bevacizumab (Genentech/Roche) in gastric cancer.

Success rates can be improved only by relying on high quality scientific evidence, by fully testing mechanism and by thoughtful planning of clinical studies with well defined end-points. This could lead to higher failure rates in early phases but at the same time would save money that would better be invested into other drug candidates, which may in the end lead to a sound pipeline containing more promising drugs.

To validate a compound it is most important to know as much as possible about its mode of action. Mode of action of compounds can be broadly split into two classes, the on- as well as the off-target effects. An on-target effect is an effect induced through the direct interaction, i.e. inhibition or activation, of the compound with the intended target molecule. On the contrary, an off-target effect is an effect induced by the unwanted interaction of the compound with a different molecule than the target. Although these effects are generally unwanted, there are examples for which off-target effects of existing drugs show potential for new indications. Example are Nelfinavir [34, 35] or Xenical [36]. However, there are of course many examples were undesired off-target or adverse effects led to failure or in the worst case even withdrawal of drugs. Examples for expensive failure in Phase 3 studies have been described previously. A prominent example for a withdrawal is rofecoxib

(Vioxx®) which has been withdrawn from the market in 2004 based on early results of the APPROVe (Adenomatous Polyp PRevention On Vioxx) trial with the primary aim of evaluating the efficacy of rofecoxib for the prophylaxis of colorectal polyps [37, 38]. This study confirmed a higher risk for cardiovascular events (i.e. heart attack and stroke) only after 18 months of chronic use of the drug. Each pharmaceutical company wants to avoid such dramatic failures. First, they potentially harm patients, second, they are bad for reputation, and patients possibly lose trust in drugs marketed by the respective companies no matter how safe they are.

1.3 Objective of the Thesis

The aim of this thesis is to develop a method that supports researchers in focusing the drug development process by revealing unpromising directions as early as possible. This could be achieved by analyzing mode of action and potential side effects by incorporating existing prior biological knowledge. By supporting the analyses of mode of action and by filtering for more promising drug candidates in early phases, the process is not only more focused but also cheapened and more humanized. Dramatic failures due to lethal adverse events as well as animal experiments for unpromising compounds can hopefully be prevented.

To analyze compounds' mode of action, one important step is the understanding of cellular mechanisms induced by the drug candidate. Possibilities to survey these cellular processes include gene expression analyses or investigation of protein levels. For the drug development process it is of importance to conduct high-throughput experiments which help to get an understanding of underlying cellular processes. Gene expression analyses are commonly conducted at the very beginning of the drug development process. By analyzing such data, researchers try to explain the biological causes and consequences of transcriptional changes.

For gene expression measurements using microarrays, normalization of the data has to be performed to minimize systematic effects that are not constant between different samples of an experiment and that are not due to the factors under investigation (e.g. treatment, time). Optimal selection of a normalization method heavily depends on the nature of the experiment. Factors like comparability and quality

of single runs play a major role. It has been shown that the normalization method used may influence further downstream analysis to a great extend [39], and thus, has to be carefully chosen based on the actual data. Therefore, the first part of this thesis is concerned with the decision on an optimal normalization procedure for the experiment under investigation. Based on the normalized data, genes differentially expressed between distinct conditions like diseased versus healthy or compound-treated versus untreated cells can be detected.

Algorithms analyzing differentially expressed genes with respect to the biological process the genes are associated with have been studied thoroughly [40–51]. Based on differentially expressed genes, these methods search for gene sets or modules of genes that are related to the biological process studied in the expression experiment. Such approaches are state-of-the-art methods that could be used to analyze compounds' mode of action based on gene expression data. In the second part of this thesis, some of these well established methods were considered in basic state-of-the-art *in silico* analyses of gene expression data. Expression profiles of several compounds are investigated with respect to their biological characteristics as potential drugs.

One challenge in the interpretation of the data is that regulation does not only occur on the transcriptional level. Additionally, compound effects are mediated by binding of compounds to proteins and subsequently influencing related regulatory networks within the organism. Discovering key proteins as well as their meaning in a broader biological context is one of the most promising ways to answer questions with respect to the mode of action. As more and more data from diverse sources become available, the integration of this knowledge is important for generating a deeper insight into biology. Existing methods for the analysis of gene expression data only make very limited use of prior knowledge. To my knowledge, so far no method exists that efficiently integrates prior knowledge with data from functional genomics to elucidate the mode of action of compounds. To help closing this gap, the third part of this work is concerned with a mathematical model for the development of a new knowledge mining approach.

In the fourth part I focus on the development of an algorithm to identify modules

of genes/proteins that help explaining the mode of action by utilizing the previously proposed knowledge mining approach.

Finally, in the last part of this work, the newly developed methods are compared to existing ones, and analyzed with respect to possible advantages and improvements.

1.4 Overview of the Thesis

After having given the basic background and objective in this chapter, I give a brief overview of the layout for the remaining main part of my thesis:

In Chapter 2, I review existing knowledge related to this thesis:

In Section 2.1, TGF-β-signaling, the biological system used throughout this work, is described. In the remaining sections of that chapter, I give an overview of work available in the field of computational biology that could be helpful for the analysis of mode of action: Section 2.2 describes existing methods for data integration, Sections 2.3 and 2.4 give an overview of gene based analyses that can give insight into underlying biological processes.

In Chapter 3, the methods used throughout this thesis are described, Chapter 4 summarizes results applying these methods:

All laboratory work described in Section 3.1 has been conducted by Dr. Patrick Baum in the scope of his PhD thesis [52]. It is mentioned here for completeness. These experimental data have been used to assess the new methods developed in the present work with respect to the analysis of mode of action.

In Section 3.2 methods used to select an optimal normalization procedure for the TGF-β gene expression data are described. Respective results are presented in Section 4.1 and discussed in Section 5.1. This part of my thesis has been published in BMC Genomics [53].

In Section 3.3, I describe how we derived on- and off-target signatures and introduce available methods for the analysis of differentially expressed genes. Results

applying these methods are presented in Section 4.2 and are part of a PLoS One publication [54].

In Section 3.4, I describe different data sources and how they can be used to measure the biological relatedness for pairs of proteins. Section 3.4.6 states the final formula (Equation 3.30) to integrate the data. In Section 3.5 modEx, the newly proposed module extraction method, is introduced. modEx is a heuristic that could be applied to solve different optimization problems. These problems are stated in Section 3.5.2, their complexity is analyzed in Section 4.5.2. Weighted protein interaction networks constitute the input to modEx. Weights of the networks used in the presented use cases have been calculated by the newly proposed data integration method (Section 3.4.6). Section 3.6 states possibilities of how to assess the significance of extracted modules. In Sections 4.3, the newly proposed approaches are first assessed using TGF-β signaling (Section 4.3.2) and finally use cases of how they can be applied to complement the analysis of MoA are given based on two exemplarily selected compounds inhibiting TGF-β signaling (Section 4.3.3).

Section 3.7 explains how the newly proposed integration method and modEx are compared to widely accepted approaches, namely STRING [55] and jActiveModule [48]. The comparisons to these approaches is described in Section 4.4.

Results are discussed in Chapter 5. Sections 5.2 and 5.3 evaluate the general findings for the newly proposed data integration method and for modEx, respectively. Assessments of the methods with respect to TGF-β signaling and analysis of mode of action are discussed in Section 5.4. This section also comments on the comparison to STRING and jActiveModule.

Finally, I conclude my work in Chapter 6 and give a perspective on possible future work.

Chapter 2

Literature Review

In this chapter, I give an overview of existing approaches related to this thesis. In Section 2.1, I describe the biological system used for the underlying experiments [52, 54]. In the remaining sections, I review work available in the field of computational biology that could be helpful for the analysis of mode of action. Section 2.2 describes existing methods for data integration, Section 2.3 gives an overview of gene expression analyses which could be applied to get insight in the underlying biological processes.

2.1 Biological System Used

2.1.1 TGF-β Signaling

The transforming growth factor β (TGF-β) family is composed of structurally related cytokines which effect processes like morphogenesis of many organs and tissues as well as proliferation, differentiation, migration and apoptosis of many different cell types [59]. Examples are TGF-β1, TGF-β2, TGF-β3, the family of activins and nodal. The basic TGF-β signaling system consists of two receptor serine/threonine protein kinases (TGF-β receptor types I and II, TGF-βR1 & R2) and the SMAD proteins. Figure 2.1 schematically displays the pathway in a simplified way, a more biological representation is given in Appendix A.1. Components of this pathway, which is briefly summarized in the following, have been studied extensively [59,61–64].

Prior to activation of TGF-β signaling, factors like SARA (SMAD Anchor for

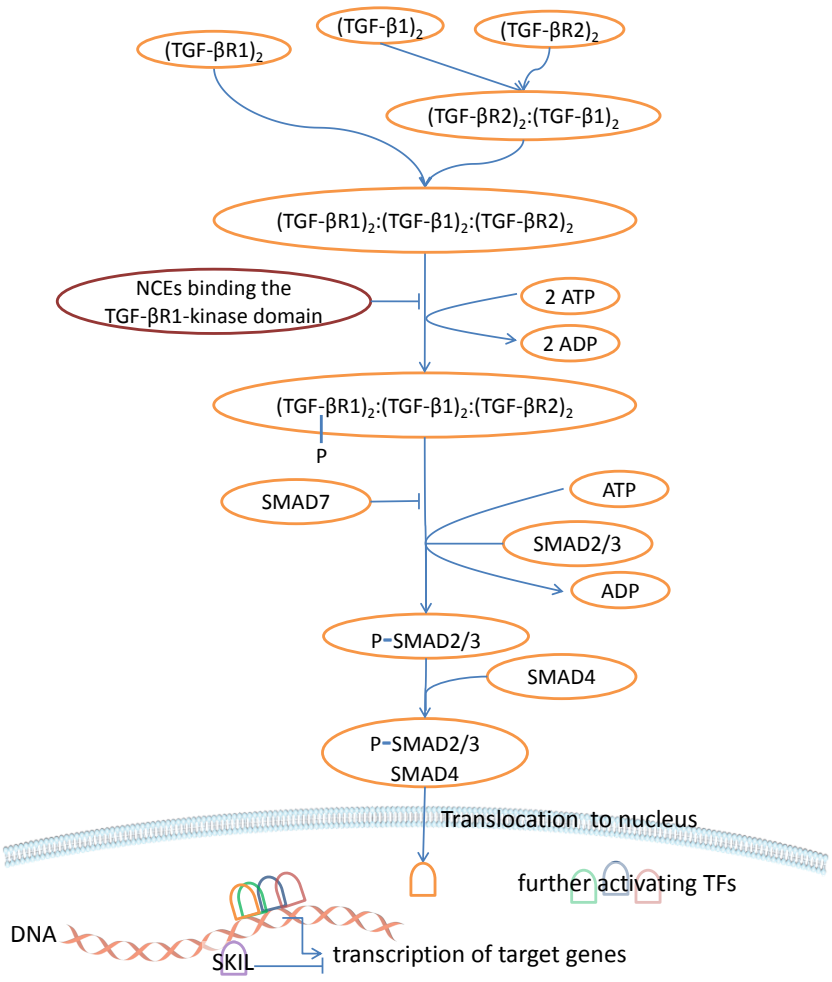


Figure 2.1: **Schematic representation of the TGF-β signaling pathway.** Basic molecular events involved in TGF-β signaling via SMAD proteins. At least three, and perhaps four to five amino acid residues of TGF-βR1 must be phosphorylated to fully activate the protein [56–58]. For simplicity, only one P is depicted per TGF-βR1 to indicate overall phosphorylation/activation. Also for the the SMAD2/3 complexe, we only depict one P, though the activation is achieved through two phosphorylations of two amino acid residues in both, SMAD2 and SMAD3 [59, 60]. A more detailed, biological representation is displayed in Appendix A.1. TF: Transcription factor, P: Indicates phosphorylation, i.e. activation.

Receptor Activation) recruit the regulated SMADs (R-SMADS) into proximity of the TGF-β receptor kinases. Signaling is initiated by binding of a TGF-β ligand dimer to TGF-βR2. The activated TGF-βR2 in turn recruits TGF-βR1 to build a

heterotetrameric complex with the ligand dimer [61]. The serine/threonine kinase region of TGF-βR2 catalyzes the phosphorylation of serine residues of TGF-βR1. Activated TGF-βR1 further propagates the signaling by phosphorylation and subsequent activation of SMAD2 and SMAD3, the R-SMADS. Phosphorylation induces a conformational change of R-SMADs which leads to dissociation from the receptor complex and SARA. The free and phosphorylated R-SMADs have a high affinity to form heteromeric complexes with common-mediator SMAD (co-SMAD), SMAD4. These phosphorylated R-SMAD/SMAD4 complexes enter the nucleus where they partner with other transcription factors resulting in cell-state specific modulation of transcription (Figure 2.1) [64].

TGF-β signaling is regulated at several levels. First, the access of the R-SMADs to activated TGF-βR1 is controlled by SARA. Second, the E3 ubiquitin ligase, SMAD ubiquitination regulatory factor-2 (SMURF2), attacks cytoplasmic R-SMADs which leads to proteasomal degradation of R-SMADs. Third, SMAD7, the inhibitory SMAD (I-SMAD), acts antagonistically and inhibits receptor mediated activation of R-SMADs. It also associates with SMURFs to form the SMAD7-SMURF complex after TGF-β stimulation and ubiquitinates the receptors on the cell surface or endosomal membranes; these are then targeted for degradation in proteasomes and lysosomes [65]. The oncoprotein c-Ski functions as a direct antagonist of TGF-βR1 [66]. Further, STRAP1 enhances the inhibitory activity of SMAD7 by binding to TGF-βR1 and SMAD7 [67]. Another level of control of the SMAD pathway is via the regulation of nuclear accumulation of SMADs, by the Ras-extracellular signal kinase (ERK) pathway. SnoN (SKIL) constitutes a self-regulatory mechanism of TGF-β signaling. Expression of SnoN is activated by TGF-β signaling. On the one hand, SnoN regulates TGF-β signaling by binding to SMADs to block transcriptional activation. On the other hand, nuclear R-SMAD-SMURF complexes recruit transcriptional repressors SnoN for ubiquitin-mediated degradation and thus down-regulate the repressor. Thus, TGF-β signaling causes degradation of SnoN, releasing SMADs to regulate transcription, but also activates expression of SnoN to down-regulate SMAD signaling at later times [67].

There are also non-SMAD mediated signaling events [68]. For instance, TGF-βR1 can mediate JNK signaling by interacting with E3 ubiquitin ligase TRAF6 and sub-

sequent activation of TAK1 and the MKK3/4, a mitogen activated protein kinase kinase complex, to trigger TAK1-p38/JNK pathway-dependent apoptosis [69,70] (Appendix A.1).

2.1.2 Diseases Related to TGF-β Signaling

Malfunctions within the TGF-β signaling pathway may result in cancer, fibrosis and diverse hereditary disorders [71–73]. During cancerogenesis the function of TGF-β has been shown to be dependent on the tumor state [74–76]. On the one hand, TGF-β acts as tumor suppressor by inhibiting the proliferation of normal epithelial, endothelial haematopoietic cells, and early epithelial cancer cells. In early cancers, for instance, cells are still subject to TGF-β-mediated growth inhibition. On the other hand, once tumorgenesis has been initiated tumor cells escape this growth control and produce high levels of TGF-β resulting in promoted tumor growth and metastasis. Dumont *et al.* [76] revealed that induction of epithelial mesenchymal transition (EMT) leads to repression of CDH1 through histone deacetylation. Longterm silencing is maintained by subsequent promoter hypermethylation. Further, Chou *et al.* [75, 77, 78] could show that such epigenetic effects contribute to the disruption of TGF-β signaling in ovarian cancer. Agents that inhibit epigenetic modifying enzymes like DNA methyltransferases and Histone deacetylases are investigated as cancer therapies [78]. Since activation of TGF-β signaling can induce EMT trough epigenetic silencing, blocking TGF-β signaling could possibly reverse the epigenetic modifications thereby preventing or even reversing EMT. Indeed this could be achieved for a mesenchymal breast cancer cell line [79].

Fibrosis, goblet cell hyperplasia and smooth muscle thickening are implications of diseases like asthma, chronic obstructive pulmonary disease (COPD), and idiopathic pulmonary fibrosis (IPF) [80,81]. Among other growth factors and cytokines, TGF-β is highly expressed in fibrotic tissues and up-regulates the expression of adhesion molecules required for the recruitment of monocytes and neutrophils which both initiate inflammatory responses. Furthermore, TGF-β plays a pivotal role in the biosynthesis and turnover of extracellular matrix (ECM) proteins like collagens, fibronectin and proteoglycans, thereby contributing to fibrosis and smooth muscle cell proliferation [82].

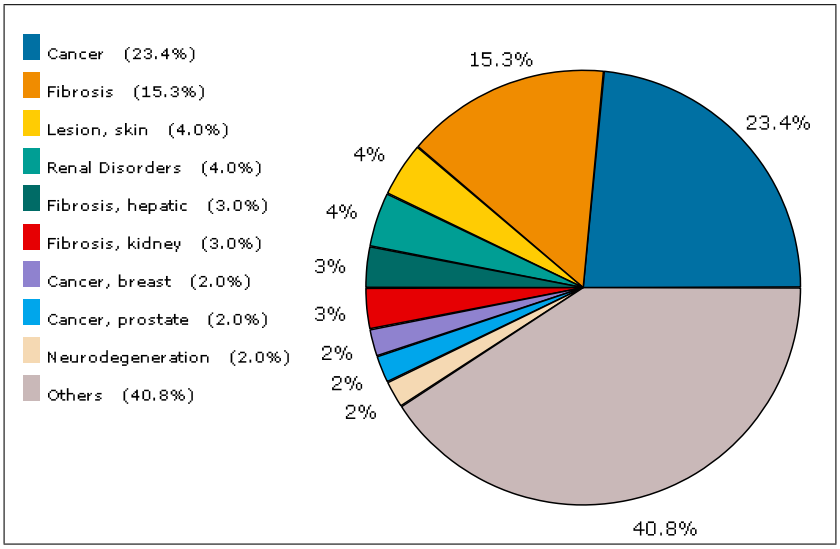


Figure 2.2: **Therapeutic groups for drugs targeting TGF-βR1 or TGF-β.** Displayed are the percent of drugs investigated in different therapeutic groups. In total, Thomson Reuters' Integrity returns 78 drugs targeting TGF-βR1 or TGF-β which are or have been tested (June 2011). 68 are reported to be in the phase of biological testing, 9 are listed under pre-clinical testing, and even one, Fresolimumab [83], a monoclonal antibody against TGF-β1-3, has entered Phase 2.

2.1.3 Conclusion

TGF-β signaling is a relatively well studied pathway with high therapeutic potential. Clearly, inhibition of TGF-βR1 holds promise for the treatment of fibrotic diseases and cancer. This is also reflected by Integrity [84] as displayed in Figure 2.2. Several small molecules inhibiting TGF-βR1 have been proposed [1]. Some of them have been sucessfully applied to prevent tumor progression in human cancer [85, 86]. Most known TGF-βR1 inhibitors occupy similar positions in the ATP pocket of the TGF-βR1 kinase domain. As these are domains conserved in several kinases [87–89], this approach holds potential for cross reactivity between off-target kinases and the compounds. Boehringer Ingelheim provided a set of 7 compounds which target the ATP binding cassette of TGF-βR1, one of the initial components of the pathway. Since patents are held for these compounds, we are granted freedom to operate. This is the ideal pre-requisite to be able to conduct an in-depth and thorough analysis of mode of action for both, on- as well as off-target effects. We do so by first using state-of-the-art methods and in a later step develop new computational methodologies that

could be applied to support and streamline necessary follow up wet-lab experiments in future projects.

2.2 Existing Methods for Data Integration

As more and more diverse and even high throughput technologies get available, more and more data is generated that could help to explain biological processes. These data usually originate from different sources like proteomics and genomics. To be able to conduct computational analyses that make use of diverse data, the integration of this knowledge is an important step. In this section, I give an overview of existing methods for data consolidation and integration.

All the methods described in this section are based on protein interactions. A protein interaction refers to two proteins that either show a direct physical interaction or that are functionally associated in a biological system. We do not refer to genetic linkage which in contrast describes the tendency of genetic loci to be inherited together. In a theoretical setting, protein interactions are commonly represented as graphs.

Definition 2.2.1. *Graph theoretical representation of protein interactions.*

> *Protein interactions can be represented as an undirected graph* $G=(V,E)$. *The set of nodes/vertices* V *refers to the proteins/genes, additionally, each pair of interacting proteins* $\{v_i, v_j\} \subseteq V$, $\{i,j\} \in \mathbb{N}$ *is represented by an undirected edge* $e_{v_i,v_j} \in E$.

As each protein is encoded by a gene, I do not strictly differentiate between these two terms and, in the graph representations, use them interchangeably throughout the thesis. Further, the terms graph and network are used interchangeably.

Based on these graphs, additional data could be integrated. This could be done by giving different weights to the edges of the graph based on prior knowledge.

Definition 2.2.2. *Edge-weighted graph.*

> *An edge-weighted graph is a pair* (G,ω), *where* $G=(V,E)$ *is an undirected graph and* $\omega : E \rightarrow \mathbb{R}$ *is a weighting function assigning a weight* $\omega(v_i,v_j)$ *to all* $e_{v_i,v_j} \in E$.

The weighting function ω could for example describe the probability of a real physical interaction or the degree of biological relatedness, i.e. the functional association, of two connected proteins based on prior knowledge.

Various approaches for the integration of heterogeneous data sources have been applied to the prediction of protein function as well as for the generation of functionally linked gene networks. The first ones have been proposed by Marcotte *et al.* [90] and Yanai *et al.* [91], applying the intersection or the union of distinct sets of evidence, respectively. More sophisticated approaches have been, for example, proposed by von Mering *et al.* [55], by Lee *et al.* [92,93], and by Linghu *et al.* [94,95]. They are described in more detail in Sections 2.2.5 to 2.2.7.

2.2.1 iRefIndex

Multiple databases/repositories exist that contain information about protein interactions. Interaction data for a single protein could be spread across these databases. Razick *et al.* [96] propose a method to consolidate the information of different protein interaction databases. They integrated data from BIND [97,98], BioGrid [99], DIP [100], HPRD [101,102], IntAct [103,104], MINT [105], MPact [106], MPPI [107] and OPHID [108]. Based on the primary sequence of the respective proteins as well as their taxonomy identifiers, a unique key for a protein interaction as well as for each participant protein is generated. Thus, the same key is only generated for identical pairs of protein sequences and taxonomy identifiers. Thereby, it is possible to filter for redundant information contained in the different protein interaction databases. The resulting interaction Reference Index (iRefIndex) is provided in PSI-MITAB 2.5 [109], via a queryable web server (iRefWeb) as well as via a plug-in (iRefScape) for Cytoscape [110] and via a web service. Based on this data, a graph G describing a consolidated set of protein iteractions can be derived.

2.2.2 Michigan Molecular Interactions (MiMI)

MiMI [111, 112] provides access to knowledge and data that was merged and integrated from numerous databases. Each repository for protein interaction data has its own data format, molecule identifier, and supplementary information. Molecules that may have different identifiers but represent the same entity are merged. Thus,

MiMI allows the user to retrieve information from different databases at once, highlighting complementary and contradictory information. Because the merge process is an automated process, and no curation occurs, any errors in the original data sources will also exist in MiMI.

MiMI gives access to the following information:

- Information on genes like Gene Ontology annotations [113], interactions, literature citations, compounds, and annotated text extracted through NLP.
- Link-outs to tools to analyze overrepresented MeSH terms for genes of interest, read additional NLP-mined text passages, and explore interactive graphics of networks of interactions
- Link-outs to PubMed and NCIBI's MiSearch interface to PubMed for better relevance rankings
- Querying by keywords, genes, lists or interactions

The MiMI database lists genes together with supplementary information like interactions, Gene Ontology information, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [114, 115] the gene occurs in and relevant publications.

Data in MiMI can be accessed in three different ways, via a web-interface, via a web-service, and as PSI-MITAB formatted flat file. The PSI-MITAB files only represent a subset of the data available in MiMI. While UniProt and RefSeq identifiers are included for each interactor, provenance is only included for parts of the interactions. Pathways and metabolomics data is not included at all. To visualize interactions of interest, a Cytoscape plug-in is available.

2.2.3 PINA

The Protein Interaction Network Analysis (PINA) [116, 117] system is an integrated platform for protein interaction network construction, filtering, analysis, visualization and management. It integrates protein-protein interaction data from six public curated databases and aims to supply a complete, non-redundant protein interaction dataset for six model organisms. PINA allows users to either edit the networks generated from the public data, or combine them with uploaded private interactions to build more complete protein-protein interaction networks. Moreover, it provides

a variety of built-in tools to filter and analyze the network for gaining more focused insight. These analyses include enriched GO term and KEGG pathways identification, topology feature calculation, identification of topologically important proteins in the interaction network, and identification of common interacting proteins. Networks can be filtered based on annotation or based on the semantic similarity score between annotated GO terms of interacting proteins.

Interaction networks can be downloaded in GraphML format, PSI-MITAB format or PINA tab-delimited format, complete and annotated lists of protein-protein interactions (PPI) for the different organisms can be obtained in PSI-MITAB format. Registered users can save protein interaction networks generated from user query or the output of the analysis tool on the server for further analysis.

2.2.4 Distributed Annotation System for Molecular Interactions (DASMI)

DASMI [118] is based on the decentralized client-server architecture of the Distributed Annotation System [119] and consists of a data exchange specification, interaction data servers, and visualization clients. DASMI provides a collection of protein-protein interaction datasets and domain-domain interaction datasets. Additionally, two systems can be used to assess the confidence of interactions: FunSimMat and Domain support. FunSimMat calculates the similarity of genes based on their GO annotations. Domain support is based on domain interactions that have been derived from crystal structures or were computationally predicted.

For deriving the data, a web server, DASMIweb [120], can be used. It is also possible to perform queries in batch mode. Finally, DASMI offers the possibility to integrate interaction data from PSI-MI XML files.

2.2.5 STRING

One of the first and most popular methods for data integration was proposed by von Mering *et al.* [121]. They introduced a Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). In the underlying database, information on possible protein-protein associations is aggregated. This database does not only

focus on direct protein-protein interactions but also links functionally associated pairs of proteins. Version 8.3 of STRING covers about 2.5 million proteins from 630 organisms. Protein interactions are scored and weighted by a quantitative integration of predictions based on genomic context, high-throughput experiments, co-expression and previous knowledge as available in protein interaction databases and literature as sources. For each individual source k, raw scores are calculated and benchmarked against a set of trusted true associations taken from KEGG as a gold standard. A predicted association $\{v_i, v_j\}$ is counted as a true positive association if the respective proteins v_i and v_j occur in the same KEGG pathway. The true positive rate referred to as confidence score $\tilde{S}_k(v_i, v_j)$ generally represents the probability of finding the linked proteins within the same KEGG pathway. Under the simplifying assumption of independence for the different sources used, the k individual scores for a pair of proteins v_i, v_j are combined in a naïve Bayesian fashion:

$$S_{v_i,v_j} = 1 - \prod_k (1 - \tilde{S}_k(v_i, v_j)),$$

were k refers to the k-th source and $\tilde{S}_k(v_i, v_j)$ to the confidence score derived based on the k-th source. S_{v_i,v_j} and the respective protein pairs $\{v_i, v_j\}$ can then be used to define an edge weighted graph $G = (V, E)$ with edge weights $\omega(e_{v_i,v_j}) = S_{v_i,v_j}$ (see Definition 2.2.2, page 20).

2.2.6 Log Likelihood Score as Method to Integrate Heterogeneous Data Sources

The method proposed by Lee *et al.* [92,93] is based on Bayesian statistics applying a log likelihood score (LLS). Based on each data set to integrate, an odds ratio is calculated. The odds ratio represents the likelihood that a pair of genes is functionally linked. If $P(L|E)$ represents the probability that two genes $\{v_i, v_j\}$ are functionally linked given a dataset E, $P(\bar{L}|E)$ gives the probability that the two genes $\{v_i, v_j\}$ are not linked, and $P(L)$ is the unconditional probability that two genes v_i, v_j are functionally linked, then the odds ratio (OR) that a given pair of proteins v_i, v_j is functionally linked is given by:

$$\mathrm{OR}(L, E) = \frac{P(L|E)/P(\bar{L}|E)}{P(L)/P(\bar{L})}$$

Here, $P(L)/P(\bar{L})$ represents the *prior* odds. It is estimated by the number of gene pairs with a shared functional annotation divided by the number of gene pairs without any shared function based on a single source of functional annotation, e.g. KEGG. $P(L|E)/P(\bar{L}|E)$ represents the *posterior* odds. It is estimated by the number of gene pairs that share functional annotation and that are supported by the given evidence E divided by the number of gene pairs that do not share functional annotation based on the given evidence E. To create an additive score such that linkage information $\mathrm{OR}(L, E)$ calculated based on different evidence E can be combined, the log likelihood score is finally calculated as $LLS = \ln(\mathrm{OR}(L, E))$, where ln is the natural logarithm.

To combine different kinds of evidence, Lee *et al.* propose to link the LLS scores calculated for each evidence using a weighted sum [92, 93]. This weighted sum is based on a Bayesian approach as it already had been used by Jansen *et al.* [122], by von Mering *et al.* [55], or by Troyanskaya *et al.* [123]. Bayesian approaches assume independence for the individual data sets to integrate. As this can not be generally assumed for biological data, in contrast to these previous methods, Lee *et al.* propose an heuristic modification of the strict Bayesian approach. The resulting weighted sum (WS) incorporates the relative weighting of the data and captures simple aspects of their relative independence either in an exponential [92]

$$WS_{v_i,v_j} = \sum_{d=1}^{n} \frac{LLS_d}{D^{d-1}},$$

or in a linear manner [93]

$$WS_{v_i,v_j} = LLS_0 + \sum_{d=1}^{n} \frac{LLS_d}{D \cdot d}, \text{ for all } LLS \geq T.$$

LLS_d represents the LLS based on a single data set d, where $d = 1, ..., n$, $D \in [1, \infty)$ represents the degree of dependence between the different data sets, and d is the rank index of the n log likelihood scores for the given gene pair v_i, v_j. D is chosen such that it optimizes the accuracy and coverage on a benchmark, e.g. KEGG. T is used as a threshold to exclude noisy low scoring LLS. The calculated WS_{v_i,v_j} can be used to derive an edge-weighted protein interaction graph $G = (V, E)$ with edge weights $\omega(e_{v_i,v_j}) = WS_{v_i,v_j}$ (see Definition 2.2.2, page 20). The authors could show that both methods perform better compared to the naïve Bayesian approach and

successfully applied them to score probabilistic functional networks for yeast.

2.2.7 Functional Linkage Network

Linghu *et al.* [94, 95] apply machine learning techniques to combine various data sources to construct so-called functional linkage networks. In these edge-weighted networks nearest neighbors are likely to be functionally related. They use such networks either to assign function to unannotated proteins [94] or to prioritize disease genes [95].

2.2.8 Conclusion

Drawbacks of MiMI, PINA and DASMI in our opinion are that all the information that can be downloaded is only listed and not used to describe the associations of pairs of genes or proteins like it is the case for STRING. They are appropriate to obtain first information on a gene but are not so much useful for deeper follow-up analyses where the association between different genes and the related biological processes are of interest. The advantage of DASMI and PINA over MiMI are the availability of confidence scores, which could be used as a basis for further data integration.

iRefIndex provides a sound basis for protein interaction data which I will use as one source of data in this thesis. It does not integrate any additional information, however, this is also out of the scope of the index. To our knowledge, STRING constitutes the most comprehensive set of protein-protein associations available. Thus, we decided to integrate it in our analyses.

The drawback of the described approaches by LLS [92, 93] and STRING [55] to score pairs of proteins are that they combine different scores in a Bayesian fashion, for which independence of the individual scores is a prerequisite. This does often not hold true for most biological data. Although Lee *et al.* [93] try to overcome this hurdle by utilizing their weighted sum approach, it is still an issue when combining evidence which possibly is not independent from each other.

For many machine learning techniques it is rather difficult to judge where the calculated scores originate from. In contrast, I focus on developing a transparent method for scoring. Thereby, it should be possible to judge how sound the score is with respect to biology. Moreover, we always need a gold standard to perform machine learning or to calculate LLS or STRING scores. Our knowledge about the role of genes/proteins is far from complete and still accumulating and evolving, thus the gold standard, a set of true positives and true negatives, is not the *complete* truth. A method independent from such a *preliminary* truth would be beneficial to be as unbiased as possible. Further, it is cumbersome to extend the described scores by personal data as for example derived by wet-lab experiments. None of the proposed methods makes it possible to easily integrate specific knowledge about the biology under investigation; expert biologists would possibly like to give higher weights to certain evidence.

In summary, we are seeking for an easy to use as well as easy to interpret scoring function to describe the pairwise relatedness of proteins. This score should be independent from a gold standard, highly flexible by allowing easy integration of newly gained knowledge and it should offer the possibility to differentially weight individual evidence.

2.3 Existing Methods to Resolve Processes Affected by Gene Expression Changes

Methods that, based on gene expression experiments, could be used to identify condition responsive subnetworks out of various types of molecular networks already exist [48–51, 124, 125]. These methods focus on the detection of subnetworks or modules that are enriched in deregulated genes. Thus, I refer to them as module detection methods. Among the most widely used are for example jActiveModule [48], the method proposed by Cabusora *et al.* [125], GXNA [50] and the methods proposed by Guo *et al.* [49] and Dittrich *et al.* [51].

Definition 2.3.1. *Subnetwork/Module.*

Let $G = (V, E)$ be an undirected graph. A graph $G' = (V', E')$ is a subgraph of G if and only if $V' \subseteq V$, $E' \subseteq E$ and for all $e_{v_i,v_j} \subseteq E' \Rightarrow$

$\{v_i, v_j\} \subseteq V'$.

A subgraph does not need to have all possible edges present in E. If a subgraph has every possible edge, it is referred to as an induced subgraph:

Definition 2.3.2. *Induced subgraph/module.*

Let $G = (V, E)$ be a graph and let $V' \subseteq V$ be a subset of vertices of G. The subgraph of G induced by V' is the subgraph $G' = (V', E')$ such that for all $\{v_i, v_j\} \subseteq V'$, $e_{v_i,v_j} \in E \Leftrightarrow e_{v_i,v_j} \in E'$. That is, G' contains all the edges of G that connect elements of $V' \subseteq V$.

Definition 2.3.3. *Node-weighted graph.*

A node-weighted graph is a pair (G, ω), where $G = (V, E)$ is an undirected graph and $\omega : V \to \mathbb{R}$ is a weighting function assigning a weight $\omega(v)$ to every node $v \in V$.

Usually the module identification tasks are formulated as optimization problems. The objective function is based on a subnetwork or module score evaluating the significance of differential expression [48,50,51,125,126] and/or co-expression [49,127]. Based on an edge- and/or node-weighted graph (Definitions 2.2.2, page 20, and 2.3.3, page 28, respectively), subnetworks optimizing the objective function are usually determined by heuristic searches or exact solutions using integer linear programming.

In this section, I briefly describe the existing approaches and summarize their pros and cons thereby concluding why further development is necessary.

2.3.1 jActiveModule

jActiveModule was proposed by Ideker *et al.* [48]. It constitutes one of the most widely used and accepted module identification methods. In this approach, nodes, i.e. proteins, of a protein-protein and protein-DNA interaction network are weighted according to p-values derived by differential expression analysis. To get an additive weight, the inverse of the normal cumulative distribution function (CDF) Φ is used to transform the p-values p_i for node v_i: $z_i = \Phi^{-1}(1 - p_i)$. Thus, lower p-values correspond to higher z_i. Based on these z_i Ideker *et al.* propose to optimize modules $G' = (V', E')$ with respect to $z_{G'} = \frac{1}{\sqrt{k}} \sum_{v_i \in V'} z_i$, where $|V'| = k$. To determine

whether the score of the network is higher than what would be expected randomly, a z-score transformation is used. Gene sets of size k are randomly sampled and $z_{G'}$ is computed for each of the sampled gene sets to estimate the mean μ_k and the standard deviation σ_k for random G's. Then the transformed score

$$s_{G'} = \frac{z_{G'} - \mu_k}{\sigma_k} \tag{2.1}$$

is $N(0,1)$ distributed and the values calculated for different modules G' are comparable.

Input: A node-weighted graph $G = (V, E)$, a number n of iterations, and a temperature function T_j which decreases with increasing j.

Output: An working subgraph $G_{\mathtt{W}}$ of G and its highest-scoring component $G' = (V', E')$.

01 randomly set every node $v \in V$ as active with probability 0.5
02 **for** $j = 1 \ldots n$ **do**
03 randomly pick a node $v \in V$ and invert its state
04 identify the highest scoring component $s_{G',j}$ of $G_{\mathtt{W}}$
05 **if** $s_{G',j} > s_{G',j-1}$ **do**
06 keep the state of v
07 **else**
08 keep the state of v with probability $p = e^{(s_{G',j} - s_{G',j-1})/T_j}$
09 **done**
10 **done**
11 **return** $G_{\mathtt{W}}$ and its highest-scoring component G'.

Figure 2.3: **Algorithm proposed by Ideker *et al.*** Brief description of the algorithm as implemented in the Cytoscape plugin jActiveModule. Throughout the algorithm, an 'active/inactive' state is associated with each node. $G_{\mathtt{W}}$ denotes the *working* subgraph induced by the active nodes. At each iteration j, $s_{G',j}$ denotes the score $s_{G'}$ (Eq. 2.1, page 29) for the highest-scoring component of $G_{\mathtt{W}}$. T_j denotes the temperature that is decreasing with increasing j.

The authors provide an NP-hardness proof for a simplified variant of their central search problem, namely for MAXIMUM WEIGHT CONNECTED SUBGRAPH. Since NP-hard problems are computationally expensive, they propose a simulated annealing algorithm to optimize their score. Simulated annealing (SA) refers to the way in which a metal cools and freezes into a minimum energy structure during the annealing process. In analogy to the annealing process, SA is a heuristic technique for trying to find the global optimum of an objective function that may possess several local optima. Probabilistically accepting worse solutions allows the algorithm to "explore" more of the possible space of solutions, thereby escaping local optima. $p = e^{-\Delta/T}$ is the probability with which worse solutions are accepted. T, by analogy with the original application known as the system temperature, is a control parameter which decreases over time, i.e. wi each iteration, and Δ is calculated as the change of the objective function. SA is used in the implementation of jActiveModule as briefly summarized in Figure 2.3.

2.3.2 Rajagopalan and Agarwal

Rajagopalan and Agarwal propose an improvement over the jActiveModule algorithm (see Section 2.3.1) [124]. They argue that about half the nodes in the network have positive scores z_i. This leads to the generation of arbitrarily large modules. By introducing a parameter β to reduce the number of nodes with positive z_i, smaller modules are generated in general. For highly connected networks, however, extracted modules stay large. According to the authors, the higher the degree of a node, the more likely one of the neighbouring nodes has a high positive z_i which in turn leads to larger modules. Thus, based on the degree of a node, the authors introduce an *edge penalty* to further reduce the scores for highly connected nodes. Thereby, they finally achieve smaller subnetworks.

In brief, Rajagopalan and Agarwal propose the following heuristic to extract subnetworks:

1. Group nodes with positive score into subnetworks using breadth-first search.
2. Merge subnetworks via non-positive nodes if this produces a higher score.
3. Apply a final pruning step. This step checks whether removal of nodes with a small positive individual score increases the overall score.

2.3.3 Cabusora *et al.*

Cabusora *et al.* [125] base their analysis on a network derived from protein interactions, metabolic reactions and co-expressed genes. As an example they use information available for *Mycobacterium tuberculosis*. Gene expression data for drug treatment and respective controls are used to score the network. In contrast to Ideker *et al.* [48], Cabusora *et al.* calculate scores for the edges and not for the nodes.

Let v_i and v_j be two gene products in the network that are connected by an edge e_{v_i,v_j}. Further, let p_i be the p-value derived for v_i or let P_i be the vector of p-values for gene v_i in case multiple p-values are available. The weight $z_{i,j}$ of edge e_{v_i,v_j} is then either calculated as the product probability $p_{i,j} = p_i \cdot p_j$ or as the empirical correlation $p_{i,j} = cor(P_i, P_j)$ in case multiple p-values are available. Following Ideker *et al.*, Cabusora *et al.* transform these values into an additive weight $z_{i,j} = \Phi^{-1}(1-p_{i,j})$. The objective function which is to be optimized for a subnetwork $G = (V', E')$ states:

$$z_{G'} = \frac{1}{\sqrt{m}} \sum_{e_{v_i,v_j} \in E'} z_{i,j} \text{ , where } m = |E'|.$$

As already proposed by previous approaches, this score is normalized towards randomly sampled subnetworks of size n: $s_{G'} = \frac{z_{G'} - \mu_n}{\sigma_n}$.

The general idea of the algorithm is to calculate the k shortest paths between selected seed nodes. Briefly, it works as follows:

1. Integers k and l as well as seed nodes (in their exemplary study, Cabusora *et al.* use the most significantly deregulated genes) are chosen as input to the algorithm.

2. Shortest, second-shortest, third-shortest,... up to the k^{th}-shortest path between all pairs of seed nodes with restricted maximal path length l are calculated.

3. G' is composed of all edges and nodes that are on the paths calculated in 2.. Based on subnetwork G', $z_{G'}$ is calculated as previously defined.

This approach guarantees a best scored subnetwork for a set of seed nodes. To obtain an optimized set of seed nodes for a better scoring sub-network Cabusora *et al.* propose a heuristic. They reduce the problem by finding best scoring pathways between pairs of seed nodes selected based on a rank-weighted random distribution. Again, the shortest pathway between these nodes is identified, subnetwork scores are calculated and the node pair recorded. After convergence to highest scoring pathways, a sub-network G' and its corresponding score $s_{G'}$ is computed using nodes with the highest pathway score.

2.3.4 ToPNet

Hanisch and Sohler [128,129] developed a framework called ToPNet. It offers several possibilities to analyze biological networks, e.g. navigating the network based on the neighborhood of a gene (iterative exploration). I will briefly describe the application of Hanisch's and Sohler's pathway queries and significant area search, since these are related to the methods I am focusing on.

Significant area search

A significant area search is performed in the following way:

1. Select seed nodes, e.g. based on a gene expression experiment.
2. Greedily expand around this seed node by including the most significant neighbors.
3. Determine the significance of the selected gene sets using Fisher's inverse χ^2 method [130].

Pathway queries

These kinds of queries are XML-based and allow for complex and at the same time specific queries which could also take into account gene expression data, gene annotation like GO, information on transcriptional regulation, and so forth. Following example is given by Sohler *et al.* [128]: "We look for kinases that are directly connected to both Fus3 and Kss1. Fus3 and Kss1 must be connected via at most one additional protein to a transcription factor that regulates genes that are differentially expressed in knockout experiments." In their publication they also display an XML template which is used to perform a similar query.

2.3.5 Graph Based Iterative Groups Analysis (GiGA)

Breitling *et al.* [131] include GO annotation as evidence into their analysis. They do this by introducing "evidence graphs", bipartite graphs which contain two kinds of nodes, one for genes and one for the associated "evidence". Such a graph is converted into a simple graph by introducing edges between all pairs of genes that share the same evidence node and subsequently removing all evidence nodes. Additional to this graph, a list of genes ranked according to absolute fold change is used to label the genes in the graph with their corresponding ranks in that list. All genes that are not contained in the list are removed from the graph. In short, the algorithm proposed by Breitling *et al.* [131] works as follows: First, nodes that are of lower rank (have higher fold change) than all their neighbours are selected as seed nodes. These seed nodes are iteratively extended to include their most significant neighbours, and p-values based on the cumulative hypergeometric distribution are calculated for the extended modules in each iteration. The extension is stopped either after all nodes reachable from the actual module are included or after a maximum size is reached. For each modules generated in this way, that module is selected as "regulated neighbourhood" that yielded the smallest p-value.

2.3.6 MATISSE and CEZANNE

In 2007 Ulitsky *et al.* proposed a method for **M**odule **A**nalysis via **T**opology of **I**N**t**eractions and **S**imilarity **SE**ts (MATISSE) [132]. The method can be based on any interaction network $G = (V, E)$. In their publication they show results based on a protein-protein and protein-DNA interaction network for *Saccharomyces cerevisiae* as well as on a protein-protein interaction network for human. Additional to the network G, the algorithm takes a symmetric similarity matrix S as input. The entries s_{ij} of S describe the similarity between genes i and j. This similarity can, for example, be calculated as the Pearson correlation between the respective gene expression patterns. S is further used to calculate an edge weight. Ulitsky *et al.* define two co-expressed genes as mates. Each edge in the network is weighted using the log-likelihood score of the adjacent genes being mates compared to not being mates. By this means, they define the *similarity* graph $G^S = (V_{sim}, E^S)$ with $V_{sim} \subseteq V$ and $E^S = (V_{sim} \times V_{sim})$. Based on this input, the algorithm tries to detect **j**ointly **a**ctive **c**onnected **s**ubnetworks (JACS). JACS are defined as disjoint

sets $U_1, U_2, ..., U_m$ that induce connected subgraphs in G and heavy subgraphs in G^S. Since this problem is NP-hard, the authors propose several heuristics which are all split into three phases:

1. Detection of small, high-scoring gene sets as *seeds*.

2. Improvement of the seed.
   For optimizing the seeds identified using one of the previous heuristics a greedy approach is used. All seeds are optimized simultaneously either by addition of an unassigned node to an existing JACS, removal of a node from a JACS, exchange of a node between JACSs, or merge of two JACSs. The algorithm keeps those moves that improve the overall score of the solution and that maintain the connectivity of the JACSs. If no such move exists, nodes $\in V \setminus V_{sim}$ are removed as long as they do not disconnect any of the JACS. As soon as no nodes can be removed, the algorithm terminates.

3. Filtering based on significance.
   Empirical similarity scores as well as empirical p-values are calculated for JACS by randomly sampling gene groups of the same size. Based on these p-values and the average similarity scores, the JACS are filtered to obtain the most relevant ones.

In 2009 the same authors proposed an improvement over MATISSE [133]. The difference between **C**o-**E**xpression **Z**one **AN**alysis using **NE**tworks, CEZANNE, and MATISSE is that for CEZANNE every edge $e \in E$ is weighted by the probability $p(e) \in [0; 1]$ that the edge, i.e. the interaction, exists. The authors applied their method to a network and confidence values for *S. cerevisiae* which were based on purification enrichment (PE) scores [134]; this score measures the likelihood of observed experimental results given the hypothesis that an interaction is genuine relative to the likelihood of the same results if the interaction is not real. The overall aim is to detect modules that are q-connected and have a maximum co-expression score. A set of vertices $U \subseteq V$ is called q-connected if for all $U' \subset U$ the probability that at least one edge connects U' with $U \setminus U'$ is $\geq q$. The problem is solved in three steps similar to the MATISSE algorithm with the difference that during the optimization step the q-connectivity has to be maintained. The authors showed that q-connectivity is fulfilled as long as the weight of every minimum cut of the module

under consideration exceeds $T = -log(1-q)$. The significance of the optimized modules is again determined based on empirically derived p-values.

2.3.7 GXNA

Gene eXpression Network Analysis (GXNA) is a method for module extraction that has been proposed by Nacu *et al.* [50]. The authors make use of gene interaction data available from EntrezGene [135] and KEGG [114] and combine it with expression data. They propose several methods to score sets of genes V', i.e. subnetworks $G' = (V', E')$ of the gene interaction network. The scores are based on the t-statistic T_{g_i} which is used to calculate the differential expression for gene g_i.

1. Averaging the test statistics:

   (a)
   $$f_1(G') = \frac{1}{k} \sum_{i=1}^{k} T_{g_i}, \text{ where } k = |V'|.$$

   (b) To take into account modules that contain both, up- and down-regulated genes the absolute value of the t-statistic can be applied
   $$f_2(G') = \frac{1}{k} \sum_{i=1}^{k} |T_{g_i}|, \text{ where } k = |V'|.$$

2. Averaging gene expression:

   (a) Calculate group expression
   $$S_j(G') = \sum_{i=1}^{k} X_{g_{ij}}$$
   where $X_{g_{ij}}$ refers to the expression level of gene i in sample j.

   (b) To take into account modules that contain both, up- and downregulated genes, the authors propose to include signs in the group expression formula
   $$S_j(G') = \sum_{i=1}^{k} \epsilon_i \cdot X_{g_{ij}}$$
   where $\epsilon_i = -1$ if the t-statistic for gene i is negative.

Based on these values, the t-statistic is calculated for G':

$$f_3(G') = (\mu_1 - \mu_0)/\sqrt{\sigma_1^2/n_1 + \sigma_0^2/n_0},$$

where μ_0 and σ_0 are the mean and the standard deviation of $S_j(G')$ for samples of the control group, respectively. Analogously, μ_1 and σ_1 are the mean and the standard deviation of the treatment group.

To extract modules, Nacu *et al.* propose two basic ideas, *the ball* $B(x, r)$ and an *adaptive subgraph search*. Both ideas are based on seed nodes that can be arbitrarily selected. In their publication, Nacu *et al.* sequentially select each gene as seed node.

The ball: Compute score $f(G')$ for $G' = B(v, r)$ where $B(v, r)$ describes the ball centered at node v with radius r. That is, the subgraph G' is composed of all nodes that are connected to v by a path of length $\leq r$ and the edges induced by these nodes.

Adaptive subgraph search: The basic idea is to start with a seed node v and gradually expand around it. The authors propose two different approaches for expansion:

1. Randomly pick a vertex with higher probability assigned to vertices that yield higher scores.
2. Pick a vertex such that the present subgraph has maximal score $f(G')$ (greedy search).

The algorithm stops either after G' reaches a predefined size or when $f(G')$ can no longer be increased.

For nodes of a high degree, $B(v, r)$ is not very selective and rather unspecific. This problem is addressed by adaptive search algorithms. The adaptive search algorithm as described in the first approach (1.) is slower, and since greedy search described in the second approach (2.) works reasonably well, the authors propose to use the latter method for speed and simplicity.

2.3.8 Guo *et al.*

In the approach proposed by Guo *et al.* [49], protein interactions from HPRD [101] and DIP [100] as well as gene expression data have been used. Proteins without

gene expression data are deleted. In this method gene expression data is used to weight the edge $e_{(x,y)}$ between directly connected proteins x and y by the covariance of the respective expression vectors X and Y: $\omega(e_{(x,y)}) = cov(X,Y) = cor(X,Y) \cdot \sigma(X) \cdot \sigma(Y)$, where $cor(X,Y)$ is the Pearson correlation of X and Y and $\sigma(X)$ and $\sigma(Y)$ their standard deviations.

Based on the edge-weighted network, Guo *et al.* use the following score for subnetworks: $T(G') = \sum_{e \in E'} \omega(e)$. After standardizing this score using a z-score transformation based on 10,000 randomly sampled edge sets E'_{rand} with $|E'_{rand}| = k$, the objective function is given by $S(G') = \frac{T(G') - \mu_k}{\sigma_k}$. Here, μ_k is the mean of the 10,000 values for $T(G'_{rand})$ and σ_k their standard deviation. Following Ideker *et al.*, they solve the optimization problem using simulated annealing. Default output is the best module contained in the optimized solution.

2.3.9 Dittrich *et al.*

Dittrich *et al.* [51] solve the module extraction problem by integer linear programming. They transform the underlying optimization problem already stated by Ideker *et al.* as the Maximum-Weight Connected Subgraph Problem (MWCS) to the well studied Price Collecting Steiner Tree Problem (PCST). Ljubić *et al.* [136] proposed an integer linear programming algorithm which is currently the fastest to solve PCST. This algorithm is applied by Dittrich *et al.* to solve the MWCS to optimality.

The input to their algorithm are protein interactions as provided by HPRD [101]. The proteins, i.e. the nodes present in the respective network are weighted by aggregated p-values derived from gene expression data. To show the applicability of their method, Dittrich *et al.* used a gene expression data set from diffuse large B-cell lymphoma [137] together with survival information from the respective patients. They aggregate p-values derived from differential expression analysis comparing two tumor subtypes and p-values derived from analysis of survival times using Cox regression [138] used to analyze the survival data. For combining the p-values, Dittrich *et al.* developed an additive score where positive values represent interesting genes and negative values denote background noise.

2.3.10 ClustEx

ClustEx proposed by Gu *et al.* [139] makes use of protein-protein interaction data taken from HPRD [101]. Additionally, they utilize gene expression data. Only genes present on the microarray are used in the analysis. The overall aim is to detect groups of differentially as well as co-expressed genes that are closely connected.

To detect the responsive gene modules, first, differentially expressed genes are identified. In a second step, the correlation of expression levels X and Y for genes x and y are used to weight the edges present in the protein interaction network:

$$\omega(e_{x,y}) = |cor(X,Y)| \,.$$

Based on this weight, the distance between two interacting proteins x, y is calculated as

$$dist(x,y) = 1 - \omega(e_{x,y}) = 1 - |cor(X,Y)| \,.$$

Based on $dist(x,y)$, shortest paths are calculated between pairs of differentially expressed genes. The differentially expressed genes are then clustered based on the lengths of the shortest paths. Based on this clustering, genes are separated into gene groups. To connect the differentially expressed genes, all genes on the shortest paths between the differentially expressed genes were added such that a connected sub-network is generated. Then the subnetwork was extended by one step, i.e. edge, in the whole gene network. To finally obtain the responsive gene-module, the genes contained in each of these groups need to be connected to a subnetwork. The authors do that by using all genes and edges on the 10 shortest paths between all the pairs of the differentially expressed genes in the extended sub-network.

2.3.11 Pandora

Although they only look at information available for yeast, Zhang *et al.* [140] are the only ones who go beyond the use of interaction networks. Similar to previous interaction networks, they make use of protein protein interactions but also of genetic interactions, information on domain-domain interactions and GO annotations [113]. Since in contrast to the previously mentioned methods they want to predict pathways and not functional modules, they want to limit the genetic interactions that occur in a pathway. Thus, genetic interactions are assigned a weight of 0. All other

kind of interactions are assigned a score of 1. By investigating the GO annotations for genes applying a method proposed by Wang *et al.* [141] (explained in more detail in Section 3.4.2, pages 64 ff.), they calculate a similarity score for pairs of gene products. Thus, for each pair of proteins, Zhang *et al.* derive scores based on different kinds of interactions as well as the GO-similarity score. Taking the average of the scores for each pair of proteins results in the final weight for the edge connecting these two proteins in the interaction network. A cutoff is applied to this score and only edges passing that cutoff are used to predict biological pathways in yeast.

The prediction is based on network topology. Topological similarity is calculated using the Jaccard coefficient [142] for neighboring proteins. Similar proteins are summarized to pathways.

2.3.12 Conclusion

Table 2.1 gives an overview of different characteristics for the methods briefly reviewed in this section.

Apart from ToPNet and Pandora, all methods described in this section focus on the detection of gene modules that are enriched in deregulated genes. Since numerous processes are regulated by, for example, post-transcriptional modifications one objective of this thesis is to expand the analysis beyond the transcriptional level. Most important when analyzing mode of action is the biological interplay of proteins. Therefore, we also want to incorporate further evidence with respect to common biological mechanisms.

ToPNet [128, 129], as an example, allows to query for GO annotations of the molecules contained in the network. Based on these queries, subnets are detected. ToPNet offers the possibility to perform specific queries that make it possible to integrate previous knowledge. To be able to formulate such queries, knowledge about what is interesting and possibly even some knowledge on the biological process of interest is needed. Thus, *de novo* analyses of mode of action and especially identification of off-target effects is probably difficult to perform using ToPNet. Further, only very specific questions are answered and it seems impossible to answer a question like “what are the off-target effects of the compound under investigation?” Apart

Table 2.1: **Comparison of module detection approaches.** Displayed are different characteristics of the individual module detection methods described in this section. Methods are characterized with respect to whether detected modules are optimized towards being enriched in deregulated genes. As indicated by the "additional data"-column, most of the methods only integrate expression data. Further, hardly any method offers the possible to add additional knowledge ("possibility to add data"-column). If it is possible to add data, this data integration is possible for information that can be reflected as p-values or significant values.

Module detection method	**module enriched in deregulated genes**	***denovo* analysis**	**software available**	**module size**	**expression data**	**additional data**	**possibility to add data**
jActiveModule (Sec. 2.3.1)	yes	yes	yes	large	yes	p-value-based	p-value-based
Rajagopalan *et al.* (Sec. 2.3.2)	yes	yes	on request	medium	yes	yes	no
Cabusora *et al.* (Sec. 2.3.3)	yes	yes	no	NA	yes	no	no
ToPNet (Sec. 2.3.4)	no	no	no	NA	no	yes	no
GiGA (Sec. 2.3.5)	yes	yes	yes	small	yes	yes	no
MATISSE/CEZANNE (Sec. 2.3.6)	no	yes	yes[#]	small	yes	no/yes	no
GXNA (Sec. 2.3.7)	yes	yes	yes	user defined	yes	no	no
Guo *et al.* (Sec. 2.3.8)	yes	yes	no	NA	yes	no	no
Dittrich *et al.* (Sec. 2.3.9)	yes	yes	yes[+]	large	yes	p-value-based	p-value-based
ClustEx (Sec. 2.3.10)	yes	yes	yes	medium	yes	no	no
Pandora (Sec. 2.3.11)	no	yes	yes	large	no	yes	no

[#] For academic use only.

[+] Depends on commercial software.

from that, ToPNet seems to be no longer available.

Some methods like those described in Sections 2.3.3, pages 31 f. and 2.3.9 try to derive only one huge module to explain the biological process reflected by the expression data. Further, Gu *et al.* state "As observed in previous studies and in our analysis, a big module usually dominates the responsive process [48, 49]" [139]. We argue that mode of action is induced by several processes that in general influence each other. Even if an analysis results in multiple networks, it is easily possible that more than one process, i.e. the on- and several off-target effects, are covered by one larger module. Thus, the individual effects are difficult to detect and interpret. By deriving small modules, it is more likely to be able to separate different effects represented in the biological system under investigation. Otherwise, it is impossible to understand the biological processes taking place and how they relate to each other.

Most of the existing approaches focus on protein/gene interaction in combination with expression data. Approaches that go beyond this and integrate further data rarely exist. Pandora integrates different kinds of data but at the same time does not make use of expression values. This is due to the fact that the purpose of the method is to detect complete pathways or networks like for example present in KEGG, but not to derive small modules explaining processes present in a gene expression experiment. GiGA does integrate expression data, protein interactions and GO annotations. But still the integrated GO information is *only* translated to edges in a network for which the methods still tries to derive modules solely based on deregulated genes.

The major drawback of the present methods is that they focus on differentially expressed genes. They neglect the vast amount of biological data that could be used to support the analyses especially with respect to the fact that not all genes are regulated on the transcriptional level. Thus, I aim at developing a method that constitutes a combination of approaches like jActiveModule, making use of protein interaction data and gene expression data, and Pandora, that integrates additional data sources. By doing so, it will be possible to extract small modules that help revealing the effects present in the expression experiment. This method will exemplarily be applied to a gene expression experiment conducted for the analyses

of compounds' mode of action.

2.4 Analyzing Groups of Genes

Given the outcome of a biological experiment like gene expression measurements, we are interested in the biological processes that are related to the experiment. One of the most prominent approaches to interpret such data is to make use of predefined gene groups. Such gene groups provide more biological knowledge compared to looking at individual genes, thereby offering the possibility of a more meaningful interpretation in the biological context. Groups of genes can be defined based on different sources, e.g., based on pathways as defined by KEGG, Reactome, or BioCarta. A second source are Gene Ontology terms. Here, groups of genes can be determined by genes annotated with the respective term or any of its children in the hierarchy (explained in more detail on pages 44 ff.). Further possibilities are, e.g. groups based on the chromosomal region in which the genes are located, genes associated with a special disease, or groups defined based on literature relevant to the biological research question under investigation.

By identifying predefined groups of genes affected by the gene expression experiment it is possible to relate the underlying experiment to a biological process like the signaling cascades of a potential off-target. In Section 2.4.1 I focus on over-representation analyses conducted using Fisher's exact test [130], GSEA [40,41], and topGO [43]. Further, I provide a brief introduction into more holistic approaches that do not separate the analysis of differential expression from gene set analysis in Section 2.4.2.

2.4.1 Gene Set Enrichment

The basic idea of gene set enrichment analyses is to first define a measure for interesting genes. This could be for example fold changes and/or p-values for genes analyzed in an expression experiment. In a second step, these interesting genes are compared to predefined groups of genes related to a certain biological process. If interesting genes are over-represented in a special predefined group, the related process will very likely be relevant for the underlying expression experiment.

One of the simplest tests to investigate over-representation is Fisher's exact test [143], which has been widely used with respect to gene expression analyses [43, 144, 145]. Fisher's exact test is a count-based method as it solely is based on a count of genes meeting a specific criterion. Given a set of K genes identified as interesting, for example genes exceeding a certain threshold of a gene-associated score; Fisher's exact test calculates the significance of the overlap between a predefined group of genes of size M and the K interesting genes with respect to the total number of genes N. A more detailed explanation is given in Section 3.6.2 (page 75.

Count based approaches require that a set of genes is selected by some definite criterion (hard thresholding). Thus, any information on the genes outside of this set is not used. In contrast, methods utilizing all gene scores or gene ranks derived based on an experiment exist. One of the most prominent ones possibly is GSEA, the Gene Set Enrichment Analysis as first applied by Mootha *et al.* [40] and described in detail by Subramanian *et al.* [41]. GSEA can be divided into two steps. First, gene-wise measures like differential expression are calculated for all N measured genes and respective genes g_i are ranked accordingly to form $L = (g_1, g_2, ..., g_N)$. In their publication, Subramanian *et al.* [41] rank genes according to their correlation r_j to a profile of interest. Second, labels are assigned to genes g_j indicating whether they belong to a gene group of interest S or not, i.e. $g_j \in S$ or $g_j \in \bar{S} = N \setminus S$, respectively. Walking down the ranked list L, two running enrichment scores (ES) are calculated:

$$ES_S(l) = \sum_{\substack{g_j \in S \\ j \leq l}} \frac{|r_j|^p}{N_S} \text{ , where } N_S = \sum_{g_j \in S} |r_j|^p$$

$$ES_{\bar{S}}(l) = \sum_{\substack{g_j \in \bar{S} \\ j \leq l}} \frac{1}{N - m} \text{ , where } m = |S|$$

The authors propose to set $p = 1$. If one wants to penalize sets of genes S for lack of coherence, $p < 1$ could be appropriate. The final ES is defined as the maximum deviation of $ES_S(l) - ES_{\bar{S}}(l)$ from 0, i.e., $ES = max\,|ES_S(l) - ES_{\bar{S}}(l)|$. If the maximum value for ES is higher than expected randomly, the group is enriched with interesting genes. The significance is calculated based on a phenotype-based permutation test. This is essentially a test for deviation from a uniform distribution. Drawback of permutation tests is their low power.

topGO is a Bioconductor package implemented by Alexa *et al.* [43]. They propose a gene set enrichment framework especially applicable to ontology structures like the Gene Ontology (GO) [113]. Figure 2.4 displays a small part of the GO hierarchy which is explained in the following.

Definition 2.4.1. *Directed acyclic graph (DAG).*

A directed acyclic graph (DAG) is a graph $G = (V, E)$ with directed edges $e_{u \to v} = (u, v)$ that contains no path that starts and ends at the same vertex. The root nodes $R \in V$ of a DAG are defined as the nodes where no edge starts, $R = \{r \in V | \nexists\, v \in V : e_{r \to v}\}$. The leave nodes $L \in V$ of a DAG are defined as all nodes where no edge ends, $L = \{l \in V | \nexists\, v \in V : e_{v \to l}\}$.

Definition 2.4.2. *Ontology.*

An ontology is a set of defined terms or vocabularies that are given hierarchical relationships to one another. It can be represented as a directed acyclic graph (DAG). GO, for instance, provides a set of terms to describe the properties of proteins.

In the case of GO, the terms are used to describe and annotate proteins with respect to their molecular function (MF), the biological processes (BP) they are involved in, and the cellular component (CC) they occur in. Each of these three classes, MF, BP and CC, builds a separate DAG exactly containing one root node r. Genes associated with special attributes are assigned to the respective ontology term which in turn is represented by a vertex/node in the DAG. The ontologies resemble a hierarchy, ancestor terms (Definition 2.4.3) are less specialized than their descendants (Definition 2.4.4). With respect to GO, whenever we refer to a *term* of the ontology, we refer to a *node/vertex* in the respective DAG. In an ontology, each gene/protein which is associated with a term is also mapped to all its ancestor terms.

Definition 2.4.3. *Ancestors of node v.*

For a DAG $G = (V, E)$, the ancestor set $V_{ancestor}(v) \subseteq V$ of any node v consists of all nodes v_i that are reachable from v via a directed path $P = (e_{v \to v_1}, e_{v_1 \to v_2}, ..., e_{v_{i-1} \to v_i})$. That is, all nodes on any path to the root node of the DAG are ancestors of v.

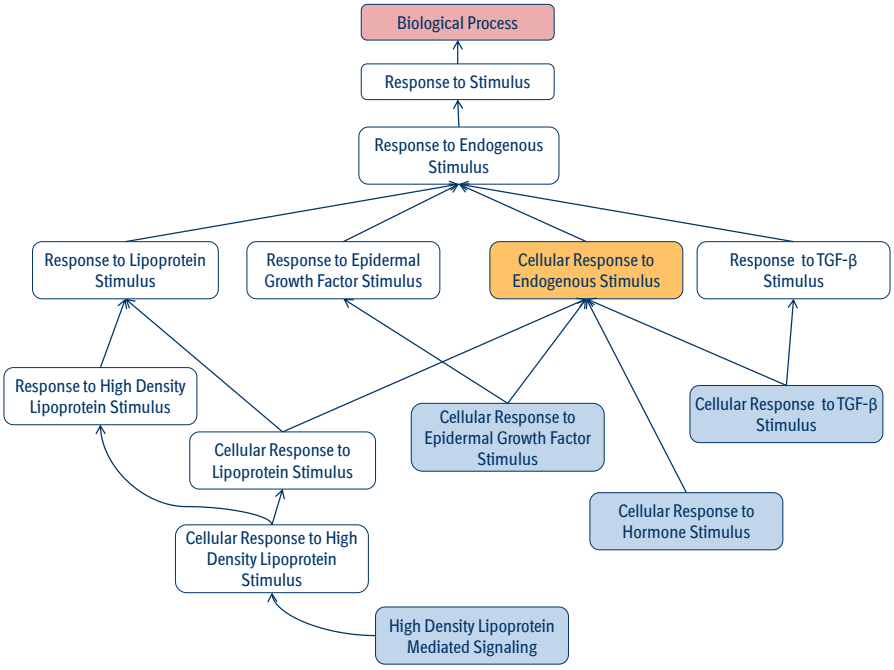


Figure 2.4: **DAG representing parts of the Gene Ontology.** Displayed is a DAG (Definition 2.4.1, page 44) representing a small part of the Biological Process hierarchy of the Gene Ontology. The node highlighted in red indicates the root node (Definition 2.4.1, page 44), blue nodes are the leaf nodes (Definition 2.4.1, page 44), and the node highlighted in orange is the lowest common ancestor (Definition 2.4.7, page 46) of, e.g. the nodes "High Density Lipoprotein Mediated Signaling" and "Cellular Response to Hormone Stimulus", two of the leaf nodes. At the same time, it is a parent node (Definition 2.4.6, page 46) of the leaf nodes "Cellular Response to Epidermal Growth Factor Stimulus", "Cellular Response to Hormone Stimulus", and "Cellular Response to TGF-β Stimulus" as well as of the node "Cellular Response to Lipoprotein Stimulus".

Analogous to an ancestor, we define the term descendant:

Definition 2.4.4. *Descendants of node v.*

> *For a DAG $G = (V, E)$ the descendant set $V_{descendant}(v) \subseteq V$ of any node v consists of all nodes v_i that v is an ancestor of. That is, all nodes on any path leading from the leaves to v are descendants of v.*

A more specific set of descendants is referred to as children:

Definition 2.4.5. *Children of node v.*

For a DAG $G = (V, E)$ the set of children $V_{children}(v)$ for a node $v \in V$ consists of all nodes $v_i \in V$ of which v is a direct ancestor. That is, $V_{children}(v) = \{v_i \in V | e_{v_i \to v} \in E\}$.

Analogous, parent term is defined:

Definition 2.4.6. *Parent of node v.*

For a DAG $G = (V, E)$ the set of parents $V_{parent}(v)$ of a node $v \in V$ consists of all nodes $v_i \in V$ of which v is direct descendant. That is, $V_{parent}(v) = \{v_i \in V | e_{v \to v_i} \in E\}$.

In later sections, the term *lowest common ancestor* is used in the context of DAGs representing ontologies:

Definition 2.4.7. *Lowest common ancestor (LCA).*

The lowest common ancestor of two nodes u, v in a DAG $G = (V, E)$ is denoted as $LCA(u, v)$. It is defined as the node with the maximum path length from the root, i.e., the lowest node in G that has both u and v as descendants (u or v are allowed to be a descendant of itself).

Not every gene is necessarily annotated to a leave node in the ontology. Due to the interleaved structure of ontologies, calculation of enrichment is statistically even more sophisticated than for other gene sets. In their proposed methods, Alexa *et al.* consider this special structure to find those terms in an ontology that show an enrichment in significant genes. Besides count-based tests like the Fisher's exact test and tests based on gene scores or ranks of genes like GSEA, they additionally implemented a test that is directly based on the gene expression data. Such tests are sometimes referred to as holistic approaches. In contrast to tests like Fisher's exact test or GSEA which at least require two separate steps, the gene-wise calculation of a score and the test for over-representation, holistic approaches perform the whole analyses in one go.

2.4.2 Holistic approaches

Methods that combine the analysis of differential expression and detection of enriched gene sets in one step are `globaltest` and `GlobalANCOVA`. These methods

are therefore also referred to as holistic approaches. Applying these methods, it is possible to consider small but consistent changes in expression. The methods try to answer the question whether the global expression pattern X of a group of genes significantly relates to some clinical variable of interest Y like disease status or survival taking into account covariates C like time, dose, age or sex.

`globaltest` has been proposed by Goeman *et al.* [42,44]. This test was developed for predicting clinical outcomes Y based on the expression pattern of a group of genes X and covariates C. The hypothesis

$$H_0 : P(Y|X, C) = P(Y|C)$$

is tested to decide whether gene expression X does improve the prediction for the phenotype Y.

`GlobalANCOVA` has been introduced by Mansmann and Meister [45, 146] and further developed by Hummel *et al.* [46, 47]. In contrast to `globaltest`, it tries to uncover the influence of the observed phenotype Y on the gene expression X:

$$H_0 : P(X|Y, C) = P(X|C).$$

2.4.3 Conclusion

Several methods exist that are capable to reveal biological processes related to a gene expression experiment. Except for Fisher's exact test, all methods are based on a measure/score present for each individual gene. The most prominent one probably is GSEA. Holistic approaches directly combine the analysis of expression with the detection of enriched gene sets. In contrast to other methods, they are capable of considering small but consistent changes in expression.

Changes in gene expression are not necessarily the only indicator to infer mode of action and, thus, should be considered in conjunction with further biological criteria. Processes mediated through protein interactions as for example frequently present in signaling cascades are essentially neglected by all enrichment methods. These methods treat networks with potentially complex topology as an unstructured set of genes. Further, they do not take into account feedback by transcriptional regulation.

Such feedback can for example also be represented by protein interactions. Existing knowledge about proteins and their relation to one another is disregarded.

Chapter 3

Methods

3.1 TGF-β Gene Expression Data

All laboratory work described in this section has been conducted by Dr. Patrick Baum in the scope of his PhD thesis [52]. It is described here for completeness, more details can be found in [52]. Primary data analysis has been performed in a cooperative manner. Further, he kindly provided the experimental data to validate the new methods developed in the present work.

3.1.1 Cell Culture and NCE Treatment

HaCaT cells were cultured under standard conditions [147]. Cells were seeded in 24-well plates and grown overnight to a confluence of approximately 70%. Cells were starved for 3 hours in DMEM without addition of fetal calf serum (FCS). Five BI compounds, in the following referred to as BI1 to BI5, and two competitor substances, Ex1 and Ex2, with inhibitory potency towards TGF-βR1 kinase were used as NCE. Details about the synthesis, design, and structure of the molecules can be found in Roth *et al.* [1]. BI1 to BI5 belong to the chemical class of indolinones, Ex1 and Ex2 to the class of pyridopyrimidinones. Cells were pre-incubated with increasing NCE concentrations (0.0032, 0.016, 0.08, 0.4, 2, 10μM) for 15 min and subsequently stimulated with 5 ng/ml of TGFβ1 (R&D Systems) and incubated for 2, 4, or 12 hours. As controls, cells where either left untreated or treated with DMSO (solvent of compounds) and either stimulated with 5ng/ml of TGF-β1 or not stimulated and incubated for 2, 4, or 12 hours. NCE treatments were conducted in triplicates, respective controls in quadruples.

3.1.2 RNA Extraction

RNA isolation was carried out using a MagMAX^{TM}Express-96 Magnetic Particle Processor and the MagMAX^{TM}-96 Total RNA Isolation Kit according to the manufacturer's protocol. Total RNA concentration was quantified by fluorescence measurement using SYBR Green II (Invitrogen) and a Synergy HT reader (BioTek) as previously described [148]. The RNA quality was characterized by the quotient of the 28S to 18S ribosomal RNA electropherogram peak using an Agilent 2100 bioanalyzer and the RNA Nano Chip (Agilent).

3.1.3 BeadChip Hybridization of RNA Samples

Illumina TotalPrep RNA Amplification Kit (Ambion) was used to transcribe 200ng total RNA according to the manufacturer's recommendation. A total of 700ng of cRNA was hybridized at 58°C for 16 hours to Illumina HumanHT-12 v3 Expression BeadChips (Illumina). BeadChips were scanned using an Illumina BeadArray Reader and the BeadScan Software (Illumina). On the HumanHT-12 v3 each gene of the human genome is represented by at least one probe. The raw data is accessible as MIAME compliant entry at Array Express [149] (E-MTAB-265).

3.1.4 qRT-PCR

Quantitative real-time polymerase chain reaction (qRT-PCR) was conducted for eight genes (CDKN1A, CDKN2B, HAND1, JUNB, LINCR, RPTN, SERPINE1, and TSC22D1) known to be deregulated at at least one time point by TGF-β stimulation. mRNA expression levels of the eight genes were determined by qRT-PCR analysis using a 7900HT Fast Real Time PCR System (Applied Biosystems) and the Universal Probe Library System (Roche). Gene specific forward and reverse primer sequences were designed using the Universal Probe Library Assay Design Center (Roche). Total RNA was transcribed into cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) according to the manufacturer's instructions. qRT-PCR was carried out in a final volume of 1μl in three replicates for each cDNA sample. Levels of RNA polymerase II were used for normalization of the data. The $\Delta\Delta$CT method [150] was used to relatively quantify mRNA levels of samples stimulated with TGF-β1 compared to untreated controls.

3.2 Normalization

For gene expression measurements using microarrays, normalization of the data has to be performed in order to minimize systematic effects that are not constant between different samples of an experiment and that are not due to the factors under investigation (e.g. treatment, time). Optimal selection of a normalization method depends on the nature of the experiment. Factors that influence the choice are for example the array technology used (e.g. one/two color, oligo or cDNA probes), the design of the experiment (e.g number of replicates or experimental conditions) and whether accuracy or stability is an issue (bias/variance trade-off). In this regard factors like comparability and quality of single runs play a major role. It has been shown that the normalization method used may influence further downstream analysis to a great extend [39], and thus has to be carefully chosen based on the actual experiment.

In this section, we describe the methods used to select an optimal normalization procedure for the TGF-β gene expression data described in the previous section. Respective results are presented in Section 4.1 and discussed in Section 5.1. This part of my thesis has been published in BMC Genomics [53].

3.2.1 Data Processing

Data has been processed with BeadStudio version 3.0 and the R Language and Environment for Statistical Computing (R) 2.7.0 [151, 152] in combination with Bioconductor 2.2 [153]. The Bioconductor `lumi` package [154] has been used for quality control. Twenty-five combinations of background correction, transformation, and normalization methods were calculated either with methods offered by BeadStudio, with methods available in the `lumi` package, or with a combination of BeadStudio and `lumi` methods. Table 3.1 on page 53 summarizes the individual pre-processing steps for each of the combinations investigated. The following two paragraphs give some details on the availability of the different methods.

BeadStudio Pre-Processing

The normalizations executed by Illumina BeadStudio were applied to the expression values on the original scale. In cases where background adjustment has been

performed, the standard background normalization offered by Bead-Studio (indicated by `bg_*`) which may lead to negative values has been used. `noBg_noNorm` and `bg_noNorm` refer to the raw and to the background corrected data, respectively. Cubic Spline, Rank Invariant, and Average methods were used for normalization and are referred to as `*_cubicSpline`, `*_rankInvariant`, and `*_Average`, respectively. Details about these normalization methods can be found in the BeadStudio Gene Expression Module User Guide [155]. Finally, expression values were log_2-transformed.

R Pre-Processing

To be able to log_2-transform the expression data, negative values which can result from BeadStudio background normalization have to be transformed to positive scale. This is achieved by `forcePositive` (`forcePos`) [154] or `rma` background adjustment [156] available through the `lumiB()` function of the `lumi` package (`bgAdjust.affy`) [154]. In case the `lumiT()` function is used for log_2 transformation, `forcePos` is automatically conducted to transform negative values. `noBg` implies that the background normalization available in BeadStudio has not been applied. For transforming the data, a simple log_2-tranformation (`log`) or variance-stabilizing transformation (`vst`) [157] was used. The latter applies a function that is asymptotically identical to $log_2(x)$, but has been shown to keep variance constant under reasonable error models [157]. Data was normalized using quantile normalization (`quantile`) [158], robust spline normalization (`rsn`) [154], local regression (`loess`) [159], or variance stabilization and normalization (`vsn`) [160]. `vst` as well as `vsn` can handle negative values in the data. Thus, neither `forcePos` nor `rma` was applied as pre-processing for any of those two methods to not unnecessarily modify the values. All methods used are implemented in the R packages `affy` [161], `vsn` [160], or `lumi` [154].

3.2.2 Statistical Measures

In the following, the statistical measures used to select an appropriate normalization method as described in Section 4.1 are briefly summarized. Unless otherwise noted, all statistical calculations were performed using R.

Table 3.1: **Summary of pre-processing steps used for the 25 different normalization procedures.** For background correction BeadStudio's background normalization was applied [162]. This can lead to negative values. To be able to log_2-transform the data, background-correction of `rma` [156] or `forcePos` [154] is used to shift the data to positive scale. Alternatively, data was transformed using (`vst`) [157] which is capable of dealing with negative values. Data was normalized using `quantile`, `loess`, or `rsn` [154]. `vsn` [160] renders transformation of the data unnecessary.

Name	Background Correction	Transformation	Normalization
bg_average	BeadStudio	log_2	average
bg_cubicSpline	BeadStudio	log_2	cubicSpline
bg_forcePos_log_loess	BeadStudio + forcePos	log_2	loess
bg_forcePos_log_quantile	BeadStudio + forcePos	log_2	quantile
bg_forcePos_log_rsn	BeadStudio + forcePos	log_2	rsn
bg_noNorm	BeadStudio + forcePos	log_2	-
bg_rankInvariant	BeadStudio	log_2	rankInvariant
bg_rma_log_loess	BeadStudio+rma	log_2	loess
bg_rma_log_quantile	BeadStudio+rma	log_2	quantile
bg_rma_log_rsn	BeadStudio+rma	log_2	rsn
bg_vsn	BeadStudio	-	vsn
bg_vst_loess	BeadStudio	vst	loess
bg_vst_quantile	BeadStudio	vst	quantile
bg_vst_rsn	BeadStudio	vst	rsn
noBg_average	-	log_2	average
noBg_cubicSpline	-	log_2	cubicSpline
noBg_log_loess	-	log_2	loess
noBg_log_quantile	-	log_2	quantile
noBg_log_rsn	-	log_2	rsn
noBg_noNorm	-	log_2	-
noBg_rankInvariant	-	log_2	rankInvariant
noBg_vsn	-	-	vsn
noBg_vst_loess	-	vst	loess
noBg_vst_quantile	-	vst	quantile
noBg_vst_rsn	-	vst	rsn

Signal-to-Noise Ratios

One aim of normalization is to minimize, for each gene, the within-group variability while maximizing the between-group variability, also referred to as mean sum of squares within (MSQ_{within}):

$$MSQ_{within} = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \tag{3.1}$$

and mean sum of squares between ($MSQ_{between}$):

$$MSQ_{between} = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2, \tag{3.2}$$

respectively. Here, k represents, for a given gene, the number of groups, n_i the size of group i, $\bar{x}_i$ the mean expression level of group i, $\bar{x}$ the total mean, N the total number of observations, and x_{ij} the jth value in group i. The aim is to maximize

$$\frac{MSQ_{between}}{MSQ_{within}} \tag{3.3}$$

which follows an F-statistic with $(k-1; N-k)$ degrees of freedom. This test has been used to derive results described in Section 4.1.1 (Figures 4.1 - 4.4, pages 82 ff.). As a reference, artificial group means $\bar{x} = 6$, $\bar{x} = 6$, and $\bar{x} = 7$, $k = 3$, $n_1 = n_2 = n_3 = 4$, and $\bar{x} = \frac{(\bar{x}_1 + \bar{x}_2 + \bar{x}_3)}{3}$ were used. This results in an $MSQ_{between}$ of 1.33 which is indicated in Figures 4.3 and 4.4 by a grey dashed line. The FDR-corrected [163] p-values for the F-statistic were summarized using their empirical cumulative distribution function (Figure 4.1, page 83).

Pseudo-ROC Curves

One of the main uses of expression arrays is the identification of genes that are differentially expressed under various experimental conditions. A typical identification rule filters genes with p-values and/or fold change exceeding a given threshold. Given a set of known true positives (TP) and false positives (FP), receiver operator characteristic (ROC) curves offer a graphical representation of both specificity and sensitivity for such a detection rule. ROC curves are created by plotting the true positive rate (sensitivity) against false positive rate (1-specificity) obtained at each possible threshold value. Since we are only sure about TPs, we made use of so-called pseudo-ROC curves [164]. 20 genes that are known to be deregulated

by TGF-β were selected as TPs (SERPINE1, SERPINE2, CDKN1A, CDKN2B, SMAD7, VASN, JUNB, BAMBI, CTGF, SOX18, TGM2, MMP10, MMP2, THBS1, TGFBI, IGFBP7, IL1R1, ACTN1, MYC, NOTCH1). True negatives (TNs) were randomly sampled from the set of transcripts remaining when subtracting the TPs from all transcripts. As threshold values, we used FDR-adjusted p-values [163] of an F-statistic (Eq. 3.3) based on the sample groups for untreated and TGF-β stimulated HaCaT cells at 2 hours.

$\log_2$ Ratios, Residual Standard Deviation, and p-Values

log_2 ratios, respective residual standard deviation, and p-values used in Chapter 4 were calculated using linear models in combination with the moderated t-statistic as supplied by the `limma` package [165].

Regression Analysis of Fold Change values and qRT-PCR Measurements

To get an overall impression of the goodness-of-fit of the fold change levels detected using the different normalization methods, an orthogonal regression, i.e. total least squares, was applied. This method is appropriate since both variables, the normalized and the qRT-PCR results, depend on each other; it is not possible to categorize them as dependent and independent variable as it would be necessary for standard linear regression. For the two dimensional case, with the normalized and the qRT-PCR results being the two dimensions, the orthogonal regression can be calculated using the `princomp()` function as available in the basic R environment. This method is applied in Section 4.1.2, pages 96 f..

3.3 Differential Expression Analysis and Gene Set Enrichment

In this section, I describe the methods used in analysis of differential gene expression data.

3.3.1 Differential Expression

Based on the results of Section 4.1 the data has been log_2-transformed and normalized using robust spline normalization (rsn) referred to as `log_rsn`. Details

are described in Section 3.2.1, pages 51f.. Linear models (Bioconductor package `limma`) [165] were used to calculate log_2 ratios, the resulting p-values were FDR-corrected [163].

3.3.2 TGF-β Signature

To define genes deregulated by TGF-β signaling, three sequential filtering steps were applied to the log_2 transformed expression values of each time point separately:

i) The first filtering is based on the comparison of TGF-β-stimulated cells against untreated cells using linear models as described in Section 3.3.1 (FDR-corrected p-value < 0.01 and $|log_2\ ratio| \geq 0.5$).

ii) A linear model was applied to the dose groups of each compound to extract all genes which are significantly deregulated (FDR-corrected p-value < 0.01) by at least one concentration compared to the respective control (cells treated with TGF-β and DMSO but no compound). Here, concentrations are treated as categorical variables.

iii) To detect genes with a dose-dependent deregulation, the likelihood ratio test statistic for monotonicity (R package IsoGene [166]) was used. Treating concentrations as ordinal variables, IsoGene performs an isotonic regression based on the replicates for each concentration resulting in regression values μ_1, μ_2, ..., μ_6 for each gene and each compound treatment. Only genes that are significantly regulated by at least one compound with $|\mu_1 - \mu_6| \geq 1$ and an FDR-corrected p-value < 0.01 for monotonicity were included in further analysis. Additionally, $\mu_1 - \mu_6$ has to be > 0 if TGF-β treatment induced down-regulation, and < 0 if TGF-β treatment induced up-regulation. Thereby, those genes that are most likely deregulated in a dose dependent manner are selected.

For each time point the genes that passed all three filters constitute the final TGF-β signature.

3.3.3 Inferring the Off-Target Signature

In order to detect transcripts that are deregulated due to off-target effects of the compounds, unstimulated cells (WOTGF class) as well as TGF-β-stimulated cells (TGF class) were considered. Since the IC_{50} of all NCEs lies between 0.08μM and 2μM,

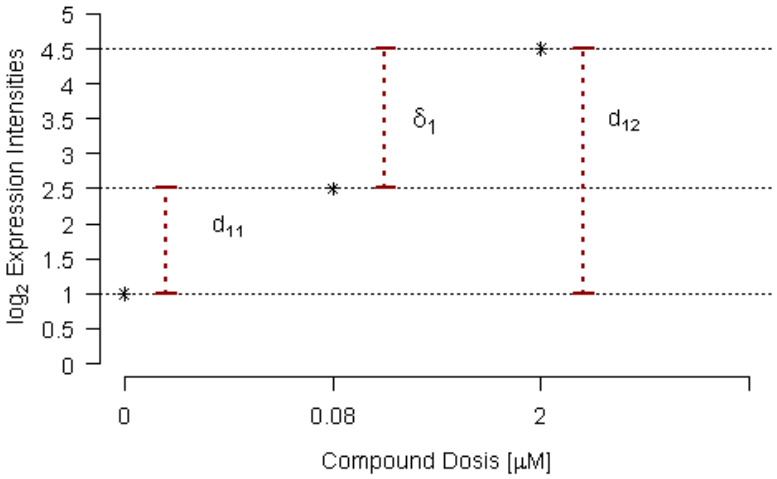


Figure 3.1: **Comparisons used to infer the off-target signatures.** Expression intensities for an arbitrary gene are indicated by $*$ for compound concentrations of $0\mu M$, $0.08\mu M$, and $2\mu M$. Horizontal dotted lines are used to indicate the different expression intensities. The WOFTG class comparisons d_{11}, d_{12}, and δ_1 are calculated based on comparisons of the indicated expression intensities (red dashed lines). Comparisons for the TGF class are calculated accordingly.

these two concentrations were considered for off-target analysis. For the WOTGF class, the NCE-treated samples at 0.08μM as well as at 2μM were compared to untreated cells. Additionally, the 2μM samples were compared to the respective 0.08μM samples. These comparisons are denoted by d_{11}, d_{12}, and δ_1, respectively (Figure 3.1). The same comparisons were made based on the TGF class and are denoted by d_{21}, d_{22} and δ_2, respectively. Significant up- and down-regulation was defined based on FDR-corrected [163] p-value (p.adj) < 0.01 and $|log_2 ratio| \geq 1$, where p-values and $log_2 ratios$ were calculated using linear models [165]. Based on these nomenclatures and values, the following boolean variables are defined:

$$d_{xy,up} = \begin{cases} 1 & \text{, if } log_2 ratio(d_{xy}) \geq 1 \text{ and } p.adj(d_{xy}) < 0.01 \\ 0 & \text{, else} \end{cases},$$

$$d_{xy,down} = \begin{cases} 1 & \text{, if } log_2 ratio(d_{xy}) \leq -1 \text{ and } p.adj(d_{xy}) < 0.01 \\ 0 & \text{, else} \end{cases},$$

$$d_{xy,sig} = \begin{cases} 1 & \text{, if } |log_2 ratio(d_{xy})| \geq 1 \text{ and } p.adj(d_{xy}) < 0.01 \\ 0 & \text{, else} \end{cases},$$

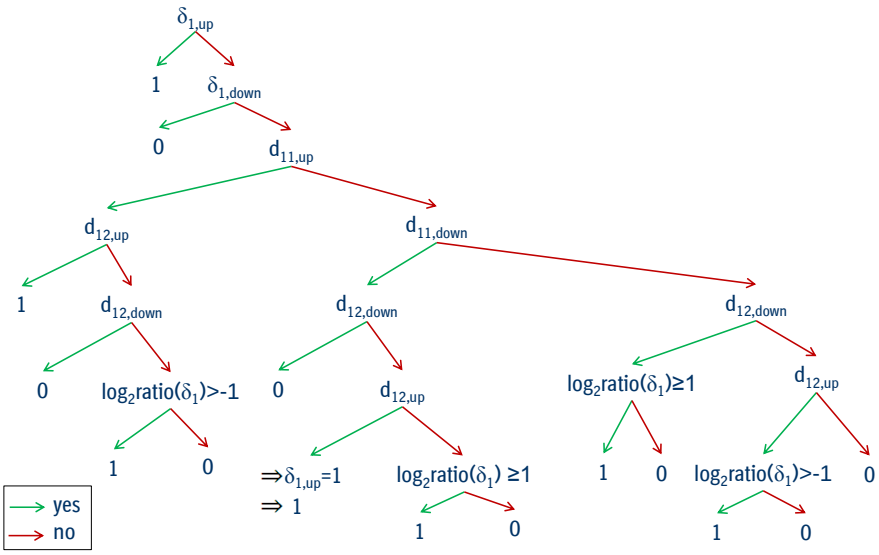


Figure 3.2: **Decision tree to calculate WOTGF$_{up}$.** Displayed is the decision tree to decide whether a gene belongs to the WOTGF$_{up}$-class or not. d_{11}, d_{12}, and δ_1 are calculated as described in Figure 3.1 and Eq. 3.4. A mathematical representation of the decision tree for WOTGF$_{up}$ is given in Eq. 3.5. WOTGF$_{down}$, TGF$_{up}$, and TGF$_{down}$ are calculated accordingly (compare Eq. 3.6, 3.7, and 3.8, respectively).

$$\delta_{x,up} = \begin{cases} 1 & \text{, if } log_2 ratio(\delta_x) \geq 1 \text{ and } p.adj(\delta_x) < 0.01 \\ 0 & \text{, else} \end{cases},$$

$$\delta_{x,down} = \begin{cases} 1 & \text{, if } log_2 ratio(\delta_x) \leq -1 \text{ and } p.adj(\delta_x) < 0.01 \\ 0 & \text{, else} \end{cases}. \tag{3.4}$$

where $x, y \in \{1, 2\}$. $x = 1$ refers to the WOTGF class, $x = 2$ to the TGF class, $y = 1$ refers to the comparison of $0.08\mu M$ to $0\mu M$, and $y = 2$ to the comparison of $2\mu M$ to $0\mu M$. Figure 3.1 indicates the comparisons based on exemplary expression intensities for an arbitrary gene.

Transcripts that are up- or down-regulated by compound treatment (WOTGF$_{up}$ and WOTGF$_{down}$, respectively) or by TGF-β stimulation together with compound treatment (TGF$_{up}$ and TGF$_{down}$, respectively) were detected based on the described comparisons as follows: a transcript belongs to the class WOTGF$_{up}$ if either δ_1 indicates significant up-regulation ($\delta_{1,up} == 1$) or if it does not indicate significant down-regulation ($\bar{\delta}_{1,down} == 1$) but d_{11}, d_{12}, and δ_1 indicate an increasing course of expression intensity for higher compound concentrations. That is, if $\bar{\delta}_{1,down}$ holds

true, five different trends render up-regulation:

1. d_{11} and d_{12} both indicate significant up-regulation ($d_{11,up} == d_{12,up} == 1$);
2. d_{11} indicates significant down-regulation but $log_2ratio(\delta_1) \geq 1$, thereby showing an increasing trend of expression for increasing compound concentrations;
3. d_{11} indicates significant up-regulated and $log_2ratio(\delta_1) > -1$, allowing for a small but not significant decreasing trend for increasing compound concentration;
4. d_{12} indicates significant down-regulated and $log_2ratio(\delta_1) \geq 1$;
5. d_{12} indicates significant up-regulated and $log_2ratio(\delta_1) > -1$.

On the one hand, as soon as one of d_{11} or d_{12} indicates significant up-regulation (cases 3 and 5), a small amount of noise, i.e. a small trend towards down-regulation, was allowed by claiming $log_2ratio(\delta_1) > -1$. On the other hand, as soon as one of d_{11} or d_{12} indicate down-regulation (cases 2 and 4), we are more strict by claiming $log_2ratio(\delta_1) \geq 1$ to call a transcript as being up-regulated.

In a more formal fashion, transcripts up-regulated within the WOTGF class are defined as follows:

$$\text{WOTGF}_{up} = \begin{bmatrix} \delta_{1,up} & \vee & \\ [\bar{\delta}_{1,down} & \wedge & ((d_{11,up} \wedge d_{12,up}) \\ & \vee & (d_{11,down} \wedge \bar{d}_{12,sig} \wedge log_2ratio(\delta_1) \geq 1) \\ & \vee & (d_{11,up} \wedge \bar{d}_{12,sig} \wedge log_2ratio(\delta_1) > -1) \\ & \vee & (\bar{d}_{11,sig} \wedge d_{12,down} \wedge log_2ratio(\delta_1) \geq 1) \\ & \vee & (\bar{d}_{11,sig} \wedge d_{12,up} \wedge log_2ratio(\delta_1) > -1))] \end{bmatrix} \tag{3.5}$$

The mirrored method was used to detect WOTGF_{down} and the analogous methods are used to detect TGF_{up} and TGF_{down} based on the cells stimulated with TGF-β. Figure 3.2 shows an exemplary decision tree based on which it is possible to decide whether a gene belongs to class WOTGF_{up}. Analgous trees can be derived for

WOTGF_{down}, TGF_{up}, and TGF_{down} based on the following formulas:

$$\text{WOTGF}_{down} = \begin{bmatrix} \delta_{1,down} & \vee & \\ [\bar{\delta}_{1,up} & \wedge & ((d_{11,down} \wedge d_{12,down}) \\ & \vee & (d_{11,down} \wedge \bar{d}_{12,sig} \wedge log_2ratio(\delta_1) < 1) \\ & \vee & (d_{11,up} \wedge \bar{d}_{12,sig} \wedge log_2ratio(\delta_1) \leq -1) \\ & \vee & (\bar{d}_{11,sig} \wedge d_{12,down} \wedge log_2ratio(\delta_1) < 1) \\ & \vee & (\bar{d}_{11,sig} \wedge d_{12,up} \wedge log_2ratio(\delta_1) \leq -1))] \end{bmatrix} \tag{3.6}$$

$$\text{TGF}_{up} = \begin{bmatrix} \delta_{2,up} & \vee & \\ [\bar{\delta}_{2,down} & \wedge & ((d_{21,up} \wedge d_{22,up}) \\ & \vee & (d_{21,down} \wedge \bar{d}_{22,sig} \wedge log_2ratio(\delta_2) \geq 1) \\ & \vee & (d_{21,up} \wedge \bar{d}_{22,sig} \wedge log_2ratio(\delta_2) > -1) \\ & \vee & (\bar{d}_{21,sig} \wedge d_{22,down} \wedge log_2ratio(\delta_2) \geq 1) \\ & \vee & (\bar{d}_{21,sig} \wedge d_{22,up} \wedge log_2ratio(\delta_2) > -1))] \end{bmatrix} \tag{3.7}$$

$$\text{TGF}_{down} = \begin{bmatrix} \delta_{2,down} & \vee & \\ [\bar{\delta}_{2,up} & \wedge & ((d_{21,down} \wedge d_{22,down}) \\ & \vee & (d_{21,down} \wedge \bar{d}_{22,sig} \wedge log_2ratio(\delta_2) < 1) \\ & \vee & (d_{21,up} \wedge \bar{d}_{22,sig} \wedge log_2ratio(\delta_2) \leq -1) \\ & \vee & (\bar{d}_{21,sig} \wedge d_{22,down} \wedge log_2ratio(\delta_2) < 1) \\ & \vee & (\bar{d}_{21,sig} \wedge d_{22,up} \wedge log_2ratio(\delta_2) \leq -1))] \end{bmatrix} \tag{3.8}$$

The final off-target signature was defined based transcripts for which the following formula holds true:

$$\begin{aligned} &(\text{TGF}_{up} \wedge \text{WOTGF}_{up}) \vee (\text{TGF}_{down} \wedge \text{WOTGF}_{down}) \vee \\ &(\text{TGF}_{up} \wedge \text{WOTGF}_{down}) \vee (\text{TGF}_{down} \wedge \text{WOTGF}_{up}). \end{aligned} \tag{3.9}$$

The profiles of the respective transcripts can be assigned to different categories:

$\text{TGF}_{up} \wedge \text{WOTGF}_{up}$: additive or bipolar on- and off-target effect or common off-target effect,

$\text{TGF}_{down} \wedge \text{WOTGF}_{down}$: additive or bipolar on- and off-target effect or common off-target effect,

$\text{TGF}_{up} \wedge \text{WOTGF}_{down}$: inverse or bipolar on- and off-target effect, and

$\mathbf{TGF}_{down} \wedge \mathbf{WOTGF}_{up}$**:** inverse or bipolar on- and off-target effect.

Bipolar on- and off-target effects describe gene expression profiles for which the direction of regulation depends on the concentration range, i.e. it is possible that a compound induces expression at lower concentrations whereas expression again decreases at higher concentrations or vice versa.

3.3.4 Gene Set Enrichment Analysis

Based on the on- and off-target signatures, standard Ingenuity Pathway Analyses (IPAs) [144] were conducted. In these analyses, gene sets defined by the Ingenuity Knowledge Base are tested for enrichment of deregulated genes using Fisher's exact test. Additionally, gene sets defined by KEGG pathways as annotated by the Bioconductor package `KEGG.db` version 2.2.11 were tested for enrichment of genes contained in the TGF-β signature (Section 3.3.2). As KEGG is not freely available to the pharmaceutical industry, no compound related analyses were conducted using KEGG. The gene sets were hierarchically clustered based on the calculated p-values using manhattan distance and complete linkage as distance measures between gene sets and clusters. Results are presented in Section 4.2.3 pages 106 ff..

3.4 A New Approach for Data Integration

In this section, I describe the individual evidence that has been used to calculate edge weights $\omega(v_i, v_j)$ for PPI networks (Definition 2.2.2). Respective networks are used in Sections 4.3 and 4.4. How individual measures are weighted is highly flexible and should be chosen according to the research objective and in close cooperation with biologists familiar with the experimental setting and background.

3.4.1 Protein Interaction Data

iRefIndex

The protein interaction network was generated based on iRefIndex [22]. PSI-MITAB formatted interactions for human were taken from the publicly available version of release 5 (9606.mitab.06042009.txt.zip), which consolidates information taken from six different Protein interaction databases (BIND [33,34], BioGRID [35,36], IntAct

[37], MINT [38], MPPI [39] and OPHID [40]). For generating the underlying protein interaction network, all binary interactions were used. This resulted in a network of 10,321 nodes and 57,811 edges. The respective proteins were represented by their UniProt Accessions.

3.4.2 Scoring Similarity of Genes Based on Gene Ontology (GO)

In contrast to most previous approaches as described in Section 2.3, our method takes into account the similarity of the GO annotation of two nodes that are adjacent in a protein interaction network. Similarity scores described in this section are calculated based on the directed acyclic graph representing the ontology. Details on GO have been given in Section 2.4.1, Definitions 2.4.2 - 2.4.4, and Figure 2.4 on pages 44 ff..

To our knowledge, Wu *et al.* [167] first integrated GO annotations to derive functional modules. Using a Bayesian approach, they combine results for the analyses of phylogenetic profiles, gene neighborhoods, and GO annotations. The combined information is used to measure the strength of gene functional relationship based on which functional modules present in *Escherichia coli* were predicted.

Let us assume two genes/proteins, g_1 and g_2 represented by nodes in an interaction graph are annotated by two sets of GO terms, GO_1 and GO_2, respectively. Below, I first describe several approaches to derive similarity scores $sim(go_1, go_2)$ for a pair of GO-terms $go_1 \in GO_1$ and $go_2 \in GO_2$ (pages 62 ff.). In a second step, I introduce different possibilities to calculate the similarity of two sets of GO-term, $sim(GO_1, GO_2)$, based on the similarity calculated for a pair of GO-terms $sim(go_1, go_2)$. $sim(GO_1, GO_2)$ can be used as a score for the similarity of two genes/proteins $Sim(g_1, g_2)$ (pages 65 f.).

Resnik's Measure [168, 169]

Resnik proposes a similarity score that is based on the information content (IC). The IC takes into account the frequency of occurrence of any ontology term in the ontology, i.e. how often the respective term is annotated to a gene/protein

compared to the maximum possible usage of that term. The more frequent a term occurs, the lower its information content is. Thus, in case of a unique root of the ontology tree, $p(root) = 1$. The respective information content IC_{root} is defined as $-log\ p(root) = 0$. Analogous, for any other term go, $IC_{go} = -log\ p(go)$. Based on the IC, the similarity of two terms go_1, go_2 is calculated. The more information two terms share, the higher is their similarity $sim(go_1, go_2)$:

$$sim_{Resnik}(go_1, go_2) = max_{go \in S(go_1, go_2)}(-log\ p(go)), \tag{3.10}$$

where $S(go_1, go_2)$ is the set of common ancestors of terms go_1 and go_2. Thus, the information content of their lowest common ancestor (LCA, Definition 2.4.7) quantifies the similarity of two terms:

$$sim_{Resnik}(go_1, go_2) = -log\ p(LCA(go_1, go_2)). \tag{3.11}$$

Lin's Measure [170]

The measure proposed by Lin is also based on the IC. In contrast to Resnik's definition of similarity, Lin normalizes the similarity of two terms go_1, go_2 with respect to the sum of the IC of the individual terms:

$$sim_{Lin}(go_1, go_2) = max_{go \in S(go_1, go_2)} \frac{2 \cdot log\ p(go)}{log\ p(go_1) + log\ p(go_2)} \tag{3.12}$$

Rel [171]

Rel, the similarity measure proposed by Schlicker *et al.* additionally weights the measure proposed by Lin according to its relevance. The relevance of a term decreases with increasing frequency of occurrences, the similarity is weighted with $1 - p(go)$:

$$sim_{Rel}(go_1, go_2) = max_{go \in S(go_1, go_2)} \left(\frac{2 \cdot log\ p(go)}{log\ p(go_1) + log\ p(go_2)} \cdot (1 - p(go)) \right) \tag{3.13}$$

Jiang's and Conrath's Measure [172]

Jiang and Conrath define a semantic distance

$$Dist(go_1, go_2) = IC(go_1) + IC(go_2) - 2 \cdot IC(LCA(go_1, go_2)). \tag{3.14}$$

The distance is an inversion of the similarity, thus, Yu [173] implements the similarity as

$$sim_{JC}(go_1, go_2) = 1 - min(1, d(go_1, go_2)), \tag{3.15}$$

where

$$d(go_1, go_2) = log\ p(go_1) + log\ p(go_2) - 2 \cdot log\ p(LCA(go_1, go_2)). \tag{3.16}$$

GOViz

GoViz was implemented by Andreas Bonertz with support from Dr. Benedikt Brors in the course of a trainee at the Computational Biology group of the former Theoretical Bioinformatics department at the DKFZ. The R-package was provided by Dr. Benedikt Brors. Making use of the GOstats package, similarity of the GO annotations of two genes is calculated as the path length of the intersection graphs induced by the annotations [174]. GoViz does not take into account the semantic similarity of the GO terms assigned to a pair of genes but only the depth of the common annotations.

Wang's Measure [141]

Wang *et al.* state that the methods described so far were developed for natural language taxonomies [168–170, 172]. All these methods base their similarity scoring on the information content present in a term. They neglect that the specificity of a GO term is determined by the information inherited from its ancestors, thus, the location within the GO graph has to be taken into account. That is why they propose a new method to measure the semantic similarity of a GO term. Based on the calculated semantic similarity scores, they developed a new algorithm to measure the similarity of genes with respect to their GO annotation.

$DAG_{go} = (T_{go}, E_{go})$ represent the GO term go. T_{go} is the set of GO terms, i.e. nodes, in DAG_{go} containing go as well as all its ancestor terms; E_{go} refers to all edges induced by T_{go}. To calculate the semantic value of go, the authors define a so-called S-value calculated for any term $t \in T_{go}$ in the ontology:

$$S_{go}(t) = \begin{cases} 1 & : t = go \\ max\{\gamma_{e_{t \to t'}} \cdot S_{go}(t') | t' \in V_{children}(t)\} & : t \in T_{go}, t \neq go \end{cases}, \tag{3.17}$$

where $\gamma_{e_{t \to t'}}$ is the semantic contribution factor for edge $e_{t' \to t} \in E_A$ linking term t' with its parent term t. Based on the S-values, the semantic value of GO term go is

given by

$$SV(go) = \sum_{t \in T_{go}} S_{go}(t). \tag{3.18}$$

The semantic similarity of GO terms go_1 and go_2 which induce $DAG_{go_1} = (T_{go_1}, E_{go_1})$ and $DAG_{go_2} = (T_{go_2}, E_{go_2})$, respectively is then calculated by

$$sim_{Wang}(go_1, go_2) = \frac{\sum_{t \in T_{go_1} \cap T_{go_2}} (S_{go_1}(t) + S_{go_2}(t))}{SV(go_1) + SV(go_2)}. \tag{3.19}$$

Calculate the similarity of genes based on their GO annotations

Similarity of genes based on GO annotations can either be calculated with respect to the cellular component (CC) in which the proteins encoded by the genes are located in, with respect to the biological process (BP) in which the genes/proteins are involved in, or with respect to the molecular function (MF) of the encoded proteins. CC, BP, and MF are described by three separate DAGs in the GO.

Wang *et al.* propose the following formula to calculate the semantic similarity $sim(go, GO)$ between GO term go and a set of GO terms $GO = \{go_1, go_2, ..., go_k\}$

$$sim(go, GO) = max_{1 \leq i \leq k}(S_{GO}(go, go_i)). \tag{3.20}$$

Now, given genes g_1, g_2 annotated by GO term sets $GO_1 = \{go_{11}, go_{12}, ..., go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22}, ..., go_{2n}\}$, respectively, their similarity $Sim(g_1, g_2)$ is defined as $sim(GO_1, GO_2)$, i.e. the similarity of GO_1 and GO_2. Wang *et al.* [141] propose the following formula to calculate $Sim(g_1, g_2)$

$$sim(GO_1, GO_2) = \frac{\sum_{1 \leq i \leq m}(sim(go_{1i}, GO_2)) + \sum_{1 \leq i \leq n}(sim(go_{2i}, GO_1))}{m + n} \tag{3.21}$$

This method is used in the Bioconductor package GOSemSim [173].

Schlicker *et al.* propose similar methods:

$$SimAvg(g_1, g_2) = \frac{1}{2}\left(\frac{\sum_{1 \leq i \leq m}(sim(go_{1i}, GO_2))}{m} + \frac{\sum_{1 \leq i \leq n}(sim(go_{2i}, GO_1))}{n}\right) \tag{3.22}$$

$$SimMax(g_1, g_2) = max\left(\frac{\sum_{1\leq i\leq m}(sim(go_{1i}, GO_2))}{m}, \frac{\sum_{1\leq i\leq n}(sim(go_{2i}, GO_1))}{n}\right) \tag{3.23}$$

In cases were the GO annotation for one gene product only matches a subset of the GO annotations of the second gene product, $SimMax$ is superior to $SimAvg$ for similar gene products. Since GO_1 and GO_2 are not necessarily equal in length and generally differ in terms, $\frac{\sum_{1\leq i\leq m}(sim(go_{1i},GO_2))}{m}$ and $\frac{\sum_{1\leq i\leq n}(sim(go_{2i},GO_1))}{n}$ are not necessarily equal. In case one of the gene products g_1 or g_2 is annotated incompletely, this could lead to artificially low $\frac{\sum_{1\leq i\leq m}(sim(go_{1i},GO_2))}{m}$ or $\frac{\sum_{1\leq i\leq n}(sim(go_{2i},GO_1))}{n}$. Taking the average as in $SimAvg(g_1, g_2)$ would then lead to a lower value than using $SimMax(g_1, g_2)$. Such situations can occur if the annotation for the first gene product are not complete or if the second gene product is multi-functional [171]. $SimMax$ is, for example, implemented in the R-package GOSim [175].

Implementations of GO Similarity Scores

R packages `GOSemSim` [173], `GOSim` [175], and `GOstats` [176] were used to weight the protein pairs according to their GO annotations. First, I made use of the `mgoSim` function implemented in the `GOSemSim` package [173], retrieved the GO annotations of the proteins using `biomaRt` [177] and filtered them by ignoring GO terms with evidence codes `NAS`, `IEA`, `ND` (non-traceable author statement, inferred from electronic annotation and no biological data available, respectively) and those which had not assigned any evidence code. Similarities for the annotations were calculated based on the methods proposed by Resnik (Eqs. 3.10 and 3.11, page 63) [178], Jiang and Conrath (Eqs. 3.14-3.16, pages 63 f.) [172], Lin (Eq. 3.12, page 63) [170], Schlicker (Eq. 3.13, page 63) [171], and Wang (Eqs. 3.17-3.19, pages 64 f.) [141]. The similarity of two genes based on their GO-annotations is calculated using Eq. 3.21.

Next, I made use of the method `mgeneSim`, also implemented in the `GOSemSim` package. Swiss-Prot Identifiers were mapped to EntrezGene IDs using the package `org.Hs.eg.db`, and similarities were calculated using the methods proposed by Resnik (Eqs. 3.10 and 3.11, page 63) and Wang (Eqs. 3.17-3.19, pages 64 f.) without filtering for specific evidence codes. Again, the similarity of two genes based on their GO-annotations is calculated using Eq. 3.21.

According to Smialowski [179], I utilized the `GOSim` package [175] which, amongst others, also implements the method from Resnik (Eqs. 3.10 and 3.11, page 63) but offers different options for comparing and summarizing GO-Terms. To compute the similarity of a pair of proteins, the method `getGeneSim` was used with default parameters but setting `similarityTerm = "Resnik"`. Thereby, Eq. 3.23 (page 66) is applied to calculate the similarity of two genes based on their GO annotations.

Finally, I used GOViz [174] that calculates the similarity score as the length of the longest path in the intersection graph induced by the GO annotations of the two genes under consideration (page 64).

3.4.3 Scoring Similarity of Genes Based on Promoter Regions

Information on transcription factor binding sites was taken from the ElDorado database using the Gene2Promoter large scale analysis [180] which is part of the Genomatix Genome Analyzer (GGA) [181]. Based on the binding site motif (weight matrix) of transcription factor TF and the sequence of the promoter region of gene G, a matrix similarity score $score_{TF}(G)$ was calculated using MatInspector [182,183]. A perfect match of the transcription factor's binding site motif to a region in the promoter gets a score of 1.0, a good match usually has a score > 0.8 [180].

A score, $score_{TF}(X, Y)$, is calculated based on the set of transcription factors $\mathbf{TF}(X)$ and $\mathbf{TF}(Y)$ predicted to regulate the protein-coding genes X and Y, respectively. The higher this score, the more similar are the promoter regions of X and Y. Using a Tanimoto-based approach, different possibilities were implemented to calculate $score_{TF}(X, Y)$:

1. Calculation of the score is based on the sum over the matrix similarity scores predicted for transcription factors regulating both, gene X and gene Y. This sum is divided by the sum over the scores for the union of the predicted transcription factor binding sites of both genes:

$$score_{TF,1}(X, Y) = \frac{\sum_{TF \in (\mathbf{TF}(X) \cap \mathbf{TF}(Y))} (score_{TF}(X) + score_{TF}(Y))}{\sum_{TF \in (\mathbf{TF}(X) \cup \mathbf{TF}(Y))} (score_{TF}(X) + score_{TF}(Y))} \quad (3.24)$$

2. $score_{TF,1}(X,Y)$ is multiplied by the number of common transcription factors:

$$score_{TF,2}(X,Y) = |\mathbf{TF}(X) \cap \mathbf{TF}(Y)| \cdot score_{TF,1}(X,Y) \tag{3.25}$$

Results based on using all predicted transcription factor binding sites were compared to those that are predicted with a matrix similarity score > 0.8 referred to as cutoff_08. This finally results in four possibilities to calculate $score_{TF}(X,Y)$.

3.4.4 Scoring Similarity of Genes Based on Literature

For each recorded protein interaction iRefIndex provides a confidence score. This score is based on the number of PubMed publications supporting the respective interaction. Within iRefIndex the confidence scores are given in the following format: *lpr/hpr/np*. *lpr* refers to the lowest number of distinct interactions that any PubMed reference supporting the respective protein interaction contains. That is, a small value of *lpr* indicates a low throughput experiment for validating an interaction, whereas larger values of *lpr* indicate high throughput experiments; *hpr* refers to the highest number of interactions that any PubMed reference supporting the respective protein interaction contains; *np* is the total number of references supporting the respective interaction. For details I refer the reader to the original publication [96] and to the README for iRefIndex MITAB 4.0. Since we want to give higher confidence to interactions that are supported by a low-throughput experiment and more publications, the calculation of the confidence score is based on *lpr* and *np*. The following formula was used to calculate a confidence for the edge representing the interaction in our network:

$$score_{conf}(X,Y) = \frac{\sqrt{np}}{lpr} \tag{3.26}$$

Since *np* and *lpr* are only available for protein interactions present in iRefIndex, $score_{conf}(X,Y)$ was only applied to iRefIndex-based but not to the STRING-based networks.

3.4.5 Scoring Similarity of Genes Based on Expression Data

In Sections 4.3.1 (pages 111 ff.) and 4.3.2 (pages 114 f.), I make use of the gene expression data derived from TGF-β1-stimulated as well as unstimulated cells mea-

sured in four biological replicates at 2, 4 and 12 hours after stimulation. In Section 4.3.3 (pages 116 f.), I considered data derived from untreated as well as treated cells. For the treated cells, I focused on cells stimulated with TGF-β1 and either treated with different concentrations of BI1 or BI4, or not treated with any compound. By applying the methods described in this section together with the algorithm described in Section 3.5 to compound-treated cells, I demonstrate how these methods can support the analysis of compounds' mode of action.

For each edge $e_{\{X,Y\}}$ in the protein interaction network $G = (V, E)$ connecting protein coding genes X and Y a score $score_{exp}(X, Y)$ is calculated based on the underlying gene expression experiment. I compared 9 different possibilities for weighting the interactions. They are either based on correlation or on co-variance (as for example also used by Guo [49]) of vectors x and y. x and y contain values derived based on two conditions either looking at three different time points (2, 4, 12 hours) or for one timepoint (2 hours). In Sections 4.3.1 (pages 111 ff.) and 4.3.2 (pages 114 f.) the two conditions are TGF-β1-stimulated and unstimulated cells, in Section 4.3.3 (pages 116 f.) data derived from untreated as well as compound-treated cells is considered.

1. Weighting based on correlation (cor): Pearson correlation is calculated based on the

   (a) log_2 ratios of two conditions,

   (b) mean of replicate values for the normalized expression values of both conditions, or

   (c) normalized expression values of both conditions for all replicates

   at the different time points measured (2, 4, 12 hours):

$$score_{exp}(X, Y) = cor(x, y) \tag{3.27}$$

2. Weighting based on covariance (cov): The covariance is calculated based on the

   (a) log_2 ratios of the two conditions for the respective gene pair (X,Y) at the different time points measured (2, 4, 12 hours),

(b) mean of the replicate values of the normalized expression values of both conditions at the different time points measured (2, 4, 12 hours), or

(c) normalized expression values of both conditions for all replicates at the different time points measured (2, 4, 12 hours) and at 2 hours.

$$score_{exp}(X,Y) = cov(x,y) = cor(x,y) \cdot sd(x) \cdot sd(y) \tag{3.28}$$

Since a negative covariance/correlation could hint at pairs of proteins exhibiting related biological contexts (e.g. inhibitor and activator), the absolute value of the correlation/covariance was taken.

The score may be weighted by a factor ω_{exp}:

$$score_{exp,weighted}(X,Y) = \omega_{exp} \cdot score_{exp}(X,Y) \tag{3.29}$$

Two different approaches have been applied to calculate 3.29: $score_{exp}(X,Y)$ was calculated based on method (1a) and two different possibilities to calculate ω_{exp} were considered:

3. For focusing on changes occurring at a specific time point ω_{exp} is set to the average of $|log_2 ratios|$ at the time point of interest (in our case 2 hours, since this is the time frame for which expression changes directly induced through TGF-β-signaling are observed).

4. To assign higher weights to genes that show a change in differential expression over time, ω_{exp} is set to the average of the standard deviations of log_2 ratios across all time points (in our case 2,4, and 12 hours) for X and Y.

To test the weighted scoring, we applied (3) and (4) using the correlation of log_2 ratios across the three time points as $score_{exp}(X,Y)$.

In total, nine different ways of calculating $score_{exp}$ were compared: Three of them are based on correlation (1a-c), four are based on covariance (2a-c, 2c was applied for measures at all 3 time points as well as for measured at 2hours), and two are based on weighting $score_{exp}$ (Eq. 3.29).

3.4.6 Combining the Individual Similarity Scores

For each edge, a linear combination over all individual scores is calculated to make up the final edge score:

$$score_{edge} = \sum_{i=1}^{n} a_i s_i, \text{ with} \tag{3.30}$$

$$a_i \in [0, \infty) \text{ and } s_i \in \{score_{GO}, score_{TF}, score_{conf}, score_{exp}, ...\}.$$

Since this weighting function is highly flexible, additional information can easily be added. All scores but those obtained based on the gene expression experiment are normalized between 0 and 1. To weight all evidence equally, $a_i = 1 \; \forall \; i$. With four different methods used for scoring the transcription factors, nine different methods scoring gene expression data, and ten different methods to score GO, 360 combinations of methods are considered. In Section 4.3.1 (pages 111 ff.), I describe how one of the possible combinations was selected. Based on this combination all further analyses presented in this thesis have been conducted. In principle, the weighting factors a_i can be adjusted according to the research objective and in close cooperation with biologists. More important scores should be given a higher weight.

3.5 modEx - A New Approach for Extraction of Protein Modules

Based on the methods presented in the previous sections, different data types are integrated into a protein interaction network to obtain an edge- and node-weighted graph $G = (V, E)$. The log_2 ratios or fold changes of a gene expression experiment are assigned as node weights. By the methods presented in Section 3.4, information on biological processes, molecular functions, cellular components, transcription factor binding sites, literature and on correlation or covariance of the gene expression is translated to edge weights $score_{edge}$ as described by Eq. 3.30 on page 71.

In Section 3.5.1, I define some graph-theoretical problems that can be posed by networks weighted as described in the previous paragraph. Solving the problems on this basis, it is possible to shed some light on the underlying biology, like for example

the mode of action of compounds. In Section 4.5.2 (pages 134 ff.) the NP-hardness of these problems is proven.

In Section 3.5.2 (pages 73 f.) I propose a heuristic, modEx, that could be used to solve the previously defined problems. After validation on a biological basis in Section 4.3.2, this method is applied in Section 4.3.3 to identify modules that help to elucidate the biological processes affected in the gene expression experiments under investigation (pages 114 f.).

3.5.1 Formal Problem Definitions

In the following, I introduce some graph-theoretical problems that could be posed to resolve biological questions based on edge weighted graphs. All problem definitions are constrained with respect to a set of vertices that have to be part of the solution and/or to the number of vertices contained in the solution. Generally speaking, all objective functions aim at maximizing the edge weight of the solution:

Definition 3.5.1. *The* VERTEX CONSTRAINED MAXIMUM EDGE WEIGHT CONNECTED GRAPH (VCMECG) *problem.*

Input: *An undirected graph* $G = (V, E)$ *with weight function* $\omega : E \to [0, \infty)$*, and a positive integer* s*.*

Task: *Find a connected subgraph* $G' = (V', E')$ *with* $|V'| = s$ *such that*

$$\Omega(G') = \sum_{e \in E'} \omega(e) \tag{3.31}$$

is maximized.

Definition 3.5.2. *The* k-VERTEX CONSTRAINED MAXIMUM EDGE-WEIGHT CONNECTED GRAPH (k-VCMECG) *problem.*

Input: *An undirected graph* $G = (V, E)$ *with weight function* $\omega : E \to [0, \infty)$*,* $V_k \subseteq V$ *with* $|V_k| = k$*, and a positive integer* s*.*

Task: *Find a connected subgraph* $G' = (V', E')$ *with* $V_k \subseteq V'$ *and* $|V'| = s$ *such that*

$$\Omega(G') = \frac{\sum_{e \in E'} \omega(e)}{|V'|} \tag{3.32}$$

is maximized.

Considering a network scored as described in Section 3.4.6, V_k could for example be chosen based on the set of k strongest deregulated genes observed in the underlying gene expression experiment. As a solution for k-vCMECG, the subgraph $G' = (V', E')$ could then help to explain and understand the biological context of the k deregulated genes. In general, any set of genes for which the biological context is in question could be used as V_k.

Definition 3.5.3. *The* k-VERTEX/EDGE CONSTRAINED MAXIMUM EDGE-WEIGHT CONNECTED GRAPH (k-veCMECG) *problem.*

> **Input:** *An undirected graph* $G = (V, E)$ *with weight function* $\omega : E \to [0, \infty)$, *and* $V_k \subseteq V$ *with* $|V_k| = k$ *and a positive integer* s.
>
> **Task:** *Find a connected subgraph* $G' = (V', E')$ *such that* $V_k \subseteq V'$, $|V'| = s$, *and*
>
> $$\Omega(G') = \frac{\sum_{e \in E'} \omega(e)}{|E'|} \tag{3.33}$$
>
> *is maximized.*

Since we are not only interested in identifying one module but in detecting as many modules as are necessary to investigate the underlying processes, we consider a special case of the k-vCMECG problem, namely the 1-vCMECG problem:

Definition 3.5.4. *The* 1-VERTEX CONSTRAINED MAXIMUM EDGE-WEIGHT CONNECTED GRAPH (1-vCMECG) *problem.*

> **Input:** *An undirected graph* $G = (V, E)$ *with weight function* $\omega : E \to [0, \infty)$, *a seed vertex* $v_i \in V$, *and a positive integer* s.
>
> **Task:** *Find a connected subgraph* $G'_i = (V', E')$ *with* $v_i \in V'$ *and* $|V'| = s$ *such that*
>
> $$\Omega(G'_i) = \frac{\sum_{e \in E'} \omega(e)}{|V'|} \tag{3.34}$$
>
> *is maximized.*

3.5.2 Heuristic Approaches to Solve the 1-vCMECG Problem

modEx starts with seed nodes selected on the basis of the gene expression experiment to analyze. Out of the significantly deregulated genes (FDR-adjusted p-value $<$ 0.01), the i mostly deregulated are selected as seed nodes. Starting

at those nodes, the networks are expanded in the direction of the heaviest edge. Depending on whether one is interested in sparse or dense networks, either only the heaviest edges are returned or the subgraph induced by the selected nodes (Definition 2.3.2, page 28) is extracted. The modules extracted in this work all represent induced subgraphs.

A number of variations with respect to how the extraction of graph G'_i around the i-th seed node (Definition 3.5.4) is performed and how the algorithm terminates can be applied. Following procedures were investigated:

(1) Apply greedy search and stop after a pre-defined number of nodes have been extracted.

(2) Apply greedy search and stop as soon as an optimum for $\Omega(G'_i)$ is reached.

(3) Use simulated annealing optimizing $\Omega(G'_i)$.

(4) Reapply (1), (2) or (3), starting with different seed nodes until a pre-defined number of connected components is extracted.

If necessary, methods (2) and (3) can be combined with (1) in order to avoid extracting undesired large networks. The results presented in this work are all based on (3) combined with (1) by stopping simulated annealing as soon as more than 50 nodes were extracted. Details on this decision are given in Section 5.3 on page 142.

3.6 Statistical Measures Used

3.6.1 Quantification of Extracted Modules

To quantitatively compare the modules extracted from the protein interaction graphs, we compared them to modules extracted from randomized graphs. Nodes of the network were permuted 500 times and edge weights were recalculated based on these permutations. Based on the permuted graphs, modules were extracted by applying approach (3) combined with (1) using 50 nodes as the limit for module sizes (page 74). For each of the 500 modules, $\Omega(G'_i)$ (Eq. 3.34, page 73), in connection with the random graphs referred to as $\Omega(G'_{i,rand})$, was calculated.

P-Value Calculation Based on Z-Score Transformation

Assuming a normal distribution for the $\Omega(G'_{i,rand})$ values obtained by the random networks, the corresponding parameters μ_{rand} and σ_{rand} were estimated by $mean(\Omega(G'_{i,rand}))$ and $sd(\Omega(G'_{i,rand}))$. Given that $\Omega(G'_i)$ of the real network follows the same distribution, a z-score transformation is performed:

$$z_i = \frac{\Omega(G'_i) - \mu_{rand}}{\sigma_{rand}} \tag{3.35}$$

Then, z_i follows a $N(0,1)$ distribution. Based on this, the probability $P(x \geq \Omega(G'_i))$ of randomly observing network scores $\geq \Omega(G'_i)$ can be calculated for the extracted modules. In the following, I refer to this probability as "z-score based p-value".

P-Value Calculation Based on Approximative Permutation Tests

As the $\Omega(G'_{i,rand})$ values not strictly follow a normal distribution, p-values were additionally calculated using an approximation-based approach. The probability of randomly observing a score $\geq \Omega(G'_i)$ is calculated as the relative frequency of how often $\Omega(G'_{i,rand}) \geq \Omega(G'_i)$ is observed, referred to as "approximative permutation test based p-value".

3.6.2 Gene Set Enrichment Using Fisher's Exact Test

Fisher's Exact Test

Given N as the total number of genes as, for example, available on a microarray or all genes present in an organism. M is the number of genes in the gene group to test for enrichment. We are interested in how probable it is to have x genes of the K most interesting genes in this group. This can be displayed in a 2×2 contingency table.

Table 3.2: 2×2 contingency table.

	$\in$ gene group	$\notin$ gene group	
$\in$ genes of interest	x	K-x	K
$\notin$ genes of interest	M-x	(N-M)-(K-x)	N-K
	M	N-M	N

Fisher showed that the probability of obtaining such a set of values under the null hypothesis is given by the hypergeometric distribution

$$P(X = x|N, M, K) = \frac{\binom{M}{x} + \binom{N-M}{K-x}}{\binom{N}{K}}$$

Thus, the p-value for the enrichment is obtained by

$$P(X \geq x|N, M, K)$$

In Section 4.4 (pages 124 f.) and 4.4.1 (pages 124 f.), gene sets are ranked according to p-values and inverse ranks of TGF-β signaling are used to compare different results.

3.7 Comparison to Existing Approaches

jActiveModule

jActiveModule [48] is a Cytoscape plugin aiming at the identification of modules (subnetworks) exhibiting significant changes in differential gene expression experiments (see Section 2.3.1, pages 28 ff.). It offers a greedy search as well as a simulated annealing approach to detect these modules. To achieve results that can be compared to results derived by our approach, we utilized the FDR-adjusted p-values obtained by comparing the TGF-β-stimulated HaCaT cells to the unstimulated cells after 2, 4, and 12 hours. Not all proteins contained in the network are represented by probes on the microarray used to measure gene expression. In case the protein coding gene is not represented, three different options have been investigated:

1. No p-values are assigned (*no-pval*).

2. 1 is assigned as p-value, 0 as log_2 ratio (*def-pval*).

3. Only the subnet of the iRefIndex-based network is kept that contains proteins for which genes are represented on the chip (*sub-net*).

Labels in parenthesis are used in Section 4.4.2 (pages 126 f.) and Section 4.4.3 (pages 127 ff.) to refer to these different approaches.

STRING Based Protein Interaction Network

STRING [55, 121, 184] is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: Genomic context, high-throughput experiments, co-expression, and previous knowledge. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. STRING version 8.3 covers 2,590,259 proteins from 630 organisms. Since the gene expression experiment under consideration is based on a human cell line (Section 3.1), I only made use of associations for human (taxonID: 9606) resulting in a network composed of 17,078 nodes and 1,236,215 edges. Associations stored in STRING have different confidence scores ranging from 0-1000. Based on filters applied to these confidence scores, three separate analyses were conducted. Using protein associations with a confidence score ≥ 400 results in a network containing 15,858 nodes and 408,619 edges, using only highly confident associations with a score > 700 in 12,692 nodes and 176,595 edges, and applying a filter with a score ≥ 848 in 10,331 nodes and 120,337 edges. Using our scoring method, we recalculated the scores for the edges contained in the STRING-based networks. They are referred to as $STRING_{mod}$. Networks based on the original scoring are referred to as $STRING_{org}$. Since a cutoff of 848 approximates the size of the iRefIndex-based network, results presented in Section 4.4 are all based on this cutoff. For easier readability, in the main text of this thesis, $STRING_{org}$ refers to the original STRING network with 848 applied as cutoff for the edge scores; analogous, $STRING_{mod}$ refers to the same network but with edge weights recalculated using the scoring method introduced in Section 3.4.6 (Eq.3.29, page 70). In the Appendix, we explicitly state the results for different cutoffs used for $STRING_{org}$ and $STRING_{mod}$.

Chapter 4

Results

This study is based on gene expression analyses described in Section 3.1. An optimal normalization method was selected for the expression data to be able to derive the TGF-β signature as well as the NCE's off-target effects. To reveal the biological context of the signatures, the gene expression profiles have been clustered, and gene set enrichment analyses were performed. Further, I applied the approaches developed in the scope of this thesis to derive compounds' mode of action. Derived *in silico* results were validated by cell biological experiments. The results are presented in [52–54], and I refer the reader to these references; here, I address the methodological aspects of this work.

This chapter is structured as follows: In Section 4.1, I describe how the optimal normalization method for our data set has been selected. This has been published in Schmid and Baum *et al.* [53]. Based on the normalized data, I describe the *in silico* analyses using state-of-the-art methods in Section 4.2. This section is part of a PLoS One publication by Baum and Schmid *et al.* [54] and of the PhD thesis of Dr. Patrick Baum [52]. In Section 4.3, I apply the new data integration method, introduced in Section 3.4, in combination with modEx, introduced in Section 3.5, to suggest the mode of action (MoA) of given compounds. In short, interaction evidence from gene expression measurements, transcription factor binding sites and information from GO annotation is integrated. This information translates into edge weighted graphs, which allow the extraction of subnetworks based on protein interactions exhibiting high confidence. Applying this approach I generated hypotheses on mechanism-related actions of compounds inhibiting TGFβ-R1. *In silico* gen-

erated hypotheses about modulation of certain signaling pathways by selected inhibitors could be confirmed by cell biological experiments. Finally, in Section 4.4, we compare our newly proposed approaches to others and show that they beneficially complement these.

4.1 Selecting an Appropriate Normalization Method

A plethora of different normalization methods has been proposed for microarray experiments [154, 156–162]. Different microarray technologies require different normalization procedures, and even for the same technology, different methods are employed depending on the number of genes being differently expressed and on the extent of their changes.

To estimate the performance of a normalization method, and to compare competing methods, one measures both, their influence on the variance of the data, e. g. the degree of heteroscedasticity or the ability to discriminate between known groups, and the bias, i. e. how far a given measure like fold change (ratio of expression in condition 2 over expression in condition 1) deviates from the truth. Typically, this is a trade-off, so methods that yield nearly unbiased estimates of the fold changes have high variance on this estimation, and vice versa.

Scores that give a measure of either variance or bias require a gold standard, i. e. that the truth on expression of certain genes or their fold change is exactly known. This can be obtained by orthogonal methods of gene expression measures, e. g. quantitative real-time polymerase chain reaction (qRT-PCR). Since it is not possible to obtain this truth for all genes, or even a significant fraction of those present on a microarray, assumptions have to be made with regard to differential expression.

In Section 4.1.1 different pre-processing methods were evaluated by analyzing the variance of the resulting gene expression intensities via various statistical measures:

- Plot of ANOVA p-values versus $\mathrm{MSQ}_{between}$ (page 84),
- Boxplots of mean sum of squares (MSQs) between and within groups (pages 85 ff.),
- Density functions of MSQs between and within groups (pages 87 ff.),

- Volcano plots for pairwise group comparisons (pages 89 f.),
- Residuals versus mean or minimal expression levels (pages 90 ff.),
- Scatterplots of pairwise replicate expression levels (pages 92 f.), and
- Pseudo-ROC curves (pages 93 f.).

Some of these have already been used in other studies [185].

In addition to the variance, in Section 4.1.2 bias of the expression intensities is investigated. Fold changes derived from resulting gene expression intensities were compared to fold changes based on quantitative measurements of RNA abundance as determined by qRT-PCR. Thereby, it is possible to evaluate the pre-processing methods with respect to their bias. To compare different normalization methods, the following scores are employed:

- Correlation of fold changes based on expression intensities and fold changes based on qRT-PCR measures (pages 95 f.) and
- Slope of regression between fold changes based on expression intensities and fold changes based on qRT-PCR measures(pages 96 f.).

Normalization corrects for two different effects: background and scaling. Background means a global (or sometimes local) signal that adds to each value and is due to light scattering, auto-fluorescence or cross-hybridization. Scaling is necessary since the amount of RNA used for hybridization, labeling rate and quantum yield cannot be as precisely controlled as required. However, all these factors reduce (or amplify) the signal by a linear factor that can be estimated. In addition, some normalization methods (e. g. variance-stabilizing-normalization (`vsn`) [160] or variance-stabilizing-transformation (`vst`) [157]) employ a variance-stabilizing transformation that will make the variance constant across the entire range of intensities, provided that the underlying model of variance-intensity relationship holds true.

Since the total setup of the expression experiment is relatively complex, the analysis has been focused on the TGF-β-stimulated and control samples measured at three time points (2 hours, 4 hours, 12 hours) in four replicates. Thereby, a consistent subset is used as representative for the whole data set based on which a normalization

method has been selected. Twenty-five different ways of pre-processing the expression data have been investigated. For a detailed overview of the normalization procedures I refer the reader to Section 3.2 and Table 3.1 on page 53. In brief, first either background normalization from BeadStudio [162] (`bg_*`) or no background modification (`noBg_*`) has been applied. In a next step, the data was transformed using either log_2-transformation (`log`) or variance-stabilizing transformation (`vst`) [157]. Since BeadStudio's background normalization can lead to negative values, the data had to be transformed to contain only positive values by using either the background correction of `rma` [156] or forcePos [154] to be able to apply log_2-transformation. In a last step, the data was normalized using `quantile`, `loess`, or `rsn` [154] normalization. Alternatively, the transformation steps were skipped and `vsn` [160] or the normalization methods supplied by BeadStudio (`average`, `rankInvariant`, `cubicSpline`) were used for normalization.

Pre-processing methods were scored from -2 to 2 based on how well they match the required criteria for the different analyses described in this section. A complete overview of the scores assigned and the final ranking is given in Section 4.1.3, Figure 4.12.

4.1.1 Analyses of Variance Based on Expression Intensities

One basic assumption of gene expression pre-processing methods is that the majority of genes do not change their expression under different conditions. Additionally, expression intensities of replicates should be very similar in contrast to the expression of transcripts between differently treated sample groups. Based on these principles, we looked at different statistical measures to identify the method best suited for our dataset with respect to variance.

Distribution of F-test Statistics

A good normalization method should minimize the variation within a biological condition, i. e. within a group of replicates. Furthermore, the variation within a group should be smaller than the variation between groups. The F-statistic measures the variation between replicates in comparison to the variation between conditions or groups [186, 187] (Eq. 3.3, page 54). Results for the F-statistic based on the gene

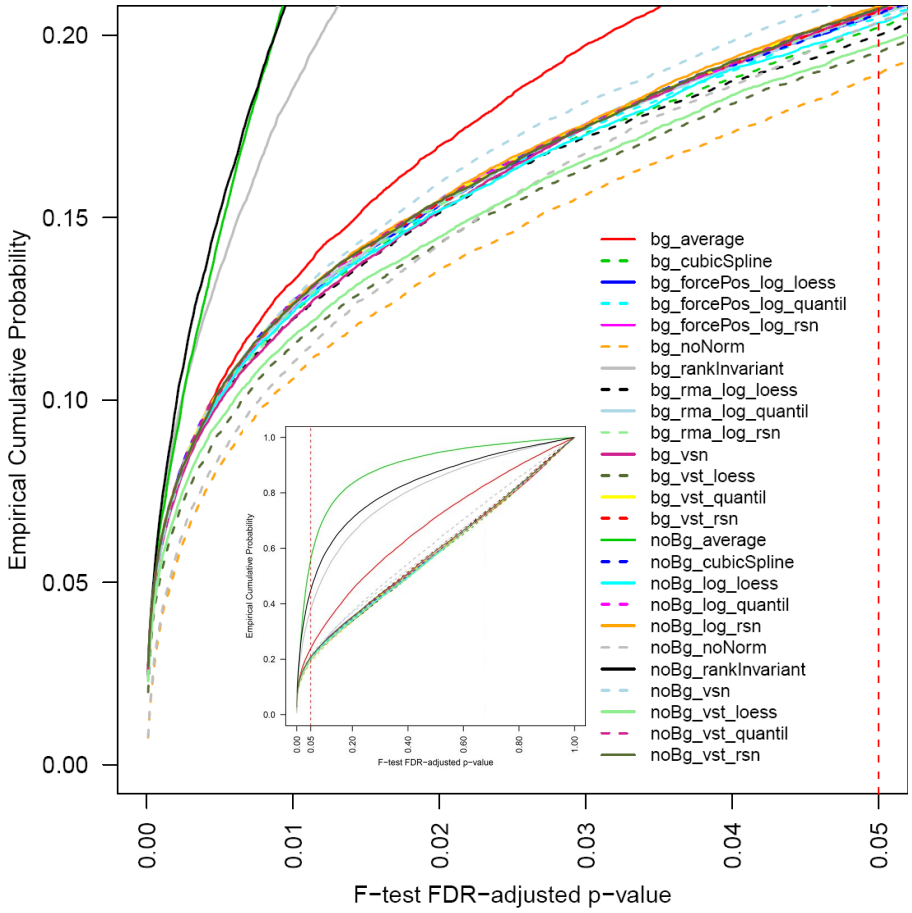


Figure 4.1: **Cumulative Distribution Functions of F-test p-values.** Cumulative Distribution Functions (CDFs) of FDR-corrected F-test p-values (Eq. 3.3, page 54) were calculated based on the gene expression measured for untreated HaCaT cells cultured for 2, 4, and 12 hours. Each of the three groups is composed of four replicates. Displayed are the results obtained for the different pre-processing methods used. The vertical red dashed line indicates the commonly chosen p-value cut-off of 0.05. The insert displays the obtained results over the whole range of values from 0 to 1 on both axes. Assuming that only few genes are differentially expressed across the different time points `bg_noNorm` (orange dashed line) outperforms the other normalization procedures.

expression measured for the untreated HaCaT cells cultured for 2 hours, 4 hours, and 12 hours are displayed in Figure 4.1.

Four BeadStudio normalization methods (`noBg_average`, `noBg_rankInvariant`, `bg_rankInvariant`, `bg_average`) show cumulative distribution functions that are

clearly above those obtained based on all other pre-processing methods. Applying neither background correction nor any normalization method (`noBg_noNorm`, Figure 4.1, light gray dashed line) results in a data set producing less transcripts with adjusted p-values < 0.02 than other pre-processing methods. With decreasing significance of the adjusted p-values, more pre-processing methods produce fewer p-values of higher significance than `noBg_noNorm`. Based on the data set used, we expect only a small subset of the transcripts to be significantly deregulated. Since for `bg_noNorm` (Figure 4.1, orange dashed line) compared to other pre-processing methods the fewest genes would be detected as being differentially expressed, i. e. showing a relatively high variation between compared to within group variability, this method seems to provide the best results. The remaining pre-processing methods for which the CDFs are running between that of `bg_noNorm` and `bg_average` perform relatively similar and equally well.

P-Values versus $\mathbf{MSQ}_{between}$

Assuming a stable variance over the within group measurements, the higher the variance between the groups (Eq. 3.2, page 54) compared to the within group variance (Eq. 3.1, page 54), the higher the respective $-log_{10}(p - value)$ should be. When plotting these parameters, an appropriate normalization method should result in smoothly increasing values with not much scattering around the fitted curve. Figure 4.2 displays the $-log_{10}(p - value)$ against the respective variance between the control groups at time points 2 hours, 4 hours, and 12 hours for three of the pre-processing methods, an overview of all results is given in [53].

Normalizations reflecting the described properties are for example `noBg_vsn`, `noBg_cubicSpline`, `noBg_log_rsn`, and `noBg_vst_rsn`. All of the normalizations performed on rma background corrected data as well as `bg_vsn` display a relatively high $-log_{10}(p-value)$ for a relatively high proportion of low between group variability values leading to a high scattering of observations in these regions. Using, for example, the rank invariant normalization of BeadStudio (`noBg_rankInvariant`) the p-values for the low between group variability tend to be relatively small. This could lead to an overestimation of differentially expressed genes when filtering solely based on p-values.

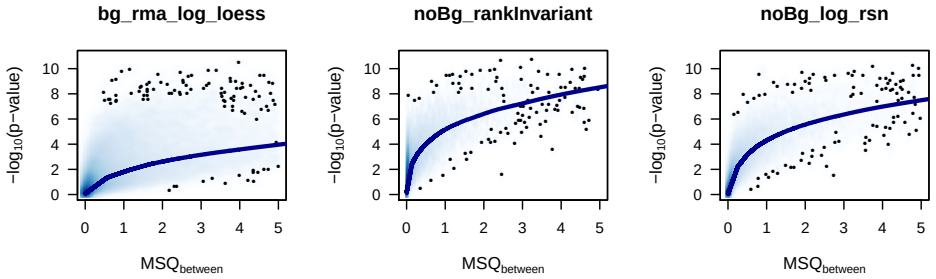


Figure 4.2: $-log_{10}(p-values)$ **against** $MSQ_{between}$ **where** $MSQ_{between} \leq 5$**.** $MSQ_{between}$ (mean sum of squares between groups, Eq. 3.2, page 3.2) was calculated based on the gene expression measured for the three control groups, namely untreated HaCaT cells cultured for 2, 4, and 12 hours. Each of the groups had been measured in four replicates. Results of three exemplary pre-processing methods of different quality are shown. `bg_rma_log_loess` exhibits the most unfavorable behavior of the three. The p-values show a high variability over the whole range of $MSQ_{between}$ and even for small $MSQ_{between}$ values there are many relatively high $-log_{10}(p-values)$. Though for `noBg_rankInvariant` the p-values show less variability in general, especially for small $MSQ_{between}$ values there are more as well as higher $-log_{10}(p-values)$. In contrast, `noBg_log_rsn` exhibits less varying p-values and does not assign as many small p-values to low $MSQ_{between}$ regions. The blue line represents a loess-curve fitted to the values. This curve takes uniformly larger values for `noBg_rankInvariant` and `noBg_log_rsn` than for `bg_rma_log_loess` indicating, on average, smaller p-values for the same $MSQ_{between}$ value. Thus, quality values of -1, 0, and 2 are assigned to `bg_rma_log_loess`, `noBg_rankInvariant`, and `noBg_log_rsn`, respectively. For an overview of all different normalization methods and their quality scores, see Figure 4.12 and [53].

Boxplots of $MSQ_{between}$ and MSQ_{within}

Distributions of between- ($MSQ_{between}$, Eq. 3.2) and within-group (MSQ_{within}, Eq. 3.1) variances and their relation to each other are further indications for normalization performance. If genes are not differentially expressed, $MSQ_{between}$ should be comparable to MSQ_{within}. For genes that are differentially expressed, $MSQ_{between}$ is supposed to be higher than MSQ_{within}. Figure 4.3 displays the boxplots for $MSQ_{between}$ (red) and MSQ_{within} (blue) values. Since we expect some genes to be differentially deregulated across the different time points under consideration, quantiles of MSQ_{within} values should lie below the corresponding quantiles of the

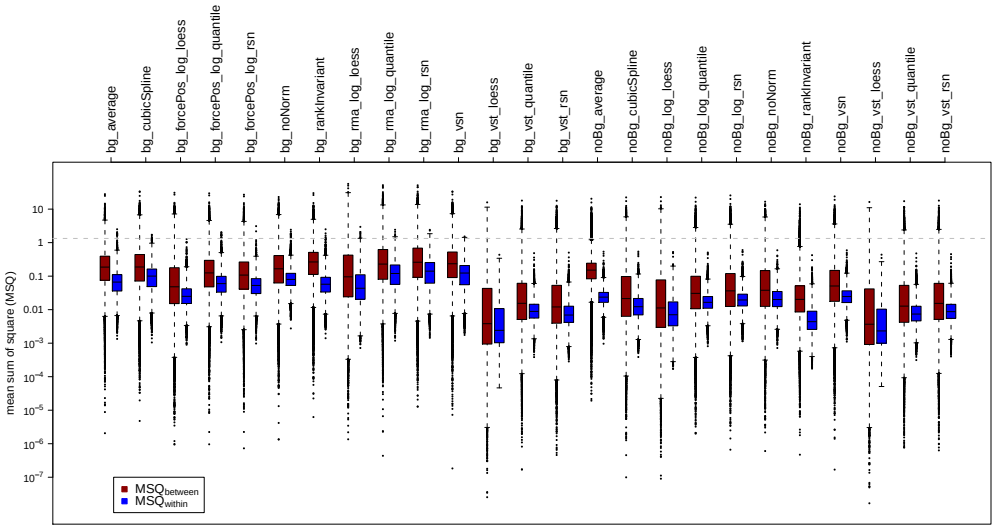


Figure 4.3: **Boxplots of MSQ_{within} (blue) and $MSQ_{between}$ (red).** Mean sum of squares ($MSQs$) were calculated based on the gene expression measured for the three sample groups analyzed, namely untreated HaCaT cells cultured for 2, 4, and 12 hours (Eqs. 3.1 and 3.2, page 54). Results obtained for the different pre-processing methods used are displayed. The gray dashed line indicates the expected value for the $MSQ_{between}$ of 1.33 based on 6, 6, and 7 as measurements for the group means of four replicates for three time points (Eq. 3.2, page 54).

$MSQ_{between}$ values. For the differentially expressed genes, within group variance should be smaller than between group variance, whereas for the genes not differentially expressed, the respective $MSQ_{between}$ and MSQ_{within} values should show no great difference. Small interquartile ranges (IQRs) of MSQ_{within} are indicative for a comparable variability between genes.

To judge the distributions, $MSQ_{between}$ was calculated for artificial group means of log_2 expression values for three time points based on four replicates. The group means used were $(6, 6, 7)$ which resulted in an $MSQ_{between}$ of 1.33 (Eq. 3.2, page 54), indicated by a dashed gray line in Figure 4.3. The mean expression values of the artificial groups have been chosen such that they exhibit a log_2 ratio of 1 when group 3 is compared to group 1 or group 2, reflecting a relevant difference between those groups. A good normalization method should result in similar expression values for replicates and thus in small MSQ_{within} values hardly crossing this arti-

ficial $MSQ_{between}$. Additionally, since we limited the whole data set to expressions measured for untreated HaCaT cells across time, we expect only few genes to be differentially expressed. Thus, only a few genes are assumed to result in an $MSQ_{between}$ above the artificial threshold of $MSQ_{between} = 1.33$.

Almost all boxplots representing MSQ_{within} of background normalized data (`bg_*`) result in outliers crossing the artificial $MSQ_{between}$, only those transformed using vst do stay below. Compared to other pre-processing methods, `noBg_vst_loess`, `noBg_log_loess`, and `bg_vst_loess` show a relatively wide IQR for both, $MSQ_{between}$ and MSQ_{within}. Methods that meet the described behavior are `bg_vst_quantile`, `bg_vst_rsn`, `noBg_log_quantile`, `noBg_log_rsn`, `noBg_noNorm`, `noBg_vsn`, `noBg_vst_quantile`, and `noBg_vst_rsn`. They show a low within group variability for which the quantiles generally exhibit lower values than the quantiles of the between group variabilities.

Density Functions of $MSQ_{between}$ and MSQ_{within}

Density functions of $MSQ_{between}$ and MSQ_{within} should exhibit clear differences. This fact renders density functions of $MSQ_{between}$ and MSQ_{within} an additional option for investigating these values. Within group variability should be smaller than between group variability and most of the genes should show a between group variability similar to the within group variability, i. e. are not differentially expressed. Thus, the mode of MSQ_{within} should be smaller than the mode of $MSQ_{between}$ and the peak of the function for MSQ_{within} is supposed to be higher than the peak for $MSQ_{between}$. Lean MSQ_{within} functions, on the one hand, reflect a comparable within group variability for many genes. On the other hand, broader $MSQ_{between}$ functions indicate that at least some of the genes, i. e. the differentially expressed ones, show a higher between than within group variability. Ideal characteristics of density functions as described here are very similar to the characteristics of ideal boxplots mentioned in the previous section. In contrast to density functions, boxplots give a very rough idea about the distribution of the values, also depicting outliers. Density functions deliver a more detailed view of how the values are distributed across different ranges.

Figure 4.4 displays density functions of MSQ_{within} (blue) and $MSQ_{between}$ (red)

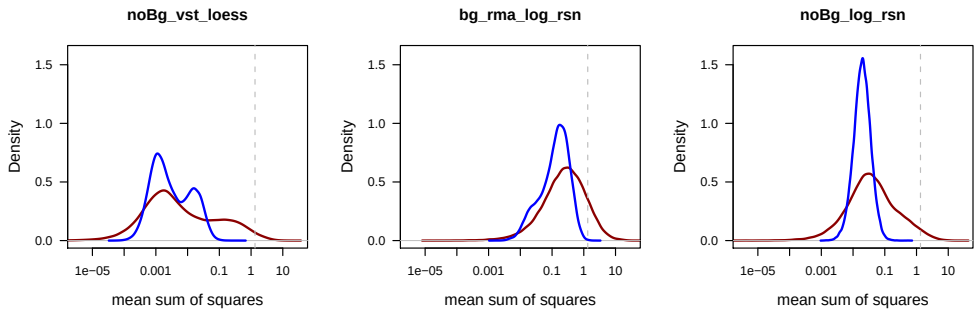


Figure 4.4: **Density plots of MSQ_{within} (blue) and $MSQ_{between}$ (red).** Mean sum of squares were calculated based on the gene expression measured for the three sample groups analyzed, namely untreated HaCaT cells cultured for 2, 4, and 12 hours (Eqs. 3.1 and 3.2, page 54). Each of the groups is composed of four replicates. The gray dashed line indicates the expected value for the $MSQ_{between}$ of 1.33 based on 6, 6, and 7 as measurements for the group means of four replicates for three time points. Three examples of different quality are shown. Based on the `noBg_vst_loess` pre-processing the MSQ_{within} values show a strong bimodal distribution, for the `bg_rma_log_rsn` pre-processing the distribution is highly skewed. The distributions for `noBg_log_rsn` based values reflect the desired behavior. The quality values assigned are -2, 0, and 2, for `noBg_vst_loess`, `bg_rma_log_rsn`, and `noBg_log_rsn`, respectively. For an overview of the distributions for all pre-processing methods and their respective plots, see [53] or Appendix B Figure B.1.

for three of the pre-processing methods, a complete overview is given in [53]. In particular density plots representing the normalization methods `noBg_log_quantile`, `noBg_log_rsn`, and `noBg_vsn` come close to the desired behavior. Unexpectedly the density functions of MSQ_{within} generated by `bg_vst_loess`, `noBG_log_loess`, and `noBg_vst_loess` are bimodal. One reason for bimodal density functions could be a group of transcripts exhibiting higher variability compared to other transcripts. In general, it is expected that the data shows a consistent variability. Having the opportunity to choose between normalization methods resulting in unimodal or bimodal density functions for MSQ_{within}, normalization methods leading to a unimodal distribution should be favored. A small overlap of the functions like for the values generated by the `noBg_average` normalization (Appendix B Figure B.1) would indicate the unlikely event that most of the genes show a higher between than within group variability, i. e. are differentially expressed. Under the assumption of constant

expression of most transcripts, this normalization method is not adequate.

Volcano Plots

Volcano plots are a standard visualisation of results for microarray differential expression analysis. They are generated by plotting $-log_{10}(p - value)$ versus the respective log_2 ratios. The former measures the significance of the change, while the latter (ratio of mean intensity in group 1 over mean intensity in group 2) measures the extent of this change. Due to the tendency of larger log_2 ratios being connected to more significant $-log_{10}(p - values)$ a volcano like shape is generated. Again using untreated HaCaT cells cultured for 2, 4, and 12 hours, pairwise comparisons (4 hours compared to 2 hours, 12 hours compared to 2 hours, and 12 hours compared to 4 hours) using a moderated t-statistic were performed to calculate log_2 ratios and p-values (Section 3.2.2, page 55, [165]). Our aim is to detect normalization procedures yielding as correct estimates of log_2 ratios as possible combined with as informative p-values as possible. As mentioned above, higher log_2 ratios should tend to have higher $-log_{10}(p - value)$. The loess fits of the log_2 ratios and $-log_{10}(p - value)$ pairs (black curves) of the volcano plots shown in Figure 4.5 shall neither be too flat nor too narrow and the scatter of the p-values for specific log_2 ratios should not be too large.

All volcano plots based on `rma` background corrected data show undesired characteristics. The fitted curves are rather flat, i. e. even for high absolute log_2 ratios the $-log_{10}(p - value)$ are relatively low. Additionally, the $-log_{10}(p - value)$ for similar log_2 ratios tend to scatter extremely. Some plots, e.g. for `bg_average`, `bg_noNorm`, `bg_rankInvariant`, and `noBg_rankInvariant`, show an asymmetrical relation between p-values for negative and positive log_2 ratios. Especially `noBg_rankInvariant` exhibits a bias towards small negative log_2 ratios for which the respective $-log_{10}(p - value)$ seems to be relatively high. In this region the fitted curve shows a very steep, linear course. Volcano plots generated for all other methods are similar to what would be expected. Still they differ in the variance of the p-values and in that some of the fitted curves show a flatter shape than others. This reflects the fact that some normalization methods generate a smaller variance than others, resulting in lower fold changes but more significant p-values. Ultimately, a method with a reasonable trade-off between fold

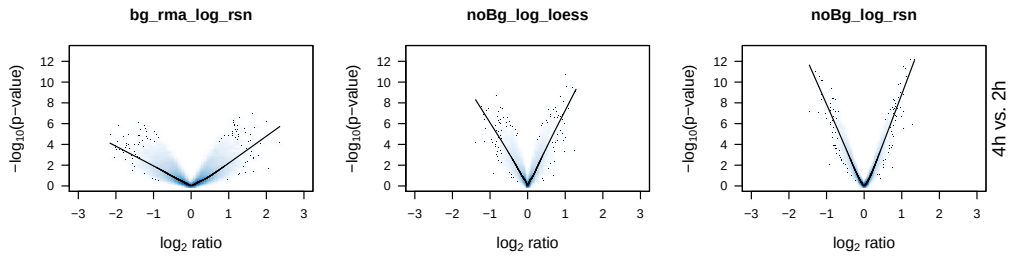


Figure 4.5: **Volcano plots.** log_2 ratios and p-values for the comparison of untreated HaCaT cells at 4 hours compared to 2 hours, 12 hours compared to 2 hours, and 12 hours compared to 4 hours were calculated based on the gene expression measured. Three examples of different qualities are displayed showing the $-log_{10}(p-value)$ against log_2 ratio comparing 4 hours to 2 hours. The black line represents a loess-curve fitted to the values. Quality values assigned to `bg_rma_log_rsn`, `noBg_log_loess`, and `noBg_log_rsn` are −2, 0, and 2, respectively. Pre-processing using `bg_rma_log_rsn` yields a very flat volcano like shape with p-values exhibiting a high degree of scattering, i. e. log_2 ratios are overestimated and at the same time p-values are not very accurate. In contrast, `noBg_log_loess` better represents the expected range of log_2 ratios (not many genes are assumed to heavily change their expression between the different time points), but compared to `noBg_log_rsn` p-values still are not very accurate, i. e. show a high degree of scattering for equivalent log_2 ratios. For a complete overview of all methods and all comparisons, see [53].

change and variance has to be chosen, and cut-off parameters for interesting genes have to be defined accordingly. Volcano plots that best reflect the desired properties in the context of our experiment were generated by `noBg_log_quantile`, `noBg_log_rsn`, and `noBg_vsn`. They show the least scattering of values around the fitted curves, but, as indicated by the steep fitted curves, they probably underestimate fold changes. Plots produced by `noBg_cubicSpline`, `noBg_log_loess`, `noBg_vst_loess`, `noBg_vst_quantile`, `noBg_vst_rsn`, `bg_forcePos_log_loess`, `bg_forcePos_log_quantile`, `bg_forcePos_log_rsn`, `bg_vst_quantile`, and `bg_vst_rsn` also fulfill the above mentioned criteria, but show more scattering.

Residual Standard Deviation Against Mean and Minimum of Gene Expression Levels

In an optimally normalized experiment the residual standard deviation of fitted gene expression intensities should be low and independent of the expression levels, i. e.

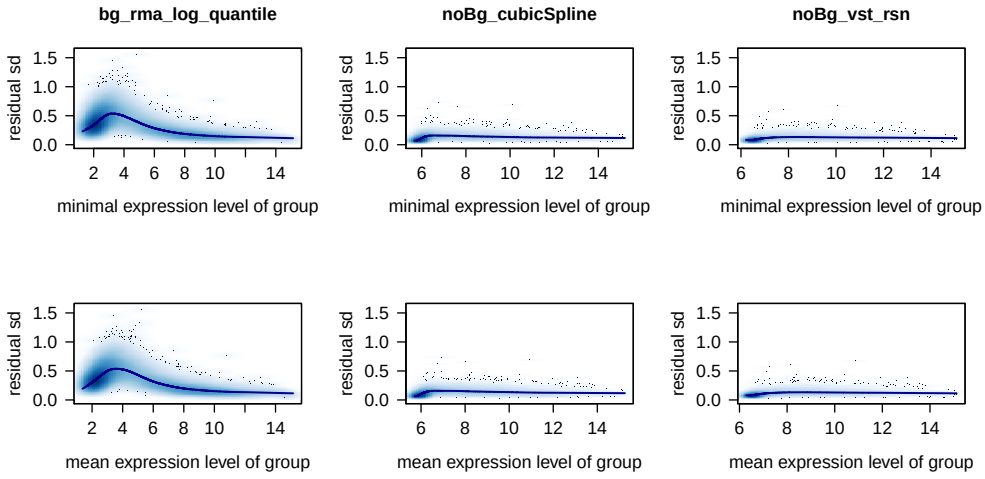


Figure 4.6: **Residual standard deviation against expression intensities.** Standard deviation of the residuals (residual sd) observed for gene expression intensities are plotted against minimum (upper row) and mean (lower row) expression intensity of each transcript. The blue line represents a loess-curve fitted to the values. `bg_rma_log_quantile` exhibits very high deviation of residuals in ranges of lower expression intensities, whereas `noBg_vst_rsn` shows homogeneous and low deviations of residuals over the whole range of expression intensities. Compared to `noBg_vst_rsn`, residual standard deviations tend to be a bit higher and less homogeneous in small ranges of expression intensities when `noBg_cubicSpline` is used for pre-processing. Thus, scores of -2, 0, and 2 are assigned for `bg_rma_log_quantile`, `noBg_cubicSpline`, and `noBg_vst_rsn`, respectively. For an overview of all methods, see [53].

the variance over the different expression levels should be stable. This is prerequisite for many statistical methods, like for example linear model fitting and moderate t-statistics [165], that are utilized for analyzing gene expression data.

As indicated in Figure 4.6, all of the methods without background normalization (`noBg_*`) show a moderate or low variance in regions of no or hardly-to-measure expression. In contrast, nearly all of the background corrected methods (`bg_*`) result in high and, compared to the other methods, instable variance in the range of low intensity values. Extreme examples especially are `rma` background corrections and `bg_vsn` normalization procedures. An exception are those methods that use background normalization in conjunction with variance-stabilizing transformation

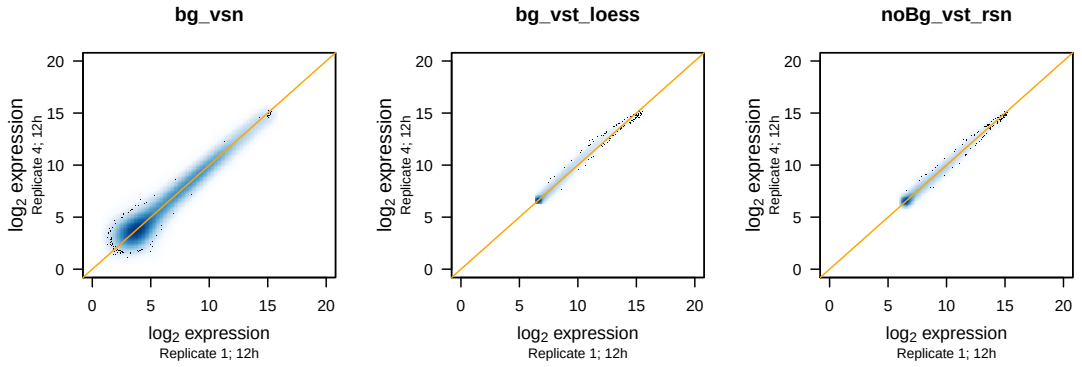


Figure 4.7: **Scatterplots between replicates.** After application of different normalization methods, expression values for the respective replicates at 12 hours are plotted against each other. `bg_vsn` as well as `noBg_vst_rsn` display a symmetrical distribution of expression values around the main diagonal (orange line), with `bg_vsn` exhibiting more scattering values especially obvious in regions of low expression. The scatterplot based on `bg_vst_loess` is slightly bended towards the upper diagonal based on a bias to higher values in the expression values for replicate 4. Scores of -2, 1, and 2 are assigned to the scatterplots based on `bg_vsn`, `bg_vst_loess`, and `noBg_vst_rsn`, respectively. For an overview of all methods, I refer the reader to [53].

(`bg_vst_*`), which in contrast to other procedures using background corrections perform especially well. Methods which perform best with respect to variance stabilization across all expression levels are `bg_vst_loess`, `bg_vst_quantile`, `bg_vst_rsn`, `noBg_vst_loess`, `noBg_vst_quantile`, and `noBg_vst_rsn`.

Scatterplots of Expression Values

Scatterplots are an easy and straightforward visualisation tool for judging the comparability of replicates. They clearly show whether high variances are to be expected and, if this is the case, in which range of the expression data. Figure 4.7 displays the expression values of replicates plotted against each other. The results confirm our previous findings. Some of the methods, for example `bg_rma_log_loess`, `bg_rma_log_quantile`, `bg_rma_log_rsn`, and `bg_vsn`, show high variance especially in the range of lower expression. Plots generated based on these procedures exhibit high variability between replicates. Some of the methods like for example `bg_noNorm`, `bg_vst_loess`, and `noBg_average` lead to asymmetric scat-

terplots indicating a bias in the expression values and a higher variability between replicates. Methods that perform well in stabilizing the variance across different expression levels, for example `bg_vst_quantile`, `bg_vst_rsn`, `noBg_vst_loess`, `noBg_vst_quantile`, and `noBg_vst_rsn`, could also be confirmed by the scatterplots. Additionally to those, `noBg_cubicSpline` and `noBg_rankInvariant` exhibit symmetric scatterplots with a very low degree of variance between replicates.

Pseudo-ROC Curves

In order to compensate for missing spike-in and dilution data, a pseudo-ROC approach [164] mimicking the presence of true negatives has been conducted. True positives were selected from bona fide target genes of TGF-β, thus they are expected to change in expression upon stimulation of TGF-β signaling (Section 3.2.2, page 54). The pseudo-ROC curve for each normalization method is a linear transformation of the true ROC curve. Common single number summaries used to score and compare ROC curves - the area under the curve (AUC) or the sensitivity at a given false positive rate - are area or distance based. They are reduced by this transformation, but to the same degree for every curve. Aiming at the validation of normalization methods with respect to their ability to generate data exhibiting a good sensitivity to specificity ratio, expression intensities derived from TGF-β-treated versus untreated cells at 2 hours were compared. Based on the AUC of the pseudo-ROC curves (Figure 4.8), all normalization methods perform relatively well in delivering values suited for separating true positives from true negatives. To assign quality values to the ROC curves, the AUC values were sorted and subsequently allocated to three bins of sizes 5, 18, and 2. Finally the bins were assigned quality values of -1, 0, and 1, respectively (Figure 4.12). `bg_rankInvariant` performs best with an AUC of 0.9102, whilst `bg_vst_loess` performs worst with an AUC of 0.8403.

4.1.2 Analyses of Bias Based on qRT-PCR

qRT-PCR has been performed for mRNAs from eight genes that are known to be deregulated by TGF-β signaling to a varying degree, namely CDKN1A, CDKN2B, HAND1, JUNB, LINCR, RPTN, SERPINE1, and TSC22D1. By this means, it is possible to compare the results of the normalization methods to values that reflect

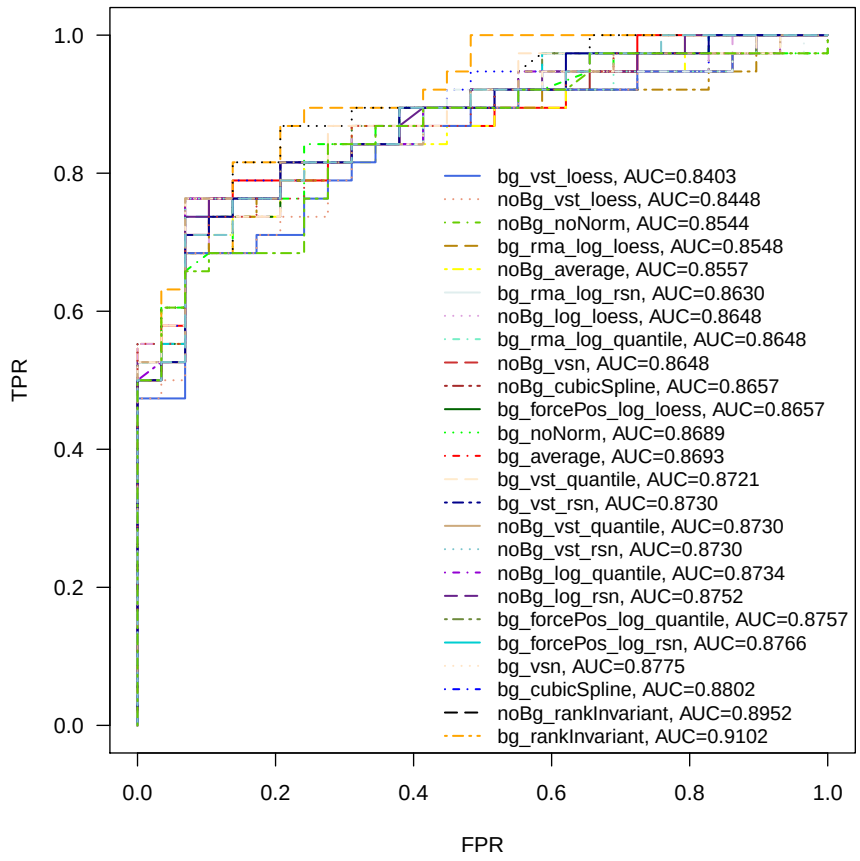


Figure 4.8: **Pseudo-ROC curves based on adjusted p-values.** Pseudo-ROC curves were calculated for the different pre-processing methods. FDR-adjusted p-values [163] of an F-statistic (Eq. 3.3, page 3.35) comparing the expression intensities measured for untreated and TGF-β-stimulated HaCaT cells cultured for 2 hours were used. Each of the two groups is composed of four replicates. (TPR: true positive rate, FPR: false positive rate, AUC: area under the curve).

the real abundance of the respective mRNA in the cells. Thus, it was possible to evaluate the accuracy of the different pre-processing methods with respect to their bias. To guarantee that the comparisons of the normalization methods are not biased towards certain intensities, the mRNAs used in qRT-PCR experiments were chosen such that the respective signals on the chips cover a broad range of expression intensities.

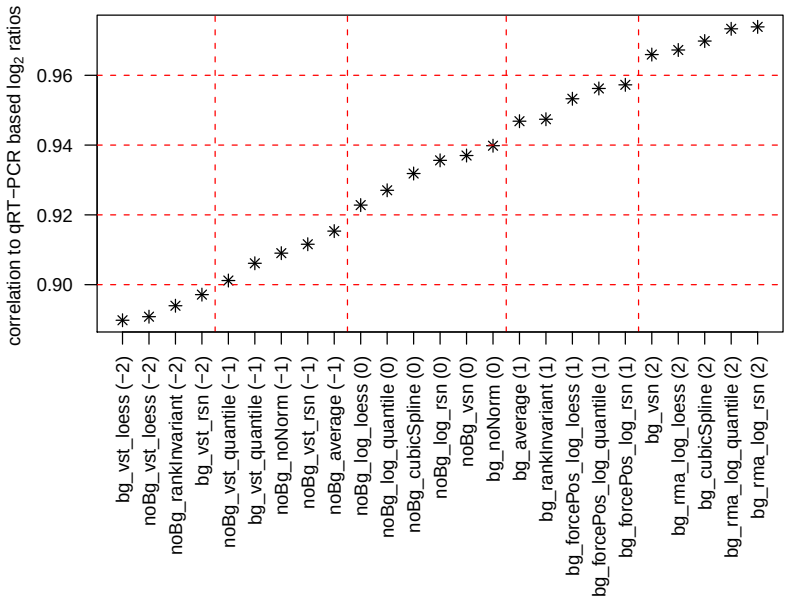


Figure 4.9: **Pearson correlation of log_2 ratios for different normalization methods and qRT-PCR.** Correlations of log_2 ratios were calculated for differently pre-processed gene expression data from BeadChip arrays and qRT-PCR based results. On the x-axis, pre-processing methods are ranked according to their correlation to qRT-PCR. The dashed red lines indicate the cut-offs used for assigning quality score between -2 (≤ 0.9) and 2 (≥ 0.96).

Correlation Analysis of Fold Changes

Based on the different normalization procedures for the gene expression experiment and based on the qRT-PCR measurements, Pearson correlations of the respective fold changes measured for TGF-β-stimulated versus untreated cells at 2, 4, and 12 hours were calculated. Figure 4.9 displays the ranked correlation coefficients describing the relation between the different normalization methods and the qRT-PCR results. Quality values were assigned based on correlation cutoffs. A value of 2 is assigned to correlation coefficients ≥ 0.96, a value of 1 to coefficients between 0.96 and 0.94, a value of 0 to coefficients ≤ 0.94 and ≥ 0.92, a value of -1 to coefficients between 0.92 and 0.9, and a value of -2 to correlation coefficients ≤ 0.9 (Figure 4.12).

Values derived from most of the methods not utilizing background correction (`noBg_*`) show a lower correlation to the qRT-PCR results than expression intensities

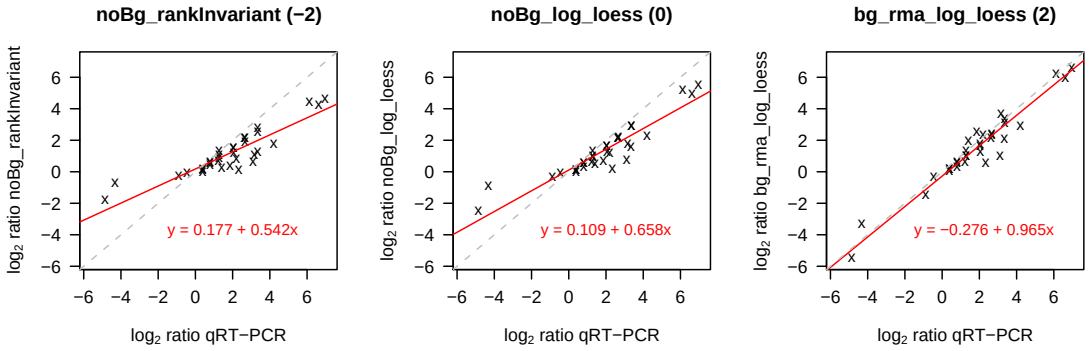


Figure 4.10: **Orthogonal regression between qRT-PCR and normalization based log_2 ratios.** Regression of log_2 ratios was conducted based on different normalization methods (y-axis) against qRT-PCR (x-axis). Equations and the respective regression lines are displayed in red. The gray dashed line indicates the main diagonal. Compared to qRT-PCR, log_2 ratios as calculated based on `noBg_rankInvariant` and `noBg_log_loess` pre-processing are overestimated in the lower and underestimated in higher ranges of log_2 ratios. This over- and underestimation is more extreme for `noBg_rankInvariant` (intercept = 0.177, slope = 0.542) than for `noBg_log_loess` (intercept = 0.109, slope = 0.658). Data pre-processed using `bg_rma_log_loess` hardly over- or underestimates the data (intercept = −0.276, slope = 0.965). This results in scores of −2, 0, and 2 for `noBg_rankInvariant`, `noBg_log_loess`, and `bg_rma_log_loess`, respectively. An overview of the results for all pre-processing methods is given in [53].

that are background corrected (`bg_*`). An exception in this regard are methods that are based on `vst` transformation (`bg_vst_*`). These three methods are amongst the six methods resulting in the lowest correlation coefficient values. Correlation coefficients exhibiting high values are delivered by methods introducing BeadStudio's background correction combined with either `rma` background correction and log_2-transformation (`bg_rma_log_*`), cubic spline normalization (`bg_cubicSpline`), or variance stabilizing normalization (`bg_vsn`).

Regression Analysis

To investigate the linear relationship between fold changes as determined by gene expression data and qRT-PCR, a linear regression analysis was performed by minimizing the sum of squares of the Euclidean distance of points to the fitted line ("orthogonal regression", Figure 4.10). This method was chosen because there is no

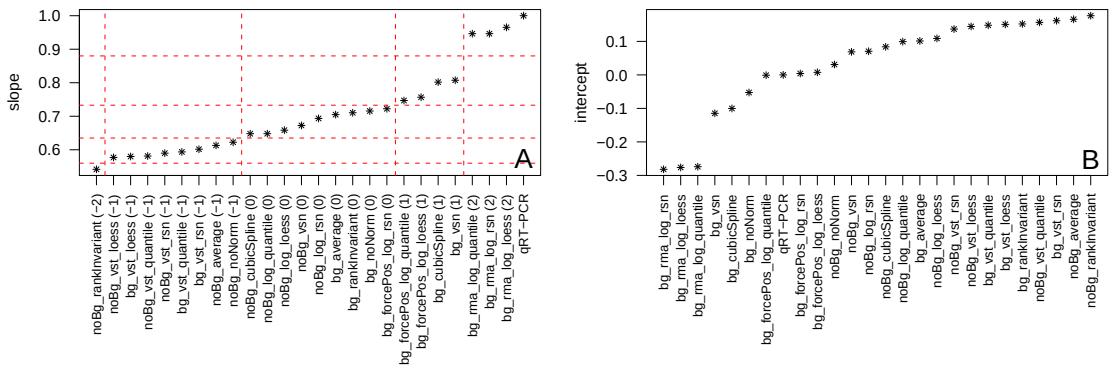


Figure 4.11: **Results of orthogonal regression.** Ranking of slope (A) and intercept (B) of the orthogonal regression lines as exemplary displayed in Figure 4.10. For the slope, quality scores are assigned from −2 to 2 as indicated by the red dashed lines.

clear assignment of dependent and independent variables (Section 3.2.2, page 55). Figure 4.11 displays the ranking of the different methods according to the slopes of the orthogonal regressions. Following rules apply for results of these analyses: The closer the slope is to 1, the better the respective normalization method reflects the qRT-PCR results in a linear manner. In this situation the deviation of the intercept from 0 indicates a constant under- or overestimation of the change of mRNA abundance across the whole range of fold changes. An intercept < 0 stands for an underestimation and an intercept > 0 for an overestimation of fold changes. In the case that the slope deviates from 1 the difference between qRT-PCR based fold changes and normalized expression based fold changes depends on the size of the fold change. Here, on the one hand, an intercept near 0 implies a continuous over- (slope > 1) or underestimation (slope < 1). Depending on the slope, an intercept deviating from 0, on the other hand, indicates overestimation for a certain range of values and underestimation for another range of values. Regardless of the intercept, the most relevant point in our case is that the scatterplots are generally linear, with low variability and a slope close to 1.

In accordance to previous results, all expression values that are transformed using `vst` together with `noBG_rankInvariant` result in slopes that exhibit the largest deviation from 1. Fold changes calculated based on `rma` background correction and log_2-transformation (`bg_rma_log_*`) best fit the qRT-PCR results (Figure 4.11).

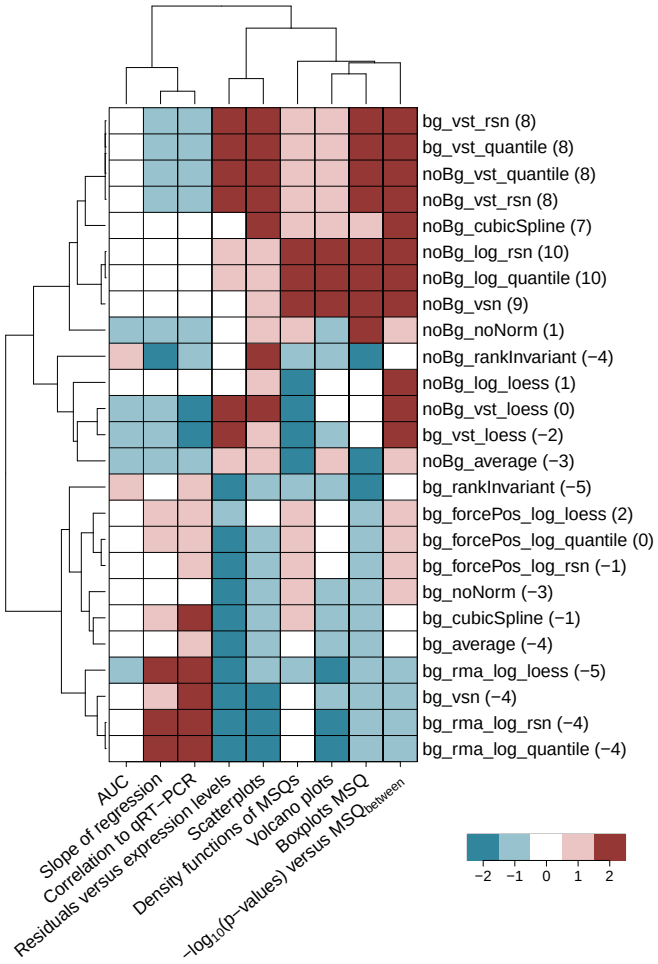


Figure 4.12: **Heatmap of quality scores assigned for the different pre-processing methods.** Displayed are the quality scores for the different pre-processing methods given for the analyses conducted. Quality scores range from −2 (bad) to 2 (good). The values in parentheses display the sum over the single quality scores for the respective pre-processing procedures. Based on this sum, the pre-processing method finally used to normalize the expression data has been chosen. Manhattan distance and complete linkage algorithm were used for clustering. Methods evaluating the bias (slope of regression, correlation to qRT-PCR) are clearly separated from methods evaluating the variance. Pre-processing procedures that perform best based on the sum over quality scores are located at the top of the heatmap.

4.1.3 Summary

After evaluating each of the twenty-five normalization procedures and assigning scores ranging from -2 to 2, the methods were clustered based on the scores as-

signed for the considered quality measures. Methods evaluating the bias (slope of regression, correlation to qRT-PCR) are clearly separated from methods evaluating the variance. Pre-processing procedures that perform best based on the sum over quality scores are located at the top of the heatmap (Figure 4.12). The bottom part of the heatmap is dominated by background normalized data sets which, on average, score worse than those that have not been background normalized. Based on the sum over the individual scores, the pre-processing method used to normalize the expression data has been chosen. As can be seen in Figure 4.12, `noBg_log_rsn` and `noBg_log_quantile` achieve the same overall score. Based on the results presented in this section, `noBg_log_rsn` was selected as the final normalization method for further analysis. A detailed discussion on this decision is given in Section 5.1.

4.2 Evaluating New Chemical Entities by State-of-the-Art Approaches

In this section, I describe how the expression data for the seven NCEs (Section 3.1.1) can be analyzed by applying state of the art approaches. Five BI compounds, BI1 to BI5, belonging to the chemical class of indolinones, and two competitor substances, Ex1 and Ex2, belonging to the chemical class of pyridopyrimidinones, with inhibitory potency towards TGF-βR1 kinase have been used. As we are interested in revealing MoA of the NCEs, two effects are of interest, the on- and the off-target effect. On-targets are those proteins, that are intended to be hit, i. e. in our case inhibited, by the compound. The compounds used in this study bind to the ATP pocket of the TGF-βR1 kinase domain. Thereby binding of ATP as well as phosphorylation of proteins downstream in the signaling cascade is prevented and, thus, signal transduction is inhibited. In contrast to on-targets, off-targets are all those proteins that are hit/modulated by the compound in addition to the intended target. ATP pockets are domains highly conserved among kinases, leading to a high potential of off-target activities [87–89]. Effects caused by on- and off-target activities of compounds are referred to as on- and off-target effects, respectively.

In Section 4.2.1 the on-target signature is derived. As the NCEs under investigation were designed to inhibit TGF-βR1, the on-target signature refers to those genes that are deregulated by TGF-β signaling. The degree of deregulation should depend

on compound treatment in a dose dependent manner. Genes of this signature should further be regulated in opposite directions for TGF-β-stimulated cells compared to unstimulated cells and compound treated plus TGF-β-stimulated cells compared to TGF-β-stimulated cells. Based on the genes contained in this signature, KEGG pathways are checked for enrichment of those genes.

In Section 4.2.2 an off-target signature is derived for each of the seven NCEs. An NCE's off-target signature is composed of that set of genes that is deregulated due to off-targets of the respective NCE. Different gene sets were checked for enrichment of off-target signature genes in Section 4.2.3.

4.2.1 Effects Related to TGF-β Signaling - On-Target Signature

In order to gain a deeper insight into the TGF-β biology we first identified genes that are differentially expressed comparing TGF-β1-stimulated to unstimulated HaCaT cells. To unravel the time-dependent effects of TGF-β treatment, HaCaT cells were stimulated with TGF-β1 for 2, 4, and 12 hours. While immediate early genes that are directly regulated by the TGF-β pathway are detected at 2 hours post stimulation, more and more secondary effects linked to TGF-β signaling are found after 4 and/or 12 hours. Additionally, genes that are directly linked to TGF-β signaling pathway should be regulated in a dose dependent manner with respect to compound treatment. Thus, to avoid arbitrary fold change cut-offs we do not only make use of the TGF-β1-stimulated compared to untreated cells, but also consider the compound treated cells to derive a TGF-β signature.

Two criteria were applied to identify TGF-β signature genes (Section 3.3.2): First, genes that were significantly deregulated (p-value ≤ 0.01) in a basic comparison of TGF-β-stimulated versus unstimulated cells were selected for further analysis (Section 3.3.2: i, page 56). 1046, 1949 and 5725 genes (6525 non-redundant genes) were found to be regulated 2, 4 and 12 hours after stimulation. In a second step, these genes were proven to be affected in a dose-dependent manner in the opposite direction to the TGF-β-effect by NCE treatment after TGF-β stimulation using a likelihood ratio test statistic for monotonicity [166] (Section 3.3.2: ii+iii, page 56). By

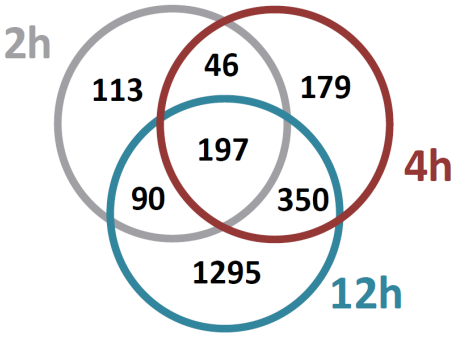


Figure 4.13: **Venn diagram for the on-target TGF-β signature** The list of genes 446, 772 and 1,932 genes were identified as the NCE-dependent on-target TGF-β signature 2, 4, and 12 hours after NCE treatment and TGF-β stimulation.

this means these genes can be separated from potential compound related off-target effects. All transcripts identified for each NCE were merged to a common signature of TGF-β dependent genes (Section 3.3.2, page 56). Thereby, the identification of a common on-target signature focused on minimizing the amount of false positive and false negative genes. The Venn diagram displayed in Figure 4.13 depicts the number of genes that were identified 2, 4, and 12 hours after stimulation: 446 genes (2 hours), 772 genes (4 hours) and 1932 genes (12 hours), respectively. A comprehensive list of genes assigned to the TGF-β signature, i.e. the on-target signature, can be found in [54].

Subsequently, the genes comprising the TGF-β signature were checked for enrichment in certain gene sets. Gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [114, 188] corresponding to 201 different pathways were used. Applying Fisher's exact test resulted in 16 different signaling pathways which were significantly enriched by genes that are deregulated upon TGF-β stimulation. By clustering the respective p-values these 16 pathways could be divided into four groups that correspond to the time frame of affectedness (Figure 4.14).

Not surprisingly, the TGF-β signaling pathway itself as well as directly affected pathways like WNT and p53 signaling were significantly regulated by the treatment of TGF-β (cluster 1). In cluster 2, signaling by MAPK, cytokines, ErbB, Shh, as well as apoptosis are strongly affected immediately early upon TGF-β stimulation. The modulation is reduced at later time points (4 and 12 hours), when more secondary

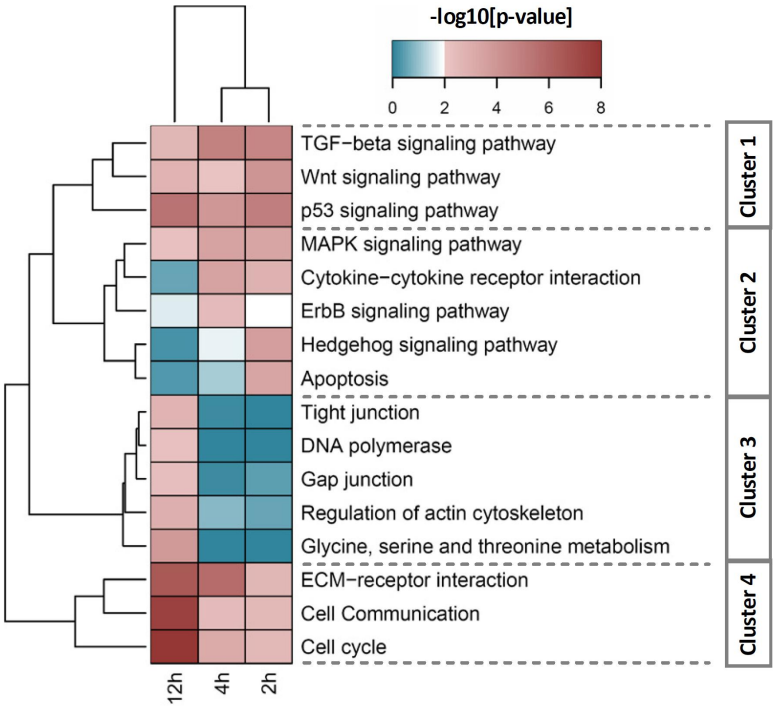


Figure 4.14: **Gene set enrichment analysis using KEGG pathways.** Enrichment analysis using Fisher's exact test resulted in 16 significantly affected gene sets/signaling pathways. Clustering of $-log_{10}$ p-values using complete linkage and manhattan distance resulted in four major clusters: immediate early affected pathways (cluster 2), permanently affected pathways with emphases at early (cluster 1) and late time points (cluster 4) or late established events (cluster 3). The color code defines the significance determined by Fisher's exact test.

effects, such as DNA polymerase, actin cytoskeleton, amino acid metabolism, gap junction, and tight junction signaling become apparent (cluster 3). The activation of these pathways in combination with the modulation of the cell cycle and cell communication activity (cluster 4) seem to relate to phenotypic consequences of TGF-β stimulation.

4.2.2 Inferring the Off-Target Effects

After the identification of the TGF-β signature (on-target signature) as well as the related pathways, we tried to identify the NCEs' off-target effects. This was done by first deriving an off-target signature solely based on gene expression as described in Section 3.3.3. Second, the genes of the respective signatures were checked for

enrichment within gene sets describing molecular functions and canonical pathways as defined by Ingenuity Pathway Analysis software [144].

Each compound treatment resulted in a unique gene expression signature of regulated genes. These signatures are composed of the cellular response to two different stimuli, i. e. TGF-β1 and NCE stimulation. Thereby, elucidating the effects based on NCE treatment is more demanding since both TGF-β, i. e. on-target as well as off-target effects occur. We also observed interaction effects of the vehicle (DMSO) with the NCEs. The effects of the different stimuli overlap and also interfere with each other constraining a clear signature resolution. The profile of a given gene may therefore be dependent on which effect prevails and thus, dose-dependency might no longer be observed [52, 54]. That is why we had to come up with a definition of the off-target signature that allows a certain degree of variability by avoiding a stringent cutoff. Details on how the final off-target signature is derived are described in Section 3.3.3. By applying the described strategy, we empirically observed a good trade-off between false positives and false negatives in the off-target signatures.

In a first approximation, the NCE treatment phenotypes were determined as the total of all regulated genes (p-value < 0.01 and $|log_2 ratio| \geq 1$) comparing NCE-treated and TGF-β-stimulated cells to DMSO control-treated and TGF-β-stimulated cells. This analysis was done separately for each of the tested compounds at each concentration. Subsequently, the different phenotypes obtained after 2 hours NCE treatment were clustered to unravel similarities between the different signatures (Figure 4.15). The early time point allowed focusing on primary affected genes that were altered as direct response to the treatment. Hierarchical clustering clearly revealed two major clusters separating the group of indolinones (BI1 to BI5) from the pyridopyrimidinones (Ex1 & Ex2). Thus, the classical concept of chemotypes determining biological profiles of NCEs holds true in this case. More details about the influence of chemotypes and different side chains are given [52]. Additional to the separation based on chemotypes, clusters also depend on high and low dose of the compound.

Identifying a particular off-target based on this approach is difficult. Further analyses were therefore performed to separate the compounds' off-target effects from the

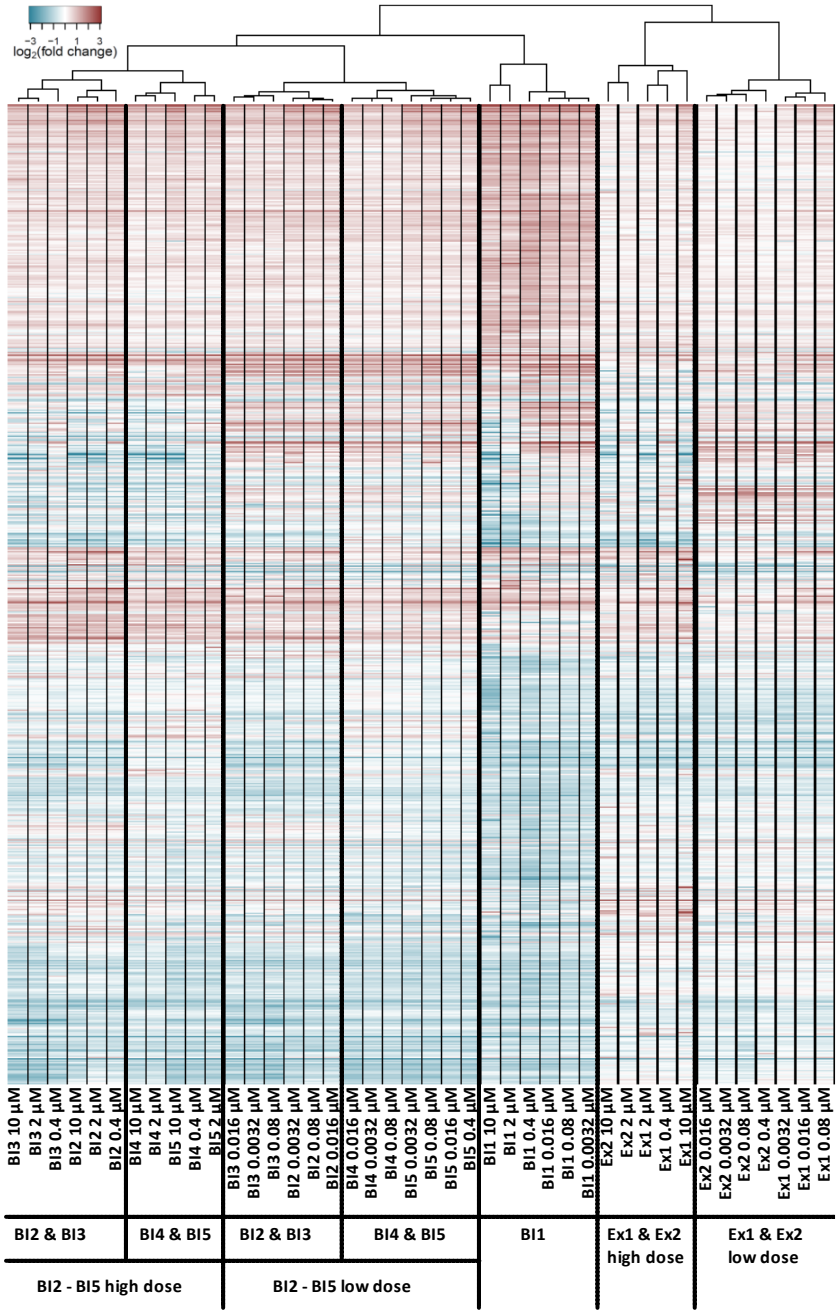


Figure 4.15: **Hierarchical clustering of genes differentially expressed due to NCE treatment.** 4,314 genes were found to be significantly deregulated ($|LR| \geq 1$ and $p - value < 0.01$) by at least one NCE. The different treatment groups were hierarchically clustered according to the correlation coefficients of the respective genes using complete linkage. The five indolinones (BI1-BI5) are grouped and separated from the two pyridopyrimidinones (Ex1 & Ex2). Additionally, clusters show grouping by high vs. low dose treatment. The indolinone BI1 separates from the other class members.

treatment signatures. As mentioned above, not all off-target effects can be identified through dose dependent correlation. Reasons for this are overlapping, inverse, and additive effects [54]. Hence, off-targets can only be identified based on NCE-treated samples in presence and absence of the TGF-β stimulus. In brief, all regulated genes (p-value < 0.01 and $|log_2 ratio| \geq 1$) identified by comparing compound-treated cells (either $0.08\mu M$ or $2\mu M$) to DMSO-treated controls were selected. Genes were considered once the regulation was observed during compound treatment upon TGF-β stimulation as well as without TGF-β stimulation. Thereby, it was ensured to select only drug off-target and TGF-β independent alterations. A more detailed description on how these genes are identified is given in Section 3.3.3 on pages 56 ff.. All genes that matched the described criteria were allocated to the off-target signature of the respective NCE after 2, 4, and 12 hours. Based on this analysis, huge differences in the amount of off-target genes were observed. While treatment with BI1 deregulated $2,752$ genes at all time points, BI3 deregulated only 973 genes. Slightly more off-target genes were identified for the indolinones BI2, BI4 and BI5 ($1,050$, $1,064$ and $1,100$, respectively). The pyridopyrimidinones regulated $1,347$ (Ex1) and $1,306$ (Ex2) genes. The largest off-target increase over time was seen for Ex1 and Ex2 with almost four times more genes being regulated comparing the 12 hours to the 2 hours time point. In contrast, the amount of off-targets for the five indolinones was at a maximum doubled within this period.

In summary, looking at the off-target signatures in general, the indolinones except for BI1 appear more favorable compared to the pyridopyrimidinones at later points in time. Among the indolinones, BI2 to BI5 deregulate fewer genes than BI1 at all points in time. This also confirms differences in structure-activity relationships observed by Roth *et al.* [1] for differences in substitution positions of certain residues of BI1 compared to BI2 to BI5. They demonstrate that indolinones such as BI1 showed a less favorable selectivity profile compared to indolinones such as BI2 to BI5. Among the indolinones, BI3 appears to be the most attractive compound when merely looking at the off-target analysis.

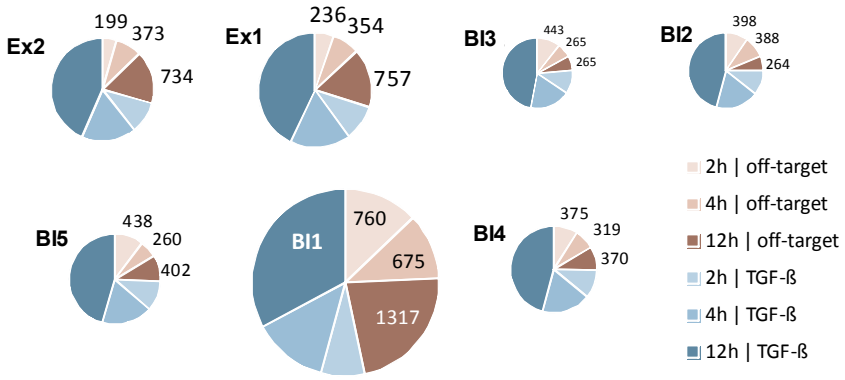


Figure 4.16: **Compound profiles.** Every circle represents one of the seven profiled compounds. The size of each circle corresponds to the number of off-target genes (in red). The relative amount of on-target gene numbers are shown in blue. As described in Section 4.2.1 and displayed in Figure 4.13, the on-target signature is composed of 446, 772, and 1,932 genes at 2, 4 and 12 hours, respectively. The relative amount of on- and off-target genes can easily be seen. Whereas BI1 has both, the highest net amount of off-targets as well as the highest relation of off- compared to on-targets, Ex1 and Ex2 lie between BI1 and BI2 to BI5. BI2 to BI5 constitute the better compounds with respect to both, number of off-targets, which should be low, as well as relation between number of off- compared to number of on-targets.

4.2.3 Mode of Action Analysis Using Existing Approaches

Different *in silico* strategies can be applied to analyze the off-target signatures of the compounds in order to generate hypothesis about their mode of action. As different criteria have been combined to derive the off-target signature, there is no homogeneous and consistent value/measure assigned to the genes contained in this signature that is appropriate to conduct enrichment analysis which rely on measures assigned for each individual gene (see Section 2.4.1). Thus, we decided to conduct enrichment analysis based on Fisher's exact test which does not depend on any measures assigned to the genes but is purely count-based (Section 3.6.2).

Gene Set Enrichment

Initially, the genes from the 12 hours off-target signatures were assigned to their molecular function using Ingenuity Pathway Analysis. Since we were interested in

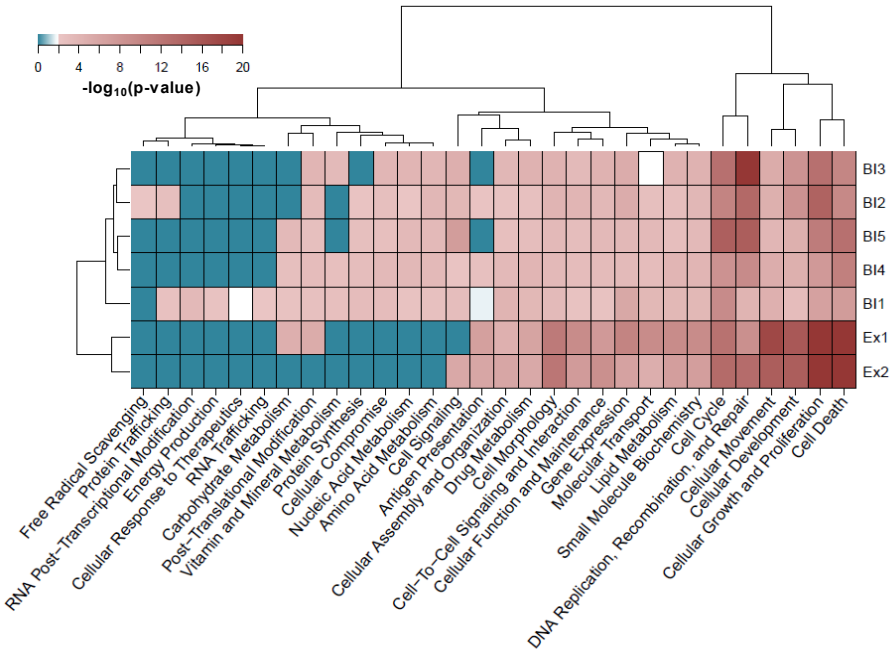


Figure 4.17: **Clustering of gene set enrichment results for off-target genes of all seven NCEs after 12 hours treatment.** Using Fisher's exact test, an enrichment test was conducted based on the molecular function gene sets as defined by the Ingenuity Knowledge Base [144]. For each gene set, a p-value was obtained. Displayed is the clustering of gene sets significant for off-targets of at least one compound. Clustering was calculated based on the $-log_{10}(p-values)$ using complete linkage and manhattan distance.

long term events, this analysis was focused on a late time point. Hierarchical clustering of the functions based on the respective $-log_{10}$(p-values) resulted in one major cluster for both pyridopyrimidinones and in one for the indolinones (Figure 4.17).

The genes regulated by the indolinones are distributed in more classes; genes involved in Vitamin and Mineral Metabolism, Cellular Compromise, Nucleic Acid- and Amino Acid Metabolism are exclusively regulated by the indolinones. In accordance with the structure-activity findings mentioned before, within the indolinone subcluster, BI1 stands apart from the four other indolinones and BI2 and BI3 as well as BI4 and BI5 are grouped in one cluster, respectively. The genes additionally regulated by BI1 are involved in RNA Post-Transcriptional Modification, Energy Production, Cellular Response to Therapeutics and in RNA Trafficking. Half of

the molecular functional classes identified are regulated by all compounds. However, this does not necessarily mean that the same genes are regulated since the categories are rather generally defined, such as Cell Cycle or Cellular Growth and Proliferation. Furthermore, different NCEs reach far higher significance scores for some categories than others caused by the higher amount of regulated genes in the respective biological process, e.g. both pyridopyrimidinones regulate a huge amount of genes involved in Cell Death and Cellular Growth and Proliferation.

In order to get a better understanding of the compounds' MoA Ingenuity Pathway Analysis was used to further analyze the off-target signatures. We found 39 (2 hours), 38 (4 hours) and 51 (12 hours) canonical signaling pathways significantly enriched with off-target genes of at least one of the NCEs (Fisher's exact test, $-log_{10}(p - value) > 2$). While immediate early affected processes can be found 2 hours after compound treatment, the effect of the compounds manifests during late phases. A hierarchical clustering of the pathway analysis representing the 12 hours results is displayed in Figure 4.18. Again the indolinones are separated from the pyridopyrimidinones, indicating that both compound classes share not only a common mode of action like TGF-β inhibition, but also generate a distinct affection of other pathways by their specific off-target function. BI1 results in 5 significantly ranked pathways and the smallest overlap among the indolinones. BI3 affects 15 signaling pathways almost exclusively involved in different cancer pathways. This could hint at carcerogenicity of this NCE. The indolinones BI2 and BI4 regulated genes that are significantly enriched in only 4 (BI2) and 2 (BI4) signaling pathways, respectively. However, pathways such as the Aryl Hydrocarbon Receptor Signaling and the LPS/IL-1 mediated inhibition of RXR function are also significantly ranked high for up to six compounds at all three time points, indicating a more general effect like a xenobiotic response to NCE treatment rather than a true compound specific effect. The highest numbers of significantly affected pathways are found for the two pyridopyrimidinones with 24 (Ex1) and 28 (Ex2). Additionally, genes involved in 29 out of the 51 signaling pathways are exclusively regulated by Ex1 or Ex2 treatment. Confirming previous findings, 12 out of the 51 identified pathways are related to toxicity and cell death. These 12 pathways reach highest significance scores for either Ex1 or Ex2 with 8 being solely affected by the two pyridopyrimidinones indicating a cytotoxic mode of action for both of them. Besides cytotoxicity, these two NCEs

deregulate genes involved in inflammatory processes like IL6 signaling, ERK/MAPK signaling and p38 MAPK signaling. A more detailed discussion also considering development of pathway significance over time is given in Baum and Schmid *et al.* [54].

Experimental Validation

Results from *in silico* analyses strongly implied different induced phenotypes after treatment with specific NCEs. The accuracy of the gene set enrichment based findings were experimentally validated. Results can be found in [52, 54].

4.3 A New Approach for the *in Silico* Analyses of Mode of Action

In Sections 3.4 and 3.5 I introduced a new method to score the protein interactions in a network along with a new module extraction method called modEx. As more and more information on genes and proteins becomes available, it is important to integrate and combine this knowledge in biological network-based analyses. By this means supporting evidence for real interactions is accumulated. At the same time the shared biological context and the functional association of a pair of molecules will be highlighted. The results presented in this section are based on iRefIndex (Section 3.4.1) as an underlying network. In addition to deriving modules or subnetworks solely based on differential expression, the proposed knowledge mining approach enables us to identify modules relevant for the experimental conditions under investigation including proteins that are not necessarily regulated on the mRNA level. The relevance of proteins is based on their biological relatedness which is reflected by a weighting scheme. Information on Gene Ontology annotation (Section 3.4.2) and predictions of common transcription factor binding sites (Section 3.4.3) are integrated for weighting the edges between pairs of proteins (Section 3.4.6). For iRefIndex-based networks, additionally a confidence score based on literature was used (Section 3.4.4) to appropriately increase the edge weights. To transfer the network into the biological context of interest, the expression experiments introduced in Section 3.1 were exemplary used as anchoring points for the

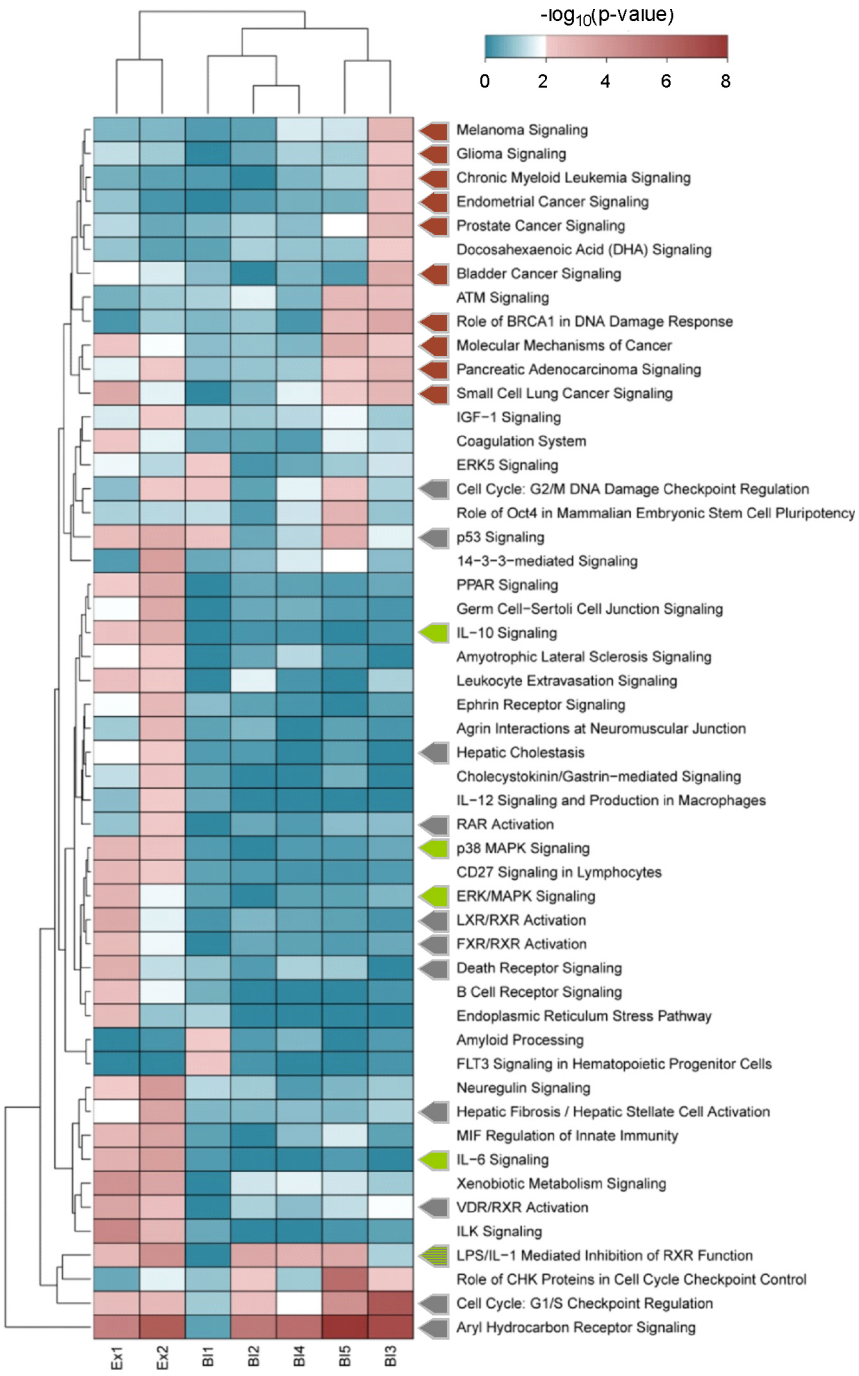


Figure 4.18: **Results of gene set enrichment based on canonical pathways as defined by the Ingenuity Knowledge Base.** Ingenuity pathway analysis (Fisher's exact test) was conducted for the off-target genes of all seven NCEs after 12 hours. Resulting $-log_{10}(p-values)$ for the 51 significantly ranked canonical signaling pathways were clustered using complete linkage and manhattan distance. Off-target genes of BI3 show strong enrichment in 10 cancer signaling pathways (red arrows). Ex1 and Ex2 off-target genes play a role in 12 pathways involved in cytotoxicity or cell death (gray arrows) and in 5 pathways involved in inflammation (green arrows).

analyses as described in Section 3.4.5.

This section is structured as follows: As there are several options to score the different measures used to calculate edge weights, in Section 4.3.1 an optimal combination of evidence is selected. IUsing these results, in Section 4.3.2 the proposed data mining approach together with modEx is evaluated on a biological basis. In Section 4.3.3, I apply the newly proposed approaches to analyze compounds' mode of action with respect to on- and off-target effects. The analyses are based on expression data for BI1 and BI4. To show the benefits of my approach in contrast to approaches only incorporating gene expression data, I compare it to jActiveModule 2.3.1 in Section 4.4. In the same Section, I additionally compare how different networks used as input affect the results obtained by modEx.

4.3.1 Choosing a Reasonable Combination of Evidence

In Section 3.4, in total 360 possible combinations were proposed that could be used to calculate an edge weight (Section 3.4.6, page 71). Since confidence scores based on literature are only available for iRefIndex based networks, the best combination of methods was selected independent of this measure. To decide which method is best suited for integration a ROC curve based approach was used. The main idea is to identify that combination of scoring methods that best distinguishes biologically more related pairs of proteins from less related ones. One would assume that proteins which physically interact are more likely biologically related than randomly sampled pairs of proteins. As true positives (TPs), i. e. biologically related proteins, all proteins that interact based on the iRefIndex network were used. As true negatives (TNs), i. e. less likely and less probable biologically related pairs of proteins, randomly sampled pairs of proteins of the iRefIndex network that were not connected by an edge were selected. This approach is comparable to a pseudo-ROC approach [53, 164]. The pseudo-ROC curve for each of the 360 combinations of scoring methods is a linear transformation of the true ROC curve. Common single number summaries used to score and compare ROC curves - the area under the curve (AUC) or the sensitivity at a given false positive rate - are area or distance based, and thus reduced by this transformation, but to the same degree for every curve. Thus, even if there are pairs selected as TN though they are TP this artifact occurs in each of the pseudo-ROC curves and it is still possible to compare them

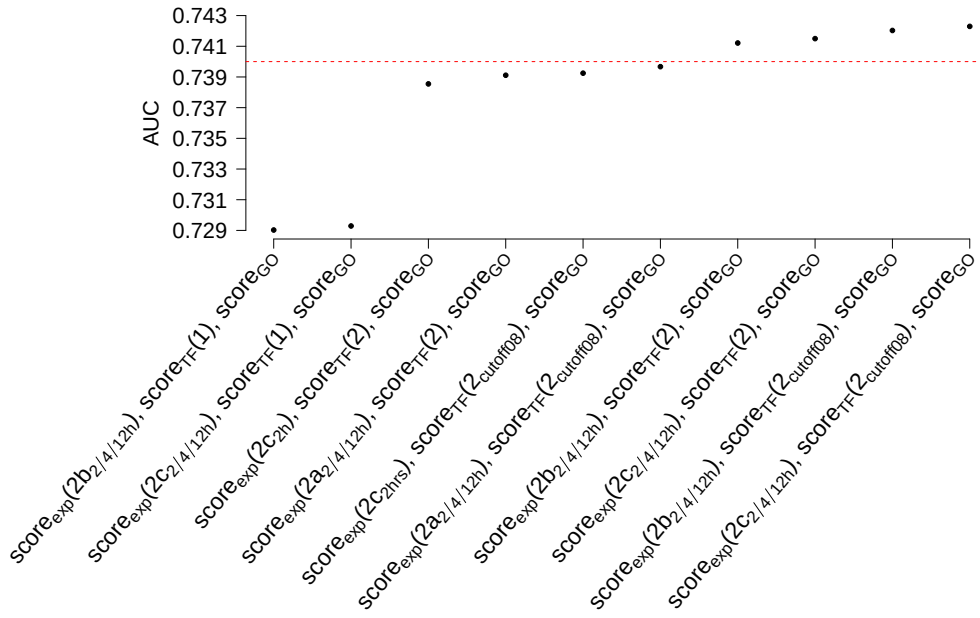


Figure 4.19: **Results of the pseudo-ROC analysis for different combinations of evidence.** Displayed are the values of the top ten Area Under the Curve (AUC) results for the comparison of the real to a node permuted iRefIndex graph based on different combinations of evidence (Section 3.4). The combination of methods that best separate scores for TP (real) from TN (random) edges and, thus, yields the highest AUC, is given by $score_{exp}(2c_{2/4/12h}), score_{TF}(2_{cutoff08}), score_{GO}$. It is obtained by combining the covariance of the replicates across the three time points ($score_{exp}(2c_{2/4/12h})$, Method 2c, page 70), the calculation of $score_{TF}$ using Eq. 3.25 (page 68) only considering matrix scores > 0.8 ($score_{TF}(2_{cutoff08})$), and $score_{GO}$ by using the Resnik method (page 62) as implemented in the GOSemSim package without filtering GO annotations for evidence codes (page 66). Following abbreviations are used for remaining individual measures used: $score_{exp}(2b_{2/4/12h})$: covariance across the three time points, Method 2b, page 70; $score_{TF}(2)$: calculated using Eq. 3.25, page 68, considering all matrix scores; $score_{exp}(2a_{2/4/12h})$: covariance of the $log_2 ratios$ of the two conditions, i. e. TGF-β-stimulated compared to untreated cells, across the three time points, Method 2a, page 2; $score_{exp}(2c_{2h})$: covariance of the replicates across measured 2 hours after stimulation, Method 2c, page 70; $score_{TF}(1)$: calculated using Eq. 3.24, page 67, considering all matrix scores.

based on their AUC. Based on the measures given in the previous section and analogous to the TPs, edge scores, $score_{edge}$, for the TNs were calculated for each of the 360 combinations. Next, the AUC for the pseudo-ROC curves for the real (TP) and

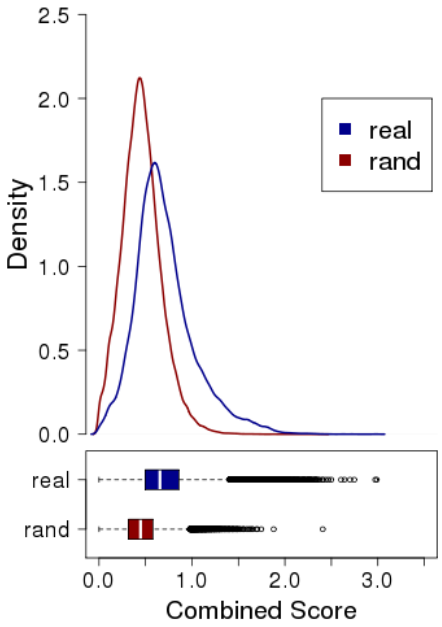


Figure 4.20: **Density distribution and boxplots of edge scores.** Comparison of edge scores as obtained for the TP (real) compared to the TN (random) edges using the best combination of methods based on the AUCs displayed in Figure 4.19. According to the density functions and boxplots, scores for the real protein interactions are, on average, higher than those for the random interactions.

random (TN) edges for each of the possible 360 combinations were calculated. Based on the ranking of the AUC values for the pseudo-ROC curves, the combination used for the final analysis was chosen.

Figure 4.19 displays AUC values for the ten highest ranking combinations of methods. The best AUC is obtained for scores based on the covariance of the expression values for the replicates over all time points (Method 2c, page 70) in combination with GOSemSim Resnik without filtering for evidence codes (Section 3.4.2, page 66) and applying $score_{TF,2}$ only considering matrix scores > 0.8 for transcription factor weighting (Eq. 3.25, page 68). A comparison of edge score distributions and boxplots for the TP (real) and TN (random) edge scores as given in Figure 4.20 clearly shows higher scores obtained for real protein interactions in contrast to random pairs of proteins.

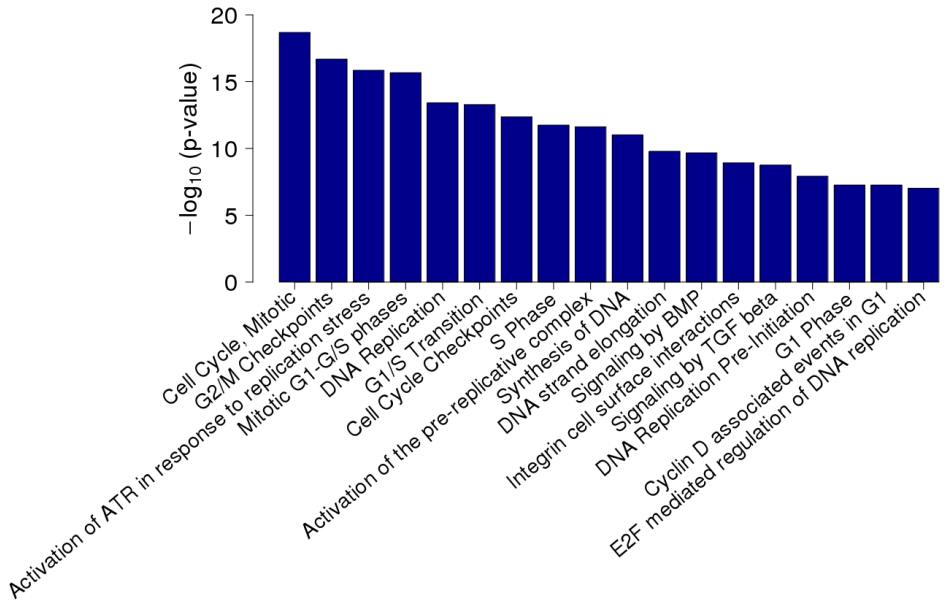


Figure 4.21: **Result of gene set enrichment based on Fisher's exact test.** Based on gene expression for TGF-β1-stimulated cells, modEx was used to extract 10 modules. Fisher's exact test was conducted on gene sets as obtained by Reactome pathways for the union of these modules. Significantly enriched gene sets showing p-values $< 1 \cdot 10^{-7}$ are displayed. Detected pathways show a clear link to TGF-β signaling.

4.3.2 Biological Evaluation by TGF-β Stimulation Experiment

In order to show the principle applicability of our approach, I first utilized an easy to interpret gene expression experiment comparing TGF-β-stimulated against unstimulated HaCaT cells at 2, 4, and 12 hours after stimulation [54]. Briefly speaking, an iRefindex based network was scored using the best combination of evidence selected in the previous section, with covariance for adjacent proteins calculated based on gene expression values for the two conditions (TGF-β-stimulated and unstimulated) across the three time points. Based on the weighted PPI network, ten networks were extracted using modEx (Section 3.5.2) and selecting the 10 strongest deregulated, significant genes as seed nodes (adjusted $p - value < 0.01$). Table 4.1 displays the probabilities for obtaining these networks by chance calculated by the statistical measures described in Section 3.6.1. For biological evaluation, gene set enrichment analysis by Fisher's exact test [143] was conducted based on the nodes (proteins)

Table 4.1: **Overview of modules extracted based on TGF-β stimulation experiment.** Displayed are different measures for modules extracted based on TGF-β-stimulated cells compared to unstimulated cells using modEx (Section 3.5.2). The p-values indicate the probability of observing the respective score $\Omega(G'_i)$ or a higher one for the i-th extracted module of the random graphs. The module ranking column indicates the order for the extracted modules according to z-score based p-values and $\Omega(G'_i)$.

Net No. i	Seed node	log_2 ratio#	Module ranking	$\Omega(G'_i)$§	z-score p-value+	approx. p-value+
1	VASN	5.4	10	2.203	0.032	0.044
2	SERPINE1	5.34	4	2.76	0.001	0.01
3	CTGF	3.94	9	2.258	0.029	0.036
4	JUN	3.48	2	3.159	<0.001	0
5	TGM2	3.39	8	2.233	0.025	0.04
6	FOSB	3.13	3	3.022	<0.001	0
7	BHLHE40	3.06	7	2.53	0.006	0.016
8	CDKN2B	3	1	4.533	<0.001	0
9	SMAD7	2.87	5	2.658	0.004	0.004
10	SOX18	2.85	6	2.507	0.006	0.014
Union*				2.786	<0.001	0

#log_2 ratio comparing TGF-β-stimulated cells to unstimulated cells after 2 hours.

§ $\Omega(G'_i)$ is calculated using Eq. 3.34.

+Approximative permutation test (approx.) and z-score based p-values refer to the probabilities calculated based on 500 random graphs (Section 3.6.1).

*Union refers to the network as obtained by the union of the ten individual nets extracted.

contained in the union graph of the ten networks extracted. Results are displayed in Figure 4.21. Proteins contained in the extracted networks show a significant enrichment in gene sets related to TGF-β signaling. It is well known that TGF-β signaling plays a major role in controlling the cell cycle as well as reorganization of the extracellular matrix. Additionally, our findings are confirmed by a significant enrichment in BMP-signaling, a signaling pathway induced by the BMP receptor which also belongs to the TGF-β receptor superfamily, and TGF-β signaling itself.

4.3.3 Analyses of Compounds' On- and Off-Target Effects

To investigate whether it is possible to detect on- and off-target effects of compounds based on the proposed method, I used expression data derived from human HaCaT cells treated with inhibitors of TGFBR1 [54]. In what follows, I focus on two compounds which are referred to as BI1 and BI4. HaCaT cells were stimulated with TGF-β and either treated with $2\mu M$ (lowest concentration above IC50, [1,54]) of the respective compound or not treated with compound. Gene expression of stimulated and compound-treated HaCaT cells was compared to the gene expression of TGF-β-stimulated cells after a time span of 2 hours. Out of the genes with FDR-adjusted p values < 0.01, ten seed nodes were chosen based on log_2 ratios. In contrast to the covariance used above, covariance between the expression profiles of two genes was calculated across 2, 4 and 12 hours for five different compound concentrations (0.0032, 0.016, 0.08, 0.4, $2\mu M$) measured in triplicates and the control group measured in quadruples. Remaining scores were calculated as described for the TGF-β stimulation experiment. By applying modEx, modules were extracted based on the ten seed nodes selected.

On-Target Effects

Considering the union of the ten modules, connected components were identified. Figure 4.22 displays one connected component extracted based on TGF-β-stimulated compared to unstimulated cells (Figure 4.22 A) and one connected component extracted based on TGF-β-stimulated cell compared to cells stimulated with TGF-β and treated with BI4 (Figure 4.22 B). Seed nodes that led to the connected components displayed were JUN and FOSB (Figure 4.22 A) or FOSB, JUN, and JUNB (Figure 4.22 B). As displayed in Tables 4.1 and 4.2, the individual extracted modules based on these seed nodes had a p-value ≤ 0.002, i. e. modules with the respective scores are very unlikely to be observed by chance. An on-target effect of compounds can be detected by inversion of the TGF-β effect. This is the case for networks extracted based on BI4 treatment (Figure 4.22). Genes that are up-regulated by TGF-β stimulation (red) are down-regulated by compound treatment (green).

Networks extracted based on BI1 treatment do not exhibit as strong on-target effects (Figure 4.23). Compared to results obtained for BI4 (Figure 4.22), for BI1

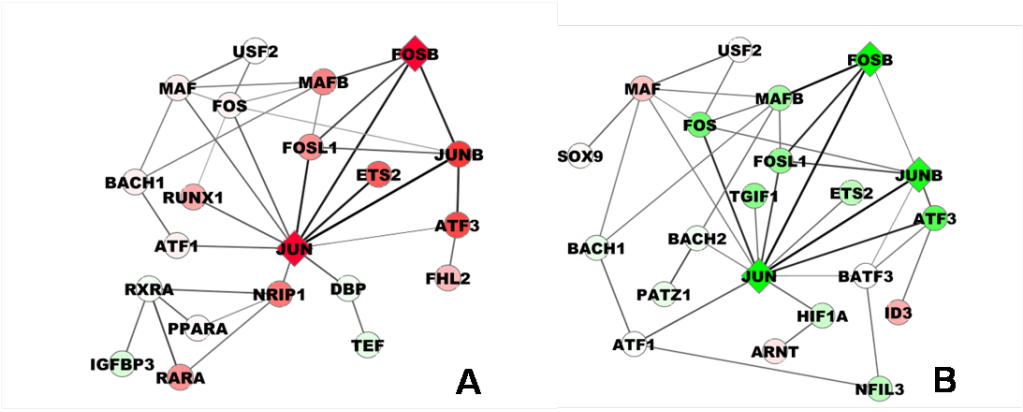


Figure 4.22: **modEx results demonstrating on-target effect of BI 4.** Comparison of networks extracted based on TGF-β-stimulated compared to unstimulated (A) and BI4-treated and TGF-β-stimulated compared to TGF-β-stimulated (B) cells. Red indicates up-, green down-regulation of the respective genes, diamond-shaped nodes indicate seed nodes. The on-target effect of BI4 inverting the TGF-β stimulation can directly be seen by the opposite regulation of many of the genes contained in the extracted modules.

an as clear inversion of the direction of deregulation cannot be observed. For BI4, two nearly identical modules have been extracted basesd on TGF-β-stimulated (Figure 4.22A) and based on compound treatment (Figure 4.22B). For BI1, the module displayed in Figures 4.24A and 4.24B shows the densest enrichment in genes which due to their direction of deregulation hint at on-target effects. But at the same time it contains genes that are neither strong nor significantly deregulated but are direct off-targets of BI1 (Appendix F). Figure 4.24C displays the module extracted based on TGF-β1-stimulation that exhibits the greatest overlap to the latter module with respect to the proteins contained. Compared to BI4 (Figure 4.22), the overlap of Figures 4.24A and 4.24C is not as striking with regard to on-target effects.

Off-Target Effects

For investigating off-target effects, first, results obtained for expression data as measured based on BI1-treated and TGF-β1-stimulated cells were analyzed. In a second step, the same analysis was done using data from BI4-treated and TGF-β1-stimulated cells. To focus the analysis towards off-target effects, only those seed nodes were used that are not significantly de-regulated (FDR-adjusted p-value $\geq$ 0.01) in the opposite direction when comparing TGF-β-stimulated to unstimulated cells

Table 4.2: **Overview of modules extracted based on BI4 treatment.** Different Measures for modules extracted based on BI4-treated and TGF-β-stimulated cells compared to just TGF-β-stimulated cells using modEx (Section 3.5.2) are summarized. The p-values indicate the probability of observing the respective average edge score $\Omega(G'_i)$ or a higher one for the i-th extracted module of the random graphs.

Net No. i	**Seed node**	log_2 **ratio**[#]	$\Omega(G'_i)$[§]	**z-score p-value**[+]	**approx. p-value**[+]
1	SERPINE1	-4.43	2.06	0.031	0.038
2	VASN	-3.88	1.955	0.074	0.072
3	FOSB	-3.29	2.41	0.004	0.016
4	CTGF	-3.23	2.034	0.045	0.044
5	JUN	-3.15	3.57	<0.001	0
6	CYR61	-3.08	2.025	0.041	0.038
7	SCGB1A1	2.87	2.428	0.003	0.01
8	SKIL	-2.76	2.372	0.003	0.006
9	KRT1	2.71	1.649	0.238	0.178
10	JUNB	-2.67	3.089	<0.001	0.002
Union*			2.359	<0.001	0

[#] log_2 ratio comparing BI4-treated and TGF-β1-stimulated cells to TGF-β1-stimulated cells after 2 hours.

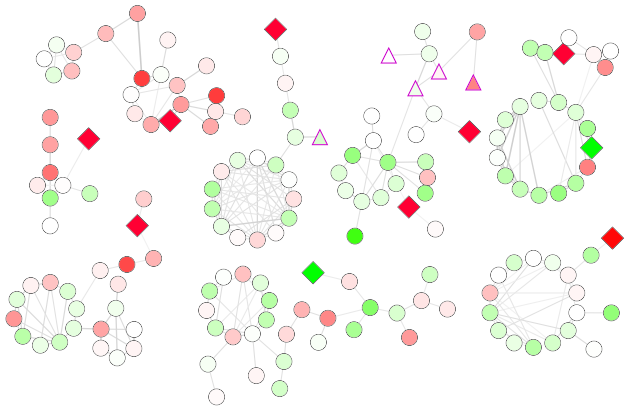
[§] $\Omega(G'_i)$ is calculated using Eq. 3.34.

[+] Approximative permutation test (approx.) and z-score based p-values refer to the probabilities calculated based on 500 random graphs. (Section 3.6.1)

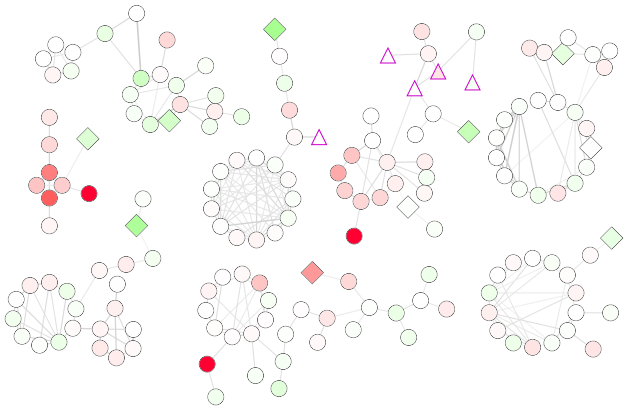
*Union refers to the network as obtained by the union of the ten individual nets extracted.

after 2 hours.

Table 4.3 summarizes the results obtained for the first ten extracted networks based on the BI1 data, i. e. based on the comparisons of BI1- and TGF-β1-treated compared to TGF-β1-treated cells; Figure 4.25 displays the results of Fisher's exact test conducted for one of the most significant networks extracted (Net No. 8). In contrast to the results described in the analysis of on-target effects (pages 116 f.)



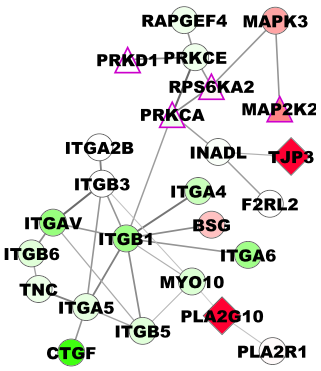
A: Nodes are colored according to log_2 ratios of expression values based on BI1 and TGF-β1 treatment compared to the respective control.



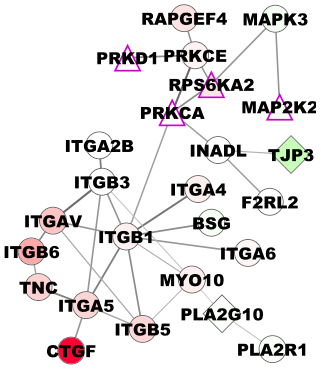
B: Nodes are colored according to log_2 ratios of expression values based on TGF-β-treated compared to untreated cells

Figure 4.23: **Union graph extracted based on BI1 treatment.** Displayed is the union graph of the 10 graphs extracted based on the mostly deregulated genes in BI1-treated and TGF-β1-stimulated compared to TGF-β1-stimulated cells. Diamond-shaped nodes represent the seed nodes. Triangular purple-framed nodes represent wet-lab validated direct off-targets of BI1. Green indicates down-, red up-regulation. The more intense the colors, the stronger the respective genes are deregulated. In contrast to Figure 4.22 which is based on BI4, for BI1 we can not see an as obvious inversion of the direction of deregulation when comparing Subfigure A and Subfigure B.

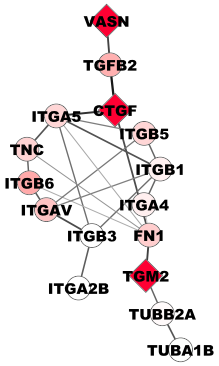
the results for BI1 do not show any direct connection to TGF-β signaling. Instead, the modules detected by modEx method are enriched in different signaling cascades



A: Nodes are colored according to log_2 ratios of expression values based on BI1- plus TGF-β1-treated cells compared to the respective control.



B: Nodes are colored according to log_2 ratios of expression values based on TGF-β1-treated compared to untreated cells.



C: Nodes are colored according to log_2 ratios of expression values based on TGF-β1-treated compared to untreated cells.

Figure 4.24: **Subgraph of union graphs extracted based on BI1 treatment (A&B) or TGF-β (C) stimulation. Subfigure A** displays the connected component out of the union graph in Figure 4.23 that contains most of the direct off-targets of BI1 and at the same time best could be linked to on-target effects compared to other connected components. Diamond-shaped nodes represent the seed nodes. Triangular purple-framed nodes in A&B represent wet-lab validated direct off-targets of BI1 (Appendix F. Green indicates down-, red up-regulation. The more intense the colors, the more the respective genes are deregulated. **Subfigure B** depicts the same subgraph but with nodes colored according to the deregulation observed in TGF-β-treated compared to untreated cells. In contrast to Figure 4.22 which is based on BI4, for BI1 we can not see an as obvious inversion of the direction of deregulation. **Subfigure C** shows one connected component of the union graph derived based on expression data for TGF-β1-treated compared to untreated cells. Out of all components in the union graph this component yields the highest overlap with genes in Subfigure A.

triggered by kinases which could be confirmed as direct BI1 off-targets in wet-lab experiments (Figure 4.25, Appendix Tables F.1-F.3, pages 170 ff.) [54]. It could be shown that BI1 inhibits all three FGF receptor kinases. In line with this observation highest significant scores are obtained for gene sets describing FGFR signaling.

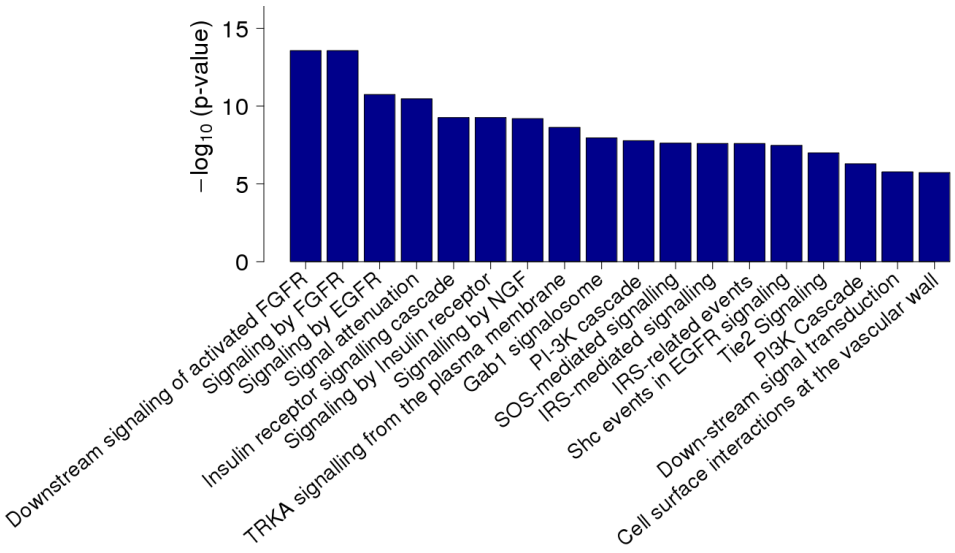


Figure 4.25: **Result of gene set enrichment based on Fisher's exact test.** Based on gene expression for BI1-treated and TGF-β1-stimulated cells, modEx was used to extract 10 modules. One of the most significant modules (Net No. 8, Table 4.3) was used in the enrichment analysis. Fisher's exact test was conducted based on gene sets as obtained by Reactome pathways for genes contained in Net No. 8. Displayed are significantly enriched gene sets with p-values $< 2 \cdot 10^{-6}$. Compared to random networks, the network extracted using our method is very unlikely to be observed randomly (p-value < 0.001, Table 4.3). Additionally, it shows highly significant enrichment for genes related to signaling pathways regulated by various off-target kinases such as FGFR signaling (Appendix Tables F.1-F.3, pages 170 ff.) [54].

Furthermore BI1 inhibits the neurotrophic tyrosine kinase, receptor 1 (NTRK1 also referred to as TRKA). Again gene sets involved in TRKA signaling itself as well as its ligand NGF were identified. In addition the signaling pathway component GAB1, a known activator of PI3 kinase signaling [189], was detected by our analysis. Finally, BI1 inhibits MEK1 (MAP2K1) that acts as a mitogen-activated protein (MAP) kinase kinase and is involved in the integration of multiple biochemical signals including insulin receptor signaling (IRS) and epidermal growth factor receptor (EGFR) signaling [190]. Both related gene sets were identified by our approach.

Compared to BI1, for BI4 off-target effects are not as evident. Again, seed nodes for modEx have been chosen such they do not exhibit an inverse deregulation compared to TGF-β1-stimulated cells. Figure 4.26 displays the results for the

Table 4.3: **Overview of modules extracted for the off-target analysis of BI1.** Displayed are results for modules extracted based on BI1-treated and TGF-β1-stimulated cells compared to TGF-β1-stimulated cells using modEx (Section 3.5.2). The p-values indicate the probability of observing the respective average edge score $\Omega(G'_i)$ score or a higher one for the i-th extracted module of the random graphs.

Net No. i	**Seed node**	***log*$_2$ ratio**$^{\#}$	$\boldsymbol{\Omega(G'_i)}^{\S}$	**z-score p-value**$^{+}$	**approx. p-value**$^{+}$
1	KRT1	3.59	1.71	0.208	0.149
2	VWF	3.52	2.78	<0.001	0.016
3	SCGB1A1	3.44	1.969	0.06	0.101
4	PLA2G10	3.12	1.814	0.124	0.112
5	HSP90AA1	-3.07	3.105	<0.001	0
6	GBP2	2.9	2.187	0.031	0.053
7	TRIM31	2.78	2.104	0.049	0.037
8	ABP1	2.77	2.987	<0.001	0
9	RGS2	-2.75	2.181	0.025	0.027
10	ARL4A	-2.75	5.043	<0.001	0
Union*			2.588	<0.001	0

$^{\#}log_2$ ratio comparing BI1-treated and TGF-β1-stimulated cells to TGF-β1-stimulated cells after 2 hours.

§ $\Omega(G'_i)$ is calculated using Eq. 3.34.

$^{+}$Approximative permutation test (approx.) and z-score based p-values refer to the probabilities calculated based on 500 random graphs. (Section 3.6.1)

*Union refers to the network as obtained by the union of the ten individual nets extracted.

Fisher's exact test conducted based on proteins contained in the module extracted for SCGB1A1. This module is one of the most significant ones (approximative p-value = 0.01, Table 4.2, page 118) as well as one of the few ones for which gene sets could be linked to direct off-targets of BI4. The respective gene sets hint at PDGF and MAPK related signaling events, PDGF as well as different MAP-kinases could be confirmed as direct BI4 off-targets (Appendix F, Table F.5, page 174). Looking at the union graphs extracted based on BI1 and BI4, enriched gene sets for BI1 hint

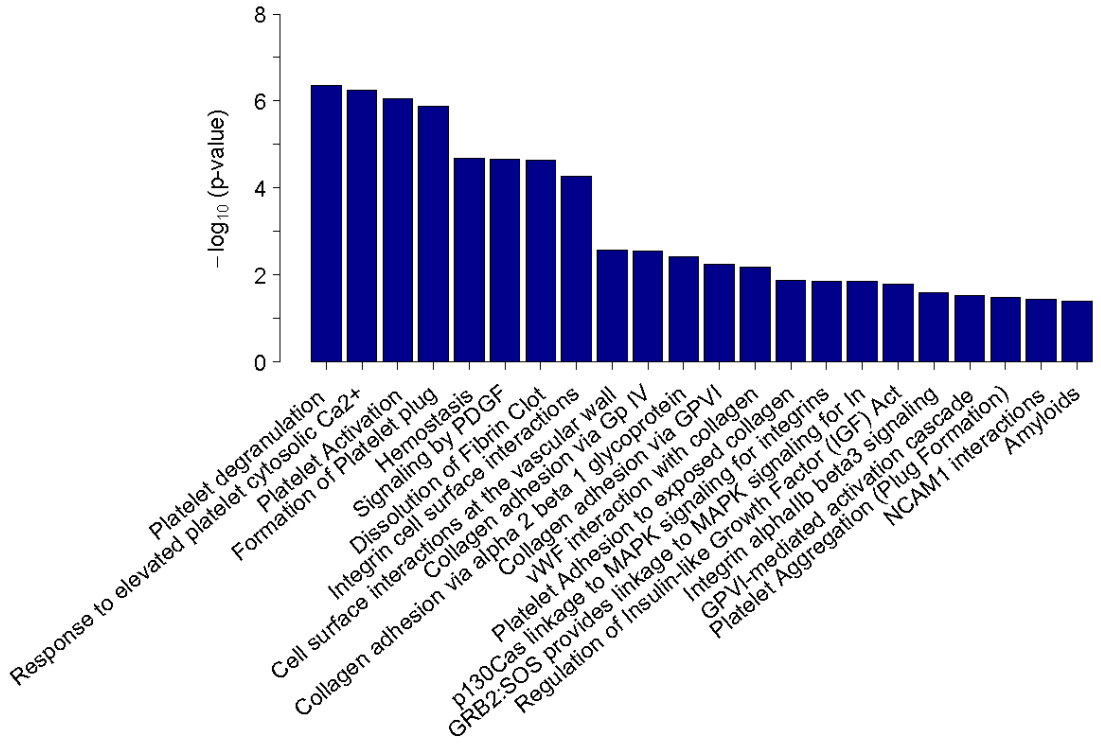


Figure 4.26: **Off-target effects observed for BI4.** Fisher's exact test was conducted based on gene sets as obtained by Reactome pathways for genes contained in a network extracted using SCGB1A1 (P11684) as seed node (p-value = 0.003, Table 4.2). Displayed are significantly enriched gene sets with p-values < 0.05, page 118. Compared to BI1, for BI4 gene sets hinting at off-target effects are much rarer to observe. Genes contained in the subnetwork extracted show enrichment in PDGF and MAPK related signaling events. These kinases have been confirmed as direct off-targets of BI4 (Appendix F, Table F.5, page 174).

at more off-target effects than those detected for BI4. For BI4, TGF-β-signaling is even the fourth most significant. In summary, this all hints at compound BI4 being much "cleaner", i. e. being more specific for TGF-βR1, than compound BI1. This could be confirmed by the kinase screen exhibiting much more direct off-targets for BI1 than for BI4 (Appendix F, pages 169 ff.).

4.4 Comparison to Existing Approaches

Results derived using the proposed approach do not only depend on the extraction procedure, modEx, but also on the underlying network and the edge scoring. Thus, to further evaluate the method, I first conducted analyses based on three different protein networks, namely iRefIndex, STRING, which I refer to as STRING$_{org}$, and STRING with recalculated edge weights (Eq. 3.30, page 71), which I refer to as STRING$_{mod}$ (see Secion 3.7, page 77 for details). Second, I compared the modules derived using our approach to modules derived using jActiveModule [48] (see Section 3.7, page 76, for more details).

To compare the results based on different networks and scoring methods, gene expression data derived from cells stimulated with TGF-β1 compared to unstimulated cells was used. Thus, the modules extracted should in the ideal case be enriched in genes linked to TGF-β signaling. Nodes of jActiveModule instances are weighted according to the FDR-corrected p-values calculated for the comparisons of TGF-β1-stimulated compared to unstimulated cells at 2, 4, and 12 hours (see Section 3.3.1, page 55). Edges of modEx instances are weighted based on the method selected in Section 4.3.1 as summarized on page 113. Gene expression data of TGF-β1-stimulated and unstimulated cells across 2, 4, and 12 hours was used to calculate $score_{exp}$ (Eq. 3.28, page 70). log_2ratios and FDR-corrected p-values (Section 3.3.1, page 55) comparing gene expression of TGF-β1-stimulated to unstimulated cells at 2 hours were used to select seed nodes. This setting is equivalent to the setting used in Section 4.3.2 were this data set was used to evaluate the new approaches.

4.4.1 Comparison Between iRefIndex and STRING

modEx was applied to different networks, a complete overview of the results is given in Appendix C, Tables C.1 - C.3 (pages 157 ff.). Based on the proteins contained in the extracted modules Fisher's exact test was conducted and gene sets were ranked according to the respective p-values (Figure 4.27). Results displayed are based on KEGG gene set. An equivalent analysis has also been conducted using Reactome gene sets. Results for the latter one are summarized in Appendix Table C.2 on page 159.

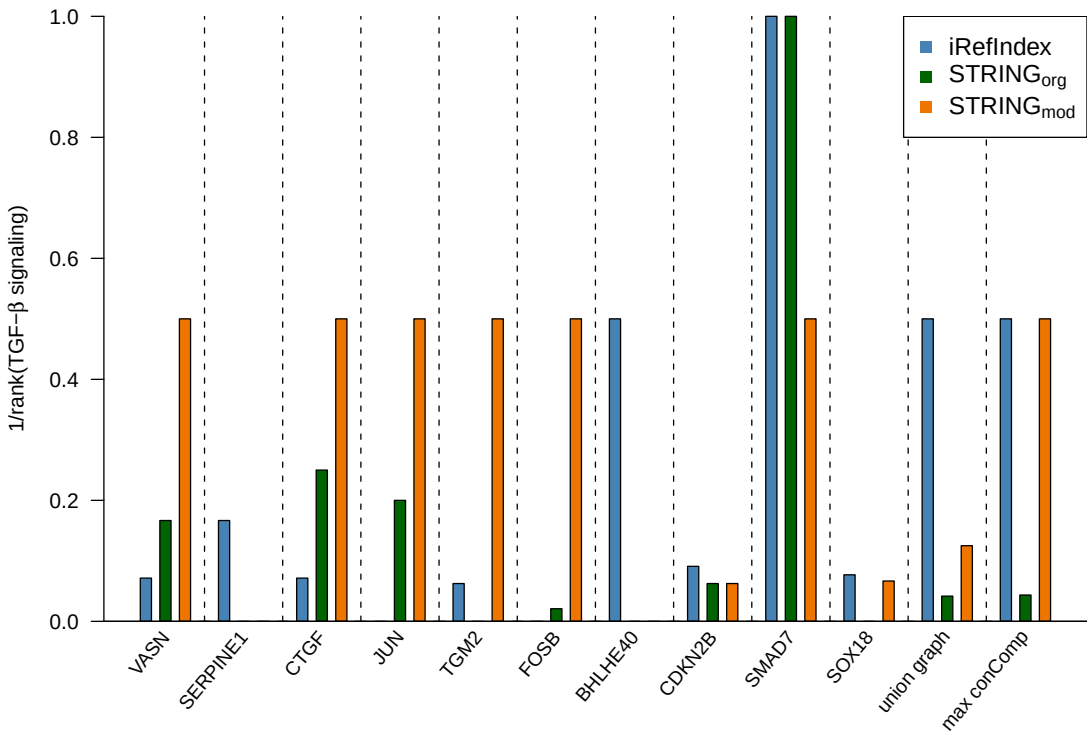


Figure 4.27: **Comparison of STRING and iRefIndex networks.** Displayed are gene set enrichment results for the modules derived when applying modEx to networks with edges weighted according to the best combination of evidence (Section 4.3.1, Figure 4.19, page 112) using expression values obtained by a comparison of TGF-β1-stimulated cells to unstimulated cells after 2, 4 and 12 hours. modEx was applied to iRefIndex, STRING$_{org}$ and STRING$_{mod}$. For the iRefIndex network, these modules have already been used in the scope of the analyses of Section 4.3.2 and are summarized in Table 4.1 on page 115. Based on KEGG defined gene sets, Fisher's exact test was conducted for the proteins contained in the extracted modules (indicated by their seed nodes on the x-axis), their union graph as well as for the maximal connected component (max conComp), i. e. that connected component of the union graph containing the most nodes. Results of this enrichment analysis are displayed in Table C.3 on page 160. For comparing the results, the reciprocal ranks of the TGF-β signaling gene set for the different modules were used. The gene sets were ranked according to p-values. Focusing on networks based on single seed nodes, STRING$_{mod}$ is superior to both, iRefIndex as well as STRING$_{org}$. When considering unions of the individually extracted modules, iRefIndex yields the best results.

On average, the best results are obtained based on STRING$_{mod}$. For six of the

ten extracted networks TGF-β signaling is ranked second when STRING$_{mod}$ is used as an underlying network (Figure 4.27, orange bars). Five of them even exceed the STRING$_{org}$ as well as the iRefIndex based results. In only two cases (SERPINE1 and BHLHE40), the iRefIndex based results (blue bars) clearly outperform the STRING based results. For three of the ten extracted networks (CDKN2B, SMAD7, SOX18), iRefIndex slightly outperforms STRING$_{mod}$. STRING$_{org}$ (green bars) outnumbers STRING$_{mod}$ in only one case (SMAD7) and iRefIndex in four cases (VASN, CTGF, JUN, FOSB). For the union graph as well as for the maximal connected component of the union graph, STRING$_{org}$ yields the worst results. Here, the best results are obtained by iRefIndex, though it only slightly outperforms STRING$_{mod}$ for the union graph (rank two versus rank eight). Thus, when focusing on networks based on single seed nodes, STRING$_{mod}$ is superior to both other networks. When considering unions of the individually extracted modules, iRefIndex yields the best results.

4.4.2 Results for jActiveModule

Since jActiveModule does not make use of any edge scores, I did not compare results based on different scoring methods. As it already has been done for the comparison between iRefIndex and STRING, first, a module search was performed; second, Fisher's exact test was conducted based on the proteins contained in the identified modules comparing them to predefined gene sets. By this means, it was possible to compare results obtained for the different extraction methods as well as for different protein networks. Again, I focus on results as obtained based on KEGG gene set here. Results for Reactome based analysis are summarized in Appendix Tables D.2 and D.5 pages 163 f..

jActiveModule offers different search options. Using default parameters, greedy search, simulated annealing, and simulated annealing with hub finding switched on were applied. Simulated annealing alone did not return any modules, simulated annealing with hub finding returned modules, but none of the returned modules showed a significant enrichment (p-value < 0.01) with genes related to TGF-β signaling (Appendix D, Tables D.2 and D.3, page 163). Thus, in what follows, I focus on results derived using greedy search (referred to as *greedyDef*).

When applying jActiveModule, a search for modules active across different time

points was performed, i. e. the method was applied to p-values derived for TGF-β1-stimulated cells compared to unstimulated cells after 2, 4, and 12 hours. Based on these p-values, jActiveModule tries to detect modules active at any of these time points. Since the five detected modules are only active for the 12 hours expression data, further analyses exclusively refer to this time point. Modules are ranked from 1 to 5 based on their scores $s_{G'}$ (Eq. 2.1, page 29).

First, I investigated results of jActiveModule based on different networks, namely STRING and iRefIndex (Figure 4.28, Appendix D Tables D.3 and D.6, pages 163 ff.). Two of the three highest ranking modules are based on iRefIndex, one is based on STRING$_{org}$. The TGF-β signaling gene set ranks sixth (module2, sub-net) and nineth (module5, no-pval) or eighth (module3, def-pval) based on iRefIndex or STRING$_{org}$, respectively. Even though the best rank is obtained by the iRefIndex network, the STRING$_{org}$ networks more often result in higher ranks. Nevertheless these ranks are very low with the highest rank of eight for module3 (def-pval) and all remaining ranks below fifteen. Thus, it is difficult to say which of the networks is superior to others when using jActiveModule.

4.4.3 Comparison Between modEx and jActiveModule

To get a more detailed impression of the difference between jActiveModule and modEx results, I next compared them based on the underlying protein networks. iRefIndex as well as STRING were used as different networks. Additionally, for STRING different cut-offs were applied to the native edge score. A complete overview of the comprehensive analysis is given in Appendix C, Tables C.1- D.6. Modules extracted using modEx are ranked from 1 to 5 according to their p-value and according to $\omega(G'_i)$ (Table 4.1, page 115, "Module ranking" column), modules derived using jActiveModule are ranked according to their scores $s_{G'}$ (Eq. 2.1, page 29). In the following, I summarize the quintessence of these analysis based on enrichment analysis conducted using KEGG pre-defined gene sets. A complete overview of the KEGG-based results is given in Appendix Table C.3, page 160 for modEx and in Tables D.3 and D.4, pages 163 f. for jActiveModule. Again, the respective results for Reactome-based analysis can be found in Appendix C, Tables C.2, page 159 for modEx, and in Appendices D, D.2 and D.5, pages 163 f. for jActiveModule.

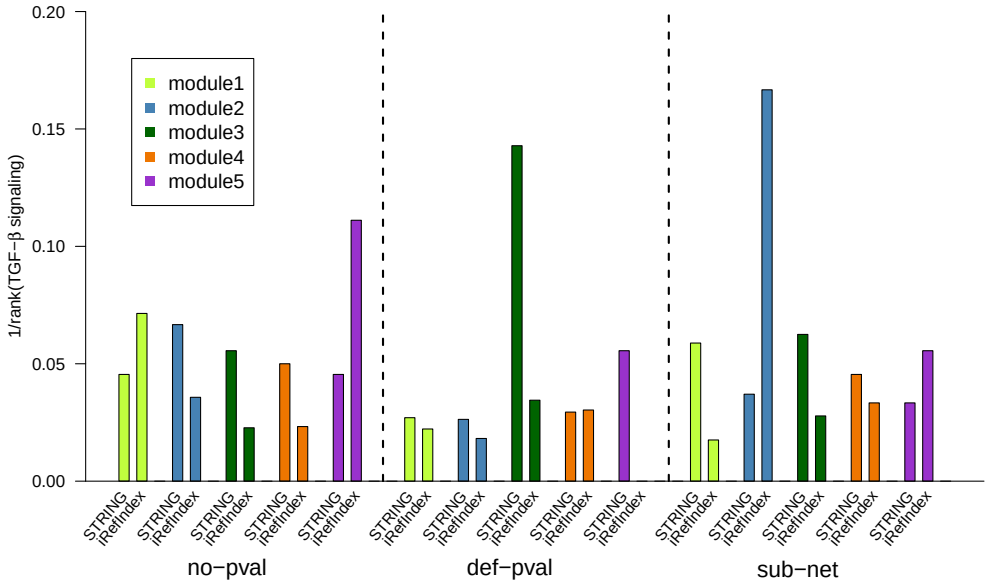


Figure 4.28: **Comparison of jActiveModule results for STRING or iRefIndex networks.** Displayed are the results for the individual modules derived when applying jActiveModule to networks with nodes weighted according to p-values obtained by a comparison of TGF-β1-stimulated cells to unstimulated cells after 2, 4 and 12 hours. To compare jActiveModule derived modules to modEx derived modules the latter ones are ranked from 1 to 5 according to their p-value and their score $\omega(G'_i)$ (Table 4.1, page 115, "Module ranking" column). In case the protein coding genes are not represented on the microarray used to measure gene expression either no p-values are assigned to the respective genes (no-pval, left part of the plot), 1 is assigned as default p-value (def-pval, middle part of the plot), or only the subnet induced by the genes represented on the microarray is used in the analysis (sub-net, right part of the plot) (Section 3.7, page 76). Gene set enrichment results are ranked according to p-values and reciprocal ranks are used for visualization (y-axis). The best rank is obtained by the iRefIndex network, but the STRING$_{org}$ networks more often result in higher ranks. These ranks are very low with the highest rank of eight for module3 (dark-green bar, def-pval) and all remaining ranks below fifteen. Nevertheless, there is no clear evidence for one network being superior to the other when using jActiveModule.

iRefIndex Used as an Underlying Network

Figure 4.29 displays the comparison of modEx to jActiveModule based on the iRefIndex. module5 extracted using modEx results in TGF-β signaling being ranked

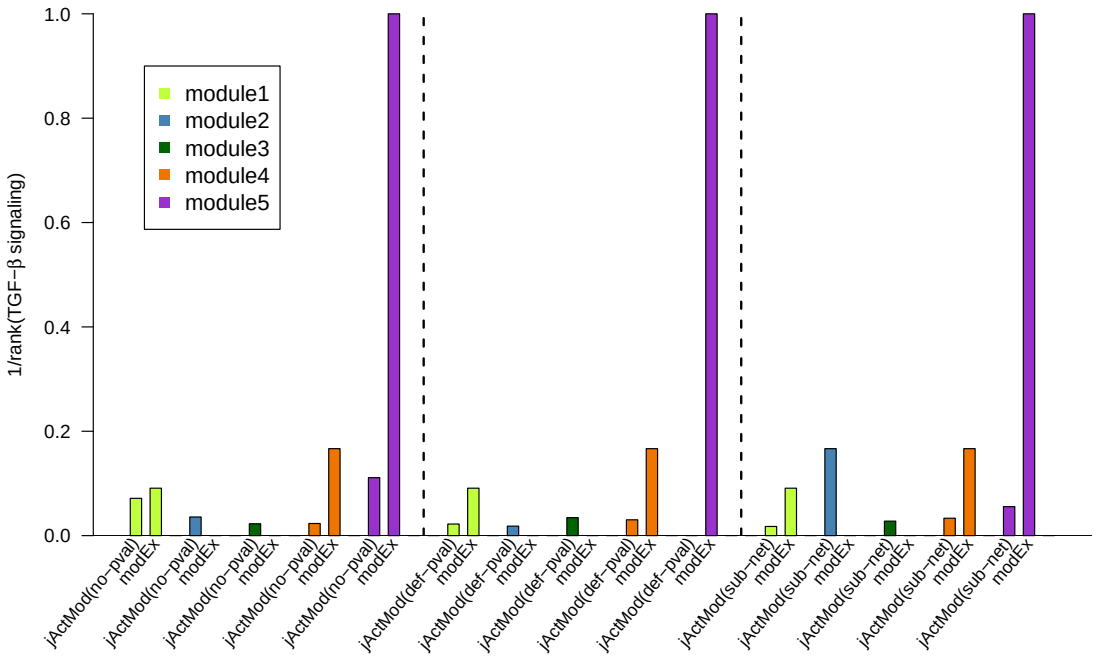


Figure 4.29: **Comparison of jActiveModule and modEx results based on iRefIndex networks.** Displayed are the results for the individual modules derived when applying the methods to networks weighted according to a comparison of TGF-β1-stimulated cells to unstimulated cells after 2, 4, and 12 hours. In case the protein coding genes are not represented on the microarray used to measure gene expression either no p-values are assigned to the respective genes (no-pval, left part of plot), 1 is assigned as default p-value (def-pval, middle part of plot), or only the subnet induced by the genes represented on the microarray is used in the analysis (sub-net, right part of plot) (Section 3.7). Enrichment analysis was conducted for the individual modules extracted and gene sets ranked according to p-values. For the five best modules extracted, the reciprocal ranks for the TGF-β signaling gene set are plotted along the y-axis. This gene set should be ranked high, as we used expression data from TGF-β1-stimulated cells for which this signaling pathway is switched on. Since modEx results in TGF-β signaling being ranked first for module5 and since modEx based ranks are consistently higher than jActiveModule ranks for three of the five modules, modEx based modules are considered more informative.

first. Thus, and since modEx based ranks are higher than jActiveModule ranks in all cases except of module2 and module3, modEx based modules are considered more informative.

STRING used as an underlying network

As it already has been shown in Section 4.4.1 that STRING$_{mod}$ is superior to STRING$_{org}$ and as the difference in edge scoring is only relevant for modEx but not for jActiveModule, STRING$_{mod}$ was used in this final comparison. With STRING$_{mod}$ as underlying protein network, the differences in performance of jActiveModule and modEx are even clearer (Figure 4.30). Still, for one module, module4 (orange), jActiveModule outperforms modEx, however the ranks for this module are very low with 20, 34, and 22 for no-pval, def-pval and sub-net, respectively. Thus, these results can hardly directly be linked to TGF-β signaling. In all other cases, modEx performs better than jActiveModule with respect to this signaling pathway. For three of the modules, module2 (blue), module3 (green), and module5 (purple) the TGF-β signaling gene set is even ranked second for modEx derived modules. This, on the one hand, confirms the results of the previous section (Figure 4.27), indicating that using STRING$_{mod}$ is superior to using iRefIndex. On the other hand, it shows that modEx is superior to jActiveModule in cases where we want to derive modules that are not only enriched with deregulated genes but also are related to the underlying biological process.

4.5 Complexity Analysis

In Section 3.5.1 I introduced some optimization problems that could be posed in the analysis of weighted protein-protein interaction networks. In this section, I will analyze these problems with respect to their computational complexity. The computational complexity of a problem can be measured by the resources needed to solve it (see [191–193] for more details). In general, computational complexity theory aims at a classification of problems into respective *complexity classes*. In the following, I briefly introduce the two complexity classes P and NP from classical complexity theory. Based on this, I subsequently prove the NP-hardness of vCMECG, k-vCMECG, k-vECMECG, and 1-vCMECG as introduced in Definitions 3.5.1 to 3.5.4 on pages 72f..

The most prominent classes in classical complexity theory are P and NP. P stands for *polynomial time*, this class contains all problems that can be solved in polynomial

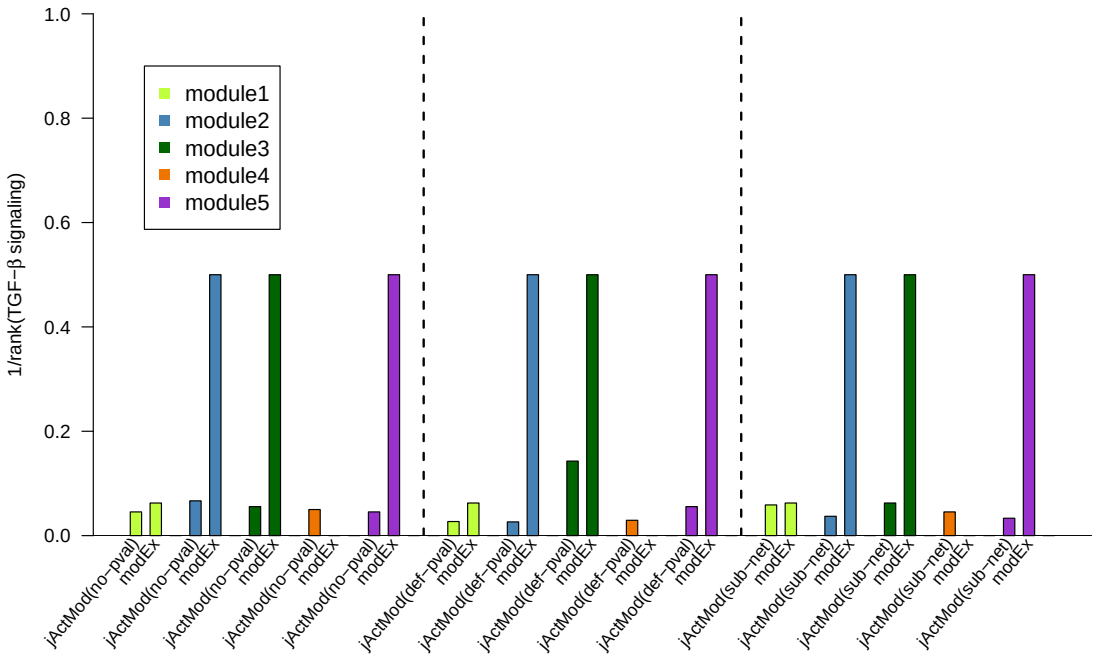


Figure 4.30: **Comparison of jActiveModule and modEx results based on STRING$_{mod}$ network.** Displayed are the results for the individual modules derived when applying the methods to networks weighted according to a comparison of TGF-β1-stimulated cells to unstimulated cells after 2, 4, and 12 hours. In case the protein coding genes are not represented on the microarray used to measure gene expression either no p-values are assigned to the respective genes (no-pval, left part of plot), 1 is assigned as default p-value (def-pval, middle part of plot), or only the subnet induced by the genes represented on the microarray is used in the analysis (sub-net, right part of plot) (Section 3.7). Enrichment analysis was conducted for the individual modules extracted and gene sets were ranked according to p-values. For the five best modules extracted, the reciprocal ranks for the TGF-β signaling gene set are plotted along the y-axis. As expression data from TGF-β1-stimulated cells was used, TGF-β signaling should be ranked high. jActiveModule outperforms modEx for only one module, namely for module4 (orange). As the ranks for this module are very low with 20, 34, and 22 for no-pval, def-pval and sub-net, respectively, these results would hardly be directly linked to TGF-β signaling. In all other cases, modEx performs better than jActiveModule. For three of the modules, module2 (blue), module3 (green), and module5 (purple) the TGF-β signaling gene set is even ranked second for modEx derived modules. Thus, modEx is superior to jActiveModule using STRING$_{mod}$ as an underlying network.

time. NP stands for *nondeterministic polynomial time* and describes a complexity class of problems for which it is possible to *guess*, i. e. nondeterministically find a solution in polynomial time. Given a solution for a problem in NP, it is possible to test its correctness in polynomial time. It is widely believed that P is not equal to NP implying that there are problems in NP that are not in P. This leads to NP-hard problems. Within the considered framework, showing that a problem is at least as hard as another problem is done by a *many-one reduction* defined as follows.

Definition 4.5.1. MANY-ONE REDUCTION.

Let A and B denote two problems. The problem A many-one reduces to B if there is a polynomial-time computable function f such that for an instance x of problem A denoted as $x \in A$

$$x \in A \Leftrightarrow f(x) \in B.$$

That means, a many-one reduction is a reduction which converts instances of a decision problem A into instances of a decision problem B, formally written as $A \leq_m B$. If $A \leq_m B$, any algorithm that solves instances of B can be applied to solve instances of A in the time needed for the algorithm to solve B plus the time needed for the reduction and with the maximum space needed for the algorithm plus the space needed for the reduction. As A is known to be NP-hard and $A \leq_m B$, B is NP-hard, otherwise it would contradict A being NP-hard. [192, 194, 195]. A problem is NP-hard if all problems from NP many-one reduce to it. An NP-hard problem belonging to NP is NP-complete. Hence, the class of NP-complete problems comprises a large set of *equivalent* problems for which presumably no polynomial-time algorithms exist.

4.5.1 Problems Related to vCMECG

In this section, I introduce some well known graph theoretical problems that are closely related to the problems I introduced in Section 3.5. These methods provide the basis for the NP-hardness proof given in Section 4.5.2.

Definition 4.5.2. COMPLETE GRAPH.

A complete graph is a graph $G = (V, E)$ where each pair of nodes $\{v_i, v_j\} \in V$ is connected by an edge $e_{\{v_i,v_j\}}$.

Definition 4.5.3. Clique Problem.

Input: *An undirected graph $G = (V, E)$ and a positive integer s.*

Task: *Is there a complete graph $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$ consisting of s vertices.*

This problem is well known to be NP-complete [196].

Definition 4.5.4. Steiner Tree Problem.

Input: *An undirected graph $G = (V, E)$, a weight function $\omega : E \to [0, \infty)$, and a set of terminal nodes $R \subseteq V$, $l \in [0, \infty)$.*

Task: *Find a tree $T = (V_T, E_T)$ with $R \subseteq V_T$ such that*

$$\Omega(T) = \sum_{e \in E_T} \omega(e) \leq l$$

The Steiner Tree Problem is known to be NP-complete [196].

The problems given in Sections 3.5 are all related to the maximum edge-weight connected graph (MECG) and the constrained maximum edge-weight connected graph (k-CMECG) problems as described by Li *et al.* [197]:

Definition 4.5.5. Maximum edge weight connected graph (MECG).

Input: *An undirected graph $G = (V, E)$, a weight function $\omega : E \to [0, \infty)$ and a positive integer s.*

Task: *Find a connected subgraph $G' = (V', E')$ with $|E'| = s$ such that*

$$\Omega(G') = \sum_{e \in E'} \omega(e)$$

is maximized.

Definition 4.5.6. k-constrained maximum edge weight connected graph (k-CMECG).

Input: *An undirected graph $G = (V, E)$, a weight function $\omega : E \to [0, \infty)$ and $E_k \subset E$ with $|E_k| = k$, and a positive integer s.*

Task: *Find a connected subgraph $G' = (V', E')$ with $|E'| = s$ and $E_k \subseteq E'$ such that*

$$\Omega(G') = \sum_{e \in E'} \omega(e)$$

is maximized.

The authors provide an NP-hardness proof of this problems and propose an integer linear programming algorithm to solve it.

The main difference between these and our problem definitions is that Li *et al.* center their analyses around edges only, while we want to incorporate both, edges and vertices. Instead of s edges (Definition 4.5.5) we look for s nodes (Definition 3.5.1, page 72), instead of fixed edges E_k (Definition 4.5.6), we consider fixed vertices V_k (Definition 3.5.2, page 72), and finally, instead for optimizing $\sum_{e \in E'} \omega(e)$ (Definition 4.5.5-4.5.6), we either optimize with respect to the number of edges $|E'|$ (Definition 3.5.3, page 73) or with respect to the number of vertices $|V'|$ (Definition 3.5.2, page 72) contained in the subgraph G'. Thus, our problem definitions are even more complex than those given by Li *et al.*.

4.5.2 NP-hardness of vCMECG and Related Problems

Using the technique of *many-one reduction* [198, 199], I provide NP-hardness proofs for vCMECG(Definition 3.5.1, page 72), k-vCMECG (Definition 3.5.2, page 72), k-vECMECG (Definition 3.5.3, page 73) and 1-vCMECG (Definition 3.5.4, page 73).

Theorem 4.5.7. vCMECG *is NP-hard.*

Proof. Clique $\leq_m$ vCMECG: Given a Clique instance (G=(V,E),s), construct an instance of vCMECG by using the same graph with weight one for every edge and let s also denote the number of vertices of the connected subgraph. Then, G has a clique of size s if and only if in the new graph there is a connected subgraph with s vertices and $s \cdot (s-1)/2$ edges, that is, with $w(G') = s \cdot (s-1)/2$.
$\Rightarrow$ Clique $\leq_m$ vCMECG.

□

Theorem 4.5.8. k-vCMECG *is NP-hard.*

Proof. Steiner Tree $\leq_m$ k-vCMECG: Given a Steiner Tree instance, $G = (V, E)$, $\omega : E \rightarrow 1$, $R \subseteq V$, and l. Construct an instance of k-vCMECG, $G' = (V', E')$, V_k, and s as follows:
Add a degree-one vertex to every terminal and set the weight of the corresponding new terminal edge $\omega(e)$ to $\omega(e) = |E| = m$. The weights of all remaining edges are

set to zero. V_k is identified with R. We obtain $G' = (V', E')$ with $|V'| = |V| + |R|$, $|E'| = |E| + |R|$, V_k=R, $s = l + |R|$.
Now, the following is easy to prove: k-vCMECG instance has a solution of size at least $mk/(l+k+1)$ if and only if the Steiner Tree instance allows for a Steiner Tree of size at most l.
Since taking a non-terminal edge to the solution the overall sum is decreased, as few of these edges are to be taken. Thus, the optimal solution of the k-vCMECG instance is given by the Steiner Tree of the terminals R plus the new terminal edges. The denominator of the solution for the k-vCMECG instance resolves to $(l+k+1)$ as this problem is optimizing the weight of the solution with respect to the number of nodes which, in trees, is equal to the number of edges + 1. Equally, the optimal solution for the Steiner Tree is given by the solution for k-vCMECG by erasing all terminal edges.
$\Rightarrow$ Steiner Tree $\leq_m$ k-vCMECG.

□

Theorem 4.5.9. *k-veCMECG is NP-hard.*

Proof. Steiner Tree $\leq_m$ k-veCMECG: Given a Steiner Tree instance, $G = (V, E)$, $\omega : E \rightarrow 1$, $R \subseteq V$, and l. Construct an instance of k-veCMECG, $G' = (V', E')$, V_k, s as follows:
Add a degree-one vertex to every terminal and set the weight of the corresponding new terminal edge $\omega(e)$ to $\omega(e) = |E| = m$. The weights of all remaining edges are set to zero. V_k is identified with R. We obtain $G' = (V', E')$ with $|V'| = |V| + |R|$, $|E'| = |E| + |R|$, V_k=R, and $s = l + |R|$.
Now, the following is easy to prove: k-veCMECG instance has a solution of size at least $mk/(l+k)$ if and only if the Steiner Tree instance allows for a Steiner Tree of size at most l.
Since taking a non-terminal edge to the solution the overall sum is decreased, as few of these edges are to be taken. Thus, the optimal solution of the k-veCMECG instance is given by Steiner Tree of the terminals R, plus the newly added terminal edges. The denominator of the solution for the k-veCMECG instance resolves to $(l+k)$ as this problem is optimizing the weight of the solution with respect to the number of edges which is given by the size of the solution of the Steiner Tree plus the number of newly added terminal edges. Equally, the optimal solution for the Steiner Tree is given by the solution for k-veCMECG by erasing all terminal

edges.
$\Rightarrow$ Steiner Tree $\leq_m$ k-vECMECG.

□

Theorem 4.5.10. 1-vCMECG *is NP-hard.*

Proof. Steiner Tree $\leq_m$ 1-vCMECG: Given a Steiner Tree instance, $G = (V, E)$, $\omega : E \rightarrow 1$, $R \subseteq V$, and l. Construct an instance of 1-vCMECG, $G' = (V', E')$, $V_1 = v_{seed}$, s as follows:
Set v_{seed} to an arbitrary terminal vertex. Add a degree-one vertex to the terminal vertex selected as v_{seed} and set the weight of the corresponding new edge $\omega(e)$ to $\omega(e) = |E| = m$. The weight of all remaining edges are set to zero. We obtain $G' = (V', E')$ with $|V'| = |V| + 1$, $|E'| = |E| + 1$, $v_{seed} \subseteq R$, and $s = l + 1$.
Now, the following is easy to prove: 1-vCMECG instance has a solution of size at least $m/(l + 1)$ if and only if the Steiner Tree instance allows for a Steiner Tree of size at most l.
Since taking a non-terminal edge to the solution the overall sum is decreased, as few of these edges are to be taken. Thus, the optimal solution of the 1-vCMECG instance is given by the Steiner Tree of the terminals R plus the new terminal edge added to v_{seed}, i.e. to one of the terminals out of $|R|$. The denominator of the solution for the 1-vCMECG instance resolves to $(l+1)$ as this problem is optimizing the weight of the solution with respect to the number of nodes which, in trees, is equal to the number of edges + 1. Equally, the optimal solution of the Steiner Tree is given by the solution for 1-vCMECG by erasing the terminal edge.
$\Rightarrow$ Steiner Tree $\leq_m$ 1-vCMECG.

□

Chapter 5

Discussion

Based on gene expression data obtained by genome-wide methods, we want to understand the biological processes affected by TGF-β1 stimulation with or without simultaneous treatment with inhibitors of the TGFBR1 kinase domain. Present methods focus on determining differentially expressed genes. Based on these genes, gene set enrichment analyses are conducted or small modules are extracted out of protein interaction networks. Such methods neglect the vast amount of prior biological knowledge available in the public domain. This data should be used to add information to the analyses especially with respect to the fact that not all genes are regulated on the transcriptional level. I presented a method, that combines approaches like jActiveModule, making use of protein interaction data and gene expression data, and Pandora, which integrates additional data sources. As a result, I have been able to extract small modules that helped revealing the effects present in a given expression experiment. I exemplarily applied this method to a gene expression experiment conducted for the analyses of compounds' mode of action. Applying my approach proved useful to reveal off-target effects, too. The results from my analysis could be confirmed by wet-lab experiments. I showed that the proposed data integration method is superior to STRING-based scoring for the biological processes and genes that are affected in our experiment. Interaction modules derived by modEx are more informative than the modules derived by jActiveModule. In this chapter, I will discuss the results presented in Chapter 4.

5.1 Normalization

In order to put subsequent analyses on a reliable basis, it is important to select appropriate pre-processing methods for a given data set based on the experimental setup used [39]. If sample sizes of the different groups are relatively small, it is crucial to achieve a homogeneous variance for the groups. On the contrary, if sample sizes are large, variances can be estimated reliably and one should focus on calculating unbiased fold changes. Since the sample sizes for the current data set are rather small (three to four replicates per group), a stable variance is more important than an exact representation of the fold change. In general, the data should be normalized without too much reducing real variations. Figure 4.12 on page 4.12 summarizes the quality measures for all methods we investigated, demonstrating the background for the final choice. Clustering of the quality scores assigned reveals two major tendencies based on background normalization.

(i) Background normalized data (`bg_*`) tend to better reflect the real fold changes, i. e. show less bias.

(ii) Pre-processing without background normalization (`noBg_*`) leads to a more homogeneous variance.

This could be explained by accurately defined, constant experimental conditions across all experiments. As they have been conducted in parallel, this possibly led to a relatively consistent background level across all samples. Thus, background correction would introduce additional variation.

Methods combining background normalization with `vst` (`bg_vst_*`) constitute an exception. Here, `vst` leads to a better stabilization of variance while introducing more bias. As `vst` estimates an additional offset for the background based on the data [157], `noBg_vst_*` and `bg_vst_*` pre-processing methods lead to approximately similar results.

As shown in Section 4.1, there are several pre-processing methods resulting in nearly equal quality. Therefore, it is not possible to give a well-defined rationale for using only one specific method. After excluding the methods that clearly violate the imposed criteria, the decision is still subjective. It depends, e. g. on whether one

would like to account for a good estimate of fold changes or a small and homogeneous variance. With the analyses and criteria described here, a recommendation on the pre-selection of appropriate methods is provided. For the data set under study, we intended to achieve a low and homogeneous variance. Therefore, I provided extensive statistics investigating variance. If focus was on a good estimate of fold change, the researcher should account more for statistics investigating this measure. Correlation to results from qRT-PCR or slope and intercept of the regression between qRT-PCR data and gene expression fold changes are examples of analyses that could be of higher interest in this context (Section 4.1.2, pages 93 ff.). With regard to variance, best suited for the data set analysed here are `noBg_log_quantile` and `noBg_log_rsn`. Although log_2-transformation in combination with quantile normalization has been reported to perform relatively well by Du *et al.* [154] and Dunning *et al.* [200,201], due to my analyses we decided to make use of robust spline normalization (`rsn`). Quantile normalization preserves the rank order of genes but intensities are transformed discontinuously since intensity values of different microarrays are forced to follow the same distribution. Spline normalization, in contrast, provides a continuous mapping. `rsn` combines the positive features of quantile normalization and spline interpolation. Due to properties of quantile normalization the rank order of genes is preserved and due to the spline normalization intensities are transformed in a continuous manner [154, 202]. Surprisingly, the use of `vst` as recommended by Dunning *et al.* [201] and by Du *et al.* [154, 157, 202] and the combination of `vst` with `rsn` as successfully used by Du *et al.* [202] did not perform as well as expected. Reasons for this could be the different experimental setups (two replicates per group in the Barnes study [203] used for validation of `vst`, compared to three to four replicates in our setup) or the use of a newer Illumina chip technology, namely HumanHT-12 v3 chips, in our experiment. `vst` has been validated based on a pre-released version of the Human-Ref-8 v1 Expression BeadChip that contained 19 (25% quantile) to 30 (75% quantile) beads per probe. On the HumanHT-12 v3 chips an average of only 15 beads per probe is available. Since `vst` makes use of those technical replicates, this could lead to a slightly worse performance on the new chip generation. In general, `vst` still performs well in stabilizing the variance but is outperformed by `noBg_log_quantile`, `noBg_log_rsn`, and `noBg_vsn` in reflecting the expression values as measured by qRT-PCR. When restricting to methods implemented in BeadStudio, in accordance with Dunning *et al.* [200, 201] who advised

against the use of background normalization, I recommend using the cubic spline method without background normalization (`noBg_cubicSpline`). As displayed in Figure 4.12, `noBg_cubicSpline` outperforms all other BeadStudio normalization methods. Spike-in or dilution data is frequently used for evaluating different normalization methods [157, 185, 200, 201, 204]. If no such data is available for the microarray chip type used, its advisable to perform qRT-PCR measurements for genes covering different spectra of expression intensities in order to obtain a measure for judging the quality of pre-processing methods. Thereby, it becomes possible to determine how well different normalization methods are able to reflect the real changes in expression intensities across different expression levels.

In summary, we provide statistical measures based on which researchers can decide on the best suited pre-processing scenario for their own experimental design. It is also possible to estimate the bias of log_2 ratios obtained from normalized data. In conjunction with the measures for the variability of the data the basis for weighing well measured changes versus low and homogeneous variance is delivered, and by this means selecting an appropriate normalization method is feasible.

5.2 Data Integration

In Section 3.4, I proposed a scoring function to describe the pairwise relatedness of proteins, which is both, easy to use and easy to interpret. This score is highly flexible by allowing integration of prior knowledge and by offering the possibility to differentially weight individual evidence. Prior knowledge of protein interactions (Section 3.4.1) is enriched by information on BP, MF, and CC as annotated by the Gene Ontology (Section 3.4.2) and the similarity of promoter regions (Section 3.4.3). Literature-based evidence (Section 3.4.4) and gene expression data (Section 3.4.5) were integrated as well. By these data sets, I claim to obtain more informative edge scores than STRING (Section 2.2.5), which does neither make use of GO nor of transcription factor binding site information. Further, neither STRING- nor LLS- (Section 2.2.6) or Pandora- (Section 2.3.11) based scoring methods offer the possibility to extend the score for data derived in an experiment, like we exemplarily did for the gene expression data (Section 3.4.5). Introducing weighting factors a_i (Equation 3.30, page 71) offers a possibility to weight individual evidence accord-

ing to biological expert knowledge. As I could show in the analyses conducted in Sections 4.3.2 and 4.3.3 meaningful results are achieved by weighting the individual scores by setting $a_i = 1$ for all evidences used. Thus, the method can also be used in an unbiased fashion, if no expert knowledge is available.

I compared different options of how the individual evidence could be integrated. By applying a pseudo-ROC based approach, I decided for the best combination with respect to a set of known physical protein-protein interactions present in iRefIndex that are most likely to represent real functional associations of proteins. This approach suffers, however, from the same drawback as all methods referring to a gold standard like STRING or LLS, namely the still limited biological knowledge (high number of false negative findings).

In the presented applications, I claim that our data integration method is superior to STRING-based scoring. This can be explained by the fact that I integrated GO, arguably the currently best annotated and most comprehensive annotation of genes and their products with respect to both, the biological processes they affect and their function. The first to integrate Gene Ontology annotations to derive functional modules were Wu *et al.* [167]. Using a Bayesian approach, they combine results for the analyses of phylogenetic profiles, gene neighborhoods and Gene Ontology annotations. The combined information is used to measure the strength of gene functional relationship based on which functional modules present in *Escherichia coli* were predicted. The method was not designed to reveal the effects present in a specific experiment under investigation. GO is also integrated in Pandora (Section 2.3.11) but the aim of Pandora is to derive complete pathways, but not small modules.

All of the approaches integrating prior knowledge depend on the quantity and quality of the available knowledge. If this knowledge is limited, their performance will be limited as well. It thus makes sense to combine different sources like STRING or LLS annotations using the presented integration approach. The scores based on STRING and LLS could easily be added as a further data source using Equation 3.30. Alternatively, modEx (Section 3.5) or other available analyses could be performed based on each of the differently scored interaction networks, and the results could

be interpreted separately to help in understanding the underlying biology.

5.3 modEx

Based on the methods presented in Section 3.4, different data types are integrated into a protein interaction network to obtain an edge- and node-weighted graph $G = (V, E)$. In Section 3.5.1, I defined some graph theoretical problems that could be posed based on networks weighted using the described approach. Solving the underlying questions on this basis, it is possible to get biological insight, for example on the mode of action of compounds. By the data integrated, I do not only focus on deregulated genes but consider the functional relatedness of the proteins based on additional previous knowledge. In contrast, most other module extraction methods discussed in Section 2.3 are heavily based on information from gene expression profiling (Table 2.1, page 2.1). In Section 3.5.2 I introduced modEx to heuristically solve the 1-VERTEX CONSTRAINED MAXIMUM EDGE-WEIGHT CONNECTED GRAPH (1-VCMECG) problem (Definition 3.5.4). modEx is applied in Section 4.3.3 to identify modules that help to elucidate the biological processes affected in the gene expression experiments under investigation. Many of the previously described methods try to derive one main module that is affected by the deregulated genes. jActiveModule, for instance, reports the highest scoring component G_w (Figure 2.3) as "signaling or regulatory circuit of high biological interest" [48]. I, in contrast, argue that changes in gene expression could be induced through different processes, thus, we want to look for several modules that could help to elucidate these processes. The actual implementation of jActiveModule returns not only the highest scoring component but all significantly active components. Still, these are rather large and for the comparison analyses conducted, only five modules were returned (Section 4.4.2, pages 127 ff.). There, I could show that modEx results are still superior in linking the biological experiment to the specific processes. Since we do not know *a priori* how many processes are affected by compound treatment (on- as well as possibly several off-target effects), we do not know how many subgraphs we are looking for. Some of the nodes in the protein interaction network may belong to several processes/modules simultaneously. This has also not been taken into account by previous approaches. The newly proposed approach looks for dense networks to more likely derive subnetworks containing proteins/genes (nodes) that are relevant

in the same biological context. Thus, the score is optimized relative to the number of nodes and the node-induced subgraph represents our result. To detect all modules necessary to investigate the underlying processes, modEx solves a special case of the k-vCMECG problem (Definition 3.5.2), namely the 1-vCMECG (Definition 3.5.4) problem. By solving this problem for different seed nodes, v_{seed}, it is possible to extract modules that are related to the different underlying processes. In the analyses described in Sections 4.3.2 and 4.3.3, the 10 strongest deregulated genes of the underlying gene expression experiment were selected as v_{seed}s. Thereby, it is possible to detect interesting processes that agree with experimental findings. It would also be possible to re-apply modEx until all genes that are defined to be differentially expressed based on a specific cutoff are contained in at least one of the extracted modules. Thereby it is possible to generate a result that very likely covers all actually affected processes by at least one of the modules.

The optimization problem, as described in the methods section, is related to the Constrained Maximum Edge-weight Connected Graph Problem, which has been shown to be NP-hard [197]. By proving the NP-hardness, a theoretical justification for the use of a heuristic is given. Additional to the theoretical justification, seen from a biological point of view, it could be reasonable to solve the problem in a non-optimal way:

- The theoretical problem would look for one subnetwork that optimizes the objective function. The biological system under investigation could be affected in several biological processes which could be independent from each other. Thus, it is more biologically relevant to look for several subnetworks.

- We do not know all protein interactions, thus, the graph underlying the problem already suffers from incorrectness. Thus, it is to question whether a correct solution for the incorrect protein interaction network stays correct knowing all correct protein interactions.

- The correctness of the complete protein interaction network does not only depend on our biological knowledge but also on the state of the biological system under consideration. Thus, components of the network may not be present under a specific condition, or may change in a time-dependent manner.

I first applied a greedy search as heuristic algorithm. Since the extracted modules were very small, simulated annealing was used instead. This resulted in modules of adequate size. Another more empirical reason why simulated annealing could be beneficial is that missing prior knowledge could lead to artificial optima. This could be due to edges that score low based on the lacking knowledge. With complete knowledge they possibly would score high and lead to a higher $\Omega(G'_i)$ (Definition 3.5.4). By applying simulated annealing instead of a simple greedy search or exact algorithms, we have the chance to overcome local optima, while at the same time tolerating a certain degree of nescience. At the same time, we take the risk of a fuzzy solution containing proteins that are not closely related to the affected biological process. Since our current knowledge is rather limited for many biological areas, I argue that it is better to allow for some false positive proteins than to miss some important relationships in an optimal solution. False negative relationships are a greater concern than false positive findings since false positives can be identified in downstream wet-lab validation. False negatives in contrast are more difficult to detect in follow-up experiments.

5.4 Analysis of Compounds' Mode of Action Using modEx

I presented the combination of an edge scoring method and an extraction method for the analysis of biological effects based on gene expression experiments. Using this approach, it is possible to detect on- as well as potential off-target effects of compounds. For the TGF-β experiment under consideration, these effects have been validated by wet-lab experiments [54]. Since we investigated effects mediated by kinase inhibitors which usually are propagated through direct protein-protein interactions, I decided to use iRefIndex as underlying network for data integration. In Section 4.4.1 I have shown that it could be beneficial to use STRING with an adjusted edge scoring. This could lead to even better results, especially in cases where the biological effects are not necessarily dependent on direct protein interactions. I could also demonstrate that for our application modEx is superior to jActiveModule. The extracted modules show higher relation to the biological process under investigation. Additionally, the modules are much smaller than the ones found by jActiveModules. Thus, they can be analyzed in more detail and, if necessary, even

manually.

By applying the annotation and extraction methods proposed, it was possible to derive protein interactions that are meaningful in the biological context under consideration. I exemplarily showed this by investigating gene expression of TGF-β1-stimulated cells compared to unstimulated cells. Modules extracted based on this experiment should be related to TGF-β signaling. When conducting Fisher's exact test based on gene sets predefined by Reactome and KEGG, the significantly enriched gene sets obtained by using the newly proposed method can directly be linked to TGF-β signaling (Figure 4.21).

In addition, I was able to detect on- as well as off-target effects of TGF-β receptor 1 kinase-inhibiting compounds that could be confirmed by wet-lab kinase-screens. Networks extracted based on BI4 expression data reveal direct on-target effects (Section 4.3.3, Figure 4.22). Off-target effect for BI4 are very rarely observed in the networks (Figure E.2). In contrast, networks extracted based on data from experiments involving other compounds are enriched with different signaling cascades. By kinase-screens it could be confirmed in wet-lab experiments that kinases triggering the detected signaling cascades are direct off-targets of the respective compounds (Appendix F). Exemplary results for these off-target effects are shown for BI1 (Section 4.3.3, Figures 4.25 and E.1). Looking at the on-target effects of BI1, it is shown in Figures 4.23 and 4.22 that these are not as pre-dominant as for BI4 (Figure 4.22). These *in silico* findings identify BI1 as the less favorable compound exhibiting more off-target effects than BI4. This is reflected by the wet-lab finding [52, 54].

The quality of modules derived by the proposed method strongly depends on the underlying network used. By comparing results derived using different networks as well as different edge scores, I could show that STRING$_{mod}$ outperforms the iRefIndex based network as well as STRING$_{org}$ (Section 4.3, Figure 4.27).

Since different methods have been proposed for identifying active modules within a protein network, I compared modEx to one of the most prominent methods, jActiveModule. Ad for modEx, findings indicate that using STRING$_{mod}$ as input to jActiveModule is superior to using iRefIndex (Section 4.4.2, Figure 4.28). Further,

I could show that modEx is superior to jActiveModule in cases were we want to derive modules that are not only enriched with deregulated genes but also are related to the underlying processes (Section 4.4.3, Figures 4.29 and 4.30). One reason for this is the fact that not all genes relevant for a biological processes are regulated on the transcriptional level. Thus, to shed light on the biological process or mechanism of interest, the benefit of my method is the integration of knowledge that goes beyond differential expression. Along these lines, the reason for STRING outperforming iRefIndex is that STRING not only considers physical protein interactions but also functionally related proteins. By weighting the resulting edges I introduced a measure for the likeliness of the relations. One probable reason for my weighting scheme outperforming the native STRING weighting is that, based on GO biological process or the cellular component, additional prior knowledge about the relatedness is integrated. Obviously, I cannot claim that my scoring performs better than the STRING scoring for processes for which we do not have any additional *p*rior knowledge. For these cases, it could be beneficial to combine the scoring method with the final STRING score by using the latter one as an additional term in the linear combination (Equation 3.30). The results derived could be used complementarily to results based on either the original STRING scoring or my scoring, or both. Thus, modEx could be applied even with limited or no evidence to elucidate the biological processes under investigation.

One major advantage of the proposed method is its flexibility with respect to weighting single pieces of evidence. Based on the research objective and in close cooperation with experts in the biological context, the scores can be individually weighted and optimized. Furthermore, it is easy to extent the score to further evidence of relatedness.

The proposed method can also be used to complement other biological analyses. Identification and validation of new drug targets may serve as an example in this regard. To identify new targets, proteins that are known to be related to the disease under investigation could serve as seed nodes for modEx. Resulting modules could then be further analyzed for interesting intervention points. In cases proteins that are already targets in a different indication area are contained in the extracted modules, target repositioning is an option. As soon as potent compounds are available, even trying to reposition the drug would be applicable. Similarly, basic *in*

silico validation of targets is possible. For validating a target, the respective protein would be selected as seed node based on which a module is subsequently extracted. This module then can be a starting point for further analyses to confirm the indication area or to check for potential side effects or drawbacks of the target under consideration.

Chapter 6

Conclusion & Future Work

In this work, I first proposed a way to select the best suited normalization procedure for the underlying expression data. Thereby, downstream analyses are founded on a sound basis. The proposed statistical measures can easily be used for other expression experiments to guide the selection of an appropriate normalization procedure.

To analyze compounds' MoA, I introduced a method that weights interactions between pairs of proteins based on different kinds of evidence. As underlying networks, iRefIndex [96] and STRING [55] were exemplarily used. In general, other types of networks could be used as additional sources for protein interaction. I did not make use of protein interactions provided by CORUM [205], DIP [206], and HPRD [102] since these were not freely available. In principle these data could be easily added.

The relevance of proteins is based on the biological relatedness to other possibly non-deregulated protein-coding genes. Thereby, I expand the analysis beyond transcriptional deregulation. To elucidate the biological relatedness, information on molecular function, biological processes and cellular compartment, information on transcription factor binding sites and literature-based confidence scores are integrated for weighting the edges between pairs of proteins. Integration of phenotypic information as, for example, taken from the mammalian phenotype ontology [207, 208], human phenotype ontology [209] or disease ontology [210–212] could be beneficial. Thereby, it could be possible to further improve my method and make it even more valuable for biological analysis. As the proposed information is

represented as ontologies, it would be possible to apply the methods used in the present work to integrate GO. Other information that could be integrated are, e. g. information on miRNA targets [213] or information on phosphorylation sites [214].

To transfer the network into the biological context of interest, expression experiments are used as anchoring points for the analyses. For all the methods proposed, instead of using correlation/covariance of gene expression over time to calculate $score_{exp}$ (Equations 3.27 - 3.29) it is also possible to use correlation/covariance of different dosages or over different treatments or combinations of both. What is used depends on the biological question.

Further, I introduced modEx, a method to extract small modules out of a weighted protein interaction network. These modules hint at the MoA of the compounds used in our expression experiment. In my analyses, ten networks were extracted based on the strongest deregulated genes. This is a very pragmatic approach. A more comprehensive analysis would determine the number of networks to extract based on the actual data. One possibility could be to extract networks until all genes that are deregulated based on a predefined cutoff (e.g. p-value < 0.01 and $|log_2 ratio| > 1$) are contained in the extracted modules. Thereby, it would be more likely to explain all effects present in the experiment under investigation.

I show that for the expression data set used, the proposed edge scoring is superior to the STRING scoring. It would be worth investigating how the results are changed if the native STRING score is used as further evidence in our edge score (Equation 3.30). Finally, I could show that modEx extracts modules that better represent the underlying mechanism than jActiveModule. In this work, modEx is only compared to jActiveModule. It would be interesting to compare this method to further available approaches like GXNA, ClustEx or Matisse. The disadvantage of ClustEx is that is has very long run times and Matisse is only available for academically funded work.

One huge challenge that holds for all methods that return a subnetwork or a set or list of genes is how to further analyze these genes. As long as the network or list is small (<20 genes), one labor-intensive possibility is to manually go through the list

and try to explain the biological process based on the individual genes in the list. I proposed to use a very standardized method, gene set enrichment based on Fisher's exact test. Drawback of gene set enrichment is that results are based on pre-defined gene sets which are very likely incomplete based on our current knowledge and derived results are at most as good as the annotations. Further, it is not possible to detect new relationships. To close this gap, further research is needed in this regard.

Scientists of different disciplines are working together to continuously improve and deepen our constantly increasing understanding of biology. By this means, results derived by analyses as described in the present work are also permanently improving. As more and more prior knowledge is available, developments in the direction of data mining and data integration has to be done to find optimal ways to make use of this tremendous knowledge. Further, new or improved methods to extract the relevant parts of information have to be developed. With the present work, I contributed to the initial development of such methods. By further advancing research in these directions, it will get more and more possible to elucidate different kinds of biological processes and to successively solve the secrets of life.

Appendix A

TGF-β Signaling

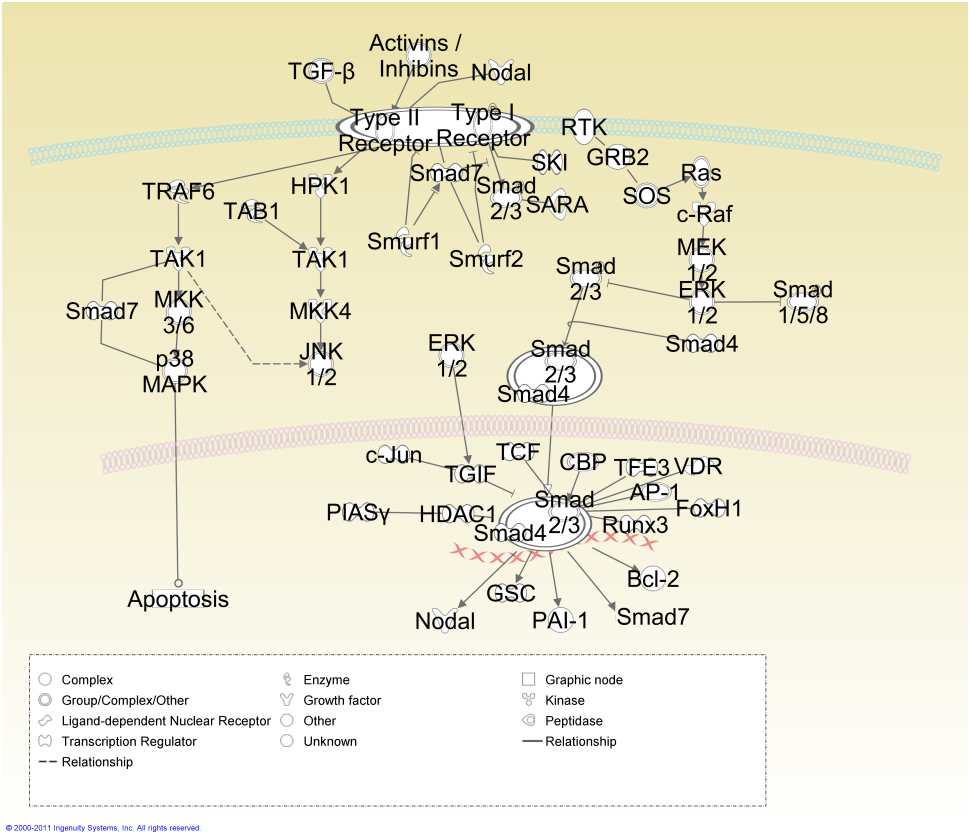


Figure A.1: **Biological representation of the TGF-β signaling pathway.** Displayed are the most important molecular events involved in TGF-β signaling via SMAD proteins and via TRAF6, an exemplary SMAD independent way. Pathway illustration was taken from Ingenuity Knowledge Base [144] library of canonical signaling pathways.

Appendix B

Comparison of Pre-Processing Methods

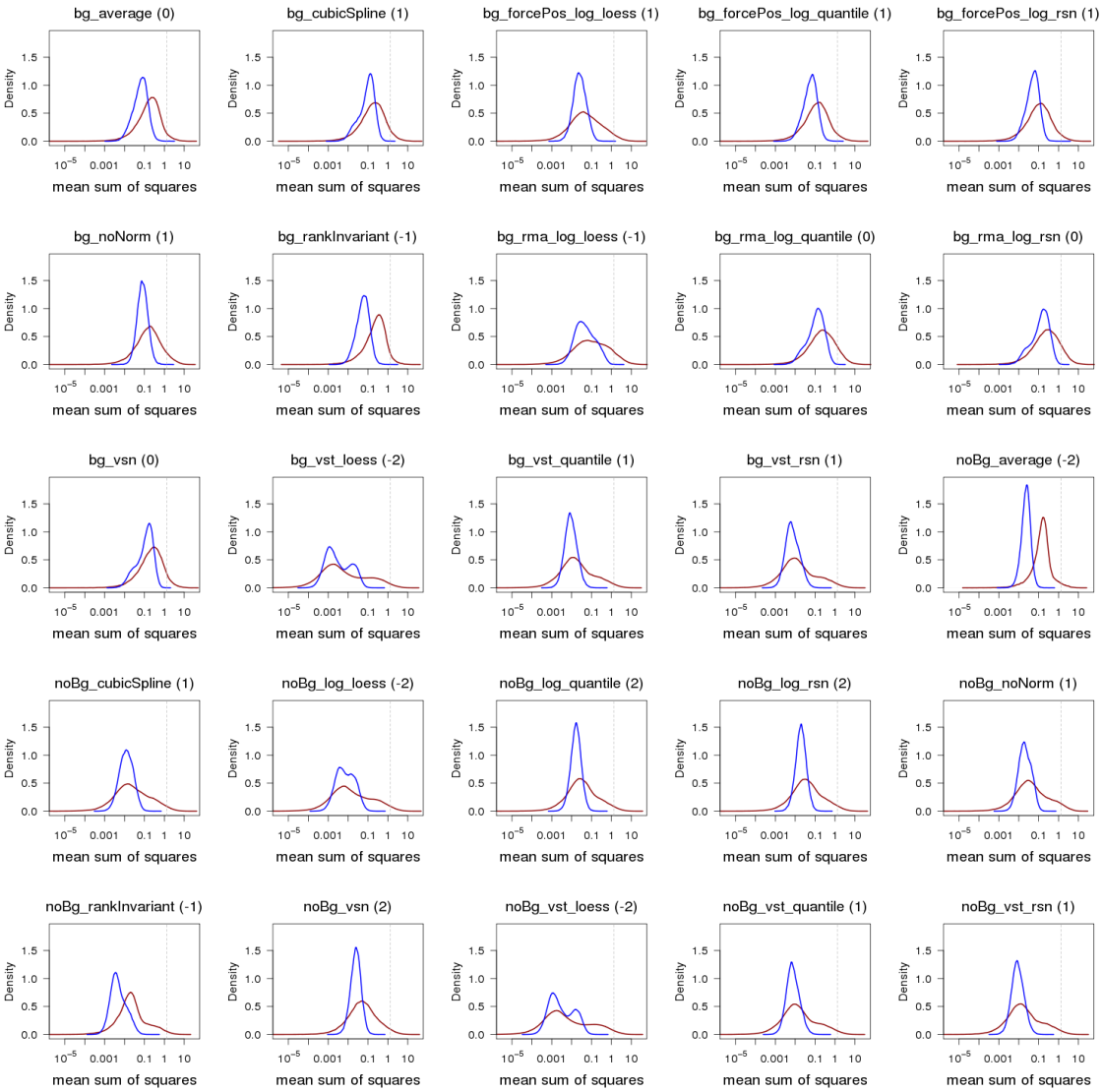


Figure B.1: Density plots of MSQ_{within} (blue) and $MSQ_{between}$ (red). MSQs were calculated based on the gene expression measured for the three sample groups analyzed, namely untreated HaCaT cells after 2, 4, and 12 hours. The grey dashed line indicates the expected value for the $MSQ_{between}$ of 1.33 based on 6, 6, and 7 as measurements for the group means of four replicates for three time points. Three examples of different quality are shown.

Appendix C

Comparison Between STRING$_{org}$, STRING$_{mod}$, and iRefIndex

Table C.1: Comparison of modules extracted using modEx based on STRING$_{org}$, STRING$_{mod}$ and iRefIndex. iRefIndex and STRING$_{mod}$ are weighted as described in Section 4.3.1, pages 111 ff.. Gene expression data of TGF-β1-stimulated and unstimulated cells across 2, 4, and 12 hours was used to calculate $score_{exp}$ (Eq. 3.28, page 70). log_2ratios and FDR-corrected p-values (Section 3.3.1, page 55) were used to select seed nodes. Additionally, for STRING$_{org}$ and STRING$_{mod}$, different cutoffs were applied to the original network (0, 400, 700, 834, 848, 900). Displayed are the number of nodes/edges that are contained in the modules extracted using modEx. Subsequently, the respective modules are used in gene set enrichment analyses using Fisher's exact test (Tables C.2 and C.2). Different cutoffs were applied to the original STRING network to approximate the size of iRefIndex and to exclude FP associations. The size of iRefIndex is best approximated using 848 as cutoff. As mentioned in Section 3.7, page 77, data referring to this cutoff was used in the main part of this thesis.

	Confidence score used as cutoff for edges								
	STRING$_{org}$						STRING$_{mod}$		iRefIndex
seed node	≥ 0	≥ 400	> 700	≥ 834	≥ 848	≥ 900	≥ 848	≥ 900	
VASN	133/ 3490	133/ 1861	133/ 1343	133/ 1061	133/ 1038	nnp	12/ 23	nnp	12/17
SERPINE1	60/ 902	60/ 602	14/ 48	14/ 44	14/ 43	14/ 41	4/ 4	10/19	23/39
CTGF	83/ 1703	83/ 997	83/ 736	83/ 595	83/ 581	83/ 527	11/ 22	6/7	11/16
JUN	91/ 2062	91/ 1122	91/ 831	91/ 663	91/ 646	91/ 567	11/ 22	10/19	6/9
TGM2	63/ 1130	63/ 766	63/ 624	63/ 506	63/ 497	63/ 439	15/ 29	14/16	16/24
KANK4	64/ 912	nnp	nnp	nnp	nnp	nnp	nnp	nnp	nnp
FOSB	65/ 1146	65/ 643	65/ 481	65/ 375	65/ 363	65/ 316	11/ 14	9/10	9/15
BHLHE40	11/ 55	11/ 55	11/ 54	11/ 53	11/ 53	11/ 45	15/ 37	13/28	18/29
CDKN2B	69/ 1255	69/ 926	69/ 751	69/ 677	69/ 670	69/ 611	21/ 75	20/63	10/23
SMAD7	60/ 828	60/ 486	60/ 360	60/ 310	60/ 307	60/ 611	7/ 9	10/19	21/42
SOX18	NA	18/ 71	18/ 58	18/ 45	8/ 11	nnp	23/ 41	nnp	24/36
IER3	NA	NA	NA	NA	NA	105/ 652	NA	14/37	NA
SKIL	NA	NA	NA	NA	NA	112/ 692	NA	11/20	NA
union graph	382/ 8599	355/ 4729	319/ 3199	319/ 2628	310/ 2541	331/ 2467	71/ 173	65/ 158	92/164
conComp	371/ 8544	344/ 4674	277/ 3039	277/ 2486	277/ 2432	306/ 2381	20/ 35	18/ 30	57/102

nnp: node not present; due to reduction of edges isolated nodes occur. These were removed since it is useless to take them as seed nodes.
NA: nodes were not treated as seed nodes, since more significant nodes are still present (not nnp).
Union graph: Union graph refers to the network as obtained by the union of the ten individual nets extracted.
conComp: Connected component containing most seed nodes.

Table C.2: Results of gene set enrichment by Fisher's exact test based on Reactome. Fisher's exact test was conducted based on Reactome predefined gene sets for modules extracted by modEx based on STRING$_{org}$, STRING$_{mod}$, and iRefIndex networks (Table C.1). Displayed are p-values/ranks of Reactome TGF-β signaling gene set for p-values < 1. Different cutoffs were applied to the original STRING network to approximate the size of iRefIndex and to exclude FP associations. The size of iRefIndex is best approximated using 848 as cutoff. As mentioned in Section 3.7, page 77, data referring to this cutoff was used in the main part of this thesis.

	Confidence score used as cutoff for edges								
	STRING$_{org}$						STRING$_{mod}$		iRefIndex
seed node	≥ 0	≥ 400	> 700	≥ 834	≥ 848	≥ 900	≥ 848	≥ 900	
VASN	<0.0001/ 11	<0.0001/ 11	<0.0001/ 11	<0.0001/ 10	<0.0001/ 9	nnp	0.0032/ 9	nnp	-
SERPINE1	-	-	-	-	-	-	-	0.0031/ 8	<0.0001/ 17
CTGF	<0.0001/ 7	<0.0001/ 7	<0.0001/ 7	<0.0001/ 7	<0.0001/3	<0.0001/ 6	0.0029/ 9	0.0018/ 9	-
JUN	<0.0001/ 5	<0.0001/ 5	<0.0001/ 5	<0.0001/ 5	<0.0001/5	<0.0001/ 5	0.0029/ 9	0.0031/ 8	-
TGM2	-	-	-	-	-	-	0.004/ 17	0.0043/ 16	-
KANK4	-	nnp	nnp	nnp	nnp	nnp	nnp	nnp	nnp
FOSB	-	-	-	-	-	-	0.0029/ 9	0.0028/ 8	-
BHLHE40	-	-	-	-	-	-	-	-	<0.0001/ 8
CDKN2B	-	-	-	-	-	-	-	-	-
SMAD7	<0.0001/ 6	<0.0001/ 6	<0.0001/ 6	<0.0001/ 6	<0.0001/5	<0.0001/ 5	0.0019/ 9	0.0031/ 8	<0.0001/ 2
SOX18	NA	-	-	-	-	nnp	0.0061/ 22	nnp	-
IER3	NA	NA	NA	NA	NA	<0.0001/ 18	NA	-	NA
SKIL	NA	NA	NA	NA	NA	<0.0001/ 17	NA	0.0034/ 8	NA
union graph	<0.0001/ 82	<0.0001/ 84	<0.0001/ 66	<0.0001/ 60	<0.0001/54	<0.0001/ 66	0.0001/ 29	0.0199/ 76	<0.0001/16
conComp	<0.0001/ 81	<0.0001/ 83	<0.0001/48	<0.0001/48	<0.0001/51	<0.0001/ 63	0.0053/ 16	0.0055/ 16	<0.0001/4

nnp: node not present; due to reduction of edges isolated nodes occur. These were removed since it is useless to take them as seed nodes.

NA: nodes were not treated as seed nodes, since more significant nodes are still present (not nnp).

Union graph: Union graph refers to the network as obtained by the union of the ten individual nets extracted.

conComp: Connected component containing most seed nodes.

Table C.3: Results of gene set enrichment by Fisher's exact test based on KEGG. Fisher's exact test was conducted based on KEGG predefined gene sets for modules extracted by modEx based on STRING$_{org}$, STRING$_{mod}$, and iRefIndex networks (Table C.1). Displayed are p-values/ranks of KEGG TGF-β signaling gene set for p-values < 1. Different cutoffs were applied to the original STRING network, first, to approximate the size of iRefIndex, and, second, to exclude FP associations. The size of iRefIndex is best approximated using 848 as cutoff. As mentioned in Section 3.7, page 77, data referring to this cutoff was used in the main part of this thesis.

	Confidence score used as cutoff for edges				
	STRING$_{org}$		STRING$_{mod}$		iRefIndex
seed node	≥ 848	≥ 900	≥ 848	≥ 900	
VASN	<0.0001/ 6	nnp	0.0002/ 2	nnp	0.0852/ 14
SERPINE1	-	-	-	0.0002/ 2	<0.0001/ 6
CTGF	<0.0001/ 4	<0.0001/ 3	0.0002/ 2	<0.0001/ 2	0.0784/ 14
JUN	<0.0001/ 5	<0.0001/ 5	0.0002/ 2	0.0002/ 2	-
TGM2	-	-	0.0003/ 2	0.0003/ 2	0.112/ 16
KANK4	nnp	nnp	nnp	nnp	nnp
FOSB	0.1065/ 48	0.1194/ 48	0.0002/ 2	0.0001/ 2	-
BHLHE40	-	-	-	0.0251/ 8	<0.0001/ 2
CDKN2B	0.0063/ 16	0.008/ 16	0.036/ 16	0.0383/ 13	0.0715/ 11
SMAD7	<0.0001/ 1	<0.0001/ 1	0.0001/ 2	0.0002/ 2	<0.0001/ 1
SOX18	-	nnp	0.039/ 15	nnp	0.0007/ 13
IER3	NA	<0.0001/ 15	NA	-	NA
SKIL	NA	<0.0001/ 6	NA	0.0002/ 2	NA
union graph	<0.0001/ 24	<0.0001/ 27	0.0003/ 8	<0.0001/ 4	<0.0001/ 2
conComp	<0.0001/ 23	<0.0001/ 26	0.0005/ 2	0.0006/ 2	<0.0001/ 2

nnp: node not present; due to reduction of edges isolated nodes occur. These were removed since it is useless to take them as seed nodes.

NA: nodes were not treated as seed nodes, since more significant nodes are still present (not nnp).

Union graph: Union graph refers to the network as obtained by the union of the ten individual nets extracted.

conComp: Connected component containing most seed nodes.

Appendix D

Comparison to jActiveModule

Table D.1: Results of jActiveModule based on iRefIndex. Nodes of jActiveModule instances were weighted according to the FDR-corrected p-values calculated for the comparisons of TGF-β1-stimulated compared to unstimulated cells at 2, 4, and 12 hours (see Section 3.3.1, page 55). Displayed are the sizes of modules detected in the iRefIndex network using jActiveModule (nodes/edges/score). Subsequently, the respective modules are used in gene set enrichment analysis using Fisher's exact test (Tables D.2 and D.3).

	*n*o-pval		*d*ef-pval		*s*ub-net	
	greedyDef	simAnnHub	greedyDef	simAnnHub	greedyDef	simAnnHub
Module1	472/1419/17.8	-	3164/18470/23.2	-	338/488/13.6	1160/1377/14.8
Module2	482/1483/17.6	-	187/616/11.1	-	377/637/13.5	2/1/5.2$^{\#}$
Module3	471/1422/17.3	-	166/820/10.9	-	377/728/13.3	9/15/4.8
Module4	493/1633/17.3	-	113/407/10.3	-	372/623/13.4	6/8/4.3$^{\#}$
Module5	427/1127/17.1	-	113/900/9.8	-	366/620/13.4	2/1/4.1*

$^{\#}$Subnet active at 2h, 4h, 12h, *active at 4h and 12h, all others only active at 12h.

Table D.2: Results of gene set enrichment by Fisher's exact test based on Reactome. P-values and ranks of Reactome TGF-β signaling gene set based on Fisher's exact test conducted for modules identified by jActiveModule in the iRefIndex network (Table D.1).

	*n*o-pval	*d*ef-pval	*s*ub-net	
	greedyDef	greedyDef	greedyDef	simAnnHub
Module1	$4 \cdot 10^{-4}/94$	$7 \cdot 10^{-4}/128$	$-/-^{+}$	$-/-^{+}$
Module2	$4 \cdot 10^{-4}/81$	0.2354/289	$1 \cdot 10^{-4}/20$	$-/-^{+,\#}$
Module3	$3.6 \cdot 10^{-3}/139$	$-/-^{+}$	0.0984/197	$-/-^{+}$
Module4	$4.3 \cdot 10^{-3}/168$	$-/-^{+}$	0.0962/242	$-/-^{+,\#}$
Module5	$2 \cdot 10^{-4}/95$	$-/-^{+}$	0.0138/117	$-/-^{+,*}$

$^{+}$ p-value$>$0.5, $^{\#}$subnet active at 2h, 4h, 12h, *active at 4h and 12h, all others only active at 12h.

Table D.3: Results of gene set enrichment by Fisher's exact test based on KEGG. P-values and ranks of KEGG's TGF-β signaling gene set based on Fisher's exact test conducted for modules identified by jActiveModule in the iRefIndex network (Table D.1).

	*n*o-pval	*d*ef-pval	*s*ub-net	
	greedyDef	greedyDef	greedyDef	simAnnHub
Module1	$1.5 \cdot 10 - 6/14$	$7.3 \cdot 10^{-3}/45$	0.24/57	0.23/36
Module2	$2.0 \cdot 10^{-4}/28$	0.16/55	$5.5 \cdot 10 - 6/6$	0.03/4
Module3	$2.4 \cdot 10^{-3}/44$	0.12/29	$1.8 \cdot 10^{-3}/36$	1/219
Module4	$3.3 \cdot 10^{-3}/43$	0.01/33	$1.6 \cdot 10^{-3}/30$	1/222
Module5	$4.2 \cdot 10 - 7/9$	1/248	$3.4 \cdot 10^{-4}/18$	0.02/6

all subnets are only active at 12h.

Table D.4: Results of jActiveModule based on STRING. Nodes of jActiveModule instances were weighted according to the FDR-corrected p-values calculated for the comparisons of TGF-β1-stimulated compared to unstimulated cells at 2, 4, and 12 hours (see Section 3.3.1, page 55). Displayed are the modules detected in the STRING network (cutoff of 848) using jActiveModule (nodes/edges/score). Subsequently, the respective modules are used in gene set enrichment analysis using Fisher's exact test (Tables D.5 and D.6).

	*n*o-pval	*d*ef-pval	*s*ub-net
	greedyDef	greedyDef	greedyDef
Module1	367/1882/19.0	998/5988/24.6	348/1628/19.1
Module2	387/2001/18.9	1009/7716/23.5	370/1844/18.8
Module3	361/1860/18.9	909/6734/23.4	373/1788/18.7
Module4	363/1825/18.8	938/6953/23.3	357/1618/18.5
Module5	395/2225/18.8	1012/7358/23.3	379/1874/18.4

all subnets are only active at 12h.

Table D.5: Results of gene set enrichment by Fisher's exact test based on Reactome. P-values and ranks of Reactome TGF-β signaling gene set based on Fisher's exact test conducted for modules identified by jActiveModule in the STRING network (Table D.4).

	*n*o-pval	*d*ef-pval	*s*ub-net
	greedyDef	greedyDef	greedyDef
Module1	$1.38 \cdot 10^{-6}$/113	$3.83 \cdot 10^{-4}$/102	$3.20 \cdot 10^{-4}$/137
Module2	$1.57 \cdot 10^{-6}$/126	$1.12 \cdot 10^{-6}$/68	$8.93 \cdot 10^{-7}$/67
Module3	$3.54 \cdot 10^{-8}$/98	$2.00 \cdot 10^{-5}$/75	$2.14 \cdot 10^{-5}$/77
Module4	$1.62 \cdot 10^{-3}$/243	$2.05 \cdot 10^{-5}$/77	$3.42 \cdot 10^{-4}$/103
Module5	$3.43 \cdot 10^{-9}$/96	$6.28 \cdot 10^{-3}$/202	$4.31 \cdot 10^{-4}$/122

all subnets are only active at 12h.

Table D.6: Results of gene set enrichment by Fisher's exact test based on KEGG. P-values and ranks of KEGG's TGF-β signaling gene set based on Fisher's exact test conducted for modules identified by jActiveModule in the STRING network (Table D.4).

	*n*o-pval	*d*ef-pval	*s*ub-net
	greedyDef	greedyDef	greedyDef
Module1	$2.08 \cdot 10^{-10}/22$	$8.52 \cdot 10^{-8}/37$	$7.60 \cdot 10^{-8}/17$
Module2	$4.28 \cdot 10^{-11}/15$	$2.07 \cdot 10^{-11}/38$	$2.23 \cdot 10^{-9}/27$
Module3	$1.76 \cdot 10^{-9}/19$	$4.81 \cdot 10^{-21}/8$	$2.21 \cdot 10^{-11}/16$
Module4	$1.93 \cdot 10^{-11}/20$	$1.12 \cdot 10^{-9}/34$	$1.16 \cdot 10^{-8}/22$
Module5	$4.09 \cdot 10^{-9}/22$	$1.04 \cdot 10^{-16}/18$	$1.95 \cdot 10^{-7}/30$

all subnets are only active at 12h.

Appendix E

Off-Target Analysis

E.1 Gene Sets Enriched for Union Graphs

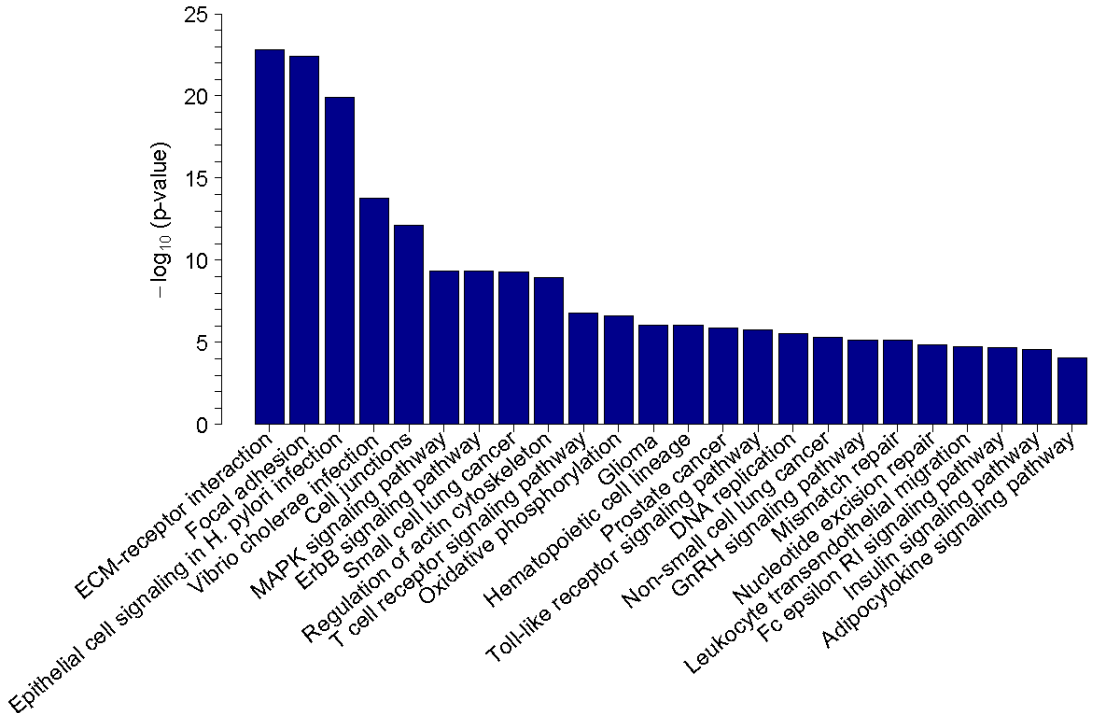


Figure E.1: Gene sets enriched for proteins contained in the union graph of BI1. Off-target analysis of BI1 was conducted as described in Section 4.3.3, pages 117ff.. Gene sets like *MAPK* and *ErbB* signaling can be directly linked to off-targets confirmed in wet-lab experiments (Appendix Table F). Though TGF-βR1 is the intended primary target of BI1, only indirect hints like *Focal adhesion* or *Cell junctions* hint at TGF-β signaling.

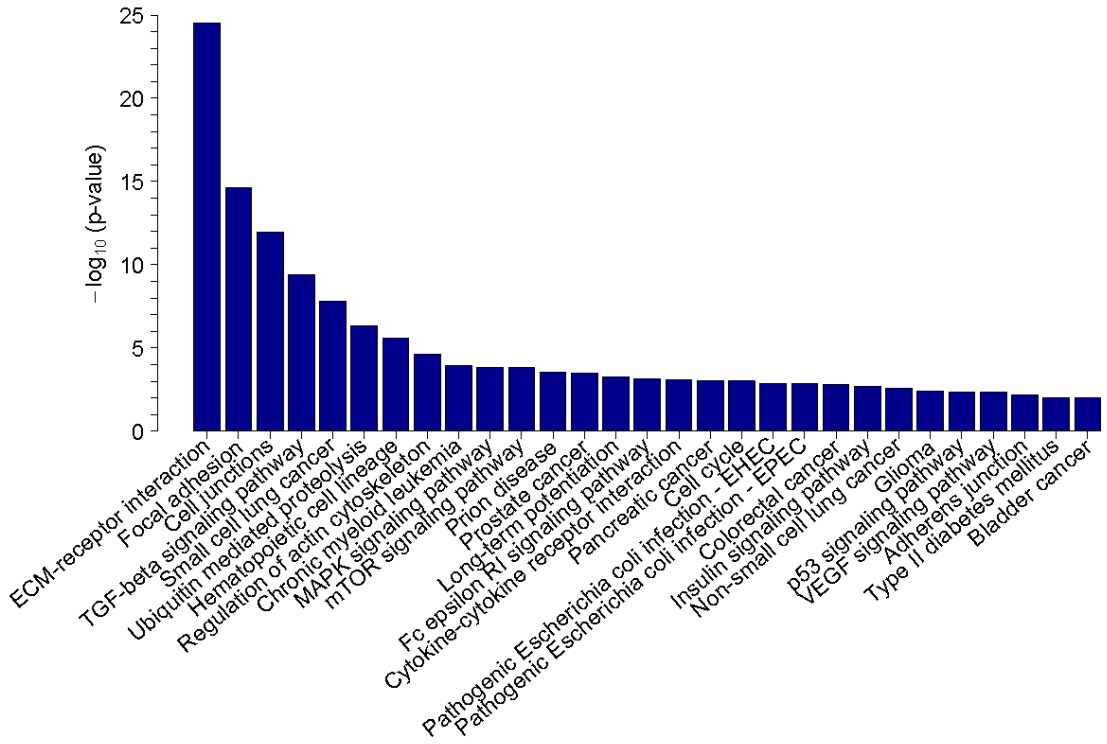


Figure E.2: Gene sets enriched for proteins contained in the union graph of BI4. Off-target analysis of BI4 was conducted as described in Section 4.3.3, pages 117ff.. Only genes which do not exhibit an inverse change compared to TGF-β1 stimulation, the desired effect of the compound, were selected as seed nodes. Still, the connection of the selected nodes to TGF-β related genes seems to be strong such that TGF-β signaling is one of the most significant gene sets (p-value $< 4 \cdot 10^{-10}$). As for BI1, additional gene set hinting at confirmed off-target effects are present. Examples are *MAPK singaling* and *VEGF singaling* (compare Appendix Table F).

Appendix F

Compounds' Wet-Lab Target Validation

F.1 Kinase Screen BI1

Results of kinase screen for BI1 are summarized in Tables F.1 to F.3

F.2 Kinase Screen BI4

Results of kinase screen for BI4 are summarized in Tables F.4 to F.6

Table F.1: Results of kinas screen for BI1 - I. Result of the screen with BI1 against a kinase panel of 239 kinases. Displayed are the percent inhibitions of BI1 for the different kinases tested.

Kinase	% Inhibition
ABL1	97.05
ABL1 Y253F	97.54
ABL2 (Arg)	89.25
ACVR1B (ALK4)	33.23
ADRBK1 (GRK2)	-6.52
ADRBK2 (GRK3)	-6.23
AKT1 (PKB alpha)	26.35
AKT2 (PKB beta)	15.63
AKT3 (PKB gamma)	2.42
ALK	97.48
AMPK A1/B1/G1	98.08
AURKA (Aurora A)	97.97
AURKB (Aurora B)	101.45
AURKC (Aurora C)	93.24
AXL	97.84
BLK	97.91
BMX	57.74
BRAF	98.11
BRAF V599E	100.34
BRSK1 (SAD1)	93.22
BTK	76.45
CAMK1 (CaMK1)	40.56
CAMK1D (CaMKI delta)	50.86
CAMK2A (CaMKII alpha)	76.64
CAMK2B (CaMKII beta)	9.16
CAMK2D (CaMKII delta)	100.00
CAMK4 (CaMKIV)	13.91
CDC42 BPA (MRCKA)	14.23
CDC42 BPB (MRCKB)	17.80
CDK1/cyclin B	83.72

Kinase	% Inhibition
CDK2/cyclin A	70.72
CDK5/p25	82.83
CDK5/p35	83.97
CDK7/cyclin H/MNAT1	96.65
CDK9/cyclin T1	54.59
CHEK1 (CHK1)	94.39
CHEK2 (CHK2)	86.66
CHUK (IKK alpha)	-2.28
CLK1	97.25
CLK2	97.96
CLK3	56.26
CSF1R (FMS)	88.65
CSK	33.28
CSNK1A1 (CK1 alpha 1)	6.32
CSNK1D (CK1 delta)	5.69
CSNK1E (CK1 epsilon)	40.95
CSNK1G1 (CK1 gamma 1)	-2.10
CSNK1G2 (CK1 gamma 2)	-0.77
CSNK1G3 (CK1 gamma 3)	1.12
CSNK2A1 (CK2 alpha 1)	10.93
CSNK2A2 (CK2 alpha 2)	29.88
DAPK1	9.12
DAPK3 (ZIPK)	36.95
DCAMKL2 (DCK2)	28.81
DYRK1A	47.83
DYRK1B	80.38
DYRK3	16.77
DYRK4	5.93
EEF2K	1.53
EPHA1	81.53

Kinase	% Inhibition
EPHA2	34.28
EPHA3	24.88
EPHA4	47.34
EPHA5	51.70
EPHA8	37.52
EPHB1	89.34
EPHB2	85.99
EPHB3	4.40
EPHB4	61.06
ERBB2 (HER2)	-7.58
ERBB4 (HER4)	18.25
FER	102.00
FES (FPS)	88.57
FGFR1	96.96
FGFR2	103.23
FGFR3	96.81
FGFR4	63.78
FGR	100.93
FLT1 (VEGFR1)	94.84
FLT3	99.70
FLT4 (VEGFR3)	102.11
FRAP1 (mTOR)	-4.39
FRK (PTK5)	86.90
FYN	97.14
GRK4	79.75
GRK5	65.82
GRK6	87.36
GRK7	92.36
GSK3A (GSK3 alpha)	99.09
GSK3B (GSK3 beta)	99.13

Table F.2: Results of kinas screen for BI1 - II. Result of the screen with BI1 against a kinase panel of 239 kinases. Displayed are the percent inhibitions of BI1 for the different kinases tested.

Kinase	% Inhibition
HCK	99.67
HIPK1 (Myak)	40.09
HIPK2	68.02
HIPK4	32.06
IGF1R	46.87
IKBKB (IKK beta)	-8.52
IKBKE (IKK epsilon)	44.19
INSR	57.16
INSRR (IRR)	77.61
IRAK4	77.50
ITK	80.39
JAK1	50.82
JAK2	91.42
JAK2 JH1 JH2	87.24
JAK3	101.29
KDR (VEGFR2)	100.09
KIT	42.82
KIT T670I	53.18
LCK	103.09
LRRK2	98.33
LTK (TYK1)	79.89
LYN A	98.71
LYN B	98.52
MAP2K1 (MEK1)	101.33
MAP2K2 (MEK2)	100.06
MAP2K6 (MKK6)	23.15
MAP3K8 (COT)	94.99
MAP3K9 (MLK1)	50.79
MAP4K2 (GCK)	102.11
MAP4K4 (HGK)	103.91

Kinase	% Inhibition
MAP4K5 (KHS1)	101.24
MAPK1 (ERK2)	1.06
MAPK10 (JNK3)	12.05
MAPK11 (p38 beta)	8.18
MAPK12 (p38 gamma)	9.52
MAPK13 (p38 delta)	6.08
MAPK14 (p38 alpha)	15.87
MAPK14 (p38 alpha) Direct	13.24
MAPK3 (ERK1)	20.09
MAPK8 (JNK1)	19.28
MAPK9 (JNK2)	22.89
MAPKAPK2	8.93
MAPKAPK3	24.27
MAPKAPK5 (PRAK)	34.10
MARK1 (MARK)	102.07
MARK2	92.44
MARK3	104.65
MARK4	102.86
MATK (HYL)	5.63
MELK	92.87
MERTK (cMER)	102.41
MET (cMet)	4.97
MINK1	103.20
MST1R (RON)	12.42
MST4	74.94
MUSK	90.27
MYLK2 (skMLCK)	28.62
NEK1	-0.74
NEK2	4.83
NEK4	-3.25

Kinase	% Inhibition
NEK6	7.06
NEK7	7.32
NEK9	-10.72
NTRK1 (TRKA)	101.30
NTRK2 (TRKB)	101.97
NTRK3 (TRKC)	103.26
PAK2 (PAK65)	-19.49
PAK3	-0.98
PAK4	71.28
PAK6	26.84
PAK7 (KIAA1264)	71.10
PASK	12.40
PDGFRA (PDGFR alpha)	97.40
PDGFRB (PDGFR beta)	97.71
PDK1	51.01
PDK1 Direct	87.05
PHKG1	90.59
PHKG2	82.52
PI4KB (PI4K beta)	8.26
PIK3C2A (PI3K-C2 alpha)	-6.35
PIK3C2B (PI3K-C2 beta)	6.37
PIK3C3 (hVPS34)	-2.11
PIK3CA/PIK3R1 (p110 alpha/p85 alpha)	-6.67
PIK3CD/PIK3R1 (p110 delta/p85 alpha)	-1.92
PIK3CG (p110 gamma)	0.29
PIM1	19.83
PIM2	-2.61
PKN1 (PRK1)	98.15
PLK1	10.41
PLK2	6.14

Table F.3: Results of kinas screen for BI1 - III. Result of the screen with BI1 against a kinase panel of 239 kinases. Displayed are the percent inhibitions of BI1 for the different kinases tested.

Kinase	% Inhibition
PLK3	3.46
PRKACA (PKA)	63.70
PRKCA (PKC alpha)	96.42
PRKCB1 (PKC beta I)	39.88
PRKCB2 (PKC beta II)	74.56
PRKCD (PKC delta)	77.58
PRKCE (PKC epsilon)	56.81
PRKCG (PKC gamma)	81.67
PRKCH (PKC eta)	13.71
PRKCI (PKC iota)	-19.16
PRKCN (PKD3)	94.90
PRKCQ (PKC theta)	99.98
PRKCZ (PKC zeta)	-0.07
PRKD1 (PKC mu)	101.71
PRKD2 (PKD2)	98.10
PRKG1	99.14
PRKG2 (PKG2)	55.28
PRKX	85.24
PTK2 (FAK)	79.76
PTK2B (FAK2)	66.45
PTK6 (Brk)	28.61
RAF1 (cRAF) Y340D Y341D	96.54
RET	101.64
ROCK1	30.30
ROCK2	24.80
ROS1	89.36
RPS6KA1 (RSK1)	99.49
RPS6KA2 (RSK3)	100.22
RPS6KA3 (RSK2)	94.34
RPS6KA4 (MSK2)	83.76

Kinase	% Inhibition
RPS6KA5 (MSK1)	64.55
RPS6KA6 (RSK4)	84.37
RPS6KB1 (p70S6K)	38.56
SGK (SGK1)	20.23
SGK2	-1.77
SGKL (SGK3)	8.65
SNF1LK2	103.15
SPHK1	-10.06
SPHK2	-29.50
SRC	97.40
SRC N1	103.31
SRMS (Srm)	32.64
SRPK1	44.06
SRPK2	72.83
STK22B (TSSK2)	40.02
STK22D (TSSK1)	105.83
STK23 (MSSK1)	46.77
STK24 (MST3)	59.26
STK25 (YSK1)	103.76
STK3 (MST2)	99.62
STK4 (MST1)	102.99
SYK	75.05
TAOK2 (TAO1)	45.36
TBK1	82.49
TEK (Tie2)	69.42
TYK2	94.81
TYRO3 (RSE)	89.12
YES1	104.26
ZAP70	4.47

Table F.4: Results of kinas screen for BI4 - I. Result of the screen with BI4 against a kinase panel of 239 kinases. Displayed are the percent inhibitions of BI4 for the different kinases tested.

Kinase	% Inhibition
ABL1	86.33
ABL1 Y253F	95.19
ABL2 (Arg)	70.63
ACVR1B (ALK4)	78.79
ADRBK1 (GRK2)	-9.38
ADRBK2 (GRK3)	-8.04
AKT1 (PKB alpha)	-0.55
AKT2 (PKB beta)	4.30
AKT3 (PKB gamma)	-1.38
ALK	43.33
AMPK A1/B1/G1	6.51
AURKA (Aurora A)	4.57
AURKB (Aurora B)	28.10
AURKC (Aurora C)	-0.80
AXL	61.81
BLK	81.06
BMX	57.58
BRAF	18.06
BRAF V599E	24.07
BRSK1 (SAD1)	3.35
BTK	54.45
CAMK1 (CaMK1)	25.71
CAMK1D (CaMKI delta)	42.61
CAMK2A (CaMKII alpha)	5.77
CAMK2B (CaMKII beta)	-3.96
CAMK2D (CaMKII delta)	14.90
CAMK4 (CaMKIV)	1.83
CDC42 BPA (MRCKA)	0.98
CDC42 BPB (MRCKB)	2.27
CDK1/cyclin B	4.90

Kinase	% Inhibition
CDK2/cyclin A	0.35
CDK5/p25	-0.49
CDK5/p35	-6.74
CDK7/cyclin H/MNAT1	12.68
CDK9/cyclin T1	23.78
CHEK1 (CHK1)	34.13
CHEK2 (CHK2)	-10.52
CHUK (IKK alpha)	4.75
CLK1	2.01
CLK2	1.16
CLK3	1.38
CSF1R (FMS)	94.63
CSK	10.82
CSNK1A1 (CK1 alpha 1)	3.94
CSNK1D (CK1 delta)	4.90
CSNK1E (CK1 epsilon)	5.32
CSNK1G1 (CK1 gamma 1)	-1.11
CSNK1G2 (CK1 gamma 2)	-1.19
CSNK1G3 (CK1 gamma 3)	-0.44
CSNK2A1 (CK2 alpha 1)	8.44
CSNK2A2 (CK2 alpha 2)	0.64
DAPK1	-2.65
DAPK3 (ZIPK)	2.43
DCAMKL2 (DCK2)	-5.96
DYRK1A	11.02
DYRK1B	10.06
DYRK3	5.54
DYRK4	3.18
EEF2K	1.04
EPHA1	10.00

Kinase	% Inhibition
EPHA2	2.65
EPHA3	-0.98
EPHA4	2.48
EPHA5	1.21
EPHA8	71.10
EPHB1	22.94
EPHB2	30.83
EPHB3	-1.36
EPHB4	12.77
ERBB2 (HER2)	-10.94
ERBB4 (HER4)	11.29
FER	64.96
FES (FPS)	7.69
FGFR1	65.75
FGFR2	61.74
FGFR3	36.22
FGFR4	16.82
FGR	93.07
FLT1 (VEGFR1)	60.93
FLT3	90.96
FLT4 (VEGFR3)	94.13
FRAP1 (mTOR)	-4.94
FRK (PTK5)	34.07
FYN	62.55
GRK4	-17.43
GRK5	-12.30
GRK6	-14.72
GRK7	-12.98
GSK3A (GSK3 alpha)	74.22
GSK3B (GSK3 beta)	64.32

Table F.5: Results of kinas screen for BI4 - II. Result of the screen with BI4 against a kinase panel of 239 kinases. Displayed are the percent inhibitions of BI4 for the different kinases tested.

Kinase	% Inhibition
HCK	60.63
HIPK1 (Myak)	-1.75
HIPK2	-3.14
HIPK4	4.53
IGF1R	3.07
IKBKB (IKK beta)	-10.81
IKBKE (IKK epsilon)	-5.44
INSR	2.22
INSRR (IRR)	3.23
IRAK4	6.35
ITK	1.17
JAK1	11.43
JAK2	63.14
JAK2 JH1 JH2	28.43
JAK3	89.09
KDR (VEGFR2)	94.74
KIT	64.40
KIT T670I	45.23
LCK	98.75
LRRK2	30.81
LTK (TYK1)	3.58
LYN A	88.71
LYN B	88.17
MAP2K1 (MEK1)	19.77
MAP2K2 (MEK2)	34.09
MAP2K6 (MKK6)	1.86
MAP3K8 (COT)	5.73
MAP3K9 (MLK1)	-1.24
MAP4K2 (GCK)	66.89
MAP4K4 (HGK)	98.64

Kinase	% Inhibition
MAP4K5 (KHS1)	99.35
MAPK1 (ERK2)	-0.43
MAPK10 (JNK3)	9.31
MAPK11 (p38 beta)	12.15
MAPK12 (p38 gamma)	5.94
MAPK13 (p38 delta)	-0.73
MAPK14 (p38 alpha)	21.70
MAPK14 (p38 alpha) Direct	13.87
MAPK3 (ERK1)	10.21
MAPK8 (JNK1)	7.71
MAPK9 (JNK2)	7.48
MAPKAPK2	6.47
MAPKAPK3	18.89
MAPKAPK5 (PRAK)	-1.44
MARK1 (MARK)	3.36
MARK2	-3.61
MARK3	26.90
MARK4	23.97
MATK (HYL)	1.18
MELK	48.28
MERTK (cMER)	66.41
MET (cMet)	-11.39
MINK1	88.41
MST1R (RON)	2.79
MST4	52.33
MUSK	0.41
MYLK2 (skMLCK)	16.53
NEK1	-1.45
NEK2	-8.04
NEK4	-8.95

Kinase	% Inhibition
NEK6	3.45
NEK7	-3.12
NEK9	-29.12
NTRK1 (TRKA)	94.84
NTRK2 (TRKB)	96.87
NTRK3 (TRKC)	102.56
PAK2 (PAK65)	-5.99
PAK3	-0.12
PAK4	-9.13
PAK6	-3.40
PAK7 (KIAA1264)	-15.06
PASK	3.32
PDGFRA (PDGFR alpha)	70.54
PDGFRB (PDGFR beta)	81.54
PDK1	11.11
PDK1 Direct	9.85
PHKG1	-18.03
PHKG2	-8.48
PI4KB (PI4K beta)	3.12
PIK3C2A (PI3K-C2 alpha)	-0.13
PIK3C2B (PI3K-C2 beta)	5.84
PIK3C3 (hVPS34)	-8.97
PIK3CA/PIK3R1 (p110 alpha/p85 alpha)	-4.99
PIK3CD/PIK3R1 (p110 delta/p85 alpha)	3.67
PIK3CG (p110 gamma)	0.76
PIM1	0.52
PIM2	-7.67
PKN1 (PRK1)	2.66
PLK1	-2.78
PLK2	-1.73

Table F.6: Results of kinas screen for BI4 - III. Result of the screen with BI4 against a kinase panel of 239 kinases. Displayed are the percent inhibitions of BI4 for the different kinases tested.

Kinase	% Inhibition
PLK3	1.26
PRKACA (PKA)	1.18
PRKCA (PKC alpha)	-14.46
PRKCB1 (PKC beta I)	-7.32
PRKCB2 (PKC beta II)	6.42
PRKCD (PKC delta)	-0.71
PRKCE (PKC epsilon)	-24.02
PRKCG (PKC gamma)	5.43
PRKCH (PKC eta)	-7.05
PRKCI (PKC iota)	-14.57
PRKCN (PKD3)	-2.54
PRKCQ (PKC theta)	-0.54
PRKCZ (PKC zeta)	-13.80
PRKD1 (PKC mu)	15.38
PRKD2 (PKD2)	0.51
PRKG1	6.87
PRKG2 (PKG2)	2.69
PRKX	4.32
PTK2 (FAK)	20.19
PTK2B (FAK2)	-1.99
PTK6 (Brk)	9.20
RAF1 (cRAF) Y340D Y341D	15.28
RET	80.46
ROCK1	-4.63
ROCK2	8.78
ROS1	6.99
RPS6KA1 (RSK1)	51.91
RPS6KA2 (RSK3)	65.17
RPS6KA3 (RSK2)	24.38
RPS6KA4 (MSK2)	35.70

Kinase	% Inhibition
RPS6KA5 (MSK1)	6.03
RPS6KA6 (RSK4)	58.86
RPS6KB1 (p70S6K)	-1.65
SGK (SGK1)	2.24
SGK2	-7.07
SGKL (SGK3)	-2.31
SNF1LK2	100.44
SPHK1	-10.94
SPHK2	-16.18
SRC	60.79
SRC N1	64.05
SRMS (Srm)	48.53
SRPK1	-2.14
SRPK2	-24.36
STK22B (TSSK2)	7.43
STK22D (TSSK1)	13.59
STK23 (MSSK1)	0.57
STK24 (MST3)	38.25
STK25 (YSK1)	30.19
STK3 (MST2)	9.53
STK4 (MST1)	71.50
SYK	-8.95
TAOK2 (TAO1)	36.75
TBK1	-0.88
TEK (Tie2)	31.61
TYK2	23.72
TYRO3 (RSE)	20.87
YES1	79.36
ZAP70	3.21

Bibliography

[1] Roth GJ, Heckel A, Brandl T, Grauert M, Hoerer S, Kley JT, Schnapp G, Baum P, Mennerich D, Schnapp A, Park JE: **Design, synthesis, and evaluation of indolinones as inhibitors of the transforming growth factor β receptor I (TGFβ RI).** *J Med Chem* 2010, **53**(20):7287–7295.

[2] Bolton EE, Wang Y, Thiessen PA, Bryant SH: **Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities**. In *J Med Chem*, *Volume 4 of* Annual Reports in Computational Chemistry, Elsevier 2008:217 – 241.

[3] *www.innovation.org*.

[4] *www.innovation.org/ drug_ discovery/ objects/ pdf/ RD_ Brochure.pdf*.

[5] Yang Y, Adelstein SJ, Kassis AI: **Target discovery from data mining approaches.** *Drug Discov Today* 2009, **14**(3-4):147–154.

[6] Yang Y, Adelstein SJ, Kassis AI: **Target discovery from data mining approaches.** *Drug Discov Today* 2012, **17 Suppl**:S16–S23.

[7] Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S, Brouwer C, Harland L: **Visualizing the drug target landscape.** *Drug Discov Today* 2010, **15**(1-2):3–15.

[8] Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S, Brouwer C, Harland L: **Visualizing the drug target landscape.** *Drug Discov Today* 2012, **17 Suppl**:S3–S15.

[9] Milligan G: **High-content assays for ligand regulation of G-protein-coupled receptors.** *Drug Discov Today* 2003, **8**(13):579–585.

[10] Mousses S, Kallioniemi A, Kauraniemi P, Elkahloun A, Kallioniemi OP: **Clinical and functional target validation using tissue and cell microarrays.** *Curr Opin Chem Biol* 2002, **6**:97–101.

[11] Lotze MT, Kost TA: **Viruses as gene delivery vectors: application to gene function, target validation, and assay development.** *Cancer Gene Ther* 2002, **9**(8):692–699.

[12] Lindsay MA: **Target discovery.** *Nat Rev Drug Discov* 2003, **2**(10):831–838.

[13] Lindsay MA: **Finding new drug targets in the 21st century.** *Drug Discov Today* 2005, **10**(23-24):1683–1687.

[14] Lessl M, Schoepe S, Sommer A, Schneider M, Asadullah K: **Grants4Targets - an innovative approach to translate ideas from basic research into novel drugs.** *Drug Discov Today* 2011, **16**(7-8):288–292.

[15] Hughes JP, Rees S, Kalindjian SB, Philpott KL: **Principles of early drug discovery.** *Br J Pharmacol* 2011, **162**(6):1239–1249.

[16] Meur GL, Stieger K, Smith AJ, Weber M, Deschamps JY, Nivard D, Mendes-Madeira A, Provost N, Péréon Y, Cherel Y, Ali RR, Hamel C, Moullier P, Rolling F: **Restoration of vision in RPE65-deficient Briard dogs using an AAV serotype 4 vector that specifically targets the retinal pigmented epithelium.** *Gene Ther* 2007, **14**(4):292–303.

[17] Fleming A: **On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae.** *Br J Exp Pathol* 1929, **10**(31):226–236.

[18] *http:// www.buscopan.com.*

[19] Tytgat GN: **Hyoscine butylbromide: a review of its use in the treatment of abdominal cramping and pain.** *Drugs* 2007, **67**(9):1343–1357.

[20] Kuhnert N: **Hundert Jahre Aspirin®. Die Geschichte des wohl erfolgreichsten Medikaments aller Zeiten.** *Chemie in unserer Zeit* 1999, **33**(4):213–220.

[21] Shelat AA, Guy RK: **Scaffold composition and biological relevance of screening libraries.** *Nature Chemical Biology* 2007, **3**(8):442–446.

[22] Thomasa GL, Johannes CW: **Natural product-like synthetic libraries.** *Current Opinion in Chemical Biology* 2011, **15**(4):516–522.

[23] Böhm HJ, Klebe G, Kubinyi H: *Wirkstoffdesign. Der Weg zum Arzneimittel.* Spektrum Verlag 1996.

[24] Smith C: **Drug target validation: Hitting the target.** *Nature* 2003, **422**:341–347.

[25] Reddy AS, Pati SP, Kumar PP, Pradeep H, Sastry GN: **Virtual Screening in Drug Discovery - A Computational Perspective.** *Current Protein and Peptide Science* 2007, **8**(4):329–351.

[26] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 2001, **46**(1-3):3–26.

[27] *www.fda.gov.*

[28] *www.ema.europa.eu.*

[29] *www.mhlw.go.jp/ english.*

[30] Barter PJ, Caulfield M, Eriksson M, Grundy SM, Kastelein JJP, Komajda M, Lopez-Sendon J, Mosca L, Tardif JC, Waters DD, Shear CL, Revkin JH, Buhr KA, Fisher MR, Tall AR, Brewer B, ILLUMINATE Investigators: **Effects of torcetrapib in patients at high risk for coronary events.** *N Engl J Med* 2007, **357**(21):2109–2122.

[31] Forrest MJ, Bloomfield D, Briscoe RJ, Brown PN, Cumiskey AM, Ehrhart J, Hershey JC, Keller WJ, Ma X, McPherson HE, Messina E, Peterson LB, Sharif-Rodriguez W, Siegl PKS, Sinclair PJ, Sparrow CP, Stevenson AS, Sun SY, Tsai C, Vargas H, Walker M, West SH, White V, Woltmann RF: **Torcetrapib-induced blood pressure elevation is independent of CETP inhibition and is accompanied by increased circulating levels of aldosterone.** *Br J Pharmacol* 2008, **154**(7):1465–1473.

[32] Clerc RG, Stauffer A, Weibel F, Hainaut E, Perez A, Hoflack JC, Bénardeau A, Pflieger P, Garriz JM, Funder JW, Capponi AM, Niesor EJ: **Mechanisms underlying off-target effects of the cholesteryl ester transfer protein inhibitor torcetrapib involve L-type calcium channels.** *J Hypertens* 2010, **28**(8):1676–1686.

[33] Arrowsmith J: **Phase III and submission failures: 2007-2010**. *Nature Reviews Drug Discovery* 2011, **10**:1.

[34] Chow WA, Jiang C, Guan M: **Anti-HIV drugs for cancer therapeutics: back to the future?** *Lancet Oncol* 2009, **10**:61–71.

[35] Xie L, Evangelidis T, Xie L, Bourne PE: **Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir.** *PLoS Comput Biol* 2011, **7**(4):e1002037.

[36] Richardson RD, Smith JW: **Novel antagonists of the thioesterase domain of human fatty acid synthase.** *Mol Cancer Ther* 2007, **6**(7):2120–2126.

[37] Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, Lines C, Riddell R, Morton D, Lanas A, Konstam MA, Baron JA, on Vioxx (APPROVe) Trial Investigators APP: **Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial.** *N Engl J Med* 2005, **352**(11):1092–1102.

[38] Baron JA, Sandler RS, Bresalier RS, Lanas A, Morton DG, Riddell R, Iverson ER, Demets DL: **Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial.** *Lancet* 2008, **372**(9651):1756–1764.

[39] Lim WK, Wang K, Lefebvre C, Califano A: **Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks.** *Bioinformatics* 2007, **23**(13):i282–i288.

[40] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES,

Hirschhorn JN, Altshuler D, Groop LC: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, **34**(3):267–273.

[41] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**(43):15545–15550.

[42] Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20**:93–99.

[43] Alexa A, Rahnenführer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22**(13):1600–1607.

[44] Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC: **Testing association of a pathway with survival using gene expression data.** *Bioinformatics* 2005, **21**(9):1950–1957.

[45] Mansmann U, Meister R: **Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach.** *Methods Inf Med* 2005, **44**(3):449–453.

[46] Hummel M, Metzeler KH, Buske C, Bohlander SK, Mansmann U: **Association between a prognostic gene signature and functional gene sets.** *Bioinform Biol Insights* 2008, **2**:329–341.

[47] Hummel M, Meister R, Mansmann U: **GlobalANCOVA: exploration and assessment of gene group effects.** *Bioinformatics* 2008, **24**:78–85.

[48] Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks.** *Bioinformatics* 2002, **18 Suppl 1**:S233–S240.

[49] Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, Rao S, Wang J: **Edge-based scoring and searching method**

for identifying condition-responsive protein-protein interaction subnetwork. *Bioinformatics* 2007, **23**(16):2121–2128.

[50] Nacu S, Critchley-Thorne R, Lee P, Holmes S: **Gene expression network analysis and applications to immunology.** *Bioinformatics* 2007, **23**(7):850–858.

[51] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach.** *Bioinformatics* 2008, **24**(13):i223–i231.

[52] Baum P: **Phenocopy - A Strategy to Qualify Chemical Compounds during Hit-to-Lead and/or Lead Optimization.** *PhD thesis*, Fakultät Energie-, Verfahrens- und Biotechnik der Universität Stuttgart 2010.

[53] Schmid R, Baum P, Ittrich C, Fundel-Clemens K, Huber W, Brors B, Eils R, Weith A, Mennerich D, Quast K: **Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3.** *BMC Genomics* 2010, **11**:349.

[54] Baum P, Schmid R, Ittrich C, Rust W, Fundel-Clemens K, Siewert S, Baur M, Mara L, Gruenbaum L, Heckel A, Eils R, Kontermann RE, Roth GJ, Gantner F, Schnapp A, Park JE, Weith A, Quast K, Mennerich D: **Phenocopy–a strategy to qualify chemical compounds during hit-to-lead and/or lead optimization.** *PLoS One* 2010, **5**(12):e14272.

[55] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res* 2003, **31**:258–261.

[56] Wieser R, Wrana JL, Massagué J: **GS domain mutations that constitutively activate T beta R-I, the downstream signaling component in the TGF-beta receptor complex.** *EMBO J* 1995, **14**(10):2199–2208.

[57] Willis SA, Zimmerman CM, Li LI, Mathews LS: **Formation and activation by phosphorylation of activin receptor complexes.** *Mol Endocrinol* 1996, **10**(4):367–379.

[58] Huse M, Muir TW, Xu L, Chen YG, Kuriyan J, Massagué J: **The TGF beta receptor activation process: an inhibitor- to substrate-binding switch.** *Mol Cell* 2001, **8**(3):671–682.

[59] Shi Y, Massagué J: **Mechanisms of TGF-β signaling from cell membrane to the nucleus.** *Cell* 2003, **113**(6):685–700.

[60] ten Dijke P, Hill CS: **New insights into TGF-beta-Smad signalling.** *Trends Biochem Sci* 2004, **29**(5):265–273.

[61] Yingling JM, Blanchard KL, Sawyer JS: **Development of TGF-β signalling inhibitors for cancer therapy.** *Nat Rev Drug Discov* 2004, **3**(12):1011–1022.

[62] Massagué J: **A very private TGF-β receptor embrace.** *Mol Cell* 2008, **29**(2):149–150.

[63] Groppe J, Hinck CS, Samavarchi-Tehrani P, Zubieta C, Schuermann JP, Taylor AB, Schwarz PM, Wrana JL, Hinck AP: **Cooperative assembly of TGF-beta superfamily signaling complexes is mediated by two disparate mechanisms and distinct modes of receptor binding.** *Mol Cell* 2008, **29**(2):157–168.

[64] Kang JS, Liu C, Derynck R: **New regulatory mechanisms of TGF-beta receptor function.** *Trends Cell Biol* 2009, **19**(8):385–394.

[65] Moustakas A, Souchelnytskyi S, Heldin CH: **SMAD regulation in TGF-β signal transduction.** *J Cell Sci.* 2001, **114**(24):4359–4369.

[66] Ferrand N, Atfi A, Prunier C: **The oncoprotein c-Ski functions as a direct antagonist of the Transforming Growth Factor-β Type I Receptor.** *Cancer Res* 2010, **70**(21):8457–8466.

[67] Moustakas A: **SMAD signalling network.** *J Cell Sci* 2002, **115**(17):3355–3356.

[68] Zhang YE: **Non-Smad pathways in TGF-β signaling.** *Cell Res* 2009, **19**:128–139.

[69] Sorrentino A, Thakur N, Grimsby S, Marcusson A, von Bulow V, Schuster N, Zhang S, Heldin CH, Landström M: **The type I TGF-β receptor engages TRAF6 to activate TAK1 in a receptor kinase-independent manner.** *Nat Cell Biol* 2008, **10**(10):1199–1207.

[70] Thakur N, Sorrentino A, Heldin C, Landström M: **TGF-β uses the E3-ligase TRAF6 to turn on the kinase TAK1 to kill prostate cancer cells.** *Nat Cell Biol* 2009, **5**:1–3.

[71] Blobe GC, Schiemann WP, Lodish HF: **Role of transforming growth factor β in human disease.** *N Engl J Med* 2000, **342**(18):1350–1358.

[72] Massagué J, Blain SW, Lo RS: **TGFβ signaling in growth control, cancer, and heritable disorders.** *Cell* 2000, **103**(2):295–309.

[73] Massagué J: **TGFβ in Cancer.** *Cell* 2008, **134**(2):215–230.

[74] Marks PA, Rifkind RA, Richon VM, Breslow R, Miller T, Kelly WK: **Histone deacetylases and cancer: causes and therapies.** *Nat Rev Cancer* 2001, **1**:194–202.

[75] Chan MW, Huang YW, Hartman-Frey C, Kuo CT, Deatherage D, Qin H, Cheng AS, Yan PS, Davuluri RV, Huang THM, Nephew KP, Lin HJL: **Aberrant transforming growth factor β1 signaling and SMAD4 nuclear translocation confer epigenetic repression of ADAM19 in ovarian cancer.** *Neoplasia* 2008, **10**:908–919.

[76] Dumont N, Wilson MB, Crawford YG, Reynolds PA, Sigaroudinia M, Tlsty TD: **Sustained induction of epithelial to mesenchymal transition activates DNA methylation of genes silenced in basal-like breast cancers.** *Proc Natl Acad Sci* 2008, **105**:14867–14872.

[77] Chou JL, Su HY, Chen LY, Liao YP, Hartman-Frey C, Lai YH, Yang HW, Deatherage DE, Kuo CT, Huang YW, Yan PS, Hsiao SH, Tai CK, Lin HJL, Davuluri RV, Chao TK, Nephew KP, Huang THM, Lai HC, Chan MWY: **Promoter hypermethylation of FBXO32, a novel TGF-β/SMAD4 target gene and tumor suppressor, is associated with poor prognosis in human ovarian cancer.** *Lab Invest* 2010, **90**(3):414–425.

[78] Chou JL, Chen LY, Lai HC, Chan MWY: **TGF-β: friend or foe? The role of TGF-β/SMAD signaling in epigenetic silencing of ovarian cancer and its implication in epigenetic therapy.** *Expert Opin Ther Targets* 2010, **14**(11):1213–1223.

[79] Papageorgis P, Lambert AW, Ozturk S, Gao F, Pan H, Manne U, Alekseyev YO, Thiagalingam A, Abdolmaleky HM, Lenburg M, Thiagalingam S: **Smad signaling is required to maintain epigenetic silencing during breast cancer progression.** *Cancer Res* 2010, **70**(3):968–978.

[80] Kang HR, Cho SJ, Lee CG, Homer RJ, Elias JA: **Transforming growth factor (TGF)-β1 stimulates pulmonary fibrosis and inflammation via a Bax-dependent, bid-activated pathway that involves matrix metalloproteinase-12.** *J Biol Chem* 2007, **282**(10):7723–7732.

[81] Zanini A, Chetta A, Olivieri D: **Therapeutic perspectives in bronchial vascular remodeling in COPD.** *Ther Adv Respir Dis* 2008, **2**(3):179–187.

[82] Rosendahl A, Checchin D, Fehniger TE, ten Dijke P, Heldin CH, Sideras P: **Activation of the TGF-β/activin-Smad2 pathway during allergic airway inflammation.** *Am J Respir Cell Mol Biol* 2001, **25**:60–68.

[83] Trachtman H, Fervenza FC, Gipson DS, Heering P, Jayne DRW, Peters H, Rota S, Remuzzi G, Rump LC, Sellin LK, Heaton JPW, Streisand JB, Hard ML, Ledbetter SR, Vincenti F: **A phase 1, single-dose study of fresolimumab, an anti-TGF-β antibody, in treatment-resistant primary focal segmental glomerulosclerosis.** *Kidney Int* 2011, **79**(11):1236–1243.

[84] *integrity.thomson-pharma.com.*

[85] Lahn M, Kloeker S, Berry BS: **TGF-β inhibitors for the treatment of cancer.** *Expert Opin Investig Drugs* 2005, **14**(6):629–643.

[86] Gaspar NJ, Li L, Kapoun AM, Medicherla S, Reddy M, Li G, O'Young G, Quon D, Henson M, Damm DL, Muiru GT, Murphy A, Higgins LS, Chakravarty S, Wong DH: **Inhibition of transforming growth factor β signaling reduces pancreatic adenocarcinoma growth and invasiveness.** *Mol Pharmacol* 2007, **72**:152–161.

[87] Knight JDR, Qian B, Baker D, Kothary R: **Conservation, variability and the modeling of active protein kinases.** *PLoS One* 2007, **2**(10):e982.

[88] Thaimattam R, Banerjee R, Miglani R, Iqbal J: **Protein kinase inhibitors: structural insights into selectivity.** *Curr Pharm Des* 2007, **13**(27):2751–2765.

[89] Patel RY, Doerksen RJ: **Protein kinase-inhibitor database: structural variability of and inhibitor interactions with the protein kinase P-loop.** *J Proteome Res* 2010, **9**(9):4433–4442.

[90] Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**(6757):83–86.

[91] Yanai I, DeLisi C: **The society of genes: networks of functional links between genes from comparative genomics.** *Genome Biol* 2002, **3**(11):research0064.

[92] Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes.** *Science* 2004, **306**(5701):1555–1558.

[93] Lee I, Li Z, Marcotte EM: **An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae.** *PLoS One* 2007, **2**(10):e988.

[94] Linghu B, Snitkin ES, Holloway DT, Gustafson AM, Xia Y, DeLisi C: **High-precision high-coverage functional inference from integrated data sources.** *BMC Bioinformatics* 2008, **9**:119.

[95] Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C: **Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network.** *Genome Biol* 2009, **10**(9):R91.

[96] Razick S, Magklaras G, Donaldson IM: **iRefIndex: a consolidated protein interaction database with provenance.** *BMC Bioinformatics* 2008, **9**:405.

[97] Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31**:248–250.

[98] Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, *et alii*: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33**(Database issue):D418–D424.

[99] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**(Database issue):D535–D539.

[100] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**(Database issue):D449–D451.

[101] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13**(10):2363–2371.

[102] Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HGM, Nagini M, Kumar GSS, Jose R, Deepthi P, Mohan SS, Gandhi TKB, Harsha HC, Deshpande KS, Sarker M, Prasad TSK, Pandey A: **Hu-**

man protein reference database–2006 update. *Nucleic Acids Res* 2006, **34**(Database issue):D411–D414.

[103] Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2004, **32**(Database issue):D452–D455.

[104] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H: **IntAct–open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35**(Database issue):D561–D565.

[105] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res* 2007, **35**(Database issue):D572–D574.

[106] Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stümpflen V: **MPact: the MIPS protein interaction resource on yeast.** *Nucleic Acids Res* 2006, **34**(Database issue):D436–D441.

[107] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D: **The MIPS mammalian protein-protein interaction database.** *Bioinformatics* 2005, **21**(6):832–834.

[108] Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21**(9):2076–2082.

[109] Razick S, Magklaras G, Donaldson IM: **iRefIndex Wiki.**

[110] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498–2504.

[111] Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, Liu B, Nandi A, Santos C, Andrews P, Athey B, States D, Jagadish HV: **Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together.** *Nucleic Acids Res* 2007, **35**(Database issue):D566–D571.

[112] Tarcea VG, Weymouth T, Ade A, Bookvich A, Gao J, Mahavisno V, Wright Z, Chapman A, Jayapandian M, Özgür A, Tian Y, Cavalcoli J, Mirel B, Patel J, Radev D, Athey B, States D, Jagadish HV: **Michigan molecular interactions r2: from interacting proteins to pathways.** *Nucleic Acids Res* 2009, **37**(Database issue):D642–D646.

[113] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29.

[114] Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27–30.

[115] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**(Database issue):D109–D114.

[116] Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP, Hautaniemi S: **Integrated network analysis platform for protein-protein interactions.** *Nat Methods* 2009, **6**:75–77.

[117] Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J: **PINA v2.0: mining interactome modules.** *Nucleic Acids Res* 2012, **40**(Database issue):D862–D865.

[118] Blankenburg H, Finn RD, Jenkinson AM, Ramírez F, Emig D, Schelhorn SE, Büch J, Lengauer T, Albrecht M: **DASMI: exchanging, annotating and assessing molecular interaction data.** *Bioinformatics* 2009, **25**(10):1321–1328.

[119] Dowell R, Jokerst R, Day A, Eddy S, Stein L: **The Distributed Annotation System**. *BMC Bioinformatics* 2001, **2**:7.

[120] Blankenburg H, Ramírez F, Büch J, Albrecht M: **DASMIweb: online integration, analysis and assessment of distributed protein interaction data.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W122–W128.

[121] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33**(Database issue):D433–D437.

[122] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302**(5644):449–453.

[123] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: **A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).** *Proc Natl Acad Sci U S A* 2003, **100**(14):8348–8353.

[124] Rajagopalan D, Agarwal P: **Inferring pathways from gene lists using a literature-derived network of biological relationships.** *Bioinformatics* 2005, **21**(6):788–793.

[125] Cabusora L, Sutton E, Fulmer A, Forst CV: **Differential network expression during drug and stress response.** *Bioinformatics* 2005, **21**(12):2898–2905.

[126] Hwang T, Park T: **Identification of differentially expressed subnetworks based on multivariate ANOVA.** *BMC Bioinformatics* 2009, **10**:128.

[127] Zhao XM, Wang RS, Chen L, Aihara K: **Uncovering signal transduction networks from high-throughput data by integer linear programming.** *Nucleic Acids Res* 2008, **36**(9):e48.

[128] Sohler F, Hanisch D, Zimmer R: **New methods for joint analysis of biological networks and expression data.** *Bioinformatics* 2004, **20**(10):1517–1521.

[129] Hanisch D, Sohler F, Zimmer R: **ToPNet–an application for interactive analysis of expression data and biological networks.** *Bioinformatics* 2004, **20**(9):1470–1471.

[130] Fisher RA: *Statistical Methods for Reasearch Workers.* Oliver and Boyd, London 1932.

[131] Breitling R, Amtmann A, Herzyk P: **Graph-based iterative Group Analysis enhances microarray interpretation.** *BMC Bioinformatics* 2004, **5**:100.

[132] Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data.** *BMC Syst Biol* 2007, **1**:8.

[133] Ulitsky I, Shamir R: **Identifying functional modules using expression profiles and confidence-scored protein interactions.** *Bioinformatics* 2009, **25**(9):1158–1164.

[134] Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ: **Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae.** *Mol Cell Proteomics* 2007, **6**(3):439–450.

[135] *ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz*.

[136] Ljubić I, Weiskircher R, Pferschy U, Klau GW, Mutzel P, Fischetti M: **An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem**. *Mathematical Programming* 2006, **105**(2):427–449.

[137] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, López-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM, Lymphoma/Leukemia Molecular Profiling Project: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.** *N Engl J Med* 2002, **346**(25):1937–1947.

[138] Andersen PK, Gill RD: **Cox's Regression Model for Counting Processes: A Large Sample Study**. *The Annals of Statistics* 1982, **10**(4):1100–1120.

[139] Gu J, Chen Y, Li S, Li Y: **Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis.** *BMC Syst Biol* 2010, **4**:47.

[140] Zhang KX, Ouellette BFF: **Pandora, a pathway and network discovery approach based on common biological evidence.** *Bioinformatics* 2010, **26**(4):529–535.

[141] Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF: **A new method to measure the semantic similarity of GO terms.** *Bioinformatics* 2007, **23**(10):1274–1281.

[142] Tan PN, Steinbach M, Kumar V: *Introduction to Data Mining.* Addison Wesley 2005.

[143] Fisher RA: **On the Interpretation of χ from Contingency Tables, and the Calculation of P**. *Journal of the Royal Statistical Society* 1922, **85**:87–94.

[144] Ingenuity Systems, Inc.: **Ingenuity Pathway Analysis**. *http://www.ingenuity.com/products/pathways_analysis.html*.

[145] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression.** *Genomics* 2003, **81**(2):98–104.

[146] Mansmann U: **Genomic profiling. Interplay between clinical epidemiology, bioinformatics and biostatistics.** *Methods Inf Med* 2005, **44**(3):454–460.

[147] Boukamp P, Petrussevska RT, Breitkreutz D, Hornung J, Markham A, Fusenig NE: **Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line.** *J Cell Biol* 1988, **106**(3):761–771.

[148] Schmidt DM, Ernst JD: **A fluorometric assay for the quantification of RNA in solution with nanogram sensitivity.** *Anal Biochem* 1995, **232**:144–146.

[149] Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A: **ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments.** *Nucleic Acids Res* 2011, **39**(Database issue):D1002–D1004.

[150] Livak KJ, Schmittgen TD: **Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method**. *Methods* 2001, **25**(4):402 – 408.

[151] R Foundation for Statistical Computing, R Development Core Team: **R: A Language and Environment for Statistical Computing**. *http://www.R-project.org*.

[152] Ihaka R, Gentleman R: **R: A Language for Data Analysis and Graphics**. *Journal of Computational and Graphical Statistics* 1996, **5**(3):299–314.

[153] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80.

[154] Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing Illumina microarray.** *Bioinformatics* 2008, **24**(13):1547–1548.

[155] Illumina Inc.: *http://www.illumina.com/*.

[156] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**(2):249–264.

[157] Lin SM, Du P, Huber W, Kibbe WA: **Model-based variance-stabilizing transformation for Illumina microarray data.** *Nucleic Acids Res* 2008, **36**(2):e11.

[158] Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**(2):185–193.

[159] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30**(4):e15.

[160] Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18 Suppl 1**:S96–104.

[161] Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy–analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**(3):307–315.

[162] Illumina Inc: **Illumina® BeadStudio**. *http://www.illumina.com/Documents/products/datasheets/datasheet_beadstudio.pdf*.

[163] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *J.Roy.Statist.Soc.Ser.B* 1995, **57**:289–300.

[164] Bourgon RW: **Chromatin immunoprecipitation and high-density tiling microarrays: a generative model, methods for analysis, and methodology assessment in the absence of a "gold standard".** *PhD thesis*, University of California, Berkeley, Department of Statistics 2006.

[165] Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article 3.

[166] Lin D, Shkedy Z, Yekutieli D, Burzykowski T, Göhlmann HWH, Bondt AD, Perera T, Geerts T, Bijnens L: **Testing for trends in dose-response microarray experiments: a comparison of several testing procedures, multiplicity and resampling-based inference.** *Stat Appl Genet Mol Biol* 2007, **6**:Article 26.

[167] Wu H, Su Z, Mao F, Olman V, Xu Y: **Prediction of functional modules based on comparative genome analysis and Gene Ontology application**. *Nucleic Acids Research* 2005, **33**(9):2822–2837.

[168] Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy**. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence* 1995:448–453.

[169] Resnik P: **Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language**. *Journal of Artificial Intellingence Research* 1999, **11**:95–130.

[170] Lin D: **An Information-Theoretic Definition of Similarity**. In *In Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann 1998:296–304.

[171] Schlicker A, Domingues FS, Rahnenführer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC Bioinformatics* 2006, **7**:302.

[172] Jiang JJ, Conrath DW: **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy**. In *Proceedings of the International Conference Research on Computational Linguistics (ROCLING)*, Taiwan 1997.

[173] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.** *Bioinformatics* 2010, **26**(7):976–978.

[174] Brors B: **Microarray annotation and biological information on function.** *Methods Inf Med* 2005, **44**(3):468–472.

[175] Fröhlich H, Speer N, Poustka A, Beissbarth T: **GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products.** *BMC Bioinformatics* 2007, **8**:166.

[176] Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23**(2):257–258.

[177] Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4**(8):1184–1191.

[178] Resnik P: **Using Information Content to Evaluate Semantic Similarity in a Taxonomy**. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence* 1995:448–453.

[179] Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A: **The Negatome database: a reference set of non-interacting protein pairs.** *Nucleic Acids Res* 2010, **38**(Database issue):D540–D544.

[180] Genomatix Software GmbH: **Gene2Promoter**. *http:// www.genomatix.de/ online_ help/ help_ gpd/ gpd_ help.html*.

[181] Genomatix Software GmbH: **Genomatix Genome Analyzer**. *http:// www. genomatix.de/ en/ produkte/ genomatix-genome-analyzer.html*.

[182] Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23**(23):4878–4884.

[183] Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21**(13):2933–2942.

[184] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**(Database issue):D412–D416.

[185] Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP: **A benchmark for Affymetrix GeneChip expression measures.** *Bioinformatics* 2004, **20**(3):323–331.

[186] Scheffé H: *The Analysis of Variance.* Wiley, John & Sons Inc 1559.

[187] Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4**(4):210.

[188] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**(Database issue):D355–D360.

[189] Reichardt LF: **Neurotrophin-regulated signalling pathways.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361**(1473):1545–1564.

[190] Skolnik EY, Batzer A, Li N, Lee CH, Lowenstein E, Mohammadi M, Margolis B, Schlessinger J: **The function of GRB2 in linking the insulin receptor to Ras signaling pathways.** *Science* 1993, **260**(5116):1953–1955.

[191] Garey MR, Johnson DS: *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co. 1979.

[192] Papadimitriou CH: *Computational complexity*. Chichester, UK: Addison Wesley 1994.

[193] Betzler N: **A Multivariate Complexity Analysis of Voting Problems**. *PhD thesis*, Fakultät f̈ur Mathematik und Informatik der Friedrich-Schiller-Universität Jena 2010.

[194] Schöning U: *Theoretische Informatik - kurzgefasst*. Spektrum Verlag 2001.

[195] *www.absoluteastronomy.com/ topics/ Computational_ complexity_ theory*.

[196] Karp RM: **Reducibility Among Combinatorial Problems**. In *Complexity of Computer Computations*, Plenum Press, New York 1972, pp:85–103.

[197] Li Z, Zhang S, Zhang X, Chen L: **Exploring the Constrained Maximum Edge-weight Connected Graph Problem**. *Acta Mathematicae Applicatae Sinica, English Series* 2009, **25**(4):697–708.

[198] Post E: **Recursively enumerable sets of positive integers and their decision problems.** *Bulletin of the American Mathematical Society* 1944, **50**:281–299.

[199] Shapiro N: **Degrees of Computability.** *Transactions of the American Mathematical Society* 1956, **82**:281–299.

[200] Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavaré S, Ritchie ME: **Statistical issues in the analysis of Illumina data.** *BMC Bioinformatics* 2008, **9**:85.

[201] Dunning MJ, Ritchie ME, Barbosa-Morais NL, Tavaré S, Lynch AG: **Spike-in validation of an Illumina-specific variance-stabilizing transformation.** *BMC Res Notes* 2008, **1**:18.

[202] Du P, Kibbe WA, Lin SM: **Using lumi, a package processing Illumina Microarray**. *http://www.bioconductor.org/packages/bioc/vignettes/lumi/inst/doc/lumi.pdf* .

[203] Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P: **Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms.** *Nucleic Acids Res* 2005, **33**(18):5914–5923.

[204] Jiang N, Leach LJ, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey MJ, Luo ZW: **Methods for evaluating gene expression from Affymetrix microarray datasets.** *BMC Bioinformatics* 2008, **9**:284.

[205] Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stümpflen V, Mewes HW: **CORUM: the comprehensive resource of mammalian protein complexes.** *Nucleic Acids Res* 2008, **36**(Database issue):D646–D650.

[206] Salwinski L, Eisenberg D: **The MiSink Plugin: Cytoscape as a graphical interface to the Database of Interacting Proteins.** *Bioinformatics* 2007, **23**(16):2193–2195.

[207] Smith CL, Eppig JT: **The mammalian phenotype ontology: enabling robust annotation and comparative analysis**. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2009, **1**(3):390–399.

[208] Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Group MGD: **The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse.** *Nucleic Acids Res* 2012, **40**(Database issue):D881–D886.

[209] Robinson PN, Mundlos S: **The human phenotype ontology.** *Clin Genet* 2010, **77**(6):525–534.

[210] Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu LJ, Danila MI, Feng G, Chisholm RL: **Annotating the human genome with Disease Ontology.** *BMC Genomics* 2009, **10 Suppl 1**:S6.

[211] Du P, Feng G, Flatow J, Song J, Holko M, Kibbe WA, Lin SM: **From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations.** *Bioinformatics* 2009, **25**(12):i63–i68.

[212] Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, Kibbe WA: **Disease Ontology: a backbone for disease semantic integration.** *Nucleic Acids Res* 2012, **40**(Database issue):D940–D946.

[213] Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, Gerangelos S, Koziris N, Dalamagas T, Hatzigeorgiou AG: **TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.** *Nucleic Acids Res* 2012, **40**(Database issue):D222–D229.

[214] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M: **PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.** *Nucleic Acids Res* 2012, **40**(Database issue):D261–D270.