Network-based analysis of gene expression data

Ludwig Geistlinger

SIGILL: UNIVERSITAT: LUDOVIC: MAXIM:

München 2016

Network-based analysis of gene expression data

Ludwig Geistlinger

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Ludwig Geistlinger
aus Leipzig

München, den 22. Januar 2016

Erstgutachter: Prof. Dr. Ralf Zimmer
Zweitgutachter: Prof. Dr. Fabian Theis
Tag der mündlichen Prüfung: 29. April 2016

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Geistlinger, Ludwig

Name, Vorname

München,

Ort, Datum

Unterschrift Doktorand/in

Formular 3.2

Contents

Abstract **xiii**

Zusammenfassung **xv**

1 Introduction: Analyzing gene expression with networks **1**

2 Construction of regulatory networks **5**

2.1 Introduction . . . 5
2.2 Compilation from scientific articles . . . 7
2.2.1 Introduction . . . 8
2.2.2 Material and Methods . . . 9
2.2.3 Results . . . 15
2.2.4 Discussion . . . 21
2.3 Inference from experimental data . . . 24
2.3.1 Introduction . . . 25
2.3.2 Material and Methods . . . 25
2.3.3 Results . . . 29
2.3.4 Discussion . . . 34

3 Detection of significant subnetworks **37**

3.1 Overview . . . 37
3.2 Gene Graph Enrichment Analysis (GGEA) . . . 39
3.2.1 Introduction . . . 40
3.2.2 Methods . . . 41
3.2.3 Results . . . 48
3.2.4 Discussion . . . 54
3.2.5 Conclusion . . . 55
3.3 Improvements of GGEA . . . 57
3.3.1 Significance approximation . . . 57
3.3.2 Edge selection . . . 60

4 Interpretation of detected subnetworks **63**

4.1 Background . . . 63

4.2 Combination and exploration of results 65
4.2.1 Background 66
4.2.2 Implementation 66
4.2.3 Results and Discussion 73
4.2.4 Conclusion 80
4.2.5 Availability 81
4.3 Composition analysis, reduction and integration 82
4.3.1 Overlap and crosstalk between subnetworks 82
4.3.2 Subnetwork reduction and integration 86

5 Conclusion and Outlook 91

A Methods of molecular biology 95
A.1 Quantitative measurement of gene expression 95
A.1.1 Microarrays 95
A.1.2 Next-generation sequencing (RNA-seq) 95
A.2 Detection of protein-DNA interactions 96
A.2.1 Chromatin immunoprecipitation (ChIP) 96
A.2.2 DNA footprinting 96

B Statistical methods 97
B.1 Hypothesis testing 97
B.1.1 Student's *t*-test 97
B.1.2 Fisher's exact test 98
B.1.3 Permutation testing 99
B.1.4 Wilcoxon's rank-sum test 99
B.1.5 Kolmogorov-Smirnov test 99
B.1.6 Multiple testing 100

C Databases 101
C.1 Genome databases 101
C.2 Gene expression databases 101
C.3 Gene set and pathway databases 101
C.3.1 Gene Ontology (GO) 102
C.3.2 Kyoto Encyclopedia of Genes and Genomes (KEGG) 102
C.4 Regulatory network databases 102
C.4.1 Cross-species resources 103
C.4.2 Species-specific resources 103

Acknowledgements 126

List of Figures

1.1 Regulatory interactions between proteins, RNA and DNA 2
1.2 Schematic illustration of a gene regulatory network 3
1.3 Workflow of gene expression data analysis 4

2.1 Approaches to construct a gene regulatory network 6
2.2 Overlaps of predicted, direct & indirect regulatory interactions in yeast . . 10
2.3 Curation approach 12
2.4 From pure text to semi-quantitative models of regulatory interactions . . . 15
2.5 The diauxic shift and its subprocesses 17
2.6 Pathway map of fatty acid metabolism 18
2.7 Current representations of PCK1 regulation 21
2.8 Outline of the recalibration approach 26
2.9 Score recalibration in network predictions 28
2.10 Assessment of predicted interactions 32
2.11 Interactions and expression profiles 33

3.1 Detection of significant subnetworks 38
3.2 Key steps of GGEA 42
3.3 Fuzzyfication of *p*-value and fold change 43
3.4 Modeling regulatory interactions using PNFL 44
3.5 Consistency of regulatory interactions 49
3.6 Explainability of expression changes 50
3.7 Edge consistency distribution 58
3.8 Significance approximation: Goodness of fit 59
3.9 Consistency threshold evaluation on the KEGGdzGEO benchmark set . . . 61

4.1 Overview: Biological interpretation 64
4.2 EnrichmentBrowser: Workflow 68
4.3 EnrichmentBrowser: Visualization and exploration 72
4.4 TCGA RNA-seq data: Method combination 75
4.5 Runtime of enrichment methods 77
4.6 GEO2KEGG evaluation: Individual methods 78
4.7 GEO2KEGG evaluation: Method combination 80

4.8 Overlap and crosstalk in GO and KEGG 83
4.9 Overlap between KEGG signaling pathways 85
4.10 Reduction and integration of enriched subnetworks 87
4.11 Outline of the Sub-GGEA reduction approach 88

List of Tables

2.1 Annotation summary 16

3.1 ORA contingency table 37
3.2 Fuzzy rule set for activation and inhibition 44
3.3 Glioma dataset: Comparison of GGEA and FiDePa 52
3.4 Neuroblastoma dataset: GGEA application 53
3.5 KEGGdzGEO benchmark set 62

4.1 ALL microarray data: Method combination 74

Abstract

The methods of molecular biology for the quantitative measurement of gene expression have undergone a rapid development in the past two decades. High-throughput assays with the microarray and RNA-seq technology now enable whole-genome studies in which several thousands of genes can be measured at a time. However, this has also imposed serious challenges on data storage and analysis, which are subject of the young, but rapidly developing field of computational biology.

To explain observations made on such a large scale requires suitable and accordingly scaled models of gene regulation. Detailed models, as available for single genes, need to be extended and assembled in larger networks of regulatory interactions between genes and gene products. Incorporation of such networks into methods for data analysis is crucial to identify molecular mechanisms that are drivers of the observed expression. As methods for this purpose emerge in parallel to each other and without knowing the standard of truth, results need to be critically checked in a competitive setup and in the context of the available rich literature corpus.

This work is centered on and contributes to the following subjects, each of which represents important and distinct research topics in the field of computational biology: (i) construction of realistic gene regulatory network models; (ii) detection of subnetworks that are significantly altered in the data under investigation; and (iii) systematic biological interpretation of detected subnetworks.

For the construction of regulatory networks, I review existing methods with a focus on curation and inference approaches. I first describe how literature curation can be used to construct a regulatory network for a specific process, using the well-studied diauxic shift in yeast as an example. In particular, I address the question how a detailed understanding, as available for the regulation of single genes, can be scaled-up to the level of larger systems. I subsequently inspect methods for large-scale network inference showing that they are significantly skewed towards master regulators. A recalibration strategy is introduced and applied, yielding an improved genome-wide regulatory network for yeast.

To detect significantly altered subnetworks, I introduce GGEA as a method for network-based enrichment analysis. The key idea is to score regulatory interactions within functional gene sets for consistency with the observed expression. Compared to other recently published methods, GGEA yields results that consistently and coherently align expression changes with known regulation types and that are thus easier to explain. I also suggest and discuss several significant enhancements to the original method that are improving its applicability, outcome and runtime.

For the systematic detection and interpretation of subnetworks, I have developed the `EnrichmentBrowser` software package. It implements several state-of-the-art methods besides GGEA, and allows to combine and explore results across methods. As part of the `Bioconductor` repository, the package provides a unified access to the different methods and, thus, greatly simplifies the usage for biologists. Extensions to this framework, that support automating of biological interpretation routines, are also presented.

In conclusion, this work contributes substantially to the research field of network-based analysis of gene expression data with respect to regulatory network construction, subnetwork detection, and their biological interpretation. This also includes recent developments as well as areas of ongoing research, which are discussed in the context of current and future questions arising from the new generation of genomic data.

Zusammenfassung

Die quantitativen Methoden der Molekularbiologie zur Erhebung von Genexpressionsdaten wurden in den letzten beiden Jahrzehnten technisch enorm weiterentwickelt. Hochdurchsatzverfahren mit der Microarray- und RNAseq-Technologie ermöglichen nunmehr Studien ganzer Genome und erlauben damit mehrere tausend Gene gleichzeitig zu erfassen. Dies stellt jedoch auch enorme Herausforderungen an die Speicherung und Analyse der Daten, welche Gegenstand der noch jungen, sich aber schnell entwickelnden Wissenschaft der Bioinformatik sind.

Um derart großskalige experimentelle Beobachtungen erklären zu können, bedarf es geeignet skalierender genregulatorischer Modelle. Dies macht die Integration bestehender Modelle, die detailliert die Regulation einzelner Gene beschreiben, in größeren Netzwerken regulatorischer Interaktionen zwischen Genen und Genprodukten erforderlich.

Resultierende Netzwerke sind ein wichtiges Hilfsmittel für die Datenanalyse, um die molekularen Mechanismen, welche die beobachteten Daten hervorrufen, identifizieren zu können. Jedoch lassen sich die Ergebnisse verschiedener Ansätze, mit denen Netzwerke in die Datenanalyse integriert werden, nur schwierig überprüfen und miteinander vergleichen, da die tatsächlichen Mechanismen zumeist nicht, oder nur unvollständig, bekannt sind. Erschwerend kommt hinzu, dass für eine Interpretation der Ergebnisse der typischerweise große Umfang zugehöriger Fachliteratur erfasst werden muss.

Die vorliegende Arbeit beschäftigt sich mit den folgenden Themen, welche jeweils wichtige und eigenständige Forschungsfelder der Bioinformatik darstellen: (i) der Konstruktion realistischer Modelle genregulatorischer Netzwerke, (ii) der Identifikation von Teilnetzwerken, welche eine signifikante Änderung in den gemessenen Daten zeigen, und (iii) der systematischen biologischen Interpretation der gefundenen Teilnetzwerke.

Für die Konstruktion genregulatorischer Netzwerke wird ein Überlick der typischen Vorgehensweisen gegeben, wobei Verfahren, die auf Kuration der Literatur bzw. Inferenz aus experimentellen Daten basieren, vertieft behandelt werden. Zunächst wird am Beispiel der Diauxie in Hefe beschrieben, wie manuelle Kuration eingesetzt werden kann, um ein regulatorisches Netzwerk für einen spezifischen Prozess zu konstruieren. Hierbei wird insbesondere die Frage behandelt, wie ein detailliertes Verständnis, wie es für die Regulation einzelner Gene existiert, auf größere Systeme übertragen werden kann. Anschließend werden Methoden zur Netzwerk-Inferenz untersucht und demonstriert, dass diese unverhältnismäßig in Richtung von Regulatoren vieler Gene verschoben sind. Zur Auflösung wird eine Kalibrierungsstrategie vorgestellt, deren Anwendung ein diesbezüglich verbessertes, genomweites regulatorisches Netzwerk für Hefe liefert.

Um Teilnetzwerke zu identifizieren, die signifikante Änderungen in den Daten zeigen, wird GGEA als eine Methode zur netzwerkbasierten Anreicherungsanalyse eingeführt. Die grundlegende Idee der Methode ist, regulatorische Interaktionen innerhalb funktionaler Gruppen von Genen bezüglich der Konsistenz mit den beobachteten Expressionsdaten zu bewerten. Verglichen mit anderen Anreicherungsmethoden, liefert GGEA Ergebnisse, welche Veränderungen in der Expression mit bekannten regulatorischen Interaktionen in Übereinstimmung bringen, wodurch sich die Ergebnisse leichter erklären lassen. Desweiteren werden verschiedene Modifikationen von GGEA diskutiert, welche Anwendbarkeit, Ergebnisse und Laufzeit der Methode signifikant verbessern.

Für die systematische Identifikation und Interpretation der gefundenen Teilnetzwerke wurde das `EnrichmentBrowser` Software-Paket entwickelt. Es implementiert verschiedene Anreicherungsmethoden, zusätzlich zu GGEA, und erlaubt die Ergebnisse über Methoden hinweg zu kombinieren und zu untersuchen. Als Bestandteil der `Bioconductor` Software-Bibliothek erlaubt das Paket einen einheitlichen Zugriff auf die verschiedenen Methoden, wodurch die Nutzung für Anwender stark vereinfacht wird. Erweiterungen des Pakets, welche die Automatisierung typischer Vorgehensweisen bei der biologischen Interpretation unterstützen, werden ebenfalls diskutiert.

Die vorliegende Arbeit trägt somit wesentlich zum Forschungsgebiet der netzwerkbasierten Analyse von Genexpressionsdaten bei, insbesondere bezüglich der Konstruktion regulatorischer Netzwerke, der Identifizierung signifikanter Teilnetzwerke und deren biologischer Interpretation. Dies beinhaltet auch neuere Entwicklungen und aktuelle Forschungsthemen, welche im Zusammenhang gegenwärtiger als auch zukünftiger Fragen, die sich aus den neuartigen genomischen Daten ergeben, diskutiert werden.

Chapter 1

Introduction: Analyzing gene expression with networks

Understanding the biology of living organisms, and especially disease-causing malfunctions occurring within them, requires a detailed understanding of how the underlying genotype is expressed into the observed phenotype. The building blocks of this translation are proteins, biochemical polymers that perform numerous fundamental cellular functions. This includes the catalysis of metabolic reactions as enzymes, the regulation of behavior and development as hormones, and their aggregation in groups to form tissue, muscles and hair [1, 2].

According to the central dogma of molecular biology, blueprints of proteins are encoded in the genome: the stable inheritance information in the DNA is first transcribed to messenger RNA (mRNA), which is then translated to functional proteins [3, 4]. Both, transcription and translation, are tightly regulated by different mechanisms assembled in networks of regulatory interactions between proteins, RNA, and the DNA (Figure 1.1).

Transcription is activated in response to extra-cellular signals that are perceived at the cell surface by specific membrane proteins. Amplified by a signaling cascade of interacting kinase proteins, signals are transduced to transcription factor proteins. Once activated, these factors enter the nucleus where they bind to regulatory DNA elements denoted as promoters, hereby activating or repressing the transcription of associated target genes (Figure 1.2). Transcriptional control is thus exerted through the complex interplay of protein-protein interactions (PPIs) between signaling proteins and transcription factors as well as gene regulatory interactions (GRIs) between transcription factors and target genes [5].

Cellular processes typically display different levels of activity in healthy and pathogenic tissue. For example, cancer cells execute processes leading to cell proliferation in higher, but processes leading to cell death in lower activity levels than their unaffected counterparts [6,7]. This is also reflected by different quantities of disease-relevant transcripts in the transcriptome denoting all mRNA transcripts in the cell. In contrast to the biochemically diverse and experimentally instable proteins, mRNA has nearly uniform experimental behavior and is thus preferentially used for quantitative measurements of gene activity. This is based on the assumption that the total cellular amount of a particular mRNA reason-

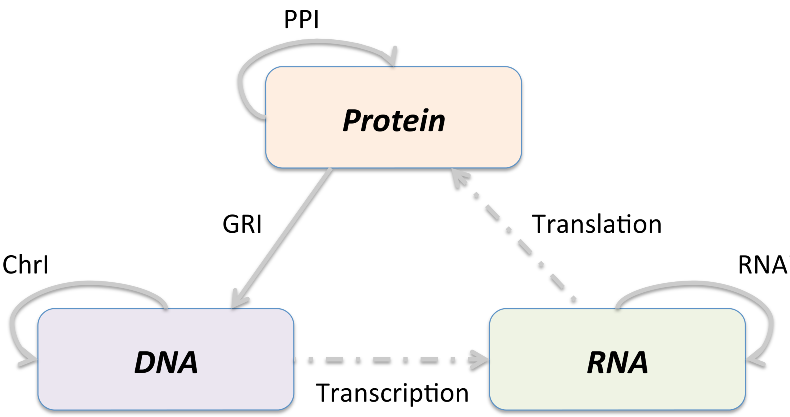


Figure 1.1: Regulatory interactions between proteins, RNA and DNA. PPI: protein-protein interaction. GRI: gene regulatory interaction. ChrI: chromatin interaction. RNAi: RNA interference.

ably approximates the amount of the corresponding protein [8, 9]. However, processes like post-transcriptional mRNA degradation can interfere with this assumption [10].

Supported by the rapid development of high-throughput devices in molecular biology, it has recently become possible to efficiently measure the transcriptome at moderate costs [11,12]. This has enabled many comparative studies designed to detect significant differences in gene expression between two groups of samples representing two different experimental conditions. In the clinical case, these are groups of healthy and diseased patients. The standard way, in which these studies are carried out, is depicted in Figure 1.3. Although all experimental steps are performed under stringent conditions, technical variation is inherent to large-scale experiments such as microarray or RNA-seq assays (see Appendix A.1 for a description of both assays). Preprocessing of the data is thus required to exclude outliers and normalize gene expression values within and between measurements [13, 14, for an overview].

Subsequently, the analysis is centered on single genes, which are significantly different in expression between the two different conditions [15, 16]. These genes indicate putative disease-causing mechanisms and are thus hotspots for medical attention. They can be also used to predict the disease status of new patients [17, 18].

However, restricting the analysis to such marker genes ignores important information. For instance, master regulators are known to affect many target genes while changing expression only slightly themselves [19]. On the other hand, minor but coherent changes of a functional group of genes are typically more meaningful than strong signals without context.

Capturing such changes is the goal of subsequent analysis of predefined gene sets representing functional categories or biochemical pathways [20, 21]. Established databases defining such gene sets (Appendix C.3) are the Gene Ontology [22, 23, GO] and the Kyoto Encyclopedia of Genes and Genomes [24, 25, KEGG]. As rarely all genes in a set show substantial

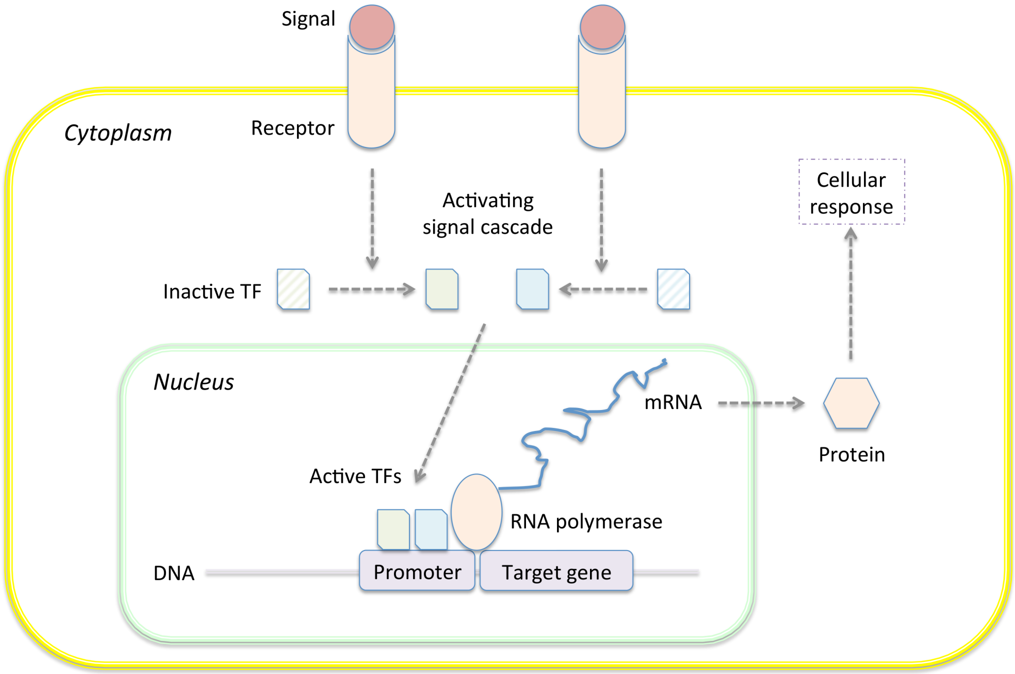


Figure 1.2: Schematic illustration of a gene regulatory network. Extra-cellular signals perceived at the cell surface are transduced to specific transcription factors (TFs) that either individually or in combination regulate the transcription of target genes by RNA polymerase. The expressed protein products execute the cell functions required to react to the perceived signal.

change in expression, the investigation is consequently centered on whether such changes are statistically overrepresented in certain gene sets [26, 27].

The detected altered gene sets are incorporated in various ways in the actual biological interpretation of the observed results. This includes a literature search for reported associations between these gene sets and the investigated phenotype as well as exploration of known interactions of genes within and between gene sets. Results can then be divided into observations that are supported by previous findings and novel hypotheses about the underlying disease-causing mechanisms.

Although network knowledge is incorporated *ad hoc* in the biological interpretation, it can be beneficial to already systematically integrate it into the gene set analysis to enhance the discovery of coherently altered functional modules of genes [28, 29]. For example, the concordant up-regulation of two proteins known to be associated in a functional complex, such as the two transcription factors in Figure 1.2, is often considered more meaningful than expression changes of two unrelated proteins. Such protein-protein interactions (PPIs) that are in agreement with the observed expression are accordingly weighted higher, and gene sets with disproportional many of these interactions are preferred.

As experimental high-throughput techniques for PPI-detection were developed early and resulting large PPI-networks became easily available, they were frequently used for network-

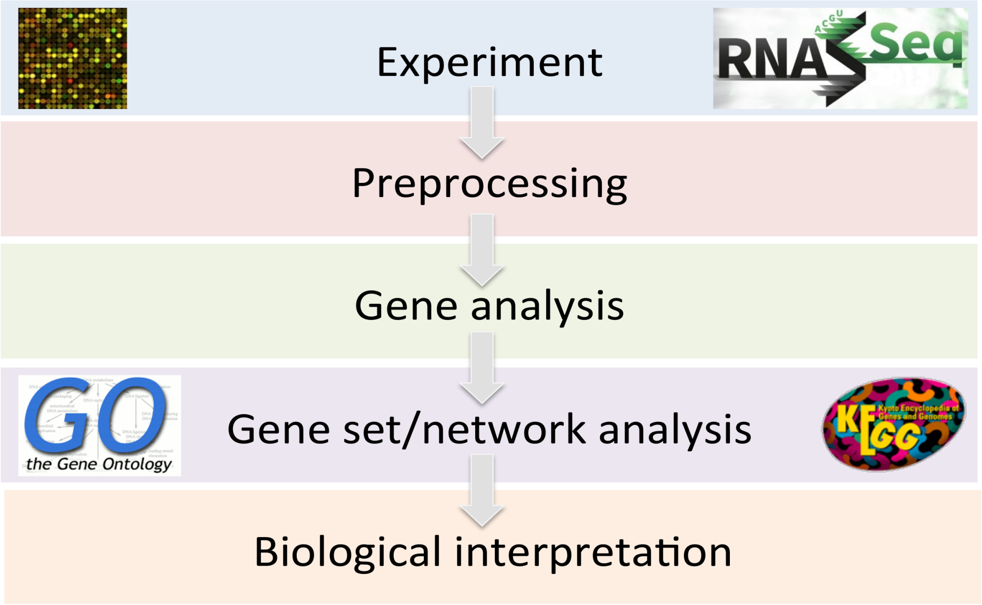

Experiment
RNA Seq
ACGU
Preprocessing
Gene analysis
Gene set/network analysis
GO
the Gene Ontology
KEGG
Kyoto Encyclopedia of Genes and Genomes
Biological interpretation

Figure 1.3: The typical workflow of gene expression data analysis.

based gene set analysis [30, 31]. However, PPIs take place post-translational and are thus not necessarily reflected by detectable quantitative changes in mRNA transcription of the interacting proteins. In contrast, such changes in mRNA transcript abundance result directly from gene regulatory interactions (GRIs) between proteins and the DNA (see Appendix A.2 for experimental procedures to detect GRIs). GRIs are thus well-suited for the analysis of transcriptomic data, as they yield straightforward explanations for observed expression changes.

In the following, I describe in detail how networks of regulatory interactions can be constructed on a large scale and used for the analysis and interpretation of gene expression data. Chapter 2 describes the construction of regulatory network models with a focus on curation and inference approaches. Chapter 3 introduces methods for the detection of subnetworks that are significantly altered in a comparative setup. The systematic biological interpretation of the detected subnetworks is subject of Chapter 4.

Chapter 2

Construction of regulatory networks

2.1 Introduction

There are well-studied model organisms for which comprehensive and well-annotated regulatory networks are available, e.g. the `RegulonDB` for *E. coli* [32] and `YEASTRACT` for *S. cerevisiae* [33]. However, for non-model and more complex organisms such a network is often missing or at least incomplete. To complement the required information, several approaches can be applied to construct a gene regulatory network (Figure 2.1). This can be done using either extrinsic or intrinsic interactions of existing gene set catalogs.

Extrinsic approaches require external resources apart from the gene set definition. Frequently used approaches are

1. Literature curation / mining,
2. Inference from experimental data, and
3. Cross-species transfer.

Literature curation is used to manually compile specific regulatory networks from relevant scientific articles. It is the method of choice of major regulatory network databases to derive reliable networks of suitable granularity. However, this typically comes at the price of many man hours of curation and rarely guarantees completeness. Thus, it is subject of ongoing research how the curation can be automated or at least supported by text-mining methods [34].

Inference from experimental data reconstructs the GRN from genome-wide expression and transcription factor binding data [35, for a comprehensive assessment of existing methods]. In case such data is available, it is an easy and fast alternative to the time-consuming curation approach. Nevertheless, it requires to deal with indirect and false-positive effects that usually accompany high-throughput data.

Cross-species transfer of regulatory networks [36, 37] is not as frequently used, but is especially suited whenever a well-annotated network of an organism related to the one investigated exists. A good example are the model organisms *S. cerevisiae* (budding yeast) and

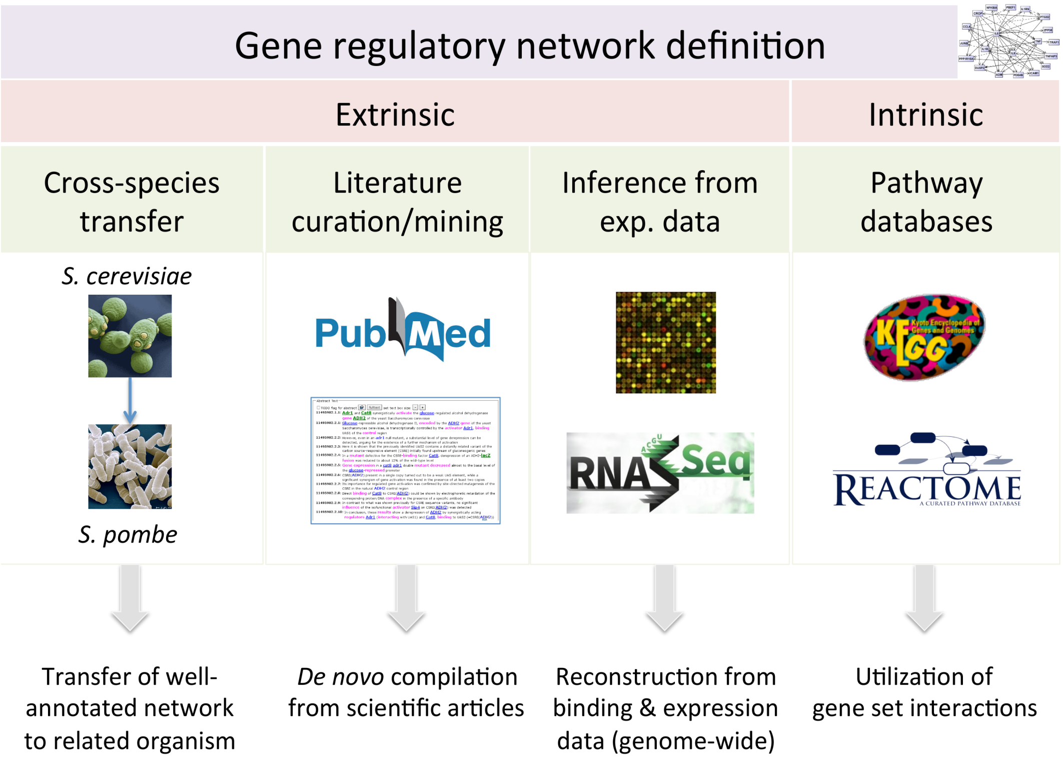

Gene regulatory network definition
Extrinsic
Intrinsic
Cross-species transfer
Literature curation/mining
Inference from exp. data
Pathway databases
S. cerevisiae
S. pombe
PubMed
RNA Seq
KEGG
REACTOME
Transfer of well-annotated network to related organism
De novo compilation from scientific articles
Reconstruction from binding & expression data (genome-wide)
Utilization of gene set interactions

Figure 2.1: Overview of approaches to construct a gene regulatory network.

S. pombe (fission yeast). Unlike budding yeast, no large-scale regulatory network has been constructed for fission yeast [38]. This suggests to transfer and incorporate the GRN in studies comprising nearly identical experimental conditions, such as the two environmental stress studies in both yeasts [39,40]. Transferring the regulatory interactions is particularly suited in this case, as stress-induced expression changes in many genes are evolutionarily conserved between these two rather distantly related yeasts [40].

An extrinsically defined network enables the incorporation of important prior knowledge. However, this requires additional input which is not always easily available. This is the reason that analysis methods relying only on expression data and gene set definition are preferably used in practice. It is thus, in this regard, advantageous to exploit intrinsic regulatory interactions of the gene sets. Although this is not applicable for set definitions without annotated interactions like `GO` [22], pathway-based gene sets defined according to databases like `KEGG` [41] and `Reactome` [42] contain usually a considerable number of such interactions.

In the following, I discuss literature curation and data inference in more detail based on recent publications in the field (Section 2.2 and 2.3). For further reading on network transfer across species, the reader is referred to the PhD thesis of Robert Pesch [43]. The integration of gene set intrinsic interactions is subject of Chapter 3.

2.2 Compilation from scientific articles

This section describes how literature curation can be used to construct a regulatory network for a specific biological process. In particular, it addresses the question how a detailed understanding, as available for the regulation of single genes, can be scaled-up to the level of larger systems. The work has been presented on August 30th, 2013, at the 26th International Conference on Yeast Genetics and Molecular Biology. The following is based on its original publication in the journal Nucleic Acids Research*, 41(18):8452-63, October 2013.*

Authors: Ludwig Geistlinger, Gergely Csaba, Simon Dirmeier, Robert Küffner, and Ralf Zimmer

Title: A comprehensive gene regulatory network for the diauxic shift in *S. cerevisiae*

Abstract: Existing machine-readable resources for large-scale gene regulatory networks usually do not provide context information characterizing the activating conditions for a regulation and how targeted genes are affected. Although this information is essentially required for data interpretation, available networks are often restricted to not condition-dependent, non-quantitative, plain binary interactions as derived from high-throughput screens. In this article, we present a comprehensive Petri net based regulatory network that controls the diauxic shift in *S. cerevisiae.* For 100 specific enzymatic genes, we collected regulations from public databases as well as identified and manually curated >400 relevant scientific articles. The resulting network consists of >300 multi-input regulatory interactions providing: (i) activating conditions for the regulators; (ii) semi-quantitative effects on their targets; and (iii) classification of the experimental evidence. The diauxic shift network compiles widespread distributed regulatory information and is available in an easy-to-use machine-readable form. Additionally, we developed a browsable system organizing the network into pathway maps, which allows to inspect and trace the evidence for each annotated regulation in the model.

Author contributions: LG, SD and RK curated and annotated the regulatory network from the literature and databases. GC implemented the annotation framework and created the front-end based on the pathway maps designed by LG. LG and RK wrote the main paper with suggestions from RZ. RK and RZ supervised the project.

2.2.1 Introduction

Gene regulatory networks (GRNs) model the effects of transcription factors (TFs) on the expression of their target genes (TGs). As large networks are collected in existing databases, such as `RegulonDB` [44], `YEASTRACT` [45], and `REDfly` [46], it is tempting to use them for the interpretation of large-scale gene and protein expression data.

However, to perform meaningful interpretation of such high-throughput transcriptomic and proteomic data, GRNs need to be modeled at least by (i) defining the conditions under which a regulation takes place or does not take place, and (ii) characterizing the effect on the expression of the regulated TG.

The first requirement results from the fact that, to adapt to changing environmental conditions, the cell usually responds with altered gene expression. For example, gene regulation in baker's yeast *S. cerevisiae* changes in response to different nutrients in the growth medium [47]. Hence, the interpretation of gene expression measured under certain conditions requires a dynamic condition-dependent definition of the enabled regulations – the active subnetwork of all possible regulations.

The second requirement is due to the fact that genes, qualitatively and quantitatively, are not regulated in a uniform way. On the one hand, again depending on the environmental conditions, relevant genes are activated or repressed to a different extent. On the other hand, combinatorial control of a TG by several TFs can have a non-trivial synergistic effect [48,49]. Thus, to understand the observed expression in the data, that is to assign observed expression changes to certain regulators, a detailed characterization of the regulatory effect on the TG expression is necessary. This includes the determination of the 'effect type' (activation or inhibition) and the 'effect strength' (weak or strong activation/inhibition) as well as an appropriate combination of multi-input effects.

While both requirements are therefore essential, such context information characterizing a regulation is often unknown or not annotated. Derived from high-throughput protein-protein interaction or TF-binding experiments [50,51], the majority of available large-scale regulatory networks consists of plain binary interactions, for example stating for a certain TF F and its TG G that F *interacts with* G. The effect of these interactions on gene expression is usually not further characterized. It is also unclear whether the interactions take place under conditions different from the setup used in the respective experiments.

In this article, we propose a model for large-scale regulatory networks satisfying both requirements and present a comprehensive realization of the model for transcriptional regulation of the diauxic shift in yeast.

S. cerevisiae is a facultative anaerobic organism preferably fermentating glucose to produce energy for fast growth. Subsequent to the depletion of glucose, fermenting yeasts switch to slower respiratory growth on a non-fermentable carbon source like ethanol, lactate, glycerol, or fatty acids. This involves a major reprogramming of gene regulation that includes the deactiviation and activation of specific TFs, which in turn activate or repress specific metabolic genes [47,52,53]. Many of the differentially regulated genes code for enzymes, which metabolize the non-fermentable carbon source available in the growth medium, and use the resulting products for the recreation of glucose via gluconeogenesis and the production of energy via the tricarboxylic acid (TCA) cycle.

2.2.2 Material and Methods

The yeast GRN

Experimental techniques

TFs activate or repress the expression of TGs in response to extra- and intracellular signals. Such gene regulatory interactions (GRIs) between TFs and TGs can be experimentally determined either by directly confirming the TF binding to the regulatory region of the TG, or indirectly inferred from TG expression changes following a TF perturbation.

Direct evidence (TF binding): Physical binding of a TF to the promoter of its TG can be determined using several techniques such as wild type versus TG promoter mutant analysis via a *lacZ*-fusion assay [54] or northern blot [55], DNA footprinting [56], electrophoretic mobility shift assay [57], and chromatin immunoprecipitation [58, ChIP].

The combination of ChIP with the microarray technology [59, ChIP-chip] allows the genome-wide identification of TF-binding sites. ChIP-chip experiments have been comprehensively performed for all yeast TFs [60, 61].

Putative TGs of a TF can be predicted based on high sequence similarity to its binding sites at the promoter of known TGs. Consensus sequences of TF-binding sites, represented as position weight matrices (PWMs), have been computed for many known yeast TFs and stored in databases like `TRANSFAC` [62] and `JASPAR` [63]. However, PWM-based GRIs are hypothetical, and only a fraction of them can be experimentally validated (Figure 2.2).

It is frequently observed that the binding of a certain TF to the promoter of its TG is ineffective, i.e. does not result in a observable quantitative expression change of the TG. This has several reasons, either other TFs might be required to bind or post-translational modifications (for example phosphorylation of the TF) or other signals might be needed to activate the regulatory function of the TF [64]. Indeed, as depicted in Figure 2.2, <10% of known direct physical bindings are associated with a subsequent quantitative fold change of the corresponding TG.

Indirect evidence (TG expression): In contrast to binding studies, regulatory effects (activation or inhibition of TG) can be derived and quantified (fold change) from gene expression studies, where certain TFs have been either knocked out, over-expressed or in other ways functionally modified. Frequently used experimental techniques include *lacZ*-fusion assays, northern blot, real-time PCR [65], and microarrays [66]. The most comprehensive series of yeast TF knockout microarrays has been performed by Hu *et al.* [67], where significant expression changes of putative TGs have been assigned to the individual deletion of almost every single yeast TF.

A large fraction of effects observed exclusively in such TF perturbation studies are assumed to be indirect, i.e. the expression change of a TG is a secondary effect, which is due to the deregulation of the TF caused by another knockout. Indeed, <12% of known indirect effects are associated with direct physical binding (Figure 2.2).

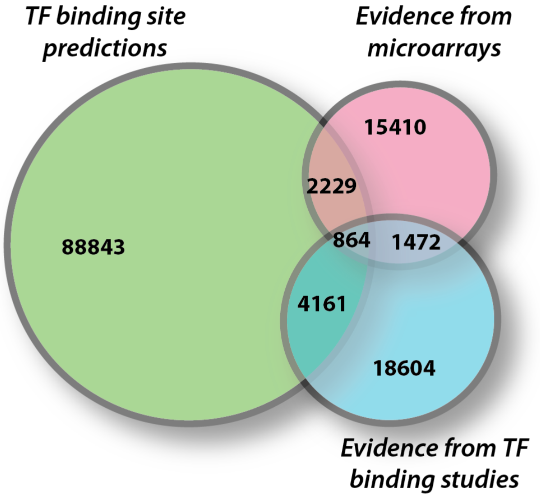


Figure 2.2: Overlaps between predicted, direct and indirect gene regulatory interactions (GRIs) in *S. cerevisiae*. GRIs with experimental evidence were taken from `YEASTRACT` (microarray and binding studies). Predictions were performed for all 160 yeast TF PWMs in JASPAR for the promoter regions of all yeast genes (using the R package *cureos*, default settings). The percentage of predictions, which have an experimental evidence for binding is 5.2% (5,025 of 96,097). On the other hand, 9.3% (2,336 of 25,101) bindings are associated with a change of TG expression.

Confidence classes of experimental evidence: Whether the confidence in reported GRIs is 'low' or 'high' depends on the available experimental evidence. Usually, combined evidence of TF binding and TG expression, i.e. the TF binds to the promoter of the TG and a perturbation of the TF results in an expression change of the TG, increases the confidence. In contrast, GRIs with evidence for either binding or expression are not highly reliable *per se* (see again Figure 2.2). The same holds for additional evidence from consensus analyses and author statements for which the experimental evidence cannot be traced. We thus discriminate in the following between 'high' confidence regulations having combined evidence for binding and expression, and 'low' confidence regulations in all other cases.

Resources

We exploited three representative resources for yeast GRIs: The Saccharomyces Genome Database [68, SGD] is the source for a variety of genomic and biological information on *S. cerevisiae* and contains regulatory information for many yeast genes (as quantified in Figure 2.3b). Besides other widespread biological facts, including post-transcriptional regulation, metabolic function and orthology to genes in other organisms, the `SGD` summary paragraph on a specific yeast gene often contains different aspects of transcriptional regulation (upstream signals, putative binding sites, validated TF binding, expression effects). However, this valuable information is not easy accessible: the gene summaries are written in free-text and the aspects described differ considerably between genes. Manual curation is thus required to extract this information.

Compared with `SGD`, `YEASTRACT` [45] is a specific database for transcriptional regulation in *S. cerevisiae*, in which GRIs are uniformly represented as binary TF-TG associations in a machine-readable format (obtainable as tabular flat file). Mainly derived from recent genome-wide TF binding and TF perturbation experiments, `YEASTRACT` aims to collect all TFs either binding to a particular TG, or show expression changes of the TG when perturbed. Although `YEASTRACT` stores a large number of GRIs (Figure 2.2), it does not provide context information characterizing under which conditions the GRIs take place and how targeted genes are affected.

In contrast to `YEASTRACT`, Herrgard *et al.* [69] have curated the nutrient-controlled regulation of yeast genes involved in metabolic pathways. Mainly derived from detailed studies with a focus on one or a few specific genes, it presumably contains significantly less false positive GRIs as compared with untargeted genome-wide experiments. Each GRI is classified (activation/inhibition), frequently assigned to a nutrient-based context (extra- and intracellular signals), and described by a boolean rule (e.g. if SIGNAL and TF then TG). Although enriched with required context information, the GRN is sparse: only a small fraction of the vast amount of articles existing on the regulation of metabolic yeast genes has been taken into account.

The diauxic shift GRN

Curation

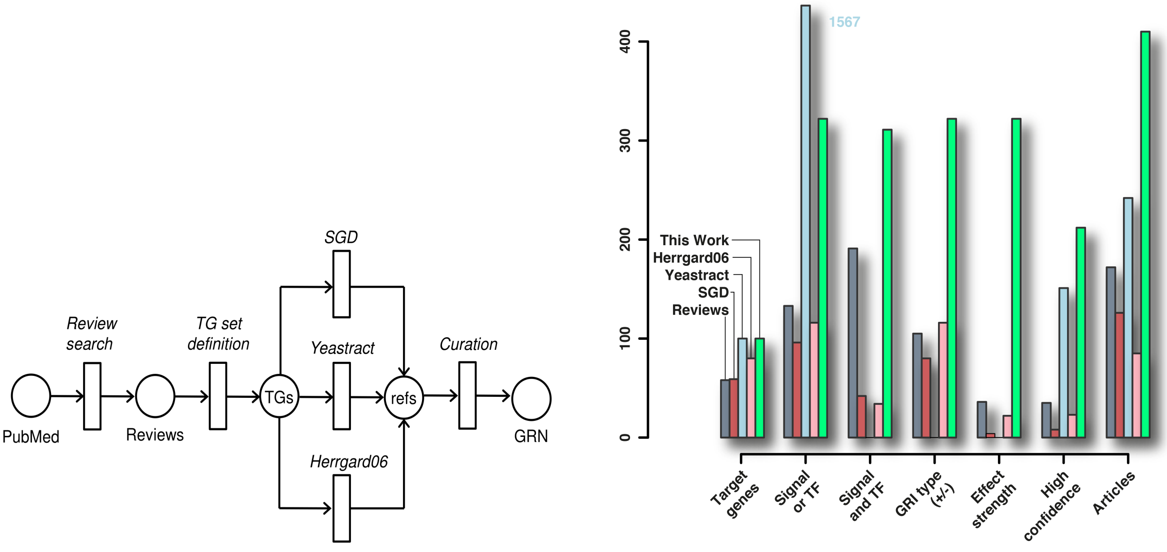
We curated a GRN that controls the diauxic shift in three steps (our approach is illustrated in Figure 2.3a and the information collected in each step is detailed in Figure 2.3b):

(1) TG set determination: We collected current reviews on transcriptional regulation of the diauxic shift to define the set of involved TGs. We concentrated on Zaman *et al.* [47], an extensive description of how *Saccharomyces* responds to different nutrients, Hiltunen *et al.* [70] and Gurvitz and Rottensteiner [71] for transcriptional regulation of fatty acid metabolism and oleate induction, and especially on Schüller [52] and Turcotte *et al.* [53], who comprehensively reviewed the transcriptional control of non-fermentative metabolism in *S. cerevisiae*. Based on this literature, we determined the involved metabolic processes and the associated enzymatic TGs.

(2) GRI collection: We systematically queried existing resources, i.e. `SGD`, `YEASTRACT` and Herrgard *et al.* [69], for information on the transcriptional regulation of the identified TGs. The representation of the information available in each of the three resources is described in more detail in the previous section.

In `SGD`, we used the `summary` site for each TG and manually screened the `Description` slot and the `Summary Paragraph` (if existing) for regulatory information. Additionally, we collected all references to the primary literature listed on that page.

In `YEASTRACT`, we retrieved for each TG all regulating TFs using the `Search for TFs` functionality. From the resulting list of binary TF-TG associations grouped by experimental evidence (direct or indirect, see previous section *Experimental techniques*), we also collected all references assigned to the associations for experimental support.



(a)

(b)

Figure 2.3: Curation approach. (a) Protocol: Based on a set of selected reviews, we defined the set of diauxic shift TGs. Each gene was queried for regulatory information in `SGD`, `YEASTRACT`, and Herrgard *et al.* [69]. The GRN was compiled from information directly retrieved from the resources and from the curation of all extracted references. The information collected in each step of our approach is detailed in **(b)**. Slots on the x-axis from left to right: (1) Number of diauxic shift TGs for which regulatory information could be annotated; (2)-(6) Number of regulations with (2) either Signal *or* TF annotated; (3) Signal *and* TF; (4) regulation type: activation or inhibition; (5) effect strength: weak, medium, strong; (6) High confidence, see *Material and Methods*; (7) Number of articles in which regulations could be annotated.

Eventually, we restricted the curation of Herrgard *et al.* for all metabolic yeast genes on the information available for the diauxic shift TGs. That yielded a list of regulatory TF-TG relationships (classified as activating or repressing) that are enabled under certain conditions, i.e. triggered by a particular extra- or intracellular signal. Again, we collected all cited references.

(3) GRN compilation: We compiled the GRN via combination of the regulatory information that was directly retrieved from the resources or curated from the references collected in Step 2.

The combination of the directly retrieved information initially required the identification of GRIs contained in two or all three resources. For such well-studied GRIs, the resources often complemented each other. For example, a binary TF-TG association from `YEASTRACT` could be characterized in more detail with features retrieved from `SGD` or Herrgard *et al.* such as regulation sign (+/-), effect strength and the enabling context. In addition, GRIs with low confidence in one resource alone frequently gained high confidence when evidence was combined from several resources (see previous section *Confidence classes of experimental evidence*).

On the other hand, curation of the collected references often allowed the annotation of additional features and a more detailed characterization, especially for poorly studied GRIs contained in only one resource. Curation was performed down to the actual experimental evidence for a GRI under investigation, i.e. references were traced iteratively until the experimental confirmation of the regulation was found. In general, we aimed at the most detailed GRI characterization possible from literature curation, for which we propose a general representation in the next section.

Representation

We integrate the curated GRIs into discrete regulation models, i.e. models in which discrete states of the regulators (TFs and signals) result in discrete quantity states of the regulated gene (for instance a low, medium, or high expression) depending on the regulation type. We use Petri net models to efficiently represent the information typically available in the literature. Petri nets are well-established graphical and mathematical models [72] and have been extensively applied to biochemical processes, like signal transduction pathways [73,74] and gene regulatory networks [75, 76]. The extension of Petri net models with fuzzy logic [77] in the PNFL approach [78] allows a more detailed semi-quantitative representation of in- and output of the Petri net transitions, which are defined by simple rule sets according to the regulation type [79,80]. Thus, we replace the frequently used representation of GRIs as *binary* TF-TG interactions by *multi-input* Petri net transitions, in which the required context knowledge (activating conditions, combinatorial control, and effects on TG expression) can be integrated by accurate parametrization of in- and output and definition of the transition type. The parametrization of such transcriptional transitions is based on a differential regulation setting, where the presence or absence of a signal induces an enhanced or reduced activity of specific TFs, which in turn regulate their TGs differentially (up or down, as compared to the corresponding opposite signal state).

Input: The input of a transcriptional transition is composed on the one hand by the context signals, which trigger the regulation, and, on the other hand, by the TFs, which perform the actual regulation of the TG under investigation. For example, the depletion of glucose and the availability of a non-fermentable carbon source (the signals) trigger the derepression of enzymes involved in non-fermentative metabolism (the TGs) by specific TFs.

Signals can be extra- or intracellular messenger molecules (such as cAMP), nutritional compositions (for instance, growth media lacking glucose), environmental and experimental conditions (such as high pH or heat stress) and even cellular states (such as retrograde regulation depending on the functional state of the mitochondria).

Based on the absence or presence of given signal(s), the TFs are classified as up- or down-regulated in discrete states 'weak', 'medium' or 'strong'. The special states 'overexpression' (for *up*) and 'knockout' (for *down*) were annotated as well (see Figure 2.4 for an example).

In general, the signal and TF assignments define the conditions under which the transition is enabled.

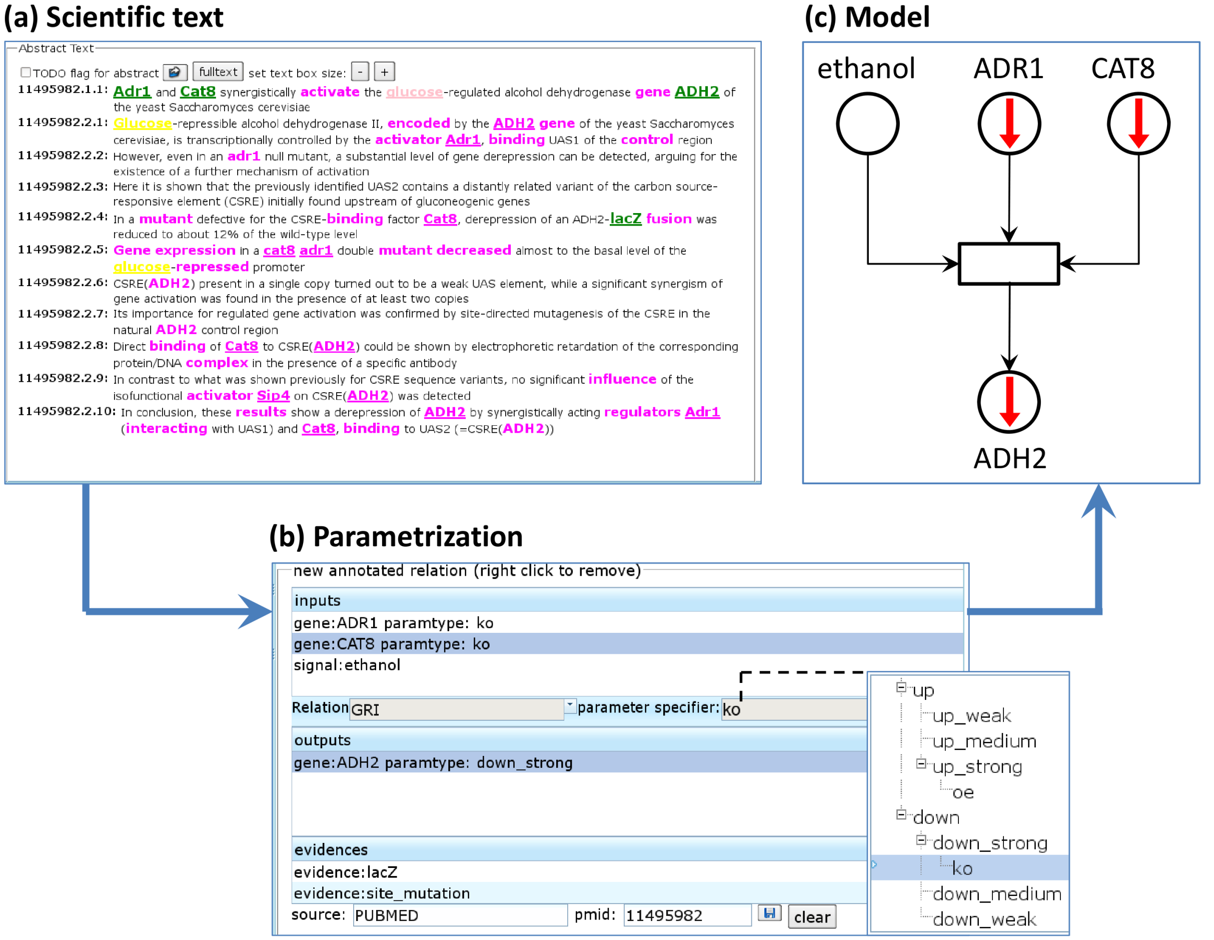
Output: Analogously, TGs are classified as up- or down-regulated by a given transition with a ’weak’, ’medium’ or ’strong’ effect strength. Intuitively, this models the fold change in the transcription of the TG. Although the effect strength might differ considerably between datasets in range and distribution, the literature often explicitly states whether a regulation has a weak or strong effect on TG expression. In cases where an exact fold change is reported, we discretize the fold change according to empiric standards: ’weak’ regulation refers to expression changes below 2-fold, ’medium’ between 2- and 5-fold, and ’strong’ above 5-fold. The transition type results immediately from a given in- and output configuration, e.g. a TF knockout, resulting in a weak upregulation of the TG, indicates a weak inhibition.

Annotation framework

In order to collect regulatory knowledge effectively from publications, we performed the curation using our in-house annotation software `RelAnn` (Csaba *et al.*, unpublished). The web-based tool was developed for general text-based annotations of different kinds of relations within a systematic framework.
The main design principles of `RelAnn` are:

- pre-indexing of defined biological entities (genes, proteins, etc.) in the literature;
- simple, click-based annotations to relate the entities to each other; and
- representation of relations as Petri net transitions.

As illustrated in Figure 2.4, we use `RelAnn` for the transformation of literature knowledge to the representation of GRIs as semi-quantitative Petri net transitions (as described in the previous section *Representation*).
Subsequent to the pre-indexing of the relevant text using a named entity search, occurrences of defined entities are used for the definition of input (regulators, i.e. TFs and signals), output (regulatees, i.e. TGs) and experimental evidence for a regulatory transition. Thus, every part of the transition (regulatory, regulatees, evidence) is linked to some phrase in a scientific article of the `PUBMED` database, thereby making the source of the knowledge traceable. In addition, in- and output specification allow the assignment of the semi-quantitative type of needed (input) or induced (output) change associated to the regulation, to wit ’up’ or ’down’, with ’weak’, ’medium’, or ’strong’ effect strength (bottom right of Figure 2.4b).
A special feature of `RelAnn` is the organization of all components (gene, signal, evidence, regulation and parameter types) in ontologies enabling powerful queries and specifications using generalization and specialization. For example, the regulation annotated in Figure 2.4 can be not only captured by searching for all regulations having ’ethanol’ as input, but also by searching for all regulations having a ’non-fermentable carbon source’ as input.



☐TODO flag for abstract set text box size:

11495982.1.1: Adr1 and Cat8 synergistically activate the glucose-regulated alcohol dehydrogenase gene ADH2 of the yeast Saccharomyces cerevisiae

11495982.2.1: Glucose-repressible alcohol dehydrogenase II, encoded by the ADH2 gene of the yeast Saccharomyces cerevisiae, is transcriptionally controlled by the activator Adr1, binding UAS1 of the control region

11495982.2.2: However, even in an adr1 null mutant, a substantial level of gene derepression can be detected, arguing for the existence of a further mechanism of activation

11495982.2.3: Here it is shown that the previously identified UAS2 contains a distantly related variant of the carbon source-responsive element (CSRE) initially found upstream of gluconeogenic genes

11495982.2.4: In a mutant defective for the CSRE-binding factor Cat8, derepression of an ADH2-lacZ fusion was reduced to about 12% of the wild-type level

11495982.2.5: Gene expression in a cat8 adr1 double mutant decreased almost to the basal level of the glucose-repressed promoter

11495982.2.6: CSRE(ADH2) present in a single copy turned out to be a weak UAS element, while a significant synergism of gene activation was found in the presence of at least two copies

11495982.2.7: Its importance for regulated gene activation was confirmed by site-directed mutagenesis of the CSRE in the natural ADH2 control region

11495982.2.8: Direct binding of Cat8 to CSRE(ADH2) could be shown by electrophoretic retardation of the corresponding protein/DNA complex in the presence of a specific antibody

11495982.2.9: In contrast to what was shown previously for CSRE sequence variants, no significant influence of the isofunctional activator Sip4 on CSRE(ADH2) was detected

11495982.2.10: In conclusion, these results show a derepression of ADH2 by synergistically acting regulators Adr1 (interacting with UAS1) and Cat8, binding to UAS2 (=CSRE(ADH2))

new annotated relation (right click to remove)

inputs
gene:ADR1 paramtype: ko
gene:CAT8 paramtype: ko
signal:ethanol

Relation GRI parameter specifier: ko

outputs
gene:ADH2 paramtype: down_strong

evidences
evidence:lacZ
evidence:site_mutation

source: PUBMED pmid: 11495982

Figure 2.4: From pure text to semi-quantitative models of gene regulatory interactions. Within our annotation framework, the pre-indexed regulatory entities in **(a)** can be easily selected and used for the parametrization in **(b)** of input and output of the Petri net model of the gene regulatory interaction in **(c)**. In the same manner, experimental evidence can easily assigned to the model.

2.2.3 Results

Descriptive analysis

We have curated a gene regulatory network for the diauxic shift in *S. cerevisiae*. As illustrated in Table 2.1, the curation yielded 1,133 text-based annotations of regulatory interactions in 410 scientific articles. The resulting 322 gene regulatory interactions cover the core processes taking place during the switch from fermentation to respiration. This includes the regeneration of fermentable glucose (gluconeogenesis), oxidation of glycolytic products (TCA cycle), and catabolism of non-fermentable carbon sources (ethanol, glycerol, lactate, acetate and fatty acids). In addition, we characterized the upstream regulation events of glucose signaling and the corresponding transcriptional regulation of the

Table 2.1: Annotation summary. Shown are the numbers for the total annotation outcome and for the corresponding subprocesses of the diauxic shift. TF: Transcription Factor.

	Genes	**TFs**	**Interactions**		**Annotations**	**Articles**
			high[1]	**all**		
Total	100	68	212	322	1133	410
Gluconeogenesis	18	37	56	77	252	117
Fatty acid metabolism	19	20	34	57	203	79
TCA Cycle	23	24	29	52	146	64
Glyoxylate cycle	5	27	16	26	102	67
Ethanol metabolism	5	17	13	16	108	76
Glycerol metabolism	3	19	9	18	36	24
Lactate metabolism	3	10	11	11	38	21
Glucose signaling	11	22	25	35	147	71
TF-TF	14	24	19	31	101	69

[1] *High*-confidence gene regulatory interactions have experimental evidence for binding and expression (*Material and Methods*). *All* interactions include *low*- and *high*-confidence interactions.

key signal proteins that act once glucose is depleted. In total, our network connects 100 TGs with 72 regulating TFs driving the transcriptional response to >50 different extra- and intracellular signal classes. The transcriptional regulation of the regulators themselves has been investigated and integrated into the network.

To estimate the completeness of our network, we extrapolated the expected number of interactions contained in an infinite number of articles relevant for the diauxic shift. Using a first order Hill equation, we estimated that our network is 71% complete (Supplementary Figure S1). The curation of further articles is expected to increase the network size only slightly. For instance, doubling the number of articles by curating 410 additional articles would increase the completeness by just 11 percentage points.

Visualization

The systematic Petri net representation of GRIs in our annotation framework is visualized in schematic flowcharts of the subprocesses of the diauxic shift (Figure 2.5), which we created using the `CellDesigner` software [81].

As exemplarily illustrated for the metabolism of fatty acids in Figure 2.6, the pathway maps are structured by a regulation, transcription and metabolic layer assigned to different cell compartments (cytoplasm, nucleus, peroxisome, and mitochondrium). Thus, the maps are not restricted to the pure illustration of the signals and TFs (regulation layer) regulating the transcription of the mostly enzymatic genes to the corresponding mRNA transcripts (transcription layer), but also visualize the metabolic reactions that are subject to the

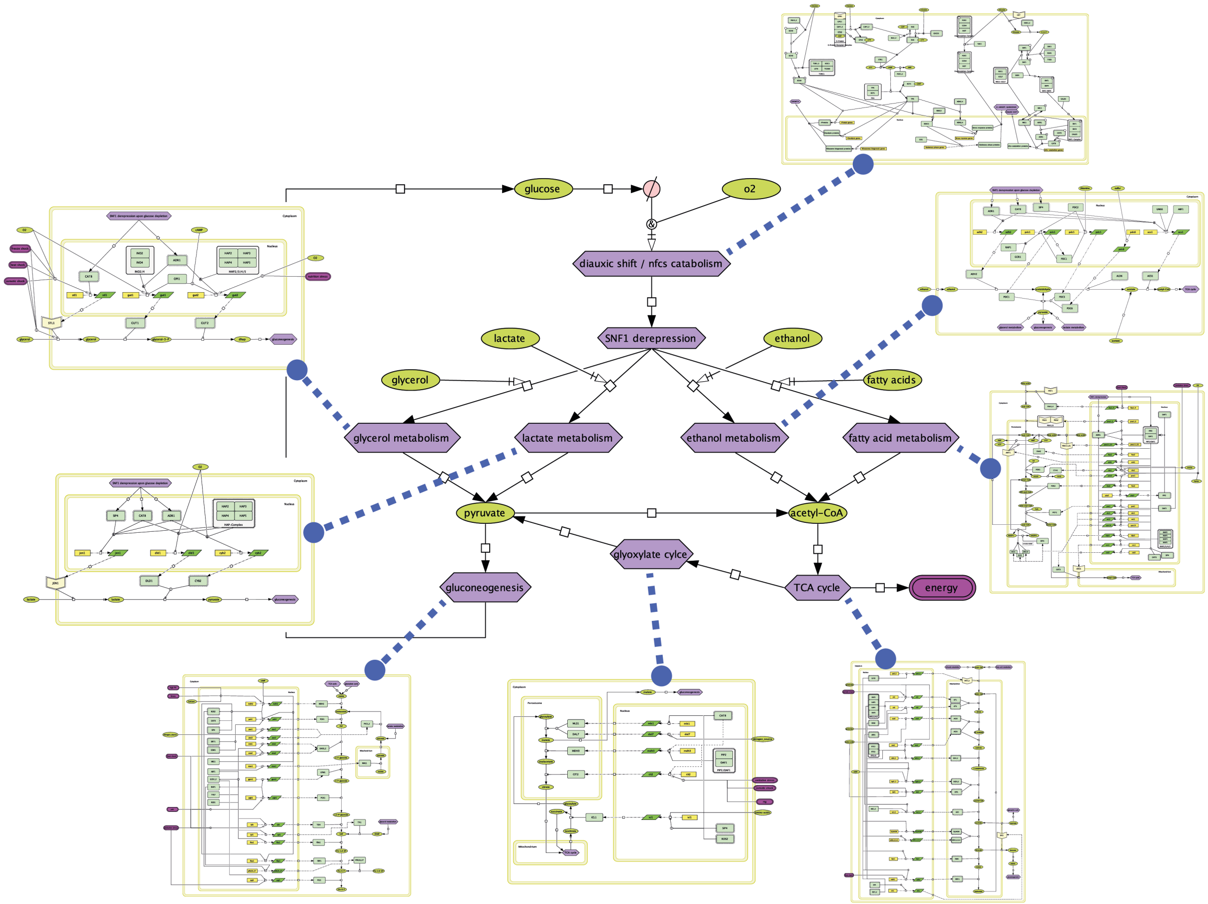


Figure 2.5: The diauxic shift and its subprocesses. On depletion of glucose, yeast switches from fermentation to respiratory growth on non-fermentable carbon sources such as glycerol, lactate, ethanol, and fatty acids. Resulting pyruvate and acetyl-CoA is used to restore glucose and produce energy via gluconeogenesis and the TCA cycle, respectively. As described in the main text, we created pathway maps for each involved subprocess using `CellDesigner` [81]. The maps are organized as exemplarily depicted in Figure 2.6 and each regulation is clickable and connected to the corresponding annotations designed in our annotation system (Figure 2.4), enabling a seamless tracing of the evidence from the schematic representation of a regulation in one of the maps down to the exact place in the curated literature.

transcriptional control.

Each transcriptional transition in the `CellDesigner` maps is clickable and connected to the corresponding annotations, enabling a seamless tracing of the evidence from the schematic representation of a regulation in one of the maps down to the exact place in the curated literature.

The interactive network can be accessed under http://services.bio.ifi.lmu.de/diauxicGRN.

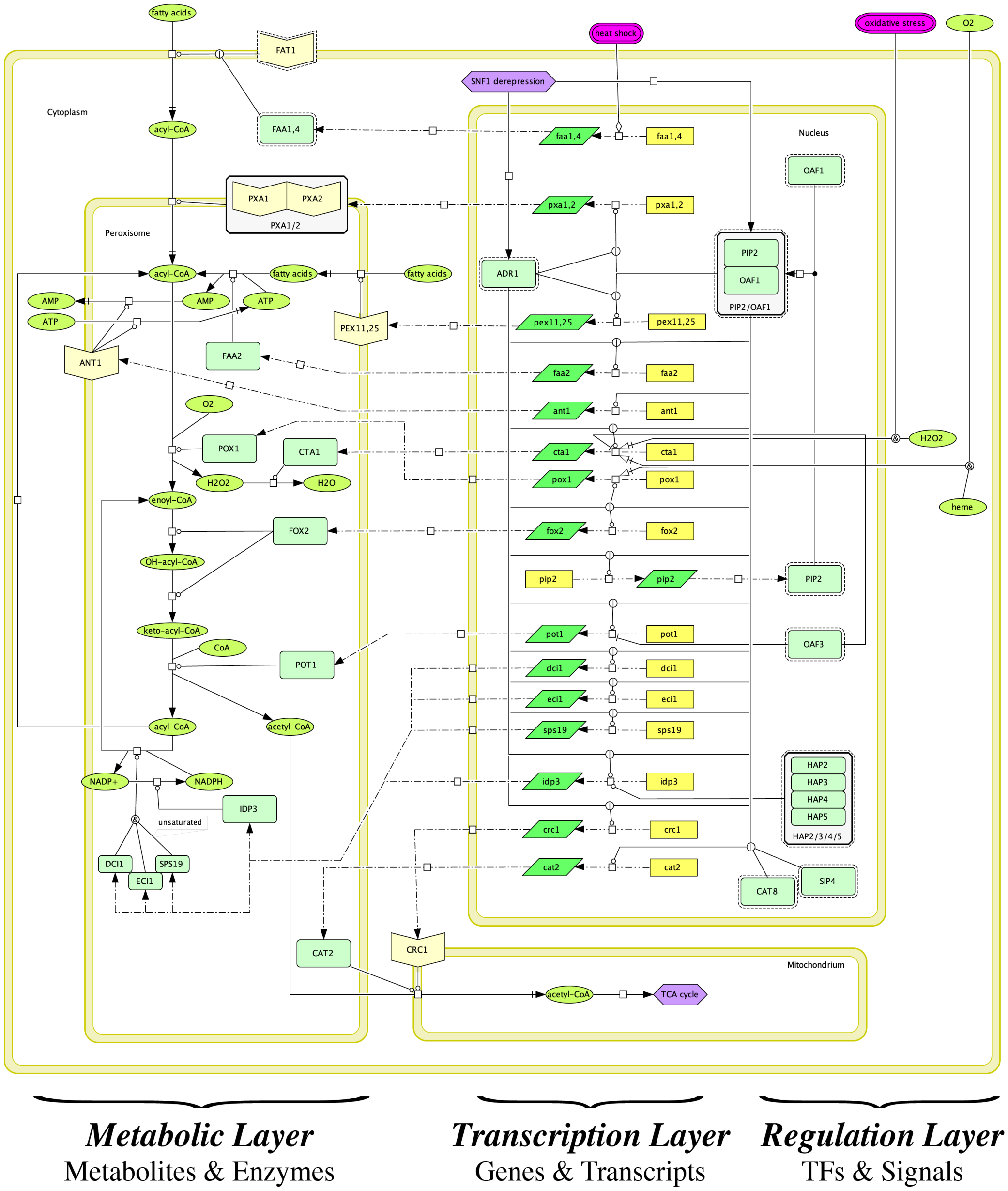


Figure 2.6: Pathway map of fatty acid metabolism. The map is compartmentalized (cytoplasm, peroxisome, mitochondrium, and nucleus) and composed from three layers: the regulation layer on the right, which contains the TFs (light green rectangles) and the signals (green and purple ellipses for metabolites and conditions, respectively) that govern the transcription of genes (yellow rectangles) to their corresponding transcripts (green rhomboids) in the middle. The metabolic layer on the left depicts the translated enzymes (light green rectangles) that catalyze the interconversion of substrates and products (green ellipses), some of which are needed or produced from other subprocesses (blue hexagons) of the diauxic shift.

Comparison to existing resources

General comparison

Existing resources on the transcriptional regulation in *S. cerevisiae* differ considerably in the way how gene regulatory interactions are represented (see *Material and Methods* for an overview). Concentrating on the diauxic shift, we combined and extended the representations in `SGD` [68], `YEASTRACT` [45], and Herrgard *et al.* [69], especially improving on three major aspects:

1. *Context*, determination of the conditions under which a regulation is enabled
2. *Effect*, characterization of regulation type and strength
3. *Evidence*, collection and classification of experimental support

Considering these aspects, the amount of information that the respective resources provided in each step of our curation approach is illustrated in Figure 2.3.

Of the 100 genes classified beforehand as relevant for the diauxic shift (see *Material and Methods*), `SGD` provides regulatory information on 59 genes, Herrgard *et al.* on 80 genes, and `YEASTRACT` on all 100 genes. A resource is defined to provide regulatory information on a gene, if it either has a regulating signal *or* TF assigned.

In `YEASTRACT`, each diauxic shift gene has a number of regulating TFs annotated, yielding in total 1,567 binary interactions (i.e. one-to-one TF-TG associations). Context information, such as extra- or intracellular signals, which turn the regulating TFs active, is not available. However, this is an essential aspect as the transcriptional response of yeast to different environmental conditions varies drastically [39], and most yeast TFs are known to change their activity in dependence on the environmental conditions [61]. In `SGD` and Herrgard *et al.* the fraction of regulations with thorough context definition (signal *and* TF) out of all regulations with signal *or* TF is small (44% and 29%, respectively). In contrast, the regulations in our network have a context annotation in >96% of the cases.

Second, we characterized the regulatory effect in more detail via annotation of the effect type and strength. That means we determined whether a regulation results in a weak, medium or strong activation or inhibition of the affected gene. This feature enables a more fine-grained interpretation and prediction of the expression change of a TG in dependence on the TF activity. Herrgard *et al.* and `SGD` typically provide regulations with an annotated effect type (activation/inhibition), whereas `YEASTRACT` does not distinguish between different interaction types. The semi-quantitative characterization of the effect strength is a novel feature of our network, and little is annotated here in other resources.

Third, we designed a classification to judge how reliable the experimental evidence of a regulatory interaction is. As defined in *Material and Methods*, a regulation with 'high' confidence is given if the corresponding TF has been experimentally determined to bind to the promoter of its TG and the TG is expressed differentially when the TF is perturbed. Our network contains 66% interactions with high confidence, compared with <10% in the other resources.

Concentrating on the diauxic shift genes, our work is based on by far the largest number of articles in which regulations of these genes could be annotated (410 articles, compared with 242, 126, and 85 articles by `YEASTRACT`, `SGD`, and Herrgard *et al.*, respectively). Although this implies that the quantity of curated articles is crucial for a comprehensive characterization, it is also important *which* articles are considered. Interestingly, we observed that the five review articles on transcriptional regulation of the diauxic shift (see *Material and Methods*) provide more regulatory information than `SGD` (see again Figure 2.3).

PCK1 example

Considering the example of PCK1 regulation, a key enzyme of gluconeogenesis, the differences in the three existing resources – with respect of the three aspects *context*, *effect*, and *evidence* elucidated in the previous section – are illustrated in Figure 2.7.

`SGD` notes, besides a variety of biological information on PCK1, putative binding sites for the TFs MIG1, CAT8, MCM1, and the HAP complex. Furthermore, it is stated that glucose represses PCK1 expression, which seems to be mediated by Ras/cAMP signaling.

`YEASTRACT` yields a relatively large number of additional TFs experimentally determined to bind to the PCK1 promoter, and TFs for which PCK1 shows a differential expression in TF mutant versus wild type analyses. Herrgard *et al.* lists that CAT8 and SIP4 activate PCK1. As explained earlier in the text, we extended the current representations of PCK1 regulation as follows.

First, we performed an accurate context assignment. In the PCK1 example, `SGD`, `YEASTRACT`, and Herrgard *et al.* indicate that CAT8 regulates PCK1. However, this regulation takes place only during growth on non-fermentable carbon sources, in particular on ethanol [53] – a crucial context information only included in our work (Figure 2.7c). As CAT8 is inactive under standard conditions (glucose medium), a CAT8 knockout would not influence PCK1 expression at all [64].

Second, we discriminate for all regulations in our network between weak, medium or strong activation and inhibition (correspondingly depicted as `+/++/+++` and `-/--/---` in Figure 2.7c).

Third, we classified the experimental evidence for a regulation to have low or high confidence in order to distinguish biological regulation *in vivo* from ineffective or indirect regulation. In the PCK1 example, all regulatory interactions from `SGD`, `YEASTRACT`, and Herrgard *et al.* have low confidence *per se*. The four putative TF-binding sites in the PCK1 promoter mentioned by `SGD` are not experimentally confirmed by a binding technique like ChIP (see *Material and Methods*). In `YEASTRACT`, one subset of TFs is shown to bind PCK1, but a regulatory effect on expression of PCK1 is not annotated. Vice versa, the subset of TFs annotated to have an expression effect lacks information on binding. Similarly, Herrgard *et al.* cites an expression study for the regulation of PCK1 by CAT8; for regulation by SIP4 no evidence is annotated.

In part, we increased the confidence in these regulations by collecting additional evidence (for CAT8, SIP4, and RDS2). We also identified new regulations with high confidence (ERT1 and GSM1) by directly querying `PUBMED` for regulation of PCK1. On the other hand, we determined regulations that are indirect. For example, the HAP2/3/4/5 activa-

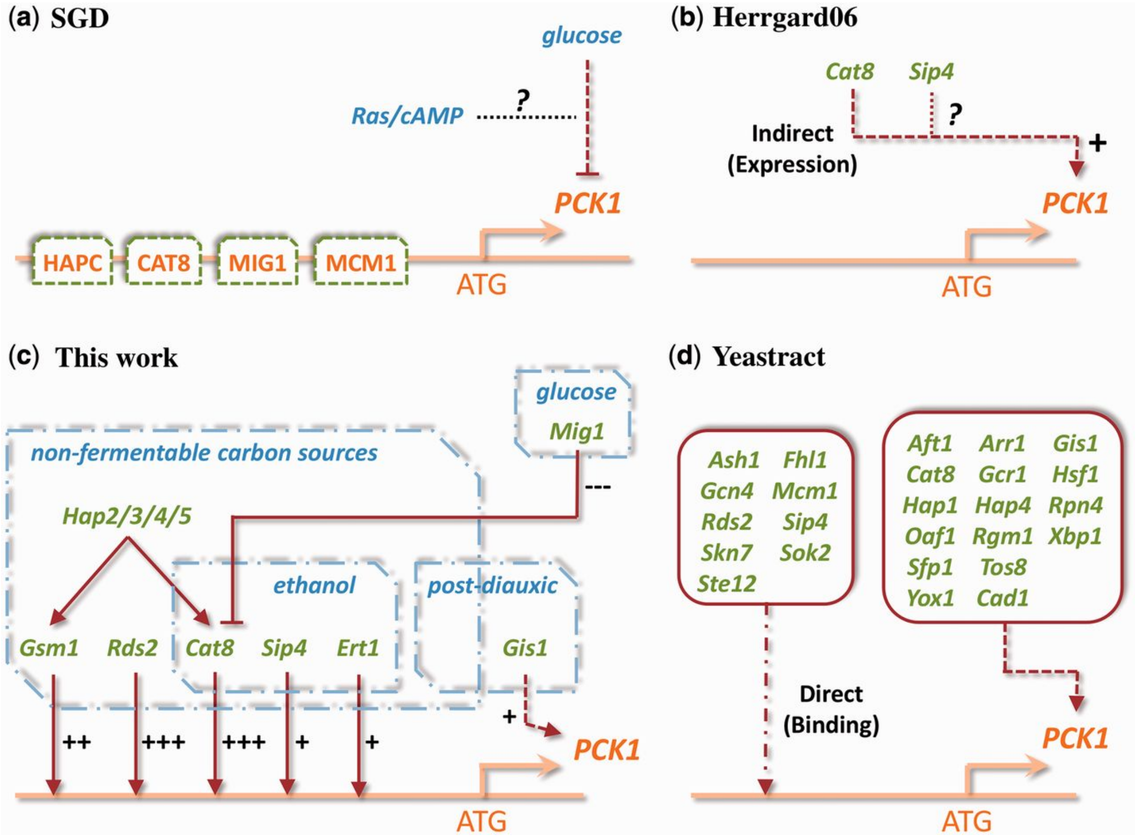


Figure 2.7: Current representations of PCK1 regulation. **(a)** `SGD` states that the PCK1 upstream region contains consensus binding sites for MIG1, the HAP complex, CAT8 and MCM1. Further, that PCK1 is glucose-repressed, which seems to be mediated by Ras/cAMP; **(b)** Herrgard *et al.* [69] state that PCK1 is CAT8/SIP4 activated, with expression evidence for the activation by CAT8, however, no annotated evidence for the activation by SIP4; **(c)** Our work extends current views by accurate context assignment and effect characterization (+/-) and quantification (weak, medium, strong) for each regulation; **(d)** `YEASTRACT` lists a variety of direct and indirect effects, which are not further detailed. In addition, combined evidence is collected for strong confidence regulations (binding & expression; continuous lines) and weak confidence regulations (binding or expression only; dashed lines).

tor complex and the MIG1 glucose repressor act indirectly via regulation of CAT8, rather than by direct regulation of PCK1 [53]. Lastly, we discarded putative regulators not biologically plausible to regulate PCK1, i.e. TFs known to exclusively regulate TGs functionally unrelated to PCK1. These are presumably false positives from high-throughput experiments (e.g. ASH1, GCN4 and STE12).

2.2.4 Discussion

The gene regulatory network of baker's yeast *S. cerevisiae* has been comprehensively studied during the past decades. To provide a machine-readable review of the current diauxic shift knowledge and to investigate how it could be represented to model the regulation of important molecular processes, we addressed the following questions:

1. Do the existing resources already fully characterize the regulation of a given process?
2. If not, how can such a comprehensive characterization be achieved?
3. Which level of granularity is best suited to represent the volume and detail of the available heterogeneous information?

For these questions, we considered different representative resources such as the `SGD` [68] that provides, for one gene at a time, a brief summary of major regulatory impacts such as extra- and intracellular signals. `YEASTRACT` [45], on the other hand, is a repository for binary GRIs (i.e. one-to-one TF-TG associations), mainly derived from published high-throughput TF binding [61] and perturbation experiments [67]. In contrast, Herrgard *et al.* [69] have manually curated the transcriptional regulation of metabolic yeast genes in more detail from the literature, annotating additional features such as the interaction type (activation or inhibition).

As all three resources have a different focus, they thus provide characteristic information on different aspects of gene regulation that we combined to obtain a more complete picture. Thus, we first evaluated to which extent the integration of the heterogeneous resources yields a comprehensive, yet detailed characterization of a process-scale gene regulatory network. As a showcase, we chose the particularly well-studied transcriptional regulation of switching from fermentation to respiration, the diauxic shift in yeast.

Based on current reviews on transcriptional regulation of the diauxic shift, we defined the set of 100 TGs whose gene products perform relevant steps of the shift such as the enzymatic conversion of metabolites. For this gene set, we aimed to retrieve details on their regulation from the three resources. That involves not only the regulators affecting a given TG, but also the conditions under which the TG is affected and whether the gene is activated or inhibited by this relationship. Although a large number of raw binary TF-TG regulatory interactions can be obtained from `YEASTRACT`, their corresponding context information necessary for a detailed understanding of the interaction could only partially annotated using information from `SGD` and Herrgard *et al.* Although each of the three resources cited a large part of the relevant literature as evidence for the regulations, they did not fully exploit the regulatory context information described in the literature. Consequently, thorough manual re-curation of the cited scientific articles, i.e. the full text including tables and figures, was necessary to obtain the activation context of the regulator(s), potential interplay between regulators, the regulation type (activation or inhibition), and the experimental evidence.

We thus dealt with the first two questions by performing a hierarchical curation approach whereby we compiled a comprehensive set of process-relevant genes, extracted and integrated the regulatory information available for these genes from current databases and resources, and finally complemented the obtained regulatory interactions by a thorough manual literature curation.

We estimate that our network, result of an exhaustive databases and literature search, captures >70% of the complete regulatory network affecting genes involved in the diauxic shift. Covering each interaction more than three times on average, we reached a saturation degree that it would, by extrapolation, need twice the number of currently considered articles to achieve 80% completeness.

Efficiently scaling-up from process-specific to organism-wide regulatory networks requires authors and data resources to accurately and uniformly annotate context information when reporting gene regulatory information. Using established machine-readable formats like SBML [82] would then allow a semi-automated processing in which expert intervention and curation is only necessary when compiling regulatory information from conflicting studies.

Addressing the third question, we compiled information on the activation context of TFs and their effect strength on their targets. The latter is often stated in terms of fold changes or discrete quantity changes of the TGs (e.g. "In a yeast strain deleted for ADR1, expression of ADH2 was found to be *strongly* decreased."). Although such semi-quantitative information was abundantly found in the literature, kinetic parameters as required for quantitative modeling with ordinary differential equations (ODEs) were only rarely reported.

We therefore suggested an intermediate representation of GRIs that is beyond current coarse-grained, purely qualitative characterization; on the other hand, of course, it does not match the fine-grained quantitative ODE models.

In such a representation, an interaction between one or more TFs and a TG is characterized in dependence on the activation context of the TFs and by the semi-quantitative effect on corresponding TGs. This seems to strike the balance between striving for a detailed model granularity, and optimally and comprehensively exploiting the available knowledge on the other hand. This also enables a model-based data view, i.e. the model can be tested whether the annotated, and thus expected, behavior of regulations agrees with the observed behavior in a particular dataset of gene expression measurements under investigation.

The suggested representation is exploited in our resulting diauxic shift network, comprising >300 multi-input regulations that also account for combinatorial control by more than one regulator. Available in a machine-readable flat format, it is readily usable in network-based approaches for the interpretation of gene expression data. As a front end, we further provide interactive pathways maps, enabling intuitive exploration of the network modules integrated into our annotation system, where the evidence for each regulation can be entered or retrieved down to the exact reference position in the primary literature. Our system can serve as a starting point to similarly annotate and incorporate additional processes, e.g. all processes subject to glucose control, as the addition of new annotations to existing transitions and pathway maps is straightforward and can be interconnected to the already existing maps.

The system and all accompanying resources are available under
http://services.bio.ifi.lmu.de/diauxicGRN.
Supplementary Figure S1 is available at NAR online.

2.3 Inference from experimental data

This section reviews existing methods for regulatory network inference and shows that they are significantly skewed towards master regulators. A recalibration strategy is introduced and applied, yielding an accurate regulatory network for yeast. The following is based on its original publication in the journal Bioinformatics*, 31(17):2836-43, September 2015.*

Authors: Tobias Petri, Stefan Altmann, Ludwig Geistlinger, Ralf Zimmer, and Robert Küffner

Title: Addressing false discoveries in network inference

Abstract: Experimentally determined gene regulatory networks can be enriched by computational inference from high-throughput expression profiles. However, the prediction of regulatory interactions is severely impaired by indirect and spurious effects, particularly for eukaryotes. Recently published methods report improved predictions by exploiting the *a priori* known targets of a regulator (its local topology) in addition to expression profiles. We find that methods exploiting known targets show an unexpectedly high rate of false discoveries. This leads to inflated performance estimates and the prediction of an excessive number of new interactions for regulators with many known targets. These issues are hidden from common evaluation and cross-validation setups, which is due to Simpson's paradox. We suggest a confidence score recalibration method (*CoRe*) that reduces the false discovery rate and enables a reliable performance estimation. *CoRe* considerably improves the results of network inference methods that exploit known targets. Predictions then display the biological process specificity of regulators more correctly and enable the inference of accurate genome-wide regulatory networks in eukaryotes. For yeast, we propose a network with more than 22,000 confident interactions. We point out that machine learning approaches outside of the area of network inference may be affected as well.

Author contributions: TP, SA, LG and RK compiled the data and conducted experiments. TP and RK developed the correction method and modular visualization of the yeast network. Result network properties were analyzed and evaluated by RK and TP. LG provided a functional interpretation and analysis of the network. TP, LG and RK wrote the paper with suggestions from RZ. RK and RZ supervised the project.

2.3.1 Introduction

Gene regulatory networks (GRNs) consist of interactions of regulators such as transcription factors (TFs) that physically bind to specific nucleotide sequences to regulate the expression of target genes. GRNs can be experimentally derived from TF-binding studies [83] such as Chromatin Immuno-Precipitation (ChIP, [84]) or DNase footprinting [85]. A large fraction of the interactions reported by these approaches are not associated with changes in target expression [86]. On the other hand, expression changes in potential TF targets can be detected from TF-knockout profiles [64]. This approach, however, is prone to indirect or spurious effects [67].

Although the number of conducted TF-binding and TF-knockout studies is growing [87] the discovery of novel regulations detected with each additional study decreases. Thus, a combination of experimental results and computational inference approaches is likely to provide more comprehensive networks.

Many inference methods use expression data exclusively. An interaction is predicted if a TF and its putative target are coexpressed. Such *expression-based* approaches infer prokaryotic networks successfully [88–91]. However, they perform hardly better than random for inference of eukaryotic networks [35, 67, 90, 92–95], although they can achieve useful results in special cases (*e.g.* for respiratory genes, [90]). Interactions in eukaryotes are difficult to infer as observable dependencies between the expression of regulator and target are weaker and context-dependent. One reason is the increased level of complexity and the combinatorial nature of the eukaryotic regulation of transcription [85].

The prediction of novel interactions can be improved for prokaryotic and *in-silico* data by exploiting *a priori* known interactions (local topology priors, [91]). This allows to determine whether a given TF is active based on the expression of its known targets [96, 97] enabling a more reliable prediction of novel targets [98–100].

Here, we investigate whether eukaryotic networks are accurately inferred by methods exploiting topology priors. First, we demonstrate that many existing performance evaluations are misleading. They are not adequate for local topology methods and overestimate network quality substantially. This effect is due to Simpson's Paradox, well-known in causal theory [101, 102]. Second, this also strongly influences the quality and composition of inferred networks. We develop a simple recalibration strategy and demonstrate how it can be applied for the inference of a confident genome-scale regulatory network in yeast.

2.3.2 Material and Methods

Network inference methods score all pairs of regulators and putative target genes to quantify the confidence that a given pair represents a true interaction. For both types of inference methods discussed here, namely expression-based methods and local topology methods, confident predictions are selected by applying a unified cutoff. Expression-based methods are based exclusively on expression data and ignore known interactions. Local topology methods use expression data and known interactions (topology priors) to train a so-called local model per regulator (Figure 2.8).

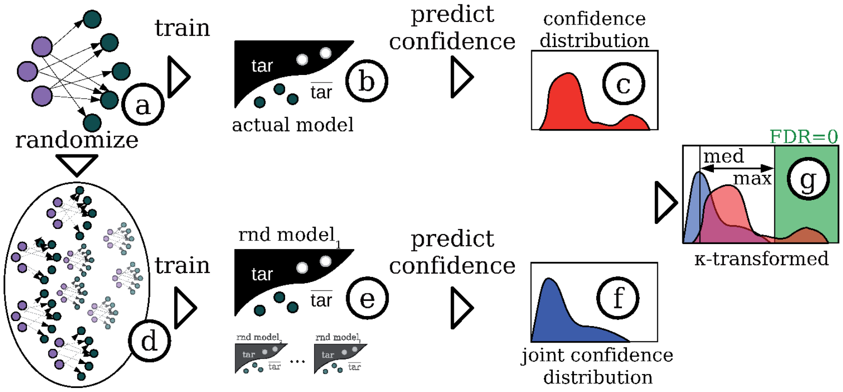


Figure 2.8: Outline of the recalibration approach. Based on the known network **(a)**, a regulator-specific model **(b)** is trained to predict potential targets for this regulator. This results in a confidence score distribution for each regulator **(c)**. Additionally, we generate random networks **(d)** maintaining in- and out-degrees from the original network and train models **(e)** for each random topology in the same way as for the original network. For each TF out-degree, we combine resulting random confidence scores into a joint distribution **(f)**. Finally, we compare the two distributions c and f based on their respective medians (med) and maxima (max). We minimize false discoveries by selecting regulations (shaded area in **(g)**) that exceed values observed for random networks.

Data

We obtained five yeast expression compendia from (1) the 5th DREAM Challenge (challenge 4, [35]), (2) the Many Microbe Microarray Database (`M3D`, [103]), (3) the study of Hu *et al.* [67], (4) the study of Chua *et al.* [64] and (5) the Gene Expression Omnibus (`GEO`, [104]). Case-control pairs were selected from 2442 yeast microarrays as described by Küffner *et al.* [93] to compute $\log_2$ fold-changes. Thereby, we obtained a matrix $M \in \mathbb{R}^{p \times n}$ with $p = 1829$ microarray pairs and $n = 5402$ genes. We normalize M by two successive z-score transformations of rows and columns, respectively.
We then collected experimentally supported interactions from the `YEASTRACT` database [45], augmented by a study of MacIsaac *et al.* [105]. We filtered genes that were not contained in the expression data. We excluded TFs regulating less than 6 known targets to enable training and cross-validation. The resulting reference standard contains 153 TFs, 4870 target genes and 24,462 interactions derived from 356 TF-target binding assays.

Training and assessment of local topologies

A regulatory interaction network of n genes G is a directed graph $N = (G, I)$, $G = R \cup T$ where R is the set of regulators, T is the set of targets and $I \subseteq R \times T$ are regulatory interactions. Each instance of I is a regulator-target pair $(r, t) \in R \times T$ that is labeled with a weight w_{rt} denoting the number of TF-binding studies that support interaction (r, t).

Machine learning models are trained to predict novel regulations $(r,t) \in R \times T$. Based on the known interactions, each putative regulation is labeled by l_{rt}, where l_{rt} is 1 if $w_{rt} \geq 1$ and 0 otherwise. The matrix of fold-changes $M = (m_{ij}) \in \mathbb{R}^{p \times n}$ represents the feature vectors. The value m_{ij} is the fold-change for gene $j \in G$ in array pair i and we denote row i by $M_{i.}$ and column j by $M_{.j}$. Then, the feature vector for (r,t) is given by $M_{.t}$.
We train $|R|$ local models, each predicting confidence estimates $\hat{c}_{rt}$ specific for a single regulator r of putative regulations (r,t):

$$s_r : \mathbb{R}^p \to \mathbb{R},\ s_r(M_{.t}) = \hat{c}_{rt}. \tag{2.1}$$

Alternatively, a single *global* model is trained for all regulators using combined feature vectors, *i.e.* feature vectors of regulator and target are concatenated to represent an interaction

$$s : \mathbb{R}^{p+p} \to \mathbb{R},\ s(M_{.r} \oplus M_{.t}) = \hat{c}_{rt}. \tag{2.2}$$

From all regulations, we build k splits for each model stratified with respect to their label distribution. Cross-validation (here: 3-CV) is performed by retaining one split at a time and training a model on the remaining $k-1$ splits, so that interactions are either used in evaluation or training, but not both. Every split results in $|R| * |T|$ confidence values $\hat{c}_{rt}$ that score all regulations $(r,t) \in R \times T$. For regulator r we denote the distribution of these confidence values as D_r (Figure 2.8b and c).
The quality of inferred networks is assessed after integrating the predictions across all regulators. Assessment compares predictions to a reference standard of *a priori* known interactions, for instance by the area under the receiver operator characteristics curve (AUC). An AUC of 1.0 indicates that the confidence scores for the true interactions are higher than those for false positives, while an AUC of 0.5 would indicate random predictions. Such a cross-validated AUC analysis is a standard approach for the assessment of inference methods [99].

Confidence recalibration (*CoRe*)

Randomized topologies are generated to share key statistics with the reference standard of known interactions (Figure 2.8a and d). We remove all regulations from the network and randomly introduce new regulations until each node k has regained its original in- and out-degree (compare [106], p.12). Further, the association of expression data and genes is shuffled by gene label permutation. For each of the q randomized networks $N^{(1)}, \ldots, N^{(q)}$ we perform a CV prediction to obtain confidence values $\hat{c}_{rt}^{(i)}$ as described above (Figure 2.8).
Let $D_r^{(i)}$ the distribution of confidence values specific to a regulator r computed using the random prior $N^{(i)}$. We then compute a joint distribution D'_r that encompasses all confidence values derived from random networks that are associated to regulators of the same out-degree (Figure 2.8f).
D'_r denotes the *randomized complement* of D_r. By comparing these two distributions we

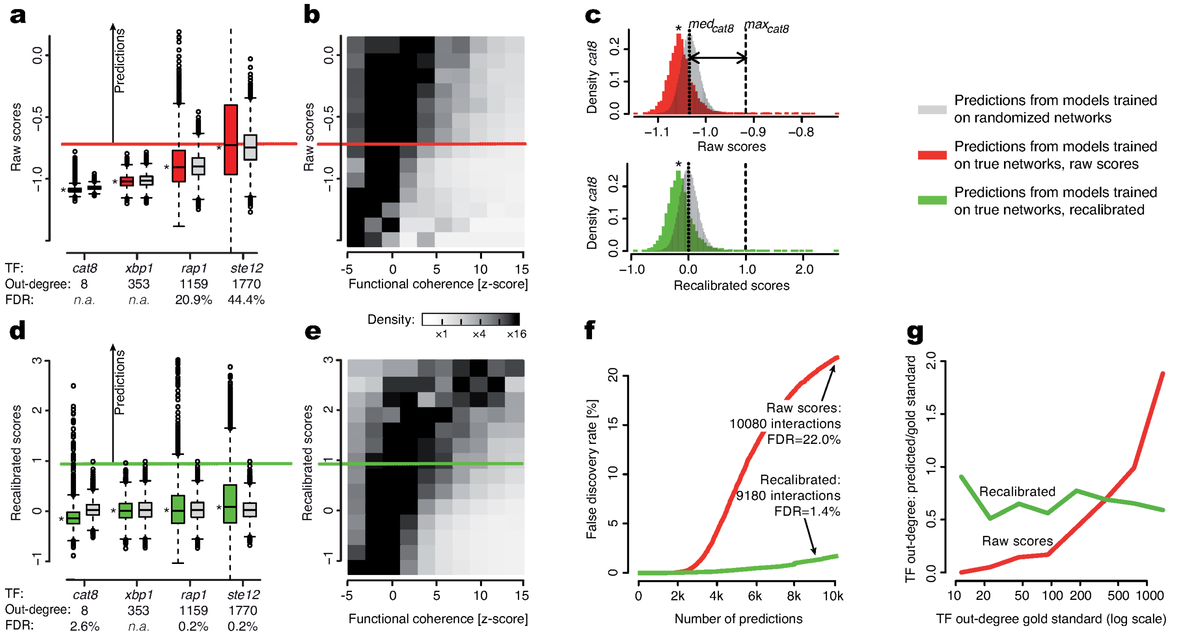


Figure 2.9: Score recalibration in network predictions. (a) We trained SVMs for each TF. Putative target genes were selected by a threshold (horizontal line) on the resulting TF-specific scores (boxplots marked by asterisks). Additional SVMs were trained on random networks (light gray) and FDRs were computed for all regulators but those such as *xbp1* where no predictions were made. **(b)** The density map displays whether predicted and known targets of the same TF overlap in their biological function. Positive z-scores (abscissa) indicate significant function overlaps for corresponding scores (ordinate). **(c)** Score distributions (marked by asterisks) were recalibrated via randomized distributions (light gray): for each TF, the median *med* (dotted line) and maximum *max* (dashed line) are mapped to 0.0 and 1.0, respectively. **(d)** and **(e)** show boxplots, threshold, and a density map of function overlap after recalibration. **(f)** plots the FDR as a function of the number of predicted interactions. Arrows indicate the number of interactions achieving a precision of at least 50% (P50). In **(g)**, the ratio of predicted to gold standard targets (ordinate) is depicted across the range of TF out-degrees (=number of gold standard targets, abscissa) before and after recalibration.

select interactions with scores higher than those observed in the randomized case. Each regulation's confidence $\hat{c}_{rt}$ is replaced by its complement κ_{rt} (Figure 2.8c and g):

$$\kappa_{rt} = \frac{\hat{c}_{rt} - med(D'_r)}{\max(D'_r) - med(D'_r)}. \tag{2.3}$$

Scores are thus recalibrated based on the median confidence $med(D'_r)$ and the distribution scale ($\max(D'_r) - med(D'_r)$). A κ value above 1.0 corresponds to a false discovery rate (FDR) of 0, *i.e.* to confidence estimates not achieved in random topologies.

2.3.3 Results

Simpson's paradox

We followed the SIRENE approach [99] and trained local models based on Support Vector Machines (SVMs) to predict confidence values for potential regulations. On a large expression data set of 2,442 yeast microarrays and a regulatory network of 24,462 interactions (Section 2.3.2) the cross-validated predictions achieved a network-wide AUC of 0.784.

However, we found this standard, cross-validated AUC analysis misleading in case of methods integrating topology priors. We demonstrated this by training the methods on randomized networks (random re-assignment of targets to regulators). The confidence scores for individual regulators are random, resulting in regulator-specific AUC values of 0.5. Strikingly, an evaluation across all regulators yielded an AUC of 0.798, a score above the AUC achieved by SIRENE.

These two results seem to be in conflict: a method that performs randomly for each regulator induced subnetwork should yield random overall performance as well. This effect resembles the Simpson's or "amalgamation" paradox [101, 102]: each of the regulator-specific distributions achieves an AUC of 0.5, while the AUC of the joint distribution suggests non-random performance.

Here, the paradox results from the fact that predicted confidence score distributions are heterogeneous across regulators and are characterized by different scale and location parameters (Figure 2.9a, light gray boxes). In particular, score distributions for regulators with many known targets (high out-degree) such as *ste12* are wider and systematically above average. We refer to this effect as *High Degree Preference* (*HDP*). These regulators contribute many true positives, *i.e.* after the integration higher scores become enriched for true positives. This in turn leads to non-random AUC values. Selected high-scoring predictions thus remain unspecific while biologically more specific signals are likely being missed [107]. Following this line of argument, the regulator out-degree confounds the integration of confidence values. This is consistent with results demonstrated for the prediction of genes involved in biological processes [108].

To examine whether the paradox is an artifact of SVMs we trained further model classes (*e.g.* decision trees and logistic regression). We observed similar effects across all examined techniques, suggesting that regulator-specific methods using topology priors are generally affected by an *HDP*.

Besides the confounding of network quality measures, the composition of predicted networks is also affected. We predicted networks by selecting high-scoring interactions using a threshold determined from the estimated size of the complete yeast network, which should be twice as large as the known network. A score threshold was chosen so that selected regulations contain 50% previously confirmed ones (the Precision-50, or P50 network).

For a regulator with out-degree d we obtained two types of score distributions: (i) from the model trained on its known targets and (ii) from models trained on the targets of randomized regulators with out-degree d (Figure 2.9a). A unified cutoff selects an excessive number of predictions for high-degree TFs that overlap with random scores. To quantify this, we computed the false-discovery rate (FDR) based on the number of interactions scored above the P50 threshold in distribution (ii) divided by the total number of interactions above that threshold in (i) and (ii). For example, the FDR is 44.4% for high-degree *ste12* and 22% across all TFs (Figure 2.9f), which is unacceptably high. In contrast to *ste12*, all predictions are rejected in case of low-degree TFs such as *cat8*, even if they substantially exceed random scores (Figure 2.9a). Only 81 of 153 TFs (53%) receive predictions. We concluded that neither cross-validation nor AUC analysis are sufficient to ensure the overall quality of networks inferred using structural priors.

We also assessed whether TFs frequently regulate the expression of targets that share similar biological functions [109]. We therefore tested whether known and predicted targets of the same TF exhibit substantial functional overlaps. We observed that the high proportion of random scores (*e.g.* for *ste12*) concealed most of the signal as interactions with higher scores hardly showed an increased functional coherence (Figure 2.9b).

Correction through score recalibration

We introduce a confidence recalibration (*CoRe*) as a wrapper for existing methods (Section 2.3.2). Based on the random networks, we derived expected location (median score) and scale (maximum score) properties for each out-degree d and used them to transform the predicted confidences into topology-corrected scores. Scores for each regulator are recalibrated by scaling the median and maximum scores to 0 and 1, respectively (Figure 2.9c). This renders score distributions comparable so that they can be integrated across TFs. The FDR is then 0 for predictions with scores above 1 as they appear only for the true but not for the randomized networks. Thus, interactions for each regulator selected after *CoRe* are scored above the random level.

To obtain a P50 network, we select interactions that achieve a corrected score of >0.92. The FDR for this network was reduced to 1.4% (as compared to 22.0% without recalibration). We observed that predictions are now balanced across TF degrees (Figure 2.9g), predicting interactions for 138 TFs *vs.* 81 without recalibration.

To gain further insight in the nature of the corrected network, we estimated the functional relationship between known and novel predicted targets. Regulatory patterns were more coherent for the corrected network (compare Figure 2.9b and e).

Application of *CoRe* to network inference

For all subsequent methods and analyses we report corrected results. To evaluate the yeast regulatory network obtained, we conducted a comparative assessment of frequently used inference approaches and a consensus approach. The approaches are roughly classified by five attributes (Figure 2.10a):

1. **method**: unsupervised expression-based [89] *vs.* supervised using a structure prior;
2. **formulation:** one-class [110] *vs.* two-class that treat unknown interactions as informative;

3. **strategy:** lazy *vs.* parameterized models;

4. **data handling:** non-integrative *vs.* integrative *e.g.* using TF binding site preferences [111];

5. **models:** global [112] *vs.* local (regulator-specific)

SIRENE [99] is a supervised, two-class, parameterized, non-integrative, local approach. For all methods, we predicted confidence scores in a 3-CV scheme and recalibrated them as described above.

Subsequently, we analyzed network motifs to capture method- and topology-specific preferences (Figure 2.10b). Unsupervised, expression-based approaches do not use topology priors but infer interactions if expression profiles of TFs and putative targets are mutually dependent. An example is CLR [89]. These methods are unable to detect auto-regulation as in this case both expression profiles would be identical. Confirming previous findings [35], expression-based approaches could hardly detect feed-forward motifs or the correct direction of interactions. In contrast, regulator-specific approaches were less affected by such difficult cases and exhibited a consistently higher performance. For cascades and low in-degree targets, a slight decrease in performance was observed. Potentially, the latter indicated the prediction of novel regulators for genes that were less well studied previously.

Next, we evaluated the performance of approaches across all interactions. Expression-based, one-class, and lazy learners performed substantially worse than the remaining methods (Figure 2.10c). We observed that integrative methods like Serend [111] suffered from false positive predictions. This is likely due to the low specificity of positional weight matrices (PWMs, [113]) predicting targets only for 6.5% of all TFs. These methods were not further analyzed. Of the remaining five methods (methods 5-12 in Figure 2.10), the best results were obtained from regulator-specific SVMs and decision trees trained on bootstrap samples (bagging).

A comprehensive yeast network

Our final yeast network includes $22,231$ interactions with 153 TFs and $3,747$ target genes. Of all predicted regulations, $12,869$ are contained in the reference standard while $9,362$ are novel predictions. The remaining $24,462 - 12,869 = 11,593$ reference standard interactions lacked an observable effect on expression and were thus not included.

The visualization and interpretation of organism-wide networks is challenging due to their size and complexity. Instead of fully depicting each regulator, target and their interactions, we employed a modular visualization. We derived regulatory modules by grouping TFs with overlapping target sets and, vice versa, target modules by grouping genes regulated by overlapping sets of TFs. We connected regulator and target modules via *meta-interactions* if more than 40% of all induced regulator-target pairs were connected. This reduced representation featured 13 meta-interactions among 9 target and 9 regulatory modules, capturing half of the final interactions ($11,232$ interactions, 50.5% of all predicted). See Figure

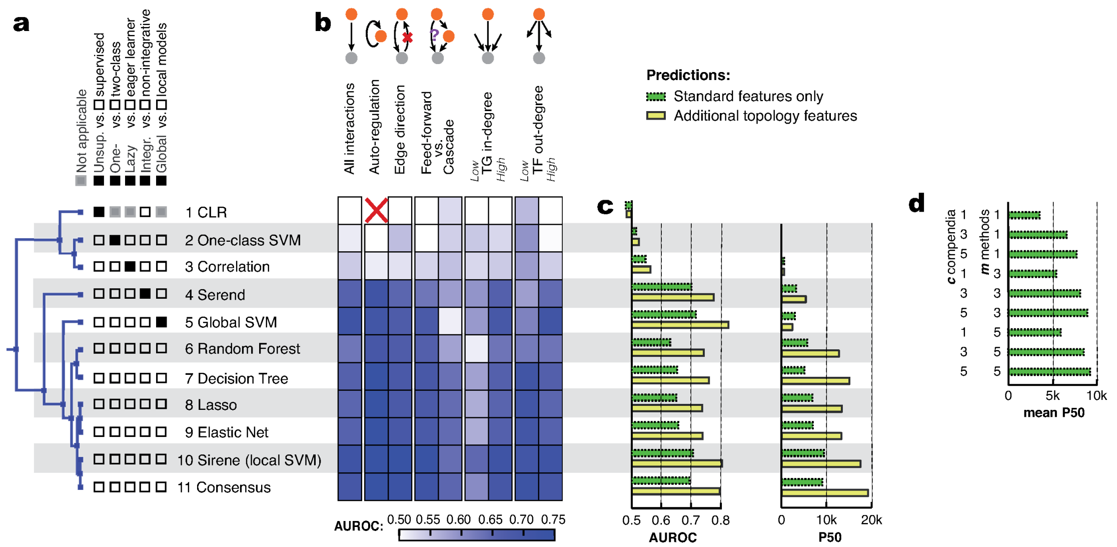


Figure 2.10: Assessment of predicted interactions. We analyzed the predictions of 11 different inference methods across five yeast gene expression compendia. **(a)** The dendrogram groups methods according to the similarity of their predictions. Properties that discriminate between different classes of methods are indicated by the check boxes. In **(b)** shows if interactions in particular network motifs are easier (dark) or harder (light) to detect in comparison to all interactions. **(c)** assesses method performance (AUC) and the number of interactions predicted at a precision of 50% or better (P50) (bars with dotted borders). Furthermore, we encoded experimentally determined targets of TFs into additional features (bars with solid borders) and integrated methods 6-10 into a consensus approach (method 11). **(d)** illustrates mean results from integrating all subsets of $c = 1..5$ compendia and $m = 1..5$ methods. All results are based on recalibrated scores.

2.11 for an excerpt (full details can be found in the supplement, available online and accessible through clickable maps).

This modular view enables an integrated display of the network as well as module-associated expression profiles. Given current data and knowledge, the respective TF-modules likely control the forming of transcriptional response patterns in the regulated target modules. Some key aspects of module-associated expression profiles are summarized below (for a comprehensive literature review on all network modules see online supplement). A representative gene was selected manually for each module.

The *hxt2* module features the most versatile regulation in our network, regulated by three different TF clusters comprising the highest total number of TFs (Figure 2.11). According to `GO` [114], most of the 190 genes of the *hxt2* cluster belong either to sugar transport (*hxt* genes) or glycogen metabolic process (*e.g. gac1*). Consequently, we observe differential expression of these genes under low *vs.* high-glucose growth conditions. When glucose is available, the sugar transporters are abundantly expressed [115], whereas under glucose starvation glycogen storage is catabolized to produce glucose preferably for fermentation

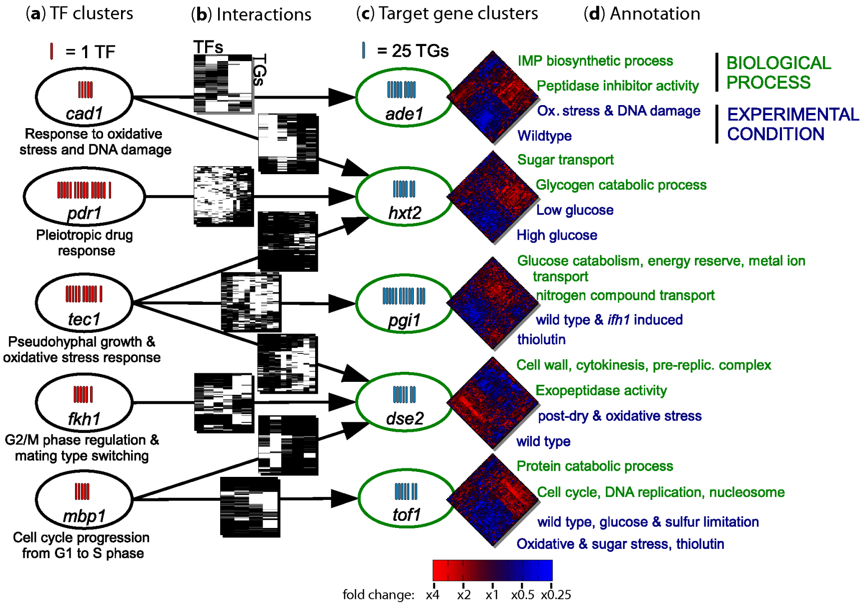

(a) TF clusters
(b) Interactions
(c) Target gene clusters
(d) Annotation
= 1 TF
TFs
TGs
= 25 TGs
cad1
Response to oxidative stress and DNA damage
pdr1
Pleiotropic drug response
tec1
Pseudohyphal growth & oxidative stress response
fkh1
G2/M phase regulation & mating type switching
mbp1
Cell cycle progression from G1 to S phase
ade1
hxt2
pgi1
dse2
tof1
IMP biosynthetic process
Peptidase inhibitor activity
Ox. stress & DNA damage
Wildtype
BIOLOGICAL PROCESS
EXPERIMENTAL CONDITION
Sugar transport
Glycogen catabolic process
Low glucose
High glucose
Glucose catabolism, energy reserve, metal ion transport
nitrogen compound transport
wild type & ifh1 induced
thiolutin
Cell wall, cytokinesis, pre-replic. complex
Exopeptidase activity
post-dry & oxidative stress
wild type
Protein catabolic process
Cell cycle, DNA replication, nucleosome
wild type, glucose & sulfur limitation
Oxidative & sugar stress, thiolutin
fold change: x4 x2 x1 x0.5 x0.25

Figure 2.11: Interactions and expression profiles. We partitioned our network of 22,231 gene regulatory interactions for visualization and identification of network modules. We derived **(a)** 9 clusters of 61 TFs that, via **(b)** 13 interactions between clusters (arrows), regulate **(c)** 9 clusters of 1758 target genes (excerpt of 5 TF and 5 target clusters connected by 9 interactions shown here, see online supplement for full figure). A representative gene is displayed for each TF and target cluster. Cluster interaction maps (black=interaction, white=no interaction) comprise a total of 11,232 (50.5%) interactions. **(d)** Thus, depicted TF-modules are likely to trigger expression responses (heatmaps: red=up-, blue=down-regulation, see online document for colors) in respective target modules and associated processes (top part of heatmap). The heatmaps display the differential expression of these target modules under the indicated knockout (KO) and other experimental conditions (bottom part).

[116].

The *pdr1* (pleiotropic drug response) cluster comprised the largest number of *hxt2* regulators. It consisted of 16 TFs, all tightly connected to the cellular response to drug and nutrition stress such as differing glucose concentrations. Despite this general response mediated by the *pdr* TFs (*stb5* and *msn1*), much of the regulation was performed by pseudohyphal growth TFs (*nrg1*, *mga1*, and *ash1*) in conditions of nitrogen limitation and abundant fermentable carbon sources like glucose [117].

Interestingly, a strong regulatory impact on the *hxt2* module was also observed for regulators of the oxidative stress response - on the one hand from the *cad1* cluster (5 TFs, also responding to resulting DNA damage), and, on the other hand, from the *tec1* cluster (11 TFs, also driving pseudohyphal growth). Oxidative stress results in cellular protection mechanisms, *e.g.* DNA repair and targeted protein degradation, which is associated with increased energy consumption [118], initiated by the *hxt2* cluster via increased glucose uptake.

2.3.4 Discussion

Gene regulatory networks are crucial to understand how regulators like transcription factors affect their target genes on the expression level. Experimentally derived networks are typically incomplete as the number of available experiments is limited. To complement them, computational inference of networks has been introduced. We revealed critical aspects but also demonstrated that inference is necessary and feasible in eukaryotes.

Even in well-studied eukaryotes such as yeast, where ∼900 publications on experimental TF-binding studies are available, current networks are far from complete and benefit from computational predictions. We found that only about half of all regulations that induce detectable expression changes ("active" interactions) are currently known. In addition, experimental techniques are prone to discover regulations without effect on the expression level. We applied computational inference both for the detection of novel active and the pruning of inactive regulations.

We reported three crucial findings based on the analysis of a wide spectrum of data-driven inference methods (for reviews see [100, 119]). First, we demonstrated that methods incorporating experimentally derived interactions as topology priors possess sufficient predictive power for the inference of eukaryotic networks. Methods using expression data alone fail here [35, 95]. We also showed that topology priors lead to Simpson's paradox [101, 102] distorting prediction and assessment of regulatory interactions. Finally, we showed how to avoid the occurrence of the paradox.

Generally, network inference methods that exploit the local topology assign an excessive number of predictions to TFs with many known targets [100, 120], and it has been doubted whether a correction is possible or sensible [108, 119]. Our analysis revealed that the number of known targets for a regulator is a confounder of regulator-target predictions. This effect is not detected by common cross-validation routines: surprisingly, the same performance reported for published network inference approaches can be achieved by guessing random regulations. We developed a confidence recalibration approach (*CoRe*) wrapping existing methods and showed that it corrected for both the over-estimation of performance and the distortion of the topology towards TFs with many known targets (*High Degree Preference*, *HDP*).

We conducted a comprehensive assessment of methods integrating topology priors and identified methods suitable to derive a corrected, accurate yeast regulatory network of active regulations. We describe disadvantages of several methods, which we excluded due to prediction performance, or the inadequate scale-up for large expression datasets. Our evaluation suggested that the selected methods detect several types of interactions successfully that are difficult to predict. For instance, auto-regulatory interactions and the assignment of directions are handled accurately, and immediate and indirect interactions could be distinguished. We then integrated the predictions from the selected methods to construct a network consisting of half novel and half experimentally-determined regulations. This choice was based on our extrapolation of the size of the complete yeast network.

Our final yeast network contains 153 TFs that regulate 3,747 target genes via 22,231 interactions. These include many novel and confident hypotheses of regulatory relationships, while we expect less than 150 false positives in total. At the same time, we reject more than half of the experimentally-determined interactions as they appear to be without observable regulatory effect.

To gain an overview of the network, we derived modules of target genes that were jointly regulated by sets of TFs. The resulting modular structure was strikingly simple featuring 13 meta-regulations that represent an index for inspecting the expression effects of interactions. A thorough literature review confirmed that the modules and their expression patterns correspond well to biological processes such as respiration, sulfate/energy metabolism, transport, stress response and cell division.

We conclude that methods integrating local topology can extend known networks substantially and at a high reliability, even in well-studied model organisms. These methods, in contrast to those using expression data alone, are well-suited for the prediction of interactions in yeast and presumably other eukaryotes. Due to Simpson's paradox however, their application was more difficult than previously acknowledged and required a correction approach. We emphasize that topology, structural priors and parameterized models are widely applied beyond network inference and encourage a review of fields that may benefit from confidence recalibration strategies such as *CoRe*.

Chapter 3

Detection of significant subnetworks

3.1 Overview

The previous chapter describes several approaches to construct a regulatory network. When using such a network for the analysis of gene expression data, the primary research question is whether certain parts of the network are significantly altered for the investigated phenotype (Figure 3.1). Such subnetworks typically correspond to functional modules that represent known or novel interplay between genes, which in turn allows conclusions about molecular mechanisms underlying e.g. a certain disease [29].

As introduced in Chapter 1, the standard approach to detect such subnetworks is to test predefined functional gene sets for overrepresentation of differentially expressed genes [26, 27]. The method assumes that all measured genes have been classified as differentially expressed or not, and is based on a simple and intuitive principle. For each predefined gene set G that represents a known functional module:

1. count how many of its genes m_G are differentially expressed (yielding m_{GD} genes, see contingency table 3.1),
2. estimate the statistical significance of observing m_{GD} genes based on the hypergeometric distribution (Fisher's exact test, Appendix B.1.2).

This yields for each gene set a p-value expressing the probability of observing at least m_{GD} genes.

Table 3.1: The 2 × 2 contingency table for assessing overrepresentation of differentially expressed (DE) genes in a gene set G.

	DE	not DE	Total
In gene set G	m_{GD}	$m_{G\overline{D}}$	m_G
Not in G	$m_{\overline{G}D}$	$m_{\overline{G}\overline{D}}$	$m_{\overline{G}}$
Total	m_D	$m_{\overline{D}}$	m

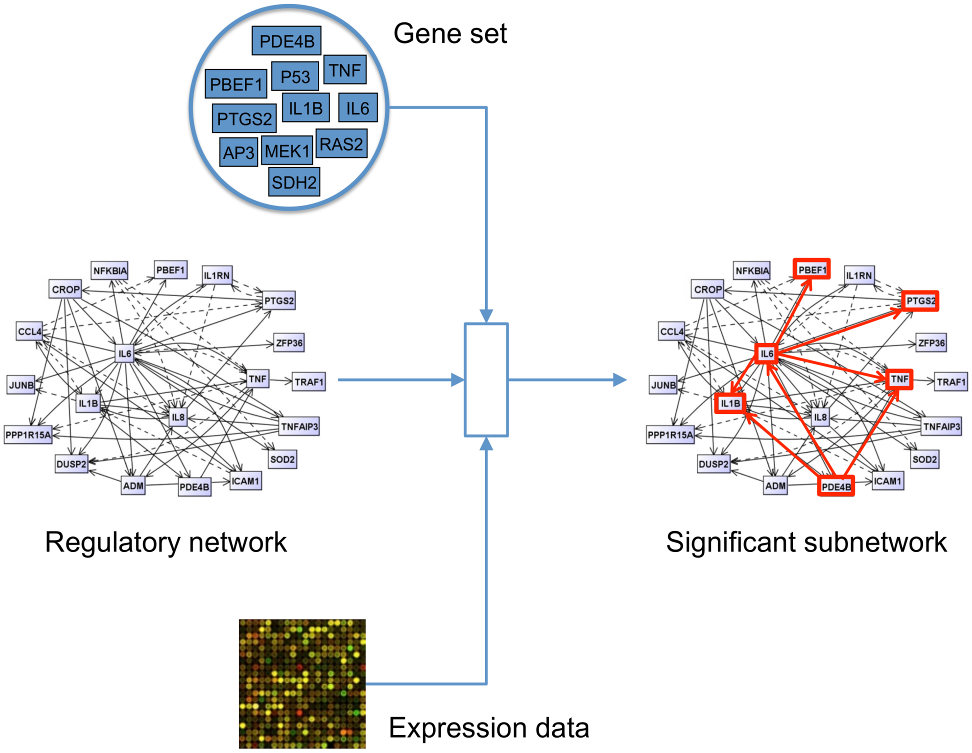


Figure 3.1: Detection of subnetworks that are significantly altered in expression and enriched for specific functional gene sets.

Although the overrepresentation test is fast and effective, it has two major drawbacks. On the one hand, the classification of genes as differentially expressed or not is usually the result of a threshold applied to the chosen statistic for differential expression. The often much smaller subset of differentially expressed genes is thus artificially separated from the remaining large number of genes not satisfying the threshold and which are therefore not further analyzed.

On the other hand, the actual search for sub-*networks* is reduced to a search for sub-*sets*, i.e. genes are treated as independently expressed and interactions between genes are not taken into account.

In the following, I discuss both issues in more detail and introduce, based on recent publications in the field, *Gene Set Enrichment Analysis* to resolve the restriction to differentially expressed genes, and *Gene Graph Enrichment Analysis* to also properly account for the network topology (Section 3.2 and 3.3). The interpretation of the resulting subnetworks is subject of Chapter 4.

3.2 Gene Graph Enrichment Analysis (GGEA)

This section describes a method for network-based enrichment analysis. The key idea is to score regulatory interactions within functional gene sets for consistency with the observed expression. The work has been presented on July 18th, 2011, at the 19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology. The following is based on its original publication in the journal Bioinformatics*, 27(13):i366-73, July 2011.*

Authors: Ludwig Geistlinger, Gergely Csaba, Robert Küffner, Nicola Mulder and Ralf Zimmer

Title: From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems

Abstract: Current gene set enrichment approaches do not take interactions and associations between set members into account. Mutual activation and inhibition causing positive and negative correlation among set members are thus neglected. As a consequence, inconsistent regulations and contextless expression changes are reported and, thus, the biological interpretation of the result is impeded. We analyzed established gene set enrichment methods and their result sets in a large-scale investigation of 1,000 expression datasets. The reported statistically significant gene sets exhibit only average consistency between the observed patterns of differential expression and known regulatory interactions. We present Gene Graph Enrichment Analysis (GGEA) to detect consistently and coherently enriched gene sets, based on prior knowledge derived from directed gene regulatory networks. Firstly, GGEA improves the concordance of pairwise regulation with individual expression changes in respective pairs of regulating and regulated genes, compared with set enrichment methods. Secondly, GGEA yields result sets where a large fraction of relevant expression changes can be explained by nearby regulators, such as transcription factors, again improving on set-based methods. Thirdly, we demonstrate in additional case studies that GGEA can be applied to human regulatory pathways, where it sensitively detects very specific regulation processes, which are altered in tumors of the central nervous system. GGEA significantly increases the detection of gene sets where measured positively or negatively correlated expression patterns coincide with directed inducing or repressing relationships, thus facilitating further interpretation of gene expression data. The method and accompanying visualization capabilities have been bundled into an R package and tied to a graphical user interface, the Galaxy workflow environment, that is running as a web server.

Author contributions: LG developed and implemented the method based on several suggestions from RZ. LG and GC validated the method, LG carried out the consistency study, GC suggested the explainability study. LG, RK and RZ wrote the manuscript. NM and RZ reviewed the manuscript and jointly supervised the project.

3.2.1 Introduction

Transcriptomic studies measure gene expression in different conditions. Striking genes, which are differentially regulated between the conditions, are of primary interest and investigated for common features and membership in group of genes, which have the same function or belong to the same biochemical pathway.

A first impression of similar behavior of genes can be achieved via clustering of genes [121]. The usually more effective overrepresentation analysis (ORA) tests the overlap of a predefined group of genes and the set of differentially expressed genes assuming the hypergeometrical distribution under the null hypothesis [26]. The method is widely accepted and has been subject to modifications of diverse visual and model related features [122], though the basic statistical principle remained unchanged. However, Goeman and Bühlmann [27] criticize that the sampling procedure of ORA is statistically invalid and leads to a hazardous interpretation of the resulting p-value. Furthermore, the concentration on the usually small group of significantly differentially expressed genes, compared to the set of all the other, usually thousands of genes analysed in the study that are ignored, is not suitable for an investigation on a global scale.

Both points of criticism are resolved in *Gene Set Enrichment Analysis* (GSEA) as it uses a valid sampling procedure and computes over the whole scope of genes [123]. A Kolmogorov-Smirnov test statistic is applied to test whether the ranks of the p-values of the genes in the gene set can be a sample from a uniform distribution. Several modifications of GSEA have been published [124].

Though ORA and GSEA are convenient in the analysis of genes that are independently expressed, a serious problem arises when these methods are applied to gene set definitions extracted from regulatory networks and metabolic pathways. The assumption of independence among set members does not hold anymore; genes are found to be correlated due to mechanisms of co-regulation and co-expression. Initial steps to deal with that problem include implicit accounting for the correlation structure [125] and integration of network topology of undirected interaction networks [126]. Based on these first efforts, Liu *et al.* [127] have proposed *Gene Network Enrichment Analysis* (GNEA) that uses ORA to test for overrepresentation of gene sets in transcriptionally affected subnetworks of a global interaction network.

As the sign of gene expression changes and the direction of regulatory interactions are so far not taken into account, substantial features of the data are still ignored and the dynamics of the transcriptomic system are not realistically reflected. Activation and inhibition are essential regulatory mechanisms in the transcriptional machinery of the cell and are causes for up- and down-regulation of particular genes. Although processes like post-translational modification and combinatorial effects between regulatory proteins impair a straightforward causal relationship between regulation and gene expression, it was shown that coexpression is correlated with functional relationships between genes [128]. Additionally, integrative analysis of transcriptome, proteome and interactome data revealed significant correlations between expression profiles and regulatory interaction on the protein level [30, 31]. Hence, we explain positive correlation in gene expression with activating edges of the transcriptional network. Vice versa, we assume inhibition to cause observed negative correlation in gene expression patterns. In our following definition of *Gene Graph Enrichment Analysis* (GGEA), we exploit both fundamental regulation types in a novel enrichment framework for signed and directed gene regulatory networks, to judge whether the topology of the network is well fitted by the expression data.

3.2.2 Methods

GGEA

Given gene regulatory information, for example extracted from biochemical pathways or a global transcriptional network, a gene set under investigation and gene expression data sampling different conditions, GGEA performs three essential steps (Figure 3.2): first, the gene set is mapped onto the underlying regulatory network, yielding an induced subnetwork. That is the affected part of the network, which consists of edges that involve members of the gene set. Second, each edge of the induced network is scored for consistency with the expression data, i.e. the signs of the expression changes of two interaction partners are evaluated for agreement with the regulation type (activation/inhibition) of the link that connects both genes. Third, the edge consistencies are summed up over the induced network, normalized and estimated for significance using a permutation procedure.

Experimental Setup

In the following, we consider the classical setup of a transcriptomic study. This incorporates a set G of usually several thousand genes g_i $(i = 1, \ldots, n)$ measured for differential expression between two conditions, each represented by a group of samples $S_1 = \{s_1, \ldots, s_k\}$ and $S_2 = \{s_{k+1}, \ldots, s_m\}$, respectively. The function

$$\mathrm{expr} : G \times (S_1 \cup S_2) \to \mathbb{R} \tag{3.1}$$

returns the expression value for a gene and a sample at a time.

Measures of Differential Expression

The most intuitive measure for expression changes of a single gene between two conditions is the fold change

$$\mathrm{fc} : G \to \mathbb{R}, \tag{3.2}$$

defined as the ratio of the estimated expression values of a particular gene in both sample groups

$$\mathrm{fc}(g_i) = \frac{\mathrm{expr}(g_i, S_1)}{\mathrm{expr}(g_i, S_2)}, \tag{3.3}$$

where $\mathrm{expr}(g, S)$ computes the mean expression level of gene g in condition S. We compute t-test derived p-values to assess the statistical significance of the expression changes [15] and correct them for multiple testing. Both measures are log-transformed

$$\tilde{\mathrm{fc}} := \log_2(\mathrm{fc}), \quad \tilde{p} := -\log_{10}(p), \tag{3.4}$$

and the significance thresholds $\alpha = -\log(0.05)$ and $\beta = 1$ (two-fold) are used as defaults for $\tilde{p}$ and $\tilde{\text{fc}}$, respectively. Such sharp thresholds are of course quite artificial and discriminate drastically between genes just over and just below α or β. In addition, noise in the data, such as imprecise and erroneous measurements of gene expression values, has to be expected and to be dealt with. Hence, we divide the range of both measures into two main categories and smooth the borders via introduction of a degree of uncertainty, according to the mathematical concept of *fuzzyfication* [77, 78]. For the fold change, we map

$$(\tilde{\text{fc}} < 0,\ \tilde{\text{fc}} > 0) \mapsto (\text{down},\ \text{up}), \tag{3.5}$$

and compute membership values for both categories via the weighting functions $w : \tilde{\text{fc}} \mapsto [0, 1]$ (displayed in Figure 3.3b), resulting in a pair

$$\langle \text{fc} \rangle \ := \ \text{fuzzy}(\tilde{\text{fc}}) = \langle\ w_{\text{down}}(\tilde{\text{fc}}),\ w_{\text{up}}(\tilde{\text{fc}})\ \rangle. \tag{3.6}$$

Analogously, we map $\tilde{p}$, using Figure 3.3a, to areas of *low* and *high* significance in the fuzzy concept

$$\langle \text{sig} \rangle \ := \ \text{fuzzy}(\tilde{p}) = \langle\ w_{\text{low}}(\tilde{p}),\ w_{\text{high}}(\tilde{p})\ \rangle. \tag{3.7}$$

For both measures, a third category can optionally be introduced to account for unspecific signals in case of very noisy data. The fold change and p-value categories are combined to

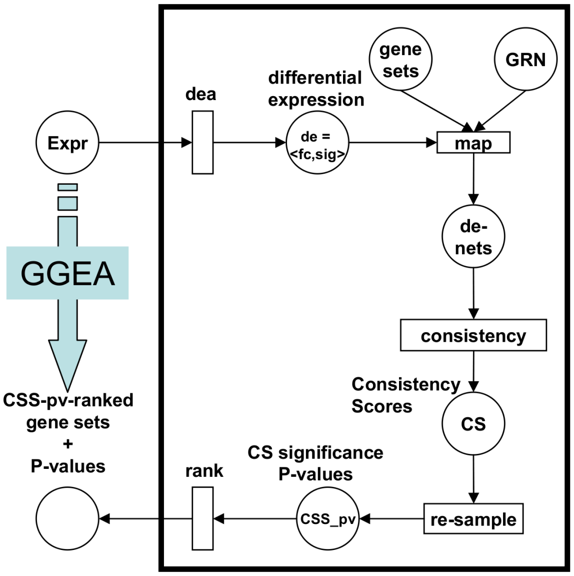
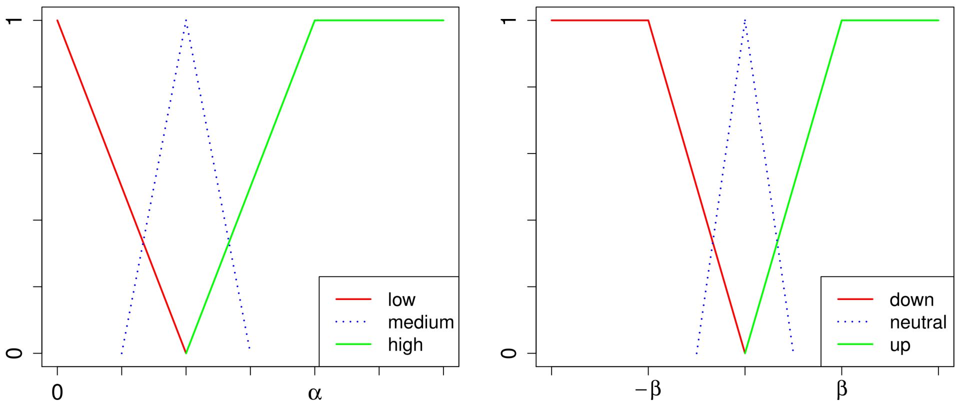


Figure 3.2: Key steps of GGEA. Subsequent to differential expression analysis *dea* of expression data *Expr*, yielding fuzzified measures *de* of differential expression, target gene sets are first mapped onto the gene regulatory network (GRN). The *de*-values are assigned to corresponding places in resulting induced nets (*de-nets*). Second, consistency scores are computed for each *de*-net and third, significance of the scores is estimated via re-sampling, and exploited to rank the gene sets.



(a) p-value (b) fold change

Figure 3.3: Fuzzyfication of p-value and fold change. Both measures are mapped onto two main categories, each having a membership function to express the uncertainty of the mapping. Additional categories, e.g. a third category *medium* and *neutral*, respectively, can be introduced for a more detailed representation.

a single measure of differential expression

$$\mathrm{de} := \langle \mathrm{fc}, \mathrm{sig} \rangle, \tag{3.8}$$

in order to simultaneously summarize and express whether the transcriptional activity of a particular gene is reduced or enhanced in one sample group, compared to the other.

Induced Gene Regulatory Networks

Enrichment analysis is the determination of significant gene sets out of a predefined universe of gene sets U, s.t. result sets accumulate differentially expressed features of the gene expression data. GGEA uses an *a priori* defined gene regulatory network (GRN), typically extracted from respective databases or compiled from the relevant literature, to introduce and exploit the interdependencies between gene set members. We model a regulatory interaction of the GRN as a transition t (see Figure 3.4) with an input place for the regulator and an output place for its target, as well as an associated effect (activation, inhibition) and the direction of the interaction. For a gene set $u \in U$, we construct the *induced* subnetwork

$$\mathrm{GRN}(u) := \{t \in \mathrm{GRN} \mid u \cap (\mathrm{in}(t) \cup \mathrm{out}(t)) \neq \emptyset\}, \tag{3.9}$$

s.t. for each gene g of the gene set u all transitions are extracted, where g is either the regulating or the regulated gene.

Gene Regulatory Networks as Petri Nets

Petri net models are well established in information theory [75, for a review] and have been extensively applied to biochemical processes, like metabolic pathways [129] and gene regulatory networks [76]. Given a GRN under investigation, we construct a corresponding

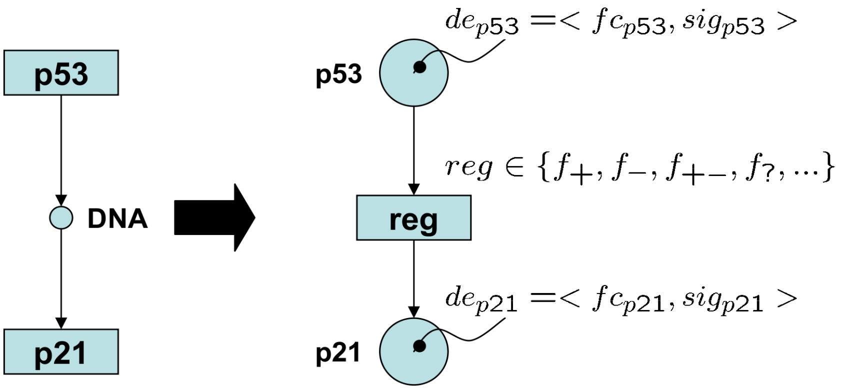

p53
DNA
p21
p53
$de_{p53} = < fc_{p53}, sig_{p53} >$
$reg \in \{f_+, f_-, f_{+-}, f_?, ...\}$
reg
$de_{p21} = < fc_{p21}, sig_{p21} >$
p21

Figure 3.4: Modeling regulatory interactions using PNFL. Shown is a `KEGG` style representation of an activation and its transformation into a PNFL transition f_+. Tokens of combined fuzzy measures *de* of differential expression assigned to Petri net (PN) places, represent the regulator and its target. The regulatory effect is defined via a specific fuzzy rule for every effect type of the GRN.

Petri net (PN) having features of fuzzy logic (FL), as it is introduced as PNFL in [79], and illustrated in Figure 3.4. The regulations of the GRN are required to be specified with direction and effect. In our model, regulator (R) and regulated target (RT) are represented via PN places holding tokens of fuzzy values for both fold change (fc) and significance of fc (sig). The variety of regulatory effects occurring in the GRN are defined by specific fuzzy rules $reg \in \{f_+, f_-, f_{+-}, f_?, ...\}$ (Table 3.2), meaning activation f_+, inhibition f_- and dual effects f_{+-}. The concept is extendable, e.g. to other effects like interactions of unknown type $f_?$. The fuzzy rules compute output tokens from given input tokens. Thus, consistency between expected (i.e. modeled) behavior and the measured values can be evaluated. Consistency takes the direction of the effect, the amount (fc) and its significance into account and is a straightforward extension of the discrete notion of consistency (e.g. R *up* and $f_+ \implies$ RT *up*). Moreover, it appropriately models noise in the actual experimental measurements.

Table 3.2: Fuzzy rule set for activation and inhibition

	⟨fc⟩		⟨sig⟩	
	down	up	low	high
f_+	down	up	low	high
f_-	up	down	low	high

Consistency of Regulatory Interactions

The major problem of set enrichment strategies, when applied to GRN-based gene sets, is that they accumulate evidence for differential expression of single genes to estimate the enrichment of the whole set. Interfering and potentially contrary constraints of the underlying GRN are ignored. For example, two significantly up-regulated genes increase the enrichment of the set, even if one gene inhibits the other. For that reason, we introduce the concept of consistency.

> **Definition** (*consistency*): a transition of a PNFL is consistent with given expression data, if the measured and the modeled expression of the regulated gene is in agreement. The modeled expression is estimated from the regulatory effect and the expression of the regulator.

Intuitively, consistency for the special case of a simple activating or inhibiting edge requires fold changes for regulator and target of the same or opposite directions, respectively. It is implied for the above example that an up-regulated inhibitor should result in reduced expression of the affected gene.
For the PN constructed above, a *consistent* transition t with fuzzy regulation function f_t between an input place i and an output place o satisfies

$$\mathrm{de}_o \approx f_t(\mathrm{de}_i), \tag{3.10}$$

i.e. the modeled predicted expression behavior agrees with the actual observed behavior.

Scoring

To determine if and to which extent t is consistent with the given expression data, we calculate the consistency

$$C(t) := \mathrm{cons}\left(\mathrm{de}_o, f_t(\mathrm{de}_i)\right), \tag{3.11}$$

where the function *cons* estimates the (fuzzy) similarity between the predicted and measured token on the output place of transition t. Consistency computation is generic, an example implementation of *cons* incorporates defuzzyfication of the fuzzy values back into real numerical values [79] and taking their reciprocal absolute difference. We compute the raw GGEA consistency score S for the subnetwork GRN(u), induced by the gene set $u \in U$, via summation over the consistencies of all transitions T_u of GRN(u)

$$S := \sum_{t \in T_u} C(t), \tag{3.12}$$

and normalize it by the number of transitions $|T_u|$

$$\bar{S} := \frac{S}{|T_u|}, \tag{3.13}$$

to adjust for the size of GRN(u).

Significance and Ranking

According to the recommendations of [27] and [130], statistical significance of the consistency score is estimated via a permutation approach based on subject sampling, which is defined in a self-contained way:

1. Permute group assignment of samples N times.
2. Recalculate differential expression measures for each permutation.
3. Recalculate consistency score for each permutation.
4. Find the consistency p-value as the proportion of permutation scores that are larger than the observed score.

We compute the consistency p-value for each gene set $u \in U$ and rank the gene sets by the adjusted p-values, i.e. p-values corrected for multiple testing (see again Figure 3.2). Gene sets below the chosen significance level are classified as *significantly and consistently* enriched.

Extensions

To apply to regulation processes involving multiple regulators and transcription complexes composed of several genes, we allow a transition t to have an arbitrary number of inputs $I_t = \{i_t^1, \ldots, i_t^k\}$ and outputs $O_t = \{o_t^1, \ldots, o_t^l\}$. This is accomplished via generalization of equation (3.10) to

$$\left(\mathrm{de}(o_t^1), \ldots, \mathrm{de}(o_t^l)\right) \approx f_t \left[\left(\mathrm{de}(i_t^1), \ldots, \mathrm{de}(i_t^k)\right)\right]. \tag{3.14}$$

We model the combined effect via computation of the average behavior of all effects, or optionally, by the effect of highest statistical significance (the effect could, of course, also be modeled as a full-blown k-dimensional fuzzy function).

Missing data, i.e. genes of the GRN, which are not measured in the study, is resolved using transitivity. By going up and down, respectively, the regulation path until a non-empty place is reached, an empty origin is filled with the found token, which is adjusted to path length of the transitive relation. The adjustment is due to the fact that the evidence for regulation weakens, as the path length increases.

Implementation and Availability

GGEA is implemented in the statistical language `R` [131] and makes use of the `Bioconductor` software suite [132]. The GGEA method and accompanying visualization capabilities have been bundled into an `R` package and tied to a graphical user interface, the `Galaxy` workflow environment [133], that is running as a web server.

Consistency and Explainability Study Setup

Data Sampling and Network Construction

Gene expression data of *E. coli* was collected and sampled from the `M3D` database [103, Many Microbe Microarrays Database]. The 1,000 datasets were designed in a two-class fashion, s.t. each class contained 15 samples. It was assured that real-world distributions of fold changes and differential expression p-values were matched. A global gene regulatory network for *E. coli* was constructed using the regulatory interactions provided in the `RegulonDB` database [44]. From the union of all stored TF/gene, TF/operon, TF/TF, σ/gene and σ/TU regulatory interactions (TF stands for *transcription factor*, TU for *transcriptional unit* and σ for the RNA polymerase σ-factor), we removed duplicated and ambiguous edges. The final network connected 2097 unique nodes by 5784 edges, which were clearly annotated as either activating or inhibiting.

Methods Collection and Gene Set Definitions

For each dataset, we applied the standard hypergeometrical overrepresentation test ORA1, and a collection of array resampling methods that correctly control false positive rates and gene correlation patterns [130]. These are the modified resampling overrepresentation test ORA2 [27], SAFE [125], GSEA [123] and SAM-GS [124]. The gene set catalog for analysis was defined on the one hand according to the `KEGG` pathway annotation [24] for *E. coli*, and, on the other hand, according to the `GO` classifications [22] of *E. coli*. We restricted both catalogs to gene sets having at minimum five and at maximum 500 set members. This yielded 83 and 446 gene sets for `KEGG` and `GO`, respectively.

Consistency Benchmark

For each method, we collected for all datasets with statistical significant outcome ($p < 0.05$) the top ranked gene sets. As not all datasets produced significant outcome for all methods, we uniformly chose 700 sets at random from these top ranked gene sets and computed the percentage of consistent relations in the corresponding induced regulatory networks. We took regulation direction, type and strength into account and distinguished respective categories. Activating relations required both interaction partners to be expressed in the same direction to be consistent, while inhibiting relations required them to be expressed in the opposite direction. Regulation strength was categorized as *weak* and *strong*, depending on the differential expression p-value of the regulator. We chose 0.5 and 0.05 as the thresholds for the weak and the strong category, respectively. To estimate the null distribution in each category, we computed the consistency of all gene sets in all datasets.

Explainability Benchmark

The selected 700 top ranked sets were restricted to differentially expressed genes of high statistical significance. The significance level was set to 0.1. Minimum spanning trees (MST) were computed for each of the reduced gene sets according to the underlying global GRN, s.t. each significant gene of a top ranked set could be reached by all other significant members of that set. Moreover, the corresponding MST for such a set minimized the number of genes not contained in the set. The direction of the regulatory link between two genes in the network (activation/inhibition) as well as the direction of the expression change of individual genes (down-/up-regulation) was ignored. We classified a restricted result set as *fully explainable* if all members were directly connected to another member in the corresponding MST. Otherwise, we counted the number x of genes in the MST, which were not a member of the set, and classified the set as *explainable with* x additional genes. As a measure of explainability achieved by a method in all its 700 top ranked sets, we calculated, for a chosen number x, the percentage of sets that were explainable with at most x additional genes.

Case Study Setup

FiDePa and Local GGEA

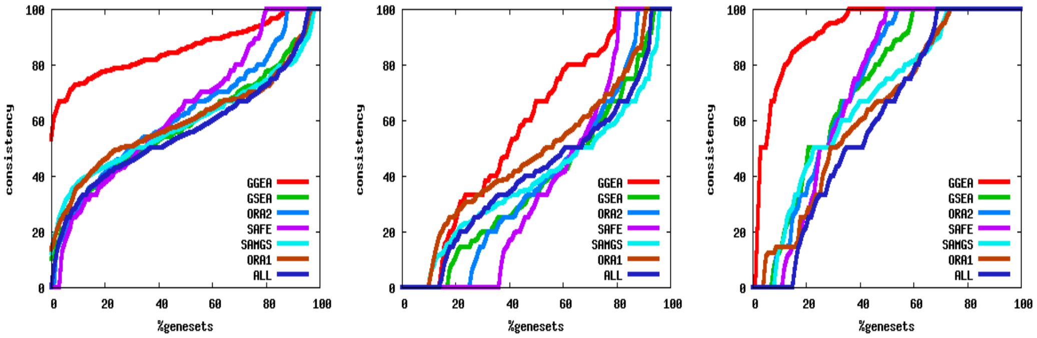
We applied GGEA to the glioma dataset that has been investigated before with the method FiDePa [134]. The method exploits GSEA first to determine differentially regulated paths of a particular length and uses the resulting paths for the construction of a consensus network, which is subsequently tested for overrepresentation of gene sets. In a similar approach, we computed consistency scores of regulatory links in all human non-metabolic `KEGG` pathways (gene regulatory and signaling pathways) and the ten edges with the highest consistency score were extracted from each of them. Duplicated edges were removed and the consensus graph was further reduced via application of a high pass consistency filter using the mean consistency score as threshold. That yielded a total of 378 edges connecting 342 unique nodes, which were tested, as in FiDePa, for overrepresentation.

3.2.3 Results

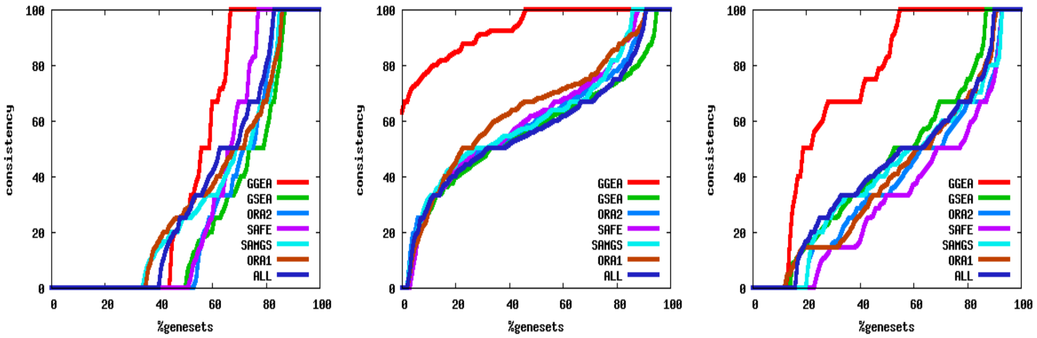
Consistency Study

We conducted a meta-analysis of 1,000 *E. coli* datasets and evaluated the consistency within results of gene set enrichment methods, based on the regulatory interactions found in the transcriptional network of *E. coli*. Details of the study setup, the consistency benchmark and the classification of interaction strength as *weak* and *strong* are described in Section 3.2.2. The results for `KEGG` gene sets are shown in Figure 3.5(a–d) and for `GO` gene sets in Figure 3.5(e–h).

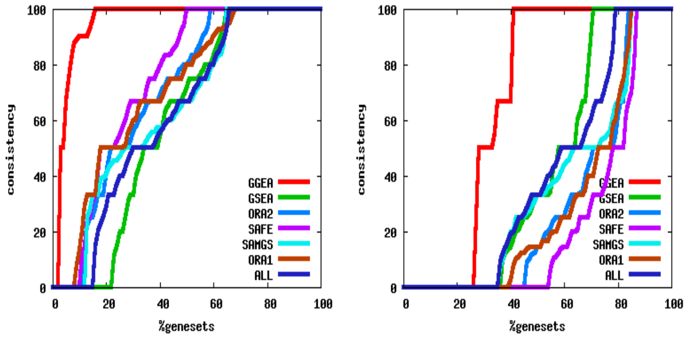
We observe that the set enrichment methods systematically neglect mutual regulation among set members. For `KEGG` gene sets, weak regulations (Figure 3.5a and Figure 3.5b) are only slightly more consistent than average (the null consistency) and the gene set with maximal consistency is frequently not reported by the set enrichment methods, regardless of activating or inhibiting links. Strong activators, with an expression change of high statistical significance, and the effects on their targets are more consistently aligned (Figure 3.5c). However, the consistency gained in strong activations is lost for strong inhibitions (Figure 3.5d). The results for `KEGG` sets are nearly replicated in `GO` gene sets (Figure 3.5e-h). In contrast, GGEA, which takes consistency into account for selecting relevant gene sets in the first place, yields the most consistent gene sets in all categories for both, `KEGG` and `GO` gene set definitions. Activations and inhibitions are similarly consistent, if adjusted to background distributions of both regulation types, and stronger signals are properly weighted in order to preserve the regulation kinetics. Although stronger signals have an higher impact on the GGEA score, weak regulations are also highly consistent in the sets found by GGEA. In general, these findings are more pronounced for `GO` sets, compared to `KEGG` gene set definitions. This is due to the fact that the `GO` catalog (446 gene sets) is nearly six times larger and contains more diverse composed gene sets than



(a) weak act, KEGG (b) weak inh, KEGG (c) strong act, KEGG

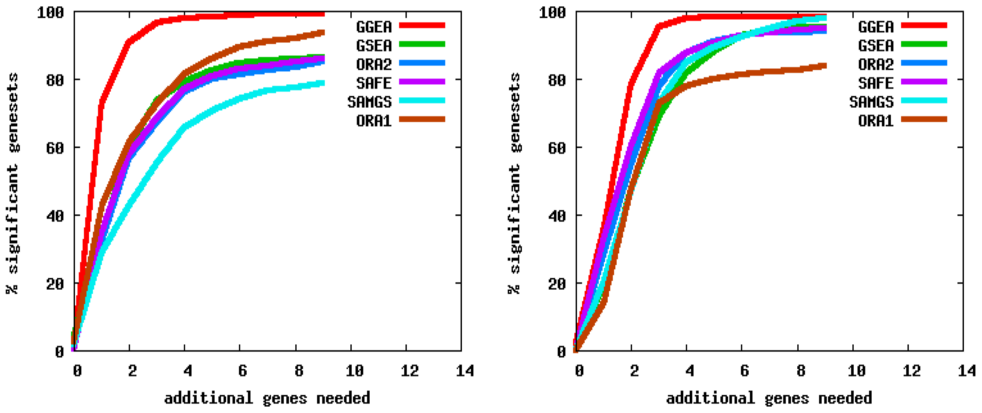


(d) strong inh, KEGG (e) weak act, GO (f) weak inh, GO



(g) strong act, GO (h) strong inh, GO

Figure 3.5: Consistency of regulatory interactions in top ranked rets. Plots (a)-(h) are explained in the main text. Each of the set enrichment methods was applied to 1,000 *E. coli* datasets using `KEGG` and `GO` gene set definitions, respectively. From datasets with statistical significant outcome, the top ranked gene sets were collected and investigated for consistency of *weak* and *strong* activation and inhibition (as described in Section 3.2.2). GGEA results are displayed in red. The plots show which fraction (x-axis) of the identified gene sets had at most a consistency of $y\%$. The y-axis shows the consistency of sets as the fraction of consistent regulatory interactions in the respective gene set. The null consistencies were estimated via the overall consistency of all gene sets in all datasets and are displayed in dark blue.



(a) GO (b) KEGG

Figure 3.6: Explainability of expression changes in top ranked sets. The 700 top ranked gene sets (introduced in the consistency study above) of each method were restricted to genes with expression changes of statistical significance. For each restricted set, we computed the minimal number of genes not in the set, but needed to connect the significantly regulated genes of that set, to a regulation network. Displayed is the percentage of gene sets, for which x or less additional genes are needed. E.g. for `GO` sets, a single additional gene makes 73% of GGEA's top ranked sets explainable, while in case of ORA1 or SAMGS a single gene makes only 42% or 29% of the top ranked sets, corresponding to each method, explainable.

the `KEGG` catalog (83 gene sets), which emphasizes differences between the set and graph enrichment methods.

Explainability Study

As the consistency is substantially incorporated in the GGEA score, we performed a second evaluation using the more independent benchmark of explainability, as described in Section 3.2.2. The main target of this investigation was to determine to which extent statistical significant expression changes of single genes can be explained by other set members. Considering that a statistical significant finding for a gene set indicates differential regulation of the corresponding biological process, it is in turn implied that a part of the global regulatory network (here a subgraph of `RegulonDB`) exists, which connects the differentially expressed genes in this set. However, it is frequently observed that important regulators or mediators are missing in a particular gene set, leaving its differentially expressed genes not connected with each other. As a result, the biological interpretation of the observed effect is impeded. Based on these considerations, we have introduced above the terms *fully explainable* and *explainable with* x additional genes, to assess how easily a result set can be interpreted. Intuitively, the less additional genes needed, the easier the interpretation: a single additional gene could possibly be a regulator or mediator not contained in the set, while the need of several additional genes requires more complex assumptions to make the outcome interpretable. For the explainabilty study, we explicitly made the input regulatory network undirected, generalizing the edges, s.t. possibly unknown inverse regulations are allowed. We enhanced this feature by additionally removing the sign of the fold change and only judged whether a gene was differentially expressed or not. The results are shown in Figure 3.6.

GGEA systematically reports more easily explainable sets than all other methods for both, `KEGG` and `GO` gene set definitions. Similar to the results of the consistency study, the gap is much bigger between the performance of GGEA and the other methods when using `GO` sets definitions, as also observed in the consistency study. For example, GGEA needs in 73% of its top ranked gene sets a single additional gene to make the differentially expressed genes in a particular set connected, whereas the best set enrichment method, ORA1, can explain only 42% of the sets with a single addition gene (SAFE: 35%, GSEA: 34%, ORA2: 32%, SAMGS: 29%). Allowing two additional genes, GGEA can explain >90% of all reported gene sets, while all other methods produce results $\leq$60%.

Case Study

In a final case study, we investigated two expression datasets of human neuronal tumors and compared results of GGEA and set enrichment strategies. Though a comparative benchmark is hard to find, due to a missing gold standard that classifies detected pathways as right or wrong in the context of the investigated expression data, we approached this matter via collection of biological evidence in the scientific literature and focused on the specificity of the findings and the sensitivity of the method used. For consistency evaluation, we used the regulatory interactions occurring in human non-metabolic `KEGG` pathways (gene regulatory and signaling pathways). In the first analysis, we applied GGEA to the glioma dataset that was investigated before by [134] with the method FiDePa (see Section 3.2.2 for details). We observe large agreement in the result lists of both methods (Table 3.3); 17 pathways listed in the FiDePa result also occur in the top 25 of the GGEA ranking. The positive control Glioma is better ranked (and has higher significance) by GGEA. Further, several unspecific and disease unrelated pathways detected by FiDePa (e.g. Type I/II diabetes mellitus, Cell cycle) are discarded by GGEA and replaced by specific, cancer related pathways (e.g. Renal cell carcinoma, Endometrial cancer). For the top rank, GGEA (Pathways in Cancer; not detected by FiDePa) gives a clear disease related hint, while FiDePa (MAPK signaling pathway) reports a general signaling process. The Neurotrophin signaling pathway, which promotes neuronal tumors via modulation of neuronal apoptosis [135], is not identified by FiDePa, but listed by GGEA on rank 4.

In the second evaluation study, we used neuroblastoma expression data that was investigated for enrichment of metabolic pathways before [136]. The application of GGEA to the neuroblastoma dataset identified 17 significantly and consistently enriched pathways (Table 3.4). Best ranked is the Neurotrophin signaling pathway, which was already detected in the glioma study to play an essential role in the development of neuronal tumors. As this pathway seemed to be particularly striking for both tumors, we determined regulations with highest consistency in that pathway, in order to get a deeper insight into the disease causing dynamics: we found that the high affinity nerve growth factor receptor, which in

Table 3.3: Result comparison of GGEA and FiDePa application to the glioma dataset. Arrows in the first column denote whether a pathway is ranked higher or lower by GGEA, compared to FiDePa.

Pathway	ORA p (GGEA)	ORA p (FiDePa)	Rank (FiDePa)
↑ Pathways in cancer	1.8e-24	–	–
↑ Focal adhesion	1.4e-18	2.5e-06	5
↑ T cell receptor signaling	1.2e-17	1.5e-05	7
↑ Neurotrophin signaling	5.5e-15	–	–
↑ Colorectal cancer	1.1e-14	9.4e-05	11
↑ Pancreatic cancer	3.8e-14	0.0001	12
↑ Renal cell carcinoma	1.3e-13	–	–
↑ VEGF signaling	1.5e-13	0.006	22
↔ Fc epsilon RI signaling	4.1e-13	1.9e-05	9
↓ Chronic myeloid leukemia	6.3e-13	1.65e-05	8
↑ ErbB signaling	8.9e-13	–	–
↑ B cell receptor signaling	4.2e-12	0.001	17
↑ Glioma	5.1e-12	0.003	20
↑ Insulin signaling	3.2e-11	0.001	18
↑ Leukocyte trans. migration	3.9e-11	0.01	24
↓ Adherens junction	4.9e-11	1.4e-05	6
↓ GnRH signaling	6.5e-11	0.0003	16
↓ Nat. killer cell med. cytotox.	6.5e-11	1.4e-11	2
↑ Wnt signaling	1.2e-10	–	–
↓ Toll-like receptor signal.	1.2e-09	5.5e-05	10
↑ Endometrial Cancer	1.6e-07	–	–
↑ Non-small cell lung cancer	3.4e-07	–	–
↑ Acute myeloid leukemia	3.9e-07	–	–
↓ mTOR signaling	1.2e-06	0.0002	15
↓ MAPK signaling	4.4e-06	1.6e-25	1
…	…	…	…
↓ Apoptosis	0.04	9.3e-11	3

Table 3.4: Result of GGEA application to the neuroblastoma dataset.

Pathway	p-value
Neurotrophin signaling	7.5e-06
Chemokine signaling	0.0004
Cell adhesion molecules (CAMs)	0.0021
Regulation of actin cytoskeleton	0.0068
Focal adhesion	0.0091
Nat. killer cell med. cytotox.	0.0092
Leukocyte trans. migration	0.0099
Pathways in cancer	0.01
T cell receptor signaling	0.016
Fc epsilon RI signaling	0.019
Long-term depression	0.023
Axon guidance	0.033
Vasc. smooth muscle contraction	0.035
p53 signaling pathway	0.035
Melanogenesis	0.039
MAPK signaling	0.043
Thyroid cancer	0.05

humans is encoded by the NTRK1 gene, is up-regulated in neuroblastoma cells and activates the adaptor protein SH2B3, the growth factor receptor-bound protein 2 (GRB2), the Abelson murine leukemia viral oncogene homolog 1 (ABL1), the phospholipase gamma 2 (PLCG2) and the SHC-transforming protein 1 (SHC1). A literature search revealed that all of the activated and associated proteins are proliferating, oncogenic and/or apoptosis influencing and thus, of cancer promoting importance [137,138]. In addition, the up-regulation of the whole NTRK1 proliferation module in neuroblastoma was experimentally validated [139] some years ago. This sensitive finding motivated a similar investigation for the other pathways in Table 3.4, which we identified to be throughout substantially involved in neuroblastoma formation. As an example: GGEA detects the Chemokine signaling pathway. We found that neuroblastoma impairs chemokine-mediated dendritic cell migration [140] and chemokines strongly promote neuroblastoma primary tumor and metastatic growth [141].

Moreover, we wanted to know whether the findings of GGEA are in concordance with the results for metabolic pathways. As a showcase, we demonstrate this via the detected Fc epsilon RI signaling pathway. In [136], only moderate attention (discussed in their supplement) is paid to the extremely significant findings for Phosphatidylinositol metabolism (p = 9e-12) and for several pathways concerning the metabolism of lipids and fatty acids, e.g. Fatty acid metabolism (p = 1.7e-9) and Glycerophospholipid metabolism (p = 3.9e-7), which are listed in Table 1 of that publication. As it can be verified in the corresponding `KEGG` pathway maps, Fc epsilon RI signaling has a regulatory impact on both -

the Phosphatidylinositol metabolism via modulation of the phospholipase (affected by the Neurotrophin pathway); and the metabolism of lipids in general via stimulation of arachidonic acid synthesis. Arachidonic acid is a polyunsaturated fatty acid that is required for membrane phospholipid synthesis. It is also involved in cellular signaling and known to activate syntaxin-3, which causes cell membrane expansion of neuronal cells [142]. Schramm *et al.* [136] explain the several revealed signals in lipid related metabolisms with TCA based energy production; the GGEA results, explaining stimulation of arachidonic acid synthesis, imply that the observed activated production of fatty acids and lipids (which is based on the latter) is rather due to the increased requirement of neuronal membrane material (i.e. specific lipids) in the fast growing and dividing neuroblastoma cells.

3.2.4 Discussion

In this work, we presented *Gene Graph Enrichment Analysis* (GGEA), a novel algorithmic framework to detect increased agreement between positively and negatively correlated expression patterns of genes, connected by activating and inhibiting edges in signed and directed transcriptional networks. The method exploits directed regulatory relations represented as fuzzy logic rules to assess and identify graphs, which maximize the consistency between the regulatory network and the expression data. GGEA is a major improvement to current gene set enrichment strategies, as we found experimentally validated regulatory interactions not to be consistent *per se* with the expression data in top ranked and statistically significant result sets of these methods. That was validated in a large-scale consistency study of 1,000 *E. coli* chips using the *E. coli* `RegulonDB`, currently the best curated regulatory network, for the investigation of consistency. As set enrichment strategies ignore mutual regulation among set members, we observed that activations and inhibitions are only average consistent with the gene expression in these result sets. Even strong causal signals, i.e. a regulator with differential expression of high statistical significance, in pairwise directed regulations were frequently not properly reflected. Inhibitions were more seriously neglected than activations. This is partly due to a data bias, as there are more activations than inhibitions in the database. Hence, more genes, and thus also more significant genes, are involved in activations just by chance. As gene set enrichment analysis mainly computes upon the leading edge of the ranked p-value vector of genewise differential expression [123], gene sets with a majority of activating genes are more likely to be reported. On the other hand, we found activations clearly better conserved than inhibitions across all experiments stored in the `M3D` database. For GGEA, we observed, under consideration of this bias, that activations were nearly optimally consistent and inhibitions were preserved in a large fraction of regulations. GGEA achieved the highest concordance between the regulation direction and the expression behavior of the incorporated regulator and regulated target gene. It should be emphasized that GGEA consistently aligned weak (only moderately differentially expressed) signals, which are usually not taken into account by set enrichment methods. That improved sensitivity enables preference of weak, but coherent regulations over strong, but contextless signals. This is expected to better reflect the nature of key cellular regulators.

As GGEA exploits the consistency for the computation of its score, we additionally carried out a more independent benchmark to investigate how well statistically significant expression changes of single genes can be explained by other set members. As a measure of explainability, we used the number of additional genes, which were needed to connect significant members of a set to a regulatory network. We found this evaluation of particular interest, as it tries to approximate the process of the human interpretation. For all set enrichment methods, only a small amount of genes could be explained by other set members in a significant result and we observed frequently that several additional genes were needed. Implied is that set enrichment indeed indicates that there is *something* striking happening in a certain result set, however, conclusions whether observed expression changes are coherent and in context with the surrounding regulators cannot be drawn. This is resolved by GGEA. It systematically reports more easily explainable sets than all other methods, and the fraction of explainable sets with a single additional gene is increased by >30% in comparison with the best set enrichment method.

Furthermore, we applied GGEA in two pilot case studies of human neuronal tumors using regulatory interactions of signaling pathways, though incorporated protein-protein regulations cannot be measured at the transcriptional level. Nonetheless, we again hypothesize that genes, annotated to be associated in a pathway, should show higher correlation patterns than arbitrary genes, which are not. On the other hand, we argue that signal cascades normally target altered gene regulation.

On the glioma dataset, GGEA discovered throughout specific and disease related pathways. Induced by increasing specificity, the fraction of false positives decreases. Unspecific and inconsistent pathways are replaced by more appropriate pathways. An example is the detection of the Neurotrophin signaling pathway that modulates neuronal apoptosis (a very specific finding), while general apoptosis is downgraded.

The Neurotrophin signaling pathway also has a major influence on the development of neuroblastoma, another neuronal tumor type. The experimentally verified connection was detected by GGEA with high significance, while GSEA failed to detect it. The discovery of such false negatives of the set enrichment analysis is due to improved sensitivity already observed in the consistency study. However, it is surprising that only GGEA is sensitive enough to detect the Neurotrophin signaling pathway, the Chemokine signaling pathway and the Fc epsilon RI signaling pathway - all of which have been shown to be of crucial importance in neuroblastoma formation - while standard GSEA does not detect them. Best ranked pathways of GSEA are: Cell cycle, Ribosome and Olfactory transduction. The connection to the disease is incomprehensible and explanations are almost arbitrary.

3.2.5 Conclusion

We showed in three independent and differently designed studies that GGEA consistently aligns regulation and expression and yields result sets where statistically significant expression changes can be explained by regulators within the set. Moreover, GGEA eases the biological interpretation of reported gene sets, as they are more coherent than sets reported by set enrichment methods. This means many more of their relevant genes are connected or can be connected by a minimum number of additional factors. In summary, GGEA is an intuitive enrichment method, which uses gene regulatory information to improve consistency and coherence of detected enriched gene sets and, thus, substantially reduces the fraction of false positive and false negative classifications of relevant gene sets. GGEA significantly improves the detection of gene sets where measured positively or negatively correlated expression patterns coincide with directed inducing or repressing relationships between the respective pairs of genes. Hence, gene set regulators, such as transcription factors, can explain a significant portion of the observed expression changes. As GGEA is as fast and easy to apply to experimental data as state-of-the-art set enrichment analysis methods, it provides an alternative for interpreting gene expression measurements and for deriving first insights into the relevant processes. The advantages of GGEA will increase in the future with the availability of better GRNs and better models for regulatory relations in these GRNs.

3.3 Improvements of GGEA

Gene Graph Enrichment Analysis (GGEA) has been defined in the previous section as a method for the detection of functional gene sets that are consistently regulated. This section introduces several significant enhancements to the original method improving its applicability, outcome and runtime.

3.3.1 Significance approximation

As described in Section 3.2.2, the statistical significance of GGEA's consistency score for a gene set under investigation is estimated via a permutation approach that is based on resampling of the subjects. Given a sufficiently large number of permutations, typically between 1,000 and 10,000, it allows for an accurate estimation of the underlying score distribution. However, according to Larson and Owen [143], such a permutation approach also has the following disadvantages:

- *Cost*: computationally expensive, as typically several statistics have to be recomputed,
- *Randomness*: random shuffling of the same setting can result in different p-values,
- *Granularity*: the smallest possible p-value is $\frac{1}{M}$ for M permutations.

Randomness and granularity are theoretical issues, becoming less pronounced with an increasing number of permutations. Cost is a practical issue as computing many permutations are time-consuming also for GGEA. In the following, I demonstrate how the significance of resulting gene sets can be estimated by adapting a statistical model that avoids permutation testing of high computational cost. Depending on the data, this can speed up the runtime of GGEA from minutes to seconds.

The raw GGEA score of a gene set is computed as the sum of edge consistencies $C(t) \in [-1, 1]$ (see again Equation 3.12 in Section 3.2.2). Assuming that the edge consistencies are normally distributed

$$C(t) \sim \mathcal{N}(0, \sigma^2) \tag{3.15}$$

with mean $\mu = 0$ and variance σ^2, the sum score S for a gene set with n edges

$$S = \sum_{i=1}^{n} C(t_i) \sim \sum_{i=1}^{n} \mathcal{N}(0, \sigma^2) \sim \mathcal{N}\left(\sum_{i=1}^{n} 0, \sum_{i=1}^{n} \sigma^2\right) \sim \mathcal{N}(0, n\sigma^2) \tag{3.16}$$

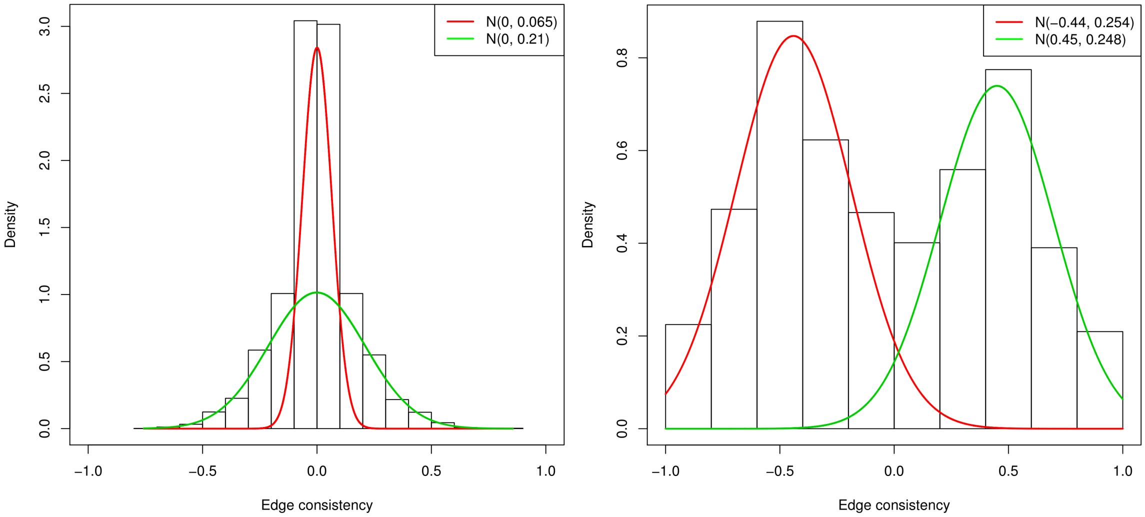
is also normally distributed with mean $\mu = 0$ and variance $n\sigma^2$.

Thus, the statistical significance of an observed score s can be estimated as

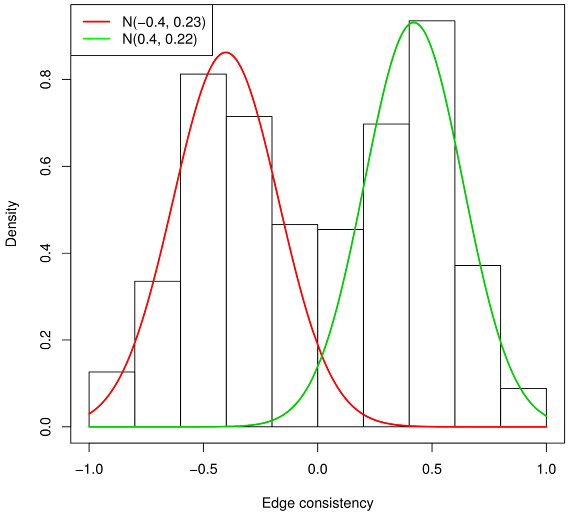
$$Pr(S > s) = 1 - Pr(S \leq s) = 1 - F\left(s, 0, n\sigma^2\right) \tag{3.17}$$

where F denotes the normal cumulative distribution function.

These considerations are based on the assumption of normality in Equation 3.15.



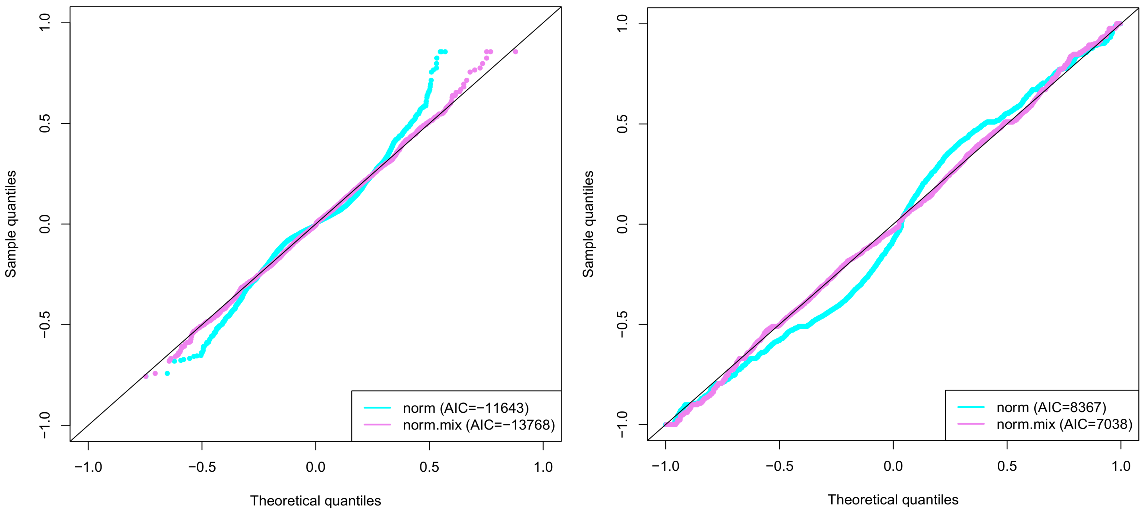
(a) *H. sapiens* (Leukemia, KEGG) (b) *E. coli* (pH shift, RegulonDB)



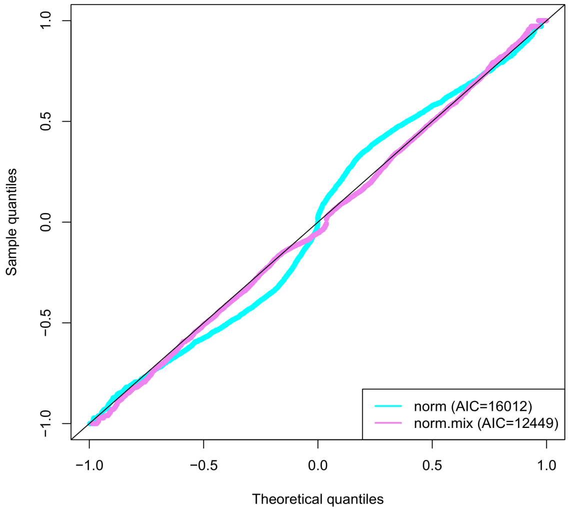
(c) *S. cerevisae* (Gentamicin, YEASTRACT)

Figure 3.7: Edge consistency distribution of different data sets and regulatory networks. The histograms in the background depict the observed edge consistencies of **(a)** 23,858 regulatory interactions annotated in human `KEGG` pathways [41] that were evaluated on acute lymphoblastic leukemia (ALL) expression data [144], **(b)** 5,784 interactions in the `RegulonDB` [32] on pH alteration data [145], and **(c)** 11,186 interactions in `YEASTRACT` [33] on gentamicin treatment data [146]. The green and the red curve show the respective fitted mixtures of two normal distributions with indicated mean and standard deviation.

The expression data was downloaded from `GEO` [147] for *E. coli* (GDS1827) and *S. cerevisae* (GDS2999). ALL data for *H. sapiens* was used as preprocessed in Bioconductor's `ALL` package [148].



(a) *H. sapiens* (Leukemia, KEGG) (b) *E. coli* (pH shift, RegulonDB)



(c) *S. cerevisae* (Gentamicin, Yeastract)

Figure 3.8: Goodness of fit for the mixtures fitted in Figure 3.7. The quantile-quantile plots show for each fitted mixture (violet) how well its distribution (theoretical quantiles, x-axis) agrees with the observed edge consistency distribution (sample quantiles, y-axis). The black diagonal indicates a perfect fit, i.e. complete agreement between fitted and observed distribution. For comparison, the cyan distribution depicts how well a fit of a single normal distribution – instead of a mixture – agrees with the observed consistency distribution. The Akaike information criterion (AIC), indicating the relative quality of both models, is shown in the respective legend.

However, inspection of different data sets and regulatory networks shows that the edge consistency distribution rather follows a mixture of two normal distributions (Figure 3.7 and 3.8). Equation 3.15 thus becomes

$$C(t) \sim \lambda_1 \mathcal{N}(\mu_1, \sigma_1^2) + \lambda_2 \mathcal{N}(\mu_2, \sigma_2^2) \tag{3.18}$$

where λ_1 and λ_2 denote the weights by which the two normal distributions are mixed. Both are roughly equal to 0.5 for the mixtures depicted in Figure 3.7.
Revisiting Equations 3.16 and 3.17 yields

$$\begin{aligned} S &= \sum_{i=1}^{n} C(t_i) \\ &\sim \sum_{i=1}^{n} \left[\lambda_1 \mathcal{N}(\mu_1, \sigma_1^2) + \lambda_2 \mathcal{N}(\mu_2, \sigma_2^2)\right] \\ &\sim \lambda_1 \sum_{i=1}^{n} \mathcal{N}(\mu_1, \sigma_1^2) + \lambda_2 \sum_{i=1}^{n} \mathcal{N}(\mu_2, \sigma_2^2) \\ &\sim \lambda_1 \mathcal{N}\left(\sum_{i=1}^{n} \mu_1, \sum_{i=1}^{n} \sigma_1^2\right) + \lambda_2 \mathcal{N}\left(\sum_{i=1}^{n} \mu_2, \sum_{i=1}^{n} \sigma_2^2\right) \\ &\sim \lambda_1 \mathcal{N}\left(n\mu_1, n\sigma_1^2\right) + \lambda_2 \mathcal{N}\left(n\mu_2, n\sigma_2^2\right) \end{aligned} \tag{3.19}$$

and

$$Pr(S > s) = 1 - \lambda_1 F\left(s, n\mu_1, n\sigma_1^2\right) - \lambda_2 F\left(s, n\mu_2, n\sigma_2^2\right). \tag{3.20}$$

The statistical significance of an observed consistency score for a gene set under investigation can thus be approximated by a mixture of two normal distributions with means $n\mu_1$, $n\mu_2$ and variances $n\sigma_1^2$, $n\sigma_2^2$. These parameters result from multiplying the number of edges in the set with corresponding means and variances from the underlying edge consistency mixture.

Figure 3.8 shows the goodness of fit for the mixtures fitted in Figure 3.7. Visual inspection of the quantile-quantile plots indicates that the mixtures closely approximate the observed edge consistency distributions. Comparing the resulting AIC, given at the bottom right of each plot, also demonstrates that the approximation is significantly better than it would be achieved by fitting a single normal distribution. Thus, in case the distribution assumption in Equation 3.18 holds, the analytical approximation represents a fast and accurate alternative to the permutation approach. However, if the observed edge consistency distribution significantly deviates from a normal mixture, the permutation approach remains the method of choice for estimating the significance of the gene set consistency score.

3.3.2 Edge selection

Edge types

The GGEA score for a gene set is originally calculated over all edges involving a gene set member, i.e. including edges that are coming from or pointing to a gene outside of the

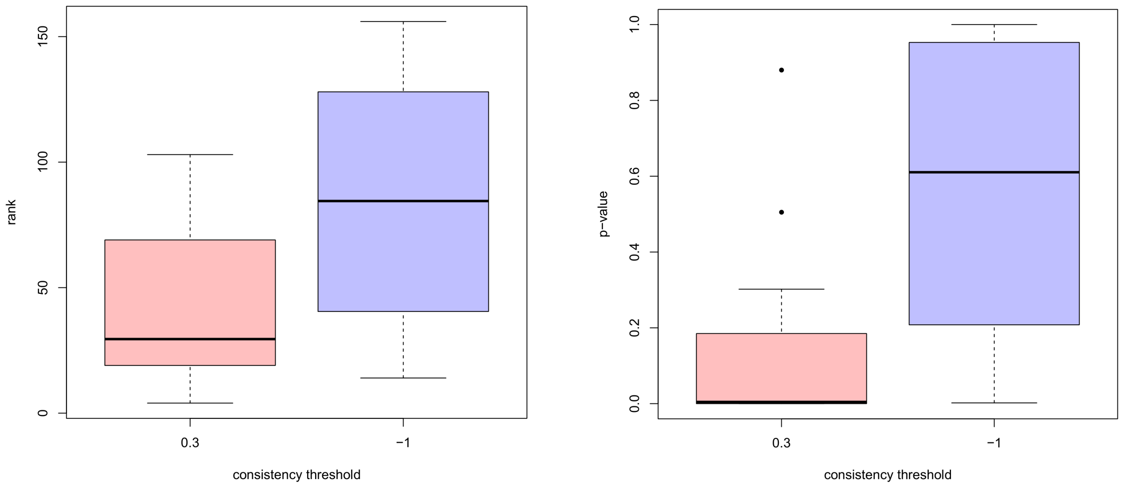


Figure 3.9: Consistency threshold evaluation on the KEGGdzGEO benchmark set. The boxplots show resulting ranks (left panel) and p-values (right panel) of the predefined target pathways, assigned to the 24 datasets listed in Table 3.5, when applying GGEA and including all edges (consistency threshold of -1, blue) or only edges above a consistency threshold of 0.3 (red).

set. Depending on the investigation scope, this is not always desired as most enrichment methods only take gene set members and edges between them into account. A similar option is thus required for GGEA to decide which edges of the GRN are considered for a gene set.

Hence, the induced subnetwork for a gene set $u \in U$ (see again Equation 3.9) is constructed in dependence on an option $o \in \{\wedge, \vee\}$ (logical AND and OR) by

$$\mathrm{GRN}(u) := \{t \in \mathrm{GRN} \mid o(\mathrm{in}(t) \in u, \mathrm{out}(t) \in u)\}. \tag{3.21}$$

This accordingly includes edges for which regulator AND / OR target gene are members of the investigated gene set.

Consistency threshold

Originally, GGEA detects sets where consistent regulation excels inconsistent regulation. A side effect is that many inconsistencies of small effect can dilute strong consistent effects. The application of a consistency threshold can thus be a reasonable alternative to filter out negligible effects.

To demonstrate the effect of a suitably chosen threshold, I apply GGEA to a benchmark set of 24 disease expression datasets from `GEO` (Appendix C.2). The datasets were collected by Tarca *et al.* [149] and a corresponding target pathway from `KEGG` has been assigned to each of them (Table 3.5). These pathways are expected to be significantly perturbed as each of them represents the disease investigated in the respective dataset.

Figure 3.9 shows how well GGEA discovers the target pathways in dependence of the

Table 3.5: The 24 datasets provided by Tarca *et al.* [149] as a benchmark set for gene set enrichment analysis. Adapted from Table 1 of the cited publication. Datasets were used as preprocessed in Bioconductor's `KEGGdzPathwaysGEO` package [150].

	Target pathway	KEGG.ID	GEO.ID
1	Alzheimer's disease	hsa05010	GSE1297
2			GSE5281 (EC)
3			GSE5281 (HIP)
4			GSE5281 (VCX)
5	Parkinson's disease	hsa05012	GSE20153
6			GSE20291
7	Huntington's disease	hsa05016	GSE8762
8	Colorectal cancer	hsa05210	GSE4107
9			GSE8671
10			GSE9348
11	Renal cancer	hsa05211	GSE14762
12			GSE781
13	Pancreatic cancer	hsa05212	GSE15471
14			GSE16515
15	Glioma	hsa05214	GSE19728
16			GSE21354
17	Prostate cancer	hsa05215	GSE6956 (AA)
18			GSE6956 (C)
19	Thyroid cancer	hsa05216	GSE3467
20			GSE3678
21	Acute myeloid leukemia	hsa05221	GSE9476
22	Non-small cell lung cancer	hsa05223	GSE18842
23			GSE19188
24	Dilated cardiomyopathy	hsa05414	GSE3585

chosen consistency threshold. Compared to using no threshold, which is equivalent to a threshold of -1, a threshold of 0.3 enables a significantly better detection of the target pathways. They are ranked significantly higher (t-test p-value $= 0.0006$; ks-test p-value $= 0.017$) and are found frequently with high statistical significance.

Chapter 4

Interpretation of detected subnetworks

4.1 Background

The previous chapter introduces GGEA as a method for network-based gene set enrichment analysis. Resulting gene sets, and interactions within and between them, correspond to significantly enriched subnetworks of the global gene regulatory network. These subnetworks typically represent functional modules of interplay between genes from which conclusions about molecular mechanisms underlying the investigated phenotype can be drawn. Subsequent basic interpretation steps typically include (1) manual inspection of these subnetworks, and (2) biological reasoning based on known associations of contained genes with the phenotype under study (see e.g. Section 3.2.3, subsection *Case Study*).

Manual inspection of the detected subnetworks is facilitated by specific tools for a network-based data view [151, for a review]. This includes general visualization and exploration tools for networks such as `Cytoscape` [152]. On the other hand, there are also tools specifically designed for particular databases such as `pathview` [153], which allows visualization of differential expression in `KEGG` pathway maps, and `GOrilla` [154], which incorporates the `GO` tree structure in the visualization. Using these tools enables an immediate identification of genes, interactions and larger modules that are affected in the expression data under investigation.

Biological reasoning with the identified network components requires sufficient expertise with respect to the investigated phenotype and comprehensive knowledge of the relevant literature [155]. Results can then be divided into observations that are supported by previous findings and novel hypotheses about the underlying mechanisms and drivers of the observed expression. However, for common diseases like cancer and diabetes there is typically a large literature corpus. Thus, reasoning with previous findings often remains incomplete and also biased towards *a priori* established investigation goals. This can be improved by using automated components of the interpretation procedure such as text mining for the selection of relevant articles [34].

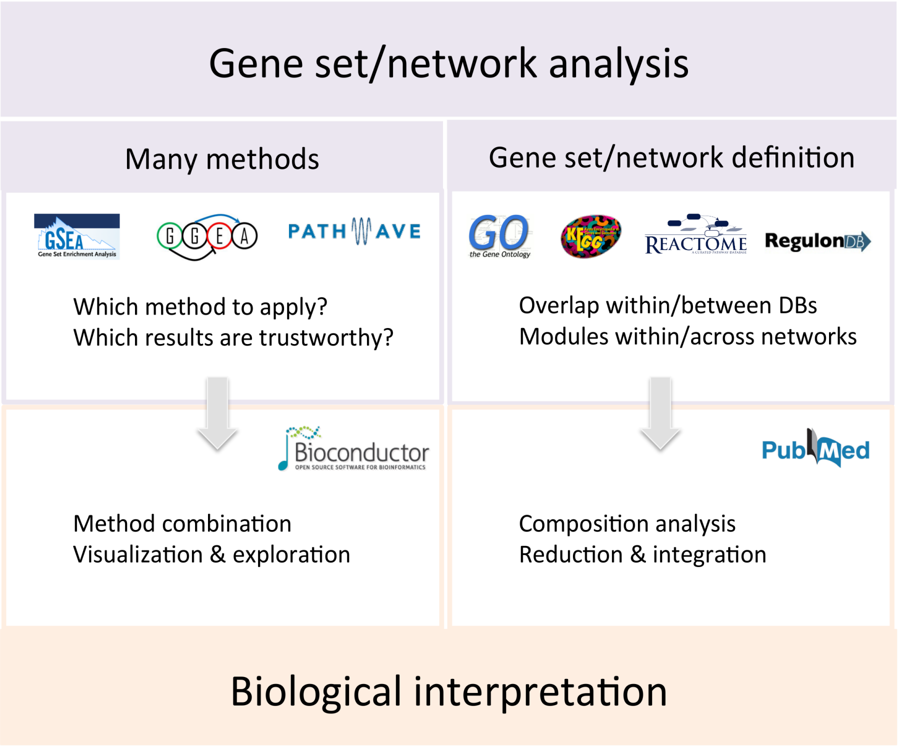

Gene set/network analysis
Many methods
Gene set/network definition
GSEA
GGEA
PATHWAVE
GO
KEGG
REACTOME
RegulonDB
Which method to apply?
Which results are trustworthy?
Overlap within/between DBs
Modules within/across networks
Bioconductor
PubMed
Method combination
Visualization & exploration
Composition analysis
Reduction & integration
Biological interpretation

Figure 4.1: Overview.

Although the described basic interpretation is valuable for deriving first insights and selecting candidates for further targeted investigations, two major issues have to be considered (Figure 4.1):

- Numerous methods are available to identify significantly altered subnetworks [21,28]. The choice for a particular method is not straightforward and can have a great impact on the result and conclusions drawn from it.
- Predefined gene set and network definitions overlap considerably [149,156]. In addition, it is frequently observed that only parts of the predefined gene sets and networks are altered. This can lead to confusingly large and redundant result networks, which are typically not easily reduced to the non-redundant and active part.

In the following, I address both of these issues and describe existing solutions based on recent publications in the field. In Section 4.2, I introduce the `EnrichmentBrowser` as a software package that allows to combine and explore results across methods. By accumulating evidence from multiple methods, the confidence in detected subnetworks is increased and biological conclusions can be drawn more robustly. Subsequently, in Section 4.3, I review methods for the composition analysis of resulting gene sets and networks. I focus on reduction and integration approaches and discuss how they can be exploited as extensions for the `EnrichmentBrowser`.

4.2 Combination and exploration of results

This section describes a software package for enrichment analysis. It implements several state-of-the-art methods and allows to combine and explore results across methods. The work has been included in the recognized Bioconductor repository on October 15th, 2014, and presented on January 13th, 2015, at the European Bioconductor Developers Meeting. The following is based on its original publication in the journal BMC Bioinformatics*, 17:45, January 2016.*

Authors: Ludwig Geistlinger, Gergely Csaba, and Ralf Zimmer

Title: Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis

Abstract: Enrichment analysis of gene expression data is essential to find functional groups of genes whose interplay can explain experimental observations. Numerous methods have been published that either ignore (set-based) or incorporate (network-based) known interactions between genes. However, the often subtle benefits and disadvantages of the individual methods are confusing for most biological end users and there is currently no convenient way to combine methods for an enhanced result interpretation. We present the `EnrichmentBrowser` package as an easily applicable software that enables: (i) the application of the most frequently used set-based and network-based enrichment methods; (ii) their straightforward combination; and (iii) a detailed and interactive visualization and exploration of the results. The package is available from the Bioconductor repository and implements additional support for standardized expression data preprocessing, differential expression analysis, and definition of suitable input gene sets and networks. The `EnrichmentBrowser` package implements essential functionality for the enrichment analysis of gene expression data. It combines the advantages of set-based and network-based enrichment analysis in order to derive high-confidence gene sets and biological pathways that are differentially regulated in the expression data under investigation. Besides, the package facilitates the visualization and exploration of such sets and pathways.

Author contributions LG implemented the package and wrote the manuscript. GC tested and evaluated the package and suggested several modifications. RZ reviewed the manuscript and supervised the project.

4.2.1 Background

Genome-wide gene expression studies with microarrays or RNA-seq typically measure several thousand genes at a time [11]. This makes biological interpretation challenging. To approach this task several statistical filters can be applied to obtain an easier tractable number of genes and to concentrate further investigation effort on genes that are differentially expressed. Subsequently, analysis focuses on whether disproportionately many of the remaining genes belong to known functional sets of genes. Such an enrichment for certain gene functions, sets or pathways immediately generates important hypotheses about underlying mechanisms of e.g. an investigated clinical phenotype.

A recent review divides existing enrichment methods into three generations [21]. The first generation of methods is centered around the traditionally used overrepresentation analysis, which tests based on the hypergeometric distribution whether genes above a predefined significance threshold are overrepresented in functional gene sets [27]. The second generation of methods resolves the restriction to the subset of significant genes, and instead scores the tendency of gene set members to appear rather at the top or bottom of the ranked list of all measured genes [123].

First and second generation methods have in common that they ignore known interactions between genes. Those methods are thus denoted as *set-based* in this manuscript. Methods that do incorporate known interactions belong to the third generation of methods and are denoted as *network-based* in the following (reviewed in [28]).

While each generation is represented by numerous published methods with individual benefits and disadvantages, there is currently no gold standard enrichment method agreed upon. This makes the decision for a particular method intricate. It also leads users, actually intending a better biological understanding of their data, to decide based on criteria not necessarily relating to biological insight such as frequency of usage and ease of application.

Combination of methods has been proven superior to individual methods in different areas of computational biology, as it increases performance [35] and statistical power [157] and biological insights often complement each other [11, 158].

In this article, we propose and implement the straightforward combination of major set- and network-based enrichment methods. We demonstrate that this filters out spurious hits of individual methods and reduces the outcome to candidates accumulating evidence from different methods. This increases the confidence in resulting enriched gene sets, and, thus, substantially enhances the biological interpretation of large-scale gene expression data.

4.2.2 Implementation

The `EnrichmentBrowser` is implemented in the statistical programming language `R` [159] and the package is included in the open-source `Bioconductor` project [132].

Overview

Given gene expression data sampling different conditions, specific functional gene sets, and optionally a gene regulatory network, the `EnrichmentBrowser` performs three essential steps: (1) chosen set- and network-based enrichment methods are executed individually, (2) enriched gene sets are combined by selected ranking criteria, and (3) resulting gene set rankings are displayed as HTML pages for detailed inspection (Figure 4.2).

Data preprocessing

The typical starting point for the `EnrichmentBrowser` is normalized gene expression data. The data are usually microarray intensity measurements or RNA-seq read counts for several thousand genes between two conditions, each represented by a group of samples.

Two inputs are required: (1) the expression matrix, in which each row corresponds to a gene and each column to a sample, and (2) a binary classification vector dividing the samples in cases and controls. In case of paired samples or sample blocks, e.g. indicating different treatments of cases and controls, a vector defining the blocks can optionally be supplied.

While each dataset typically shows individual characteristics that need to be specifically normalized for, the `EnrichmentBrowser` provides several well-established standard routines for that purpose. This includes within-array/-lane and between-array/-lane normalization for microarray and RNA-seq data based on functionality from the `limma` [160] and `EDASeq` [161] package, respectively.

In case of microarray data, once it has been read in, the data is transformed from probe to gene level. This incorporates a mapping from probe to gene identifiers, which is automatically done for recognized platforms (i.e. a corresponding Bioconductor annotation package such as `hgu95av2.db` [162] exists) or required as a user input otherwise. An important parameter is the summarization method determining, in case of multiple probes for one gene, whether an average value is computed or the probe that discriminates the most between the two sample groups is kept.

Differential expression

Differences in gene expression between the two sample groups are computed using established functionality from the `limma` package [160], involving the voom transformation [163] when applied to RNA-seq data. Alternatively, differential expression analysis for RNA-seq data can also be carried out based on the negative binomial distribution with `edgeR` [164] and `DESeq2` [165]. Resulting log2 fold changes and derived p-values for each gene can be inspected in several ways (Figure 4.3). This includes a gene report that lists the respective measures of differential expression for each gene, and several overview graphics such as:

- Heatmap: clustered overview of gene expression between the two groups for all genes, and separately, only for significant genes satisfying predefined thresholds of fold change and p-value.

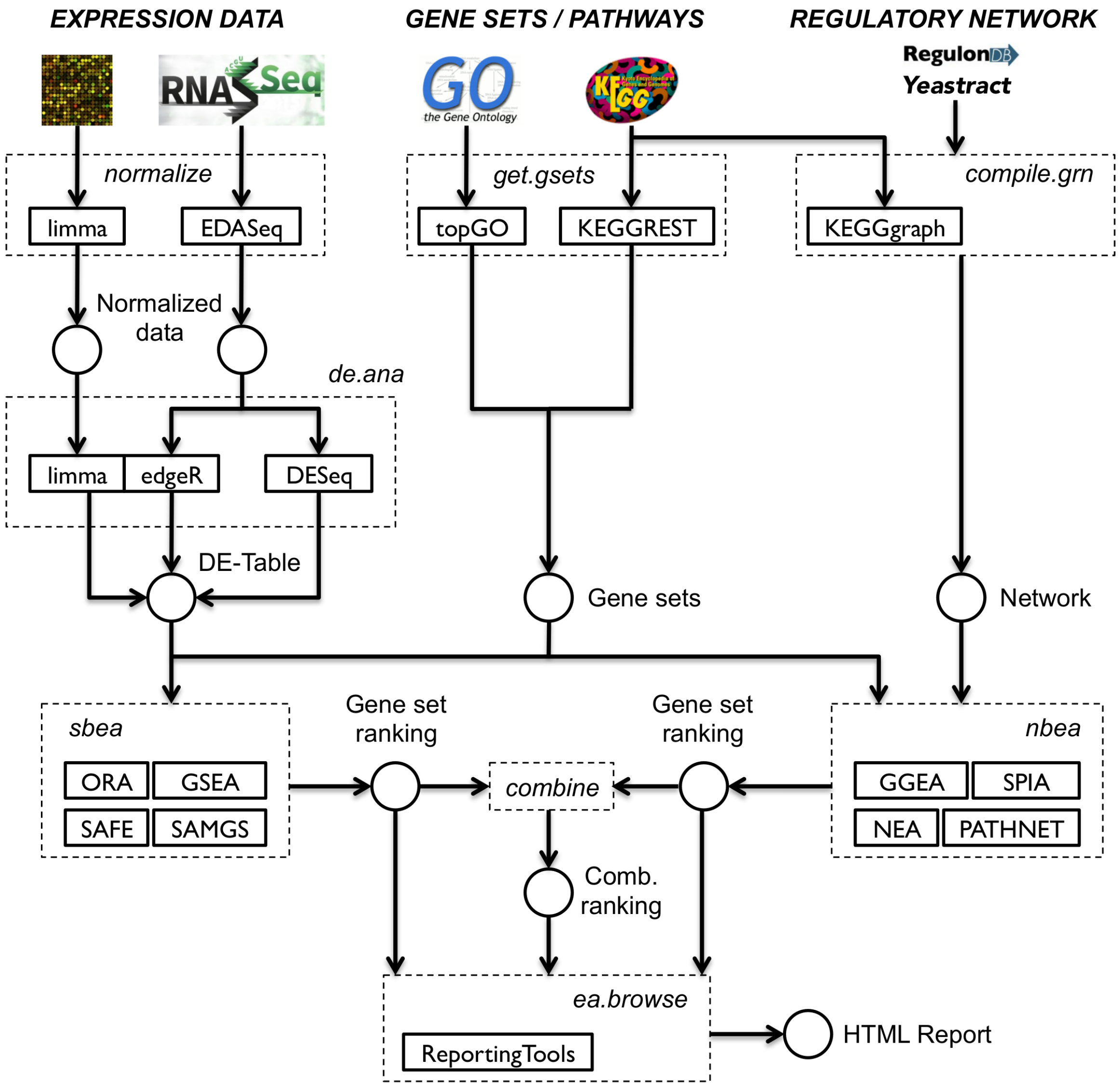


Figure 4.2: Workflow. Expression data as measured with microarrays or RNA-seq is tested for enrichment of specific functional gene sets e.g. as defined in the Gene Ontology or the `KEGG` pathway annotation. Additional information from regulatory networks annotated in specific databases such as the RegulonDB or Yeastract can be exploited. Implemented methods can be carried out individually and combined by selected ranking criteria. Resulting gene set rankings can be browsed as HTML pages allowing detailed inspection (as illustrated in Figure 4.3).

- Volcano plot: fold change versus p-value plot that illustrates the correspondence of amount, direction and statistical significance of expression changes (supporting immediate identification and exploration of significant genes by mouse-over and linking to the corresponding gene entries).
- P-value distributions: histogram of raw and adjusted p-values.

Multiple testing correction is performed using the `p.adjust` function from the `stats` package, which implements several frequently used corrections from which the user can choose (reviewed in [166]). This includes the stringent Bonferroni correction and the less conservative method of Benjamini and Hochberg.

Enrichment analysis

Set-based enrichment analysis

The `EnrichmentBrowser` implements several ways to assemble the gene sets that should be tested for enrichment. User-defined gene sets can be parsed from suitable formats such as GMT [167] or extracted from pathway XML format [168]. Frequently used organism-specific gene sets from `GO` [22] and `KEGG` [41] can be downloaded exploiting functionality from the `topGO` [169] and `KEGGREST` [170] package, respectively.

Currently supported are the following major set-based enrichment methods:

- ORA: Overrepresentation Analysis, simple and frequently used test based on the hypergeometric distribution (reviewed in [27]),
- SAFE: Significance Analysis of Function and Expression, implements a resampling version of ORA, includes other test statistics such as Wilcoxon's rank sum, and allows to estimate the significance of gene sets by sample permutation [125],
- GSEA: Gene Set Enrichment Analysis, frequently used and widely accepted, uses a Kolmogorov-Smirnov statistic to test whether the ranks of the p-values of genes in a gene set resemble a uniform distribution [123],
- SAMGS: Significance Analysis of Microarrays on Gene Sets, extending the SAM method for single genes to gene set analysis [171].

ORA is a first generation method, whereas SAFE, GSEA, and SAMGS belong to the second generation of enrichment methods. The `EnrichmentBrowser` uses its own implementation of ORA, while it integrates SAFE as implemented in the `safe` package. Implementations of GSEA and SAMGS are adapted from [172, 173], respectively.

SAFE, GSEA, and SAMGS use sample permutation for estimating the gene set significance, which involves recomputation of their individual local t-like statistics for each gene. As this is not *per se* suitable for RNA-seq read count data, the `EnrichmentBrowser` provides specific local statistics based on the `limma/voom`-transformed t-statistic, the LR-statistic from `edgeR`, and the Wald-statistic from `DESeq`. Global statistics for each gene set are accordingly chosen as the KS-statistic (for GSEA), Wilcoxon's rank sum (for SAFE), and Hotelling's T^2 (for SAMGS). Permutation testing with selected local and global statistics is carried out using the general framework implemented in the `safe` package.

Network-based enrichment analysis

Gene regulatory networks represent known interactions between genes as derived from specific experiments or compiled from the literature [174]. There are well-studied processes

and organisms for which comprehensive and well-annotated regulatory networks are available, e.g. the `RegulonDB` for *E. coli* [32] and `Yeastract` for *S. cerevisiae* [33]. While it is recommended to use these specific networks, and the `EnrichmentBrowser` supports their download and formatting, there are also cases where such a network is not easily available. For these cases the `EnrichmentBrowser` implements the possibility to compile a network from regulatory interactions annotated in the `KEGG` database. This incorporates downloading and parsing of the pathways for a selected organism making use of the `KEGGREST` and `KEGGgraph` package [175], respectively.
Currently integrated network-based enrichment analysis methods are

- GGEA: Gene Graph Enrichment Analysis, evaluates consistency of known regulatory interactions with the observed expression data [80],
- SPIA: Signaling Pathway Impact Analysis, combines ORA with the probability that expression changes are propagated across the pathway topology [176],
- NEA: Network Enrichment Analysis, applies ORA on interactions instead of genes [177],
- PathNet: Pathway analysis using Network information, applies ORA on combined evidences of the observed signal and the signal implied by connected neighbors in the network [178].

GGEA is the default network-based enrichment method of the `EnrichmentBrowser` and is also incorporated in the network-based visualization of gene sets. SPIA, NEA, and PathNet are integrated as implemented in the `SPIA`, `neaGUI`, and `PathNet` package, respectively.

Generic plug-in of additional methods

The goal of the `EnrichmentBrowser` is to provide the most frequently used enrichment methods. However, it is also possible to exploit its functionality with additional methods not among the currently implemented ones. This requires to implement a function that takes the characteristic arguments `eset` (expression data), `gs` (gene sets), `alpha` (significance level), and in case of network-based enrichment also `grn` (gene regulatory network). In addition, it must return a vector storing the resulting p-value for each gene set in `gs`.

Combining results

Different enrichment analysis methods usually result in different gene set rankings for the same dataset. To compare results and detect gene sets that are supported by different methods, the `EnrichmentBrowser` package allows to combine results from the different set- and network-based enrichment methods. The combination of results yields a new ranking of the gene sets under investigation according to a defined ranking and combination function. The ranking function determines by which statistic the individual gene set rankings are sorted and which type of ranks are computed. The ranking statistic is typically chosen to be the gene set p-value or score (sorted in ascending and descending order, respectively). Predefined rank types include:

- *Absolute* ranks r_A are assigned from 1 to n according to the sorting of the ranking statistic. Intuitively, n is identical to the number of gene sets N_{GS} if the ranking statistic takes a different value for each gene set. As ties can occur, which yields the *same* rank for gene sets with equal value, n corresponds to the number of distinct values of the ranking statistic (denoted as N_D).
- To account for a differing number of gene sets in the individual rankings, *relative* ranks r_R can be derived from absolute ranks via $r_A/n \cdot 100$.
- Although frequently used to rank gene sets, absolute and relative ranks can be misleading in case of extensive presence of ties. Especially, when comparing a coarse-grained ($N_D \ll N_{GS}$) and a fine-grained ranking ($N_D \approx N_{GS}$). Here, similar absolute/relative ranks imply a very different meaning. To resolve this, we introduce *competitive* ranks r_C calculated as the percentage of gene sets with a value of the ranking statistic at least as extreme as observed for the gene set to be ranked.

The default ranking function returns competitive ranks based on gene set p-values.
The combination function determines how ranks are combined across methods and can be chosen from predefined functions such as `mean`, `median`, `min`, and `sum` (default). User-defined ranking and combination functions can also be plugged in.

Visualization and exploration

The standard output of existing enrichment tools is a ranking of the gene sets by the corresponding p-value. The `EnrichmentBrowser` package provides additional visualization and interactive exploration of resulting gene sets far beyond that point. Based on functionality from the `ReportingTools` package [179], the resulting flat ranking can be accompanied by a HTML report from which each gene set can be inspected in detail (Figure 4.3).
Instead of providing visualization capabilities for each method, the `EnrichmentBrowser` implements general set- and network-based visualizations (SBEA and NBEA page). They represent results of methods of the corresponding class, but can also be incorporated independent of the enrichment method executed. It is thus possible to carry out e.g. set-based methods, while including a network-based visualization of significant gene sets in the result report.
The SBEA page is composed as described for the global differential expression report (the set of all measured genes). Thus, a gene set under study is visualized with an interactive volcano plot alongside 2 heatmaps for all and only differentially expressed set members.
The composition of the NBEA page depends on the gene set source and whether a regulatory network is available. For `KEGG` gene sets, differential expression is visualized directly on the pathways by overplotting the original pathway maps with `pathview` [153]. In addition, connected subgraphs within a pathway are displayed separately and can be inspected by mouse-over (involves the `imageMap` function from `biocGraph` [180]).
In case a regulatory network has been provided, gene sets can also be viewed as GGEA graphs. Such a graph displays for a gene set of interest the consistency of each interaction

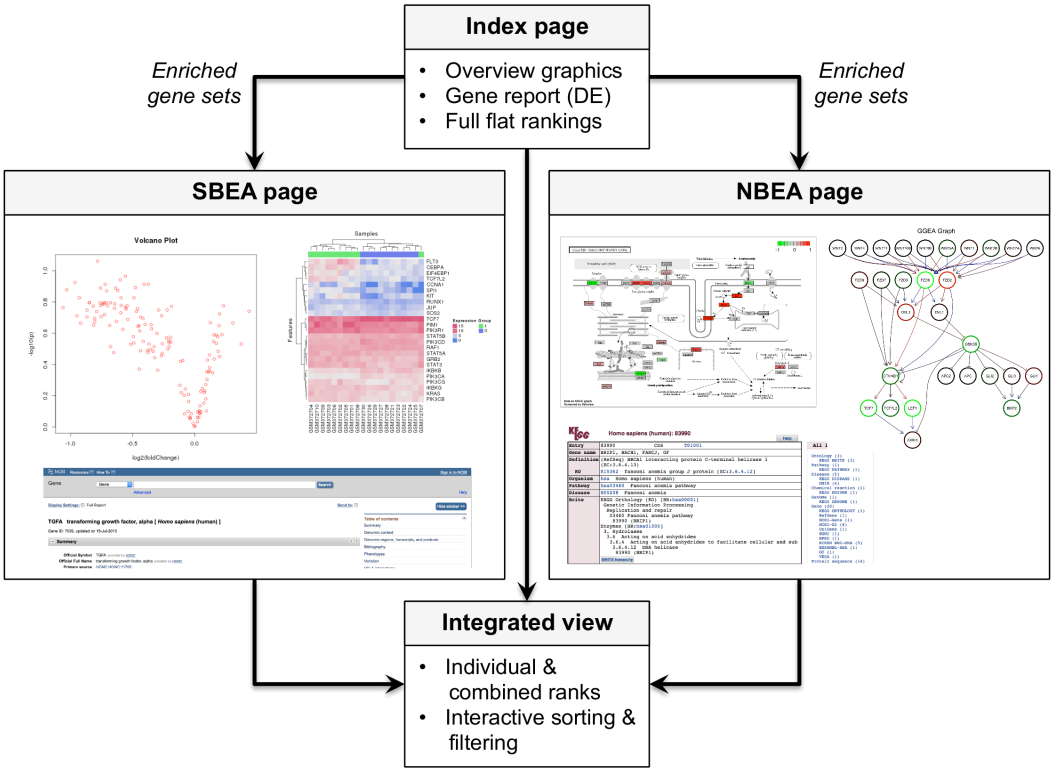


Figure 4.3: Navigation. Structured access to enrichment analysis results is provided by an index page that links overview graphics, a gene report that includes measures of differential expression for each gene under investigation, and the full flat gene set rankings for each executed method (and, optionally, their combination). In addition, a top table for each method is linked containing detailed set-based (SBEA page) and network-based (NBEA page) views of significant gene sets. The SBEA page for a gene set is composed of (1) an interactive volcano plot (fold change vs. DE p-value) allowing immediate identification of significant genes by mouse-over, and (2) two heatmaps displaying the expression of all set genes and the subset of significant genes. The NBEA page illustrates interactions within a gene set by projecting it onto the underlying regulatory network, and for `KEGG` gene sets by additionally highlighting corresponding pathway maps. Results of method combination are linked in a combined page that displays the combined ranking alongside the individual rankings, and which can be interactively sorted and filtered according to user-selected criteria. See Additional file 2-4 for several application examples and the vignette of the EnrichmentBrowser package for details of the various options.

in the network that involves a gene set member. Nodes (genes) are colored according to expression (up-/down-regulated) and edges (interactions) are colored according to consistency, i.e. how well the interaction type (activation/inhibition) is reflected in the correlation of the observed expression of both interaction partners (see the legend in Additional file 1, Figure S1). Although GGEA graphs have been originally implemented for illustrating gene sets according to GGEA, they are apparently useful for depicting mechanisms exploited by network-based methods in general.

The combined result view additionally enables an interactive ranking either based on the combined ranks across methods, or with respect to one of the chosen methods.

4.2.3 Results and Discussion

In the following, we demonstrate the application of the `EnrichmentBrowser` to microarray and RNA-seq data. Subsequently, we systematically evaluate the individual methods integrated in the package and the effect of combining methods. See Additional file 1 for supplementary material and methods. A comparative evaluation to existing `Bioconductor` packages and stand-alone tools such as `SegMine` [181] and `graphite web` [182] can also be found in Additional file 1.

Application example 1: ALL microarray data

To demonstrate the functionality of the package for microarray data, we consider expression values of patients suffering from acute lymphoblastic leukemia [144]. A frequent chromosomal defect found among these patients is a translocation, in which parts of chromosome 9 and 22 swap places. This results in the oncogenic fusion gene BCR/ABL created by positioning the ABL1 gene on chromosome 9 to a part of the BCR gene on chromosome 22.
The data is available from Bioconductor in the `ALL` data package [148] and contains normalized intensity measurements on a log-scale for 12,625 probes across 79 patients. Case and control group were defined according to presence or absence of the BCR-ABL gene fusion. We use functionality of the `EnrichmentBrowser` for transformation from probe to gene level and differential expression analysis (see Implementation, section *Data preprocessing* and *Differential expression*). Human `KEGG` pathways were downloaded as gene sets, i.e. ignoring interactions between genes.
We apply ORA to detect overrepresented `KEGG` pathways using the default significance level α of 0.05 (Table 4.1; and Additional file 2 for the detailed HTML summary). Resulting pathways can be divided in three categories: (1) clearly linked to the phenotype such as transcriptional misregulation in cancer, apoptosis and basal cell carcinoma, (2) unknown and secondary effects of phenotype or treatment like myocarditis, which can be caused by cancer radiation therapy, and (3) clearly irrelevant such as legionellosis (drinking water contamination) and shigellosis (foodborne illness).
We investigate next whether these findings can be explained by known regulatory interactions. This means, whether regulators such as transcription factors and their target genes are expressed in accordance to the connecting regulations. Therefore, we apply GGEA using a network of regulations compiled from the `KEGG` database. For comparison, we select the same number of gene sets as for ORA from the top of the GGEA ranking (Table 4.1; and Additional file 2 for the detailed HTML summary).
To identify relevant pathways reported by both methods, we combine the rankings of ORA and GGEA by rank sum, including only gene sets in the intersection of both rankings. This yields a new ranking in which irrelevant pathways such as legionellosis and shigellosis are filtered out (Table 4.1; and Additional file 2 for the detailed HTML summary). Thus,

Table 4.1: Combination of top ranked gene sets of ORA and GGEA by rank sum (ALL microarray data). Shown are the absolute ranks returned by ORA and GGEA and the resulting rank sum in the last column (see Implementation, section *Combining results*).

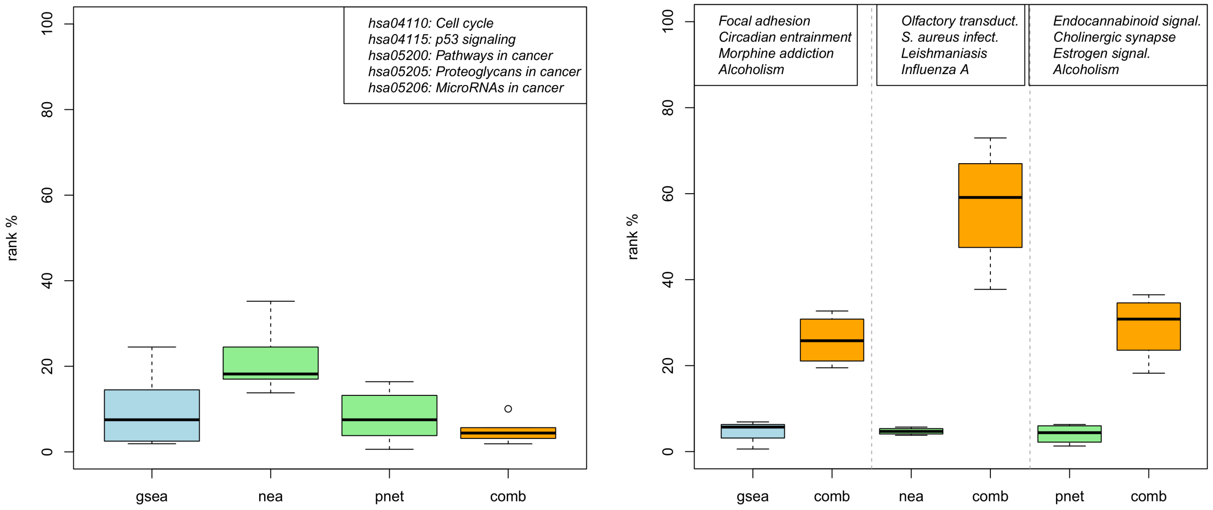
ID	Title	ORA	GGEA	$\sum$
hsa05416	Viral myocarditis	1	1	2
hsa04520	Adherens junction	4	2	6
hsa05217	Basal cell carcinoma	9	3	12
hsa04622	RIG-I-like receptor	2	12	14
hsa04210	Apoptosis	6	10	16
hsa05202	Transcript. misreg. in cancer	7	13	20
hsa05130	Pathogenic E. coli infection	3	-	-
hsa05134	Legionellosis	5	-	-
hsa05131	Shigellosis	8	-	-
hsa05412	Arrhytm. cardiomyopathy	10	-	-
hsa05100	Invasion of epithelial cells	11	-	-
hsa04670	Leukocyte trans. migration	12	-	-
hsa05206	MicroRNAs in cancer	13	-	-
hsa04350	TGF-β signaling	-	4	-
hsa04550	Pluripotency of stem cells	-	5	-
hsa05211	Renal cell carcinoma	-	6	-
hsa04310	Wnt signaling	-	7	-
hsa04660	T cell receptor	-	8	-
hsa05144	Malaria	-	9	-
hsa04514	Cell adhesion	-	11	-

combining ORA with GGEA yields a ranking reduced to the most plausible pathways, which are supported by several mechanistic explanations in the GGEA graphs.

Application example 2: TCGA RNA-seq data

The `EnrichmentBrowser` integrates specific methods for preprocessing and differential expression analysis of RNA-seq data. Accordingly, enrichment methods that rely on sample permutation are adapted to incorporate specific local statistics for recomputation of per-gene differential expression (see Implementation, section *Set-based enrichment analysis*). To demonstrate the functionality, we apply the package for the analysis of RNA-seq data from The Cancer Genome Atlas (TCGA, [183]). We consider here uterine corpus endometrial carcinoma (UCEC), which is one of the most common cancers of the female reproductive system [184].

The data is available from `GEO` under accession GSE62944 [185] and contains integer sequencing read counts for 554 UCEC tumor and 35 adjacent normal samples. We apply the `limma/voom`-based differential expression analysis and make use of the `KEGG` gene set catalogue and regulatory network described in the previous section.



(a) cancer (b) unspecific

Figure 4.4: Method combination consolidates cancer-specfic pathways and downgrades unspecific pathways (TCGA RNA-seq data). (a) Competitive rank distributions of selected cancer-specific pathways (listed top right) for GSEA, NEA, and PathNet when applied to the UCEC RNA-seq data from TCGA. The orange rightmost boxplot depicts the corresponding rank distribution when combining the individual rankings by rank sum. Analogously, **(b)** depicts the rank distributions of selected unspecific pathways (listed at the top of the respective panel) that were top ranked by the 3 individual methods.

For set-based enrichment analysis, we choose GSEA as it is among the methods specifically adapted in the `EnrichmentBrowser` for the analysis of RNA-seq data. In addition, we apply PathNet and NEA for network-based enrichment analysis (full rankings can be found in Additional file 3). To investigate the effect of method combination, we combine the 3 individual gene set rankings by rank sum as for the ALL example.

We find cancer-specific pathways such as *p53 signaling pathway* and *Pathways in cancer* clearly consolidated in the combined ranking (Figure 4.4a). On the other hand, unspecific pathways such as *Olfactory transduction* and *Morphine addiction*, which were top ranked by the individual methods, are distinguishably downgraded in the combined ranking (Figure 4.4b).

Thus, independent of the expression data type under study (microarray/RNA-seq) and the enrichment methods combined (previously: ORA/GGEA, here: GSEA/NEA/PathNet), the combination has shown to improve individual rankings by increasing confidence in specific target pathways and removing irrelevant pathways from the top of the ranking.

Systematic evaluation: GEO2KEGG benchmark set

We have observed beneficial effects of combining enrichment methods at the example of specific microarray and RNA-seq datasets. We investigate next whether these effects can be observed systematically when applied to many datasets.

For that purpose, we use a compendium of 27 `GEO` datasets derived from the *KEGGdzPathwaysGEO* and the *KEGGandMetacoreDzPathwaysGEO* benchmark sets [149,186]. See Additional file 1 for details. These datasets have been specifically selected as they investigate a certain human disease for which a corresponding `KEGG` pathway exists (e.g. Alzheimer's disease). These pathways are thus denoted as the target pathways in the following.

We investigate first how well the individual set- and network-based methods detect the target pathways and, subsequently, whether the detection can be improved by combining methods.

Individual methods

When applying the 8 methods to the 27 datasets of the `GEO2KEGG` benchmark set, an issue of practical relevance is runtime (Figure 4.5). As expected, runtime of the methods depends mainly on whether permutation testing is used to estimate gene set significance, and whether this is efficiently implemented.

ORA, applying the hypergeometric test without permutation, can thus be performed with almost no effort, displaying a constant runtime of around half a second per dataset. GSEA, applied with a default of 1,000 permutations, is slower by two orders of magnitude taking around 1 *min* 40 *sec* per dataset. It should be mentioned that the original GSEA `R`-script [172], which has been straightforward translated from `Java`, is considerably slower. The version integrated in the `EnrichmentBrowser` has been substantially optimized by making use of vectorized calculations. SAFE and SAMGS, taking typically 5-10 *sec* depending on the dataset, although methodically similar to GSEA are much faster as they do not rely on the computationally intensive cumulative KS-statistic. However, using the npGSEA permutation approximation [143] reduces the runtime of GSEA to $\approx$2 *sec* per dataset.

Concerning the network-based methods, SPIA and PathNet display similar runtime as observed for permutation-based GSEA. NEA seems to be inefficiently implemented, requiring already for 10 permutations $\approx$13 *min* on average and up to 2 1/2 days for 1,000 permutations (Additional file 1, Figure S2). On the other hand, the code of GGEA has been highly optimized and yields short computation times. The permutation-based version takes $\approx$4 *sec* per dataset. Using a similar permutation approximation as for GSEA, reduces the runtime of GGEA to $\approx$2 *sec* per dataset.

Resulting gene sets returned by the enrichment methods are typically ranked by gene set p-value. However, given that a method can return the same p-value for more than one gene set impairs a straightforward ranking. This applies especially to permutation p-values, which typically lack a suitable granularity [143]. We have thus introduced competitive ranks, defined as the percentage of gene sets with a p-value at least as extreme as observed for the gene set to be ranked (see Implementation, section *Combining results*).

Competitive rank distributions of the target pathways when applying the 8 methods to the 27 datasets of the `GEO2KEGG` benchmark set are shown in Figure 4.6a. With the exception of SAMGS and, to a lesser extent, NEA, p-value based rankings of the remaining 6 methods appear to well discover the relevance of the target pathways. Their rank distributions are

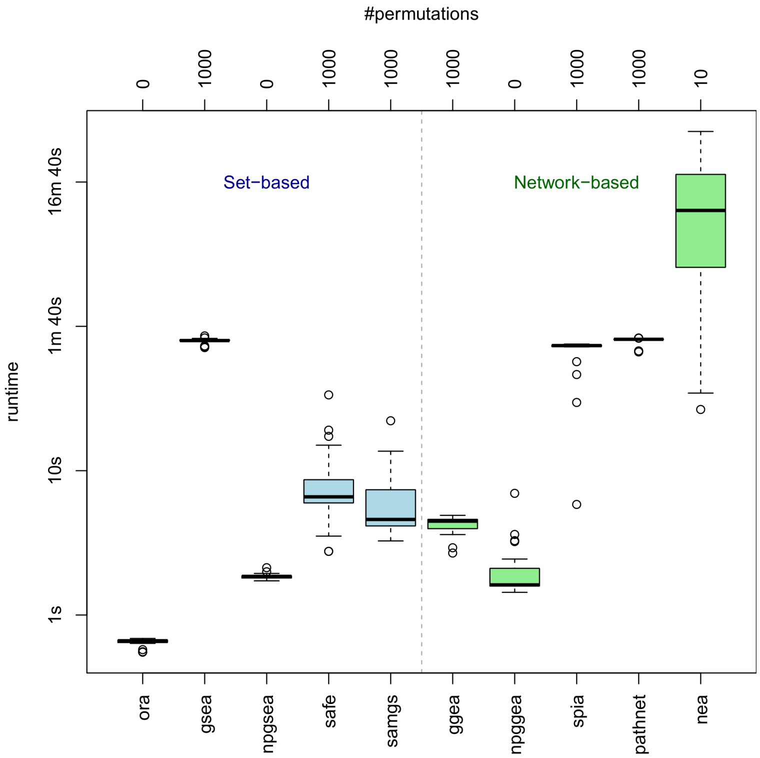
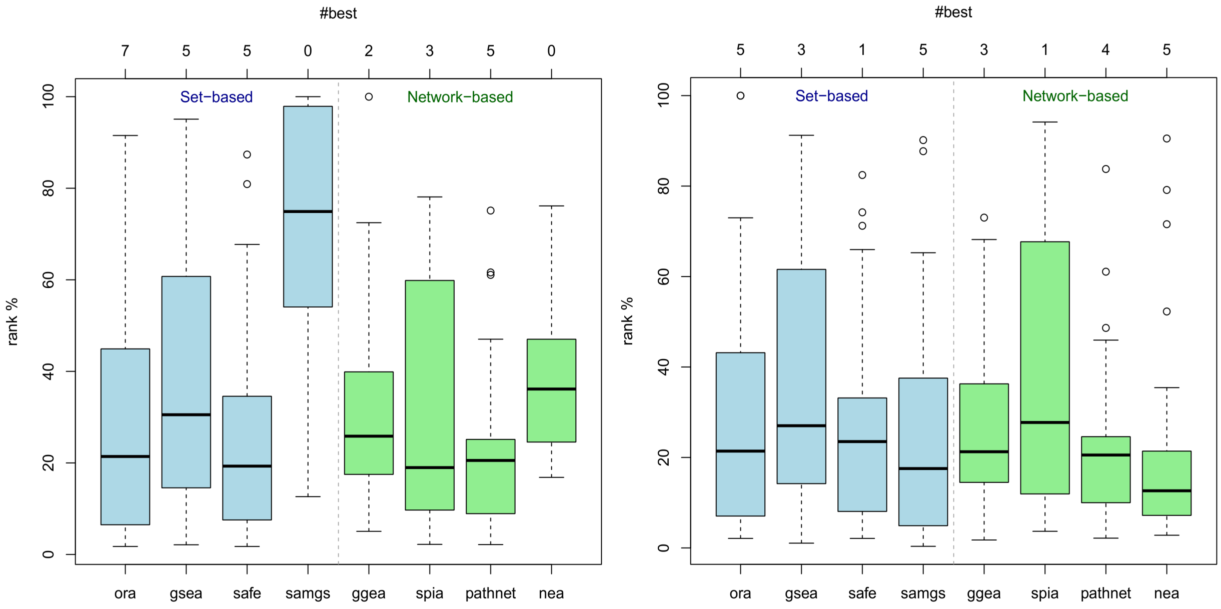


Figure 4.5: Runtime. Shown are the distributions of the elapsed processing times (y-axis, log-scale) when applying the enrichment methods indicated on the bottom x-axis to the 27 datasets of the `GEO2KEGG` benchmark set. The x-axis on top of the plot indicates the number of permutations that have been used to estimate gene set significance. Elapsed runtime of NEA when using 100 and 1,000 permutations, respectively, are shown in Additional file 1, Figure S2.

clearly shifted towards the top of the ranking (median ranging from 19% for SPIA to 30% for GSEA). This can be interpreted as a clear sign for relevance of the target pathways for the corresponding datasets. However, this also shows that there is no clearcut relation between target pathway and dataset as it would be indicated by throughout top rankings of the target pathways. This is presumably due to interfering issues inherent to `KEGG` such as incompleteness of the pathway definition as well as overlap and crosstalk between pathways [149,156]. Nevertheless, there are several notable observations that can be made here:

(1) GSEA, typically assumed superior to ORA by incorporating all measured genes, does not display an increased potential for discovering the target pathways. This indicates that most of the variance observed for these sets is explained by genes that are significantly differentially expressed. (2) Similarly, rank distributions of the network-based methods do not deviate significantly from the set-based methods, although they are typically assumed to better reflect the regulatory mechanisms within sets. This indicates that the `KEGG` network used here is of limited suitability. As it predominantly contains protein-protein interactions and only a small fraction of transcriptional regulatory interactions, quantitative changes



(a) p-value (b) score

Figure 4.6: Evaluation of individual methods on the GEO2KEGG benchmark set. Depicted are competitive rank distributions of the `KEGG` target pathways according to **(a)** gene set p-value and **(b)** gene set score of the individual methods when applied to the 27 `GEO` datasets. The x-axis on top of both plots indicates the number of datasets for which the corresponding method resulted in the best ranking (among methods) of the target pathway. As an example, the leftmost blue boxplot in **(a)** shows for ORA a median rank of ≈20%, i.e. ORA returned for half of the datasets a competitive rank below 20%. Depicted on top, $\#best = 7$ means that ORA returned for 7 datasets the best ranking (among methods) for the target. Detailed rank distributions for each dataset can be found in Additional file 1, Figure S3 and S4.

in transcriptomic data reflecting effects of interactions are presumably rare.
(3) Detailed inspection of the rank distributions shown in Figure 4.6a reveals that none of the methods is best suited for all datasets (Additional file 1, Figure S3). There are ≥2 datasets for each method yielding the best ranking (among methods) of the target pathway. (4) SAMGS returns typically a p-value of zero for 70-80% of the gene sets tested, rendering this p-value a measure not suitable for ranking. As permutation p-values should never be zero [187], usage of this p-value is also not recommended for expressing statistical significance. Similarly, permutation p-values reported by NEA, although obtained with large computational effort (see runtime discussed earlier), appear also not suitable for ranking.
Given the observed issues for NEA and SAMGS when ranking results by gene set p-value, we also ranked the target pathways by gene set score (Figure 4.6b). While ranking of the other 6 methods remained almost invariant (GGEA slightly better, SAFE and SPIA slightly worse), this substantially improved rankings returned by SAMGS and NEA. How-

ever, inspecting the rank distributions for each dataset in more detail (Additional file 1, Figure S4) showed again that no single method consistently returned best rankings. We observed at least one dataset for each method with the best ranking (among methods) of the corresponding target pathway.

Method combination

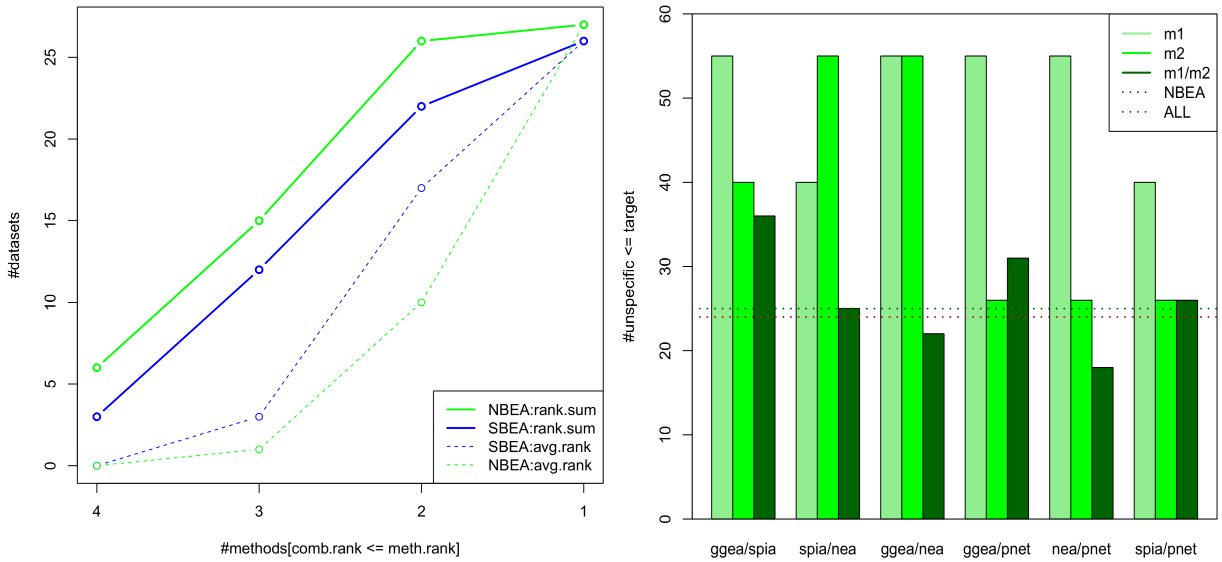
Motivated by the results observed for individual methods in the previous section, we investigated next the effect of combining results of methodically similar methods. Therefore, we computed combined ranks by rank sum for the 4 set- (SBEA) and the 4 network-based (NBEA) methods (Figure 4.7a).

The NBEA-combination yielded for 6 of the 27 datasets (SBEA: 3 datasets) a ranking of the corresponding target pathway, which was at least as good as obtained for all 4 individual methods. Importantly, the combination returned for almost all datasets (SBEA: 26, NBEA: all 27 datasets) a ranking of the corresponding target pathway, which was at least as good as obtained for 1 of the 4 methods. This indicates that combining methods is typically safe, i.e. is rarely worse than the worst individual ranking. On the other hand, the combination resulted in many cases in improved rankings of relevant target pathways. In addition, re-ranking by rank sum yielded significantly better ranks of the target pathways as obtained by simply averaging individual ranks across methods (compare dashed and solid lines in Figure 4.7a).

As observed for the microarray and RNA-seq application example, combination allowed to filter out, i.e. downgrade irrelevant pathways reported by individual methods. Therefore, we counted for all pairwise combinations of the 4 network-based methods the total number of unspecific pathways ranked at least as good as the target pathway (Figure 4.7b).

A pathway was denoted as unspecific, if it did not share any genes with the target pathways, and the pathway title suggested no relevance for the diseases studied in the `GEO2KEGG` benchmark set (such as *Synaptic vesicle cycle* and *Vitamin digestion*; see Additional file 1, Table S1). We found that all 6 pairwise combinations considerably reduced the number of unspecific pathways ranked as good or better than the target. Considering the GGEA/NEA-combination the number of unspecific pathways was reduced by >50% for both methods. On the other hand, combination with PathNet that displayed the least unspecific pathways, allowed to downgrade up to 70% unspecific pathways for NEA (while decreasing the number for PathNet even further). We also computed all pairwise combinations of the 4 set-based methods, the effect was however not as pronounced as observed for the network-based methods (Additional file 1, Figure S5).

In summary, given the heterogeneous individual rankings observed for the `GEO2KEGG` benchmark set, combining methods can, of course, not in all situations be expected to be superior to applying individual methods. However, we observed that the combination rarely results in loss of crucial information, but rather yielded in many cases a gain in confidence of relevant pathways while reducing the fraction of unspecific pathways.



(a) (b)

Figure 4.7: Combination of methods improves individual rankings on the GEO2KEGG benchmark set. (a) Combined ranks by rank sum (solid lines) were computed for the SBEA-combination of the 4 set-based methods (blue) and the NBEA-combination of the 4 network-based methods (green). Depicted is the number of `GEO` datasets (y-axis) for which the combination yielded a ranking of the corresponding target pathway, which was at least as good as obtained from x of the individual methods. As an example, the green point at $x = 3$ and $y = 15$ indicates that the NBEA-combination returned for 15 of the 27 datasets (55.6%) a ranking of the target as good or better as obtained for 3 of the 4 individual methods (i.e. only one method yielded a ranking better than the combination). For comparison, the dashed lines depict corresponding results when, instead of re-ranking by rank sum (solid lines), ranks are averaged across methods. **(b)** shows the total number of unspecific pathways ranked at least as good as the target pathway (y-axis) for the 4 network-based methods and each pairwise combination. Unspecific pathways are defined in the main text and listed in Additional file 1, Table S1. Corresponding values for the NBEA-combination of the 4 network-based methods and the ALL-combination of all 8 methods (4 set- and 4 network-based) are indicated with the green and the brown dotted line, respectively.

4.2.4 Conclusion

The ongoing development of individual enrichment methods impairs a straightforward decision for the method of choice. The `EnrichmentBrowser` offers a pragmatic solution by enabling the execution and combination of several major set- and network-based enrichment methods. Whereas no single method is best suited for all application scenarios, this allows to use them all at the same time facilitating a simple direct comparison of the results. It seamlessly displays inconsistencies reported by the applied methods, which makes the user aware that interpretation is needed and has to be done with care in order to derive valid conclusions. The combination can help to avoid misleading results of individual methods by removal of irrelevant gene sets, thus, reducing the outcome to candidates accumulating evidence from different methods. Of course, such consensus combinations come at the cost of less sensitivity but the `EnrichmentBrowser` does not prohibit that the user accepts non-consensus results from individual methods after careful assessment nevertheless. Detailed investigation of obtained gene sets and pathways is supported by accompanying comprehensive visualization and exploration capabilities. This exceeds considerably the functionality of available tools and we expect users and developers to likewise benefit from it.

4.2.5 Availability

- Project name: `EnrichmentBrowser`
- Project home page:
  http://www.bioconductor.org/packages/release/bioc/html/EnrichmentBrowser.html
- Operating system(s): Platform independent
- Programming language: `R`
- Other requirements: Bioconductor
- License: Artistic-2.0
- Any restrictions to use by non-academics: none

4.3 Composition analysis, reduction and integration

The previous section introduces the `EnrichmentBrowser` *as a general framework for combining and exploring results of different enrichment methods. In this section, I analyze the composition of predefined gene set and network definitions that are frequently used for the enrichment analysis. I discuss several aspects such as overlap and crosstalk between subnetworks, which impair a straightforward biological interpretation. I review recently published approaches to deal with these aspects and describe how they can be utilized as extensions for the* `EnrichmentBrowser`*.*

4.3.1 Overlap and crosstalk between subnetworks

The goal of the enrichment methods described in the previous section is to transform information for several thousand measured genes into information for a much smaller number of well-defined gene sets. This reduction is carried out to make biological interpretation feasible and to translate the observed expression changes into known biological functions and processes. However, the enriched processes are then typically interpreted in isolation (Section 4.1), which neglects that they are subnetworks of a global network. Crosstalk and overlap between them can thus lead to correlated and redundant findings. Incorporation of the relationships between subnetworks in analysis and visualization is therefore crucial for an accurate and comprehensive biological interpretation [188].

`GO` and `KEGG` annotations are most frequently used for the enrichment analysis of functional gene sets. Despite an increasing number of gene set and pathway databases, they are typically the first choice due to their long-standing curation and availability for a large range of species (Appendix C.3). However, overlap and crosstalk is well documented for `GO` terms [189, 190] and `KEGG` pathways [149, 156], which can substantially bias the results of the enrichment analysis when naively applied. In the following, I discuss at the example of the 2 major databases how overlap and crosstalk can be taken into account, resulting in a substantially improved biological interpretation (Figure 4.8). Similar adjustments are recommended for other gene set and pathway databases, typically of comparable structure as `GO` and `KEGG`, as these issues are likely to arise for them as well.

GO

Originally constructed in 1998, `GO` has been a popular target for enrichment tests from early on [22]. In particular, the overrepresentation test (Section 3.1) is routinely applied with `GO` in gene expression studies and a plethora of enrichment tools have been specifically designed for that purpose [20, 122].

However, frequently overlooked aspects of `GO` can result in an undesired misuse of `GO` [189]. For instance, treating `GO` terms as independent gene sets in the enrichment analysis ignores the parent-child structure of `GO` (Appendix C.3.1). As all genes annotated to a term are per definition also annotated to its parent(s), enrichment of the child inevitably results in an increased chance for enrichment of the parent term(s). Extensive redundancy in the

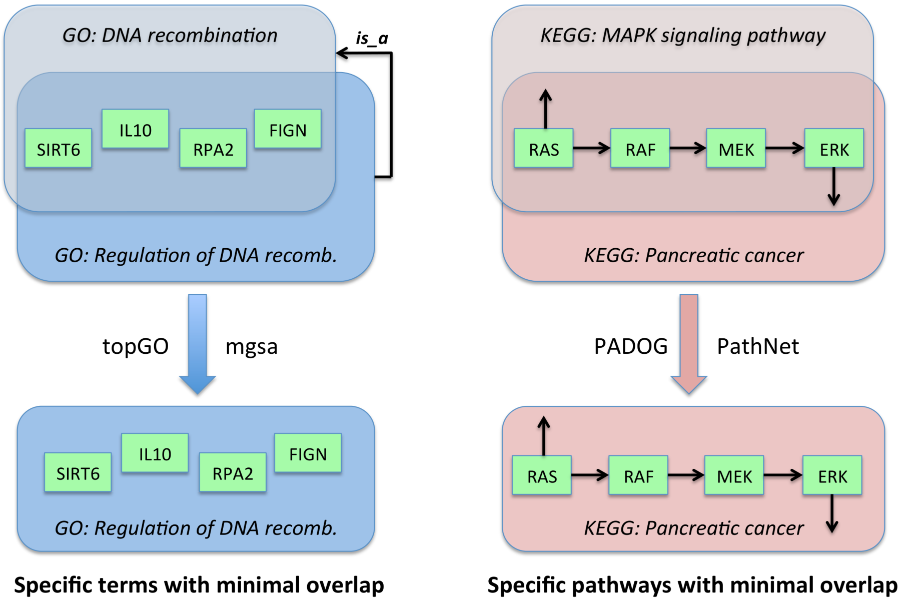

GO: DNA recombination
is_a
SIRT6
IL10
RPA2
FIGN
GO: Regulation of DNA recomb.
KEGG: MAPK signaling pathway
RAS
RAF
MEK
ERK
KEGG: Pancreatic cancer
topGO
mgsa
PADOG
PathNet
SIRT6
IL10
RPA2
FIGN
GO: Regulation of DNA recomb.
RAS
RAF
MEK
ERK
KEGG: Pancreatic cancer
Specific terms with minimal overlap
Specific pathways with minimal overlap

Figure 4.8: Overlap and crosstalk in enriched GO terms and KEGG pathways: solutions for the EnrichmentBrowser. Shown on the left in blue are two overlapping `GO` terms *DNA recombination* and *Regulation of DNA recombination*, which are connected by a parent-child relation (*is_a*). The main text describes how existing approaches implemented in packages such as `topGO` [169] and `mgsa` [194] can be exploited to resolve the overlap. Analogously, shown on the right in red is the overlap between a general and a specific `KEGG` pathway, namely *MAPK signaling* and *Pancreatic cancer*, respectively. Approaches to deal with such overlaps by down-weighting overlapping genes (implemented in the `PADOG` package [149]) or by taking contextual association into account (`PathNet` package [178]) are described in the main text. The result is, in both cases, decreased redundancy in the list of enriched gene sets and pathways, which simplifies the biological interpretation.

enriched `GO` terms is the consequence, which impairs a concise functional description of the results.

Several approaches have been suggested to remove this redundancy by incorporating the hierarchical structure of `GO` in the enrichment analysis [190–193]. In the following, approaches available in `Bioconductor` are described in more detail as they can be straightforward integrated into the `EnrichmentBrowser`.

The `topGO` package implements several algorithms to integrate the `GO` graph structure [169]. Among them are the `elim` and `weight` algorithms [195] as well as the `parentchild` algorithm [196]. The basic idea of the `elim` algorithm is to analyze terms by removing genes of significant child terms. The goal is to report the most specific terms and exclude general terms, which are predominantly reported due to the enrichment of their children. However, a side effect is that a parent term of originally higher significance than its child might be missed due to the gene removal. The `weight` algorithm addresses this by comparing significance of parent and child, and downweights genes in the less significant of both terms. Similarly, the `parentchild` algorithm not only considers the overlap of the differentially expressed genes with the genes annotated to a specific term, but also with the genes annotated to its parent(s). A combination of `elim` and `weight` has been shown to work best in practice and is used as the default algorithm of `topGO`. These algorithms can be combined with frequently used enrichment tests, including the hypergeometric test and the Kolmogorov-Smirnov test (Appendix B.1).

A different approach is implemented in the `mgsa` package [194]. Noteworthy, it differs from the typical sequential set-by-set analysis by incorporating a multiset statistic, thus working on all sets simultaneously. Using a Bayesian approach, it assumes that differential expression of a gene results from the activation of a particular gene set it belongs to. It thereby aims for gene sets that are explaining the largest number of differentially expressed genes, and takes overlaps into account by penalizing the report of more than one gene set per differentially expressed gene.

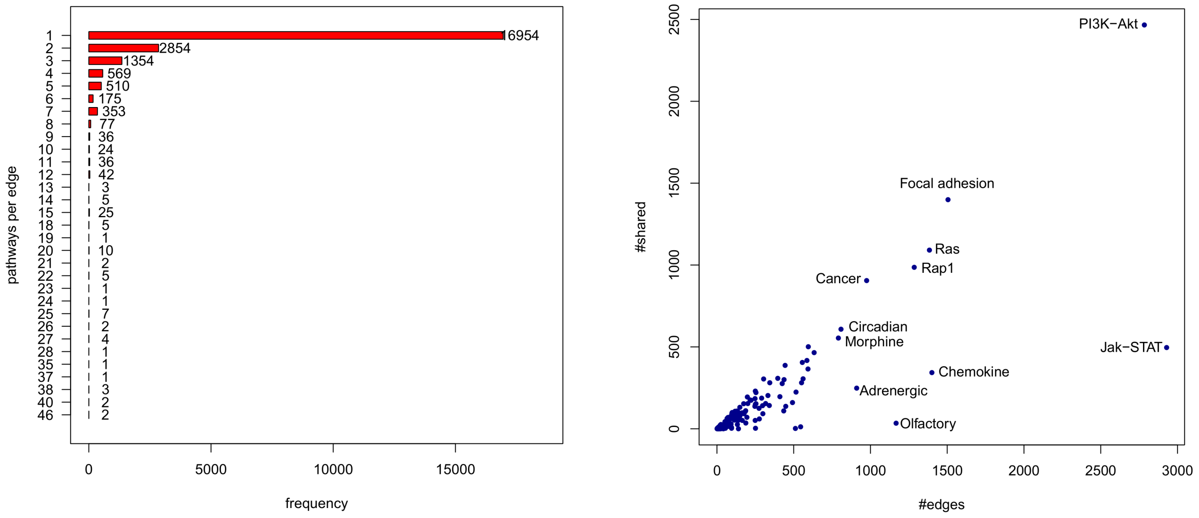
Two recent studies compare the approaches of `topGO` and `mgsa` and find that both have benefits and disadvantages [190,193]. The `topGO` approach detects highly enriched and specific terms that protrude from their parental background. Although it thus resolves overlaps between parent and child terms, overlaps between non-related terms are not taken into account. It is also exclusively applicable to `GO`. In contrast, `mgsa` can be applied to arbitrary gene set definitions. It typically detects a small number of sets with minimal overlap, which jointly best explain the observed differential expression. This produces parsimonious results, but has been observed to miss joint sets of non-additive nature, i.e. with distinguishably more information than implied by the individual sets. The `weight` algorithm of `topGO` is superior in this regard, as it allows to prefer parent terms of higher significance than their children.

KEGG

Initiated by the Japanese Human Genome Program in 1995, `KEGG` has originally been recognized as a comprehensive collection of manually curated maps of metabolic pathways [24]. Complementary to the biological process (BP) ontology of `GO`, these pathways were frequently incorporated in the enrichment analysis as gene sets of specific enzymes, which catalyze the interconversion of biochemical metabolites (Appendix C.3.2).

The inclusion of signaling pathways, illustrating molecular interactions between genes and gene products in specific small-scale gene regulatory networks, considerably added to that. Such interactions between genes are absent from `GO`. This gave rise to methods specifically designed for incorporating the topology of `KEGG` signaling pathways such as SPIA, but also general network-based methods such as GGEA.

Although `KEGG` is not organized in an ontology as `GO`, and thus does not display overlaps caused by inheritance, similar issues arise due to shared processes and crosstalk between pathways. For example, the general signaling cascade from *RAS* to *ERK* shown in Figure 4.8 is part of 40 out of 160 signaling pathways currently contained in `KEGG`. When analyzing all edges contained in `KEGG` signaling pathways, the edges between *RAS*, *RAF*,



(a) (b)

Figure 4.9: Overlap between KEGG signaling pathways. (a) shows the number of `KEGG` edges (x-axis) that occur in y pathways. For example, there are two edges that occur in 46 out of 160 investigated pathways. This includes the edge from *MEK* to *ERK*, which is part of the signaling cascade shown in Figure 4.8. **(b)** shows for each `KEGG` signaling pathway the total number of edges (x-axis) and the corresponding fraction of edges that are shared with at least one additional pathway (y-axis). For example, the *Jak-STAT* signaling pathway has in total ≈3000 edges from which ≈500 are shared with other pathways (i.e. the pathway contains 2500 unique edges).

MEK, and *ERK* are among the edges with maximal number of pathways per edge (shown at the bottom left of Figure 4.9a). A different perspective on that matter is shown in Figure 4.9b, demonstrating that many signaling pathways share a considerable fraction of their edges with other pathways.

As for `GO`, several approaches have been suggested to incorporate overlap between pathways in the enrichment analysis [149,156,178,197]. Two similar approaches are *GSEA with integration of gene Appearance Frequency*, GSEA-AF [197], and *Pathway Analysis with Down-weighting of Overlapping Genes*, PADOG [149]. Both strengthen the impact of genes that are unique to a pathway by downweighting genes that appear in several pathways. Therefore, GSEA-AF integrates the absolute appearance frequency $f(g) \in \{1, \dots, N_{GS}\}$, for a gene g and the total number of investigated gene sets N_{GS}, in the weighted KS-statistic of the original GSEA enrichment score. Similarly, PADOG multiplies the differential expression t-score for each gene with a weight $w \in [1, 2]$. This weight decreases monotonically from $w = 2$, for genes that are specific for a gene set, to $w = 1$ for genes with maximal $f(g)$. While an implementation of GSEA-AF is not availabe at present, PADOG is implemented in Bioconductor's `PADOG` package.

In addition to static overlap, pathways have also been shown to influence each other depending on the expression data under investigation; a phenomenon denoted as crosstalk [156] or contextual association [178]. Donato *et al.* [156] again attribute this to overlapping genes, with the strength of the crosstalk effect depending on whether these genes are differentially expressed. To correct for this effect, overlapping genes are similarly resolved as for `mgsa` by selecting the pathway which is maximally impacted by differential expression of those genes. On the other hand, Dutta *et al.* [178] also consider crosstalk to be caused by regulatory interactions between genes of non-overlapping pathways. Such an interaction of the global network, e.g. derived from pooling all pathways together, is assumed to be active if both interaction partners are differentially expressed. While Donato *et al.* do not provide an implementation of their crosstalk correction, Dutta *et al.* have integrated the contextual association analysis in Bioconductor's `PathNet` package.

4.3.2 Subnetwork reduction and integration

The previous section describes how to resolve overlap between subnetworks at the time of the enrichment analysis. This typically results in a rearranged and non-redundant collection of subnetworks, thereby alleviating the biological interpretation. This section deals with the actual interpretation by reducing subnetworks to active modules and integrating findings across them (Figure 4.10). In analogy to Section 4.2.1, there are *set*-based and *network*-based reduction approaches.

Set-based reduction

It has been early observed that rarely all genes of an enriched gene set show significant expression changes [66, 198]. In fact, denoting a gene set as *overrepresented* or *enriched* already indicates by the name that not all, but rather disproportionately many set genes are differentially expressed. Thus, to understand which part of an enriched functional category drives the observed expression, it is necessary to reduce the findings to the active subsets within them.

Already the original GSEA publication [123] suggests an approach to reduce an enriched gene set s to the subset that accounts for the enrichment signal. In what the authors denote as *leading edge analysis*, this subset corresponds to the part of the ranked genes in s that appear before the running sum statistic reaches its maximum. The leading edge analysis is implemented in the GSEA `R`-script [172] that has been adapted for the `EnrichmentBrowser` (Section 4.2.2, subsection *Set-based enrichment analysis*).

Two similar approaches for determining the leading edge of s are Sub-GSE [199] and SAM-GSR [124]. Sub-GSE tests subsets of s based on a seed of highly significant genes in s. This seed is gradually extended by additional genes of s, sorted by significance of differential expression, as long as a subset significance threshold is satisfied. SAM-GSR [124], a reduction approach based on SAMGS [171], differs from Sub-GSE only in the choice of the local (gene) and global (gene set) statistic. By gradually partitioning s, SAM-GSR selects the core subset of genes in s with largest SAM t-like statistic for which testing of the SAMGS set statistic satisfies a predefined significance threshold. As for GSEA, the implementation of SAM-GSR is part of the SAMGS `R`-script [173] that has

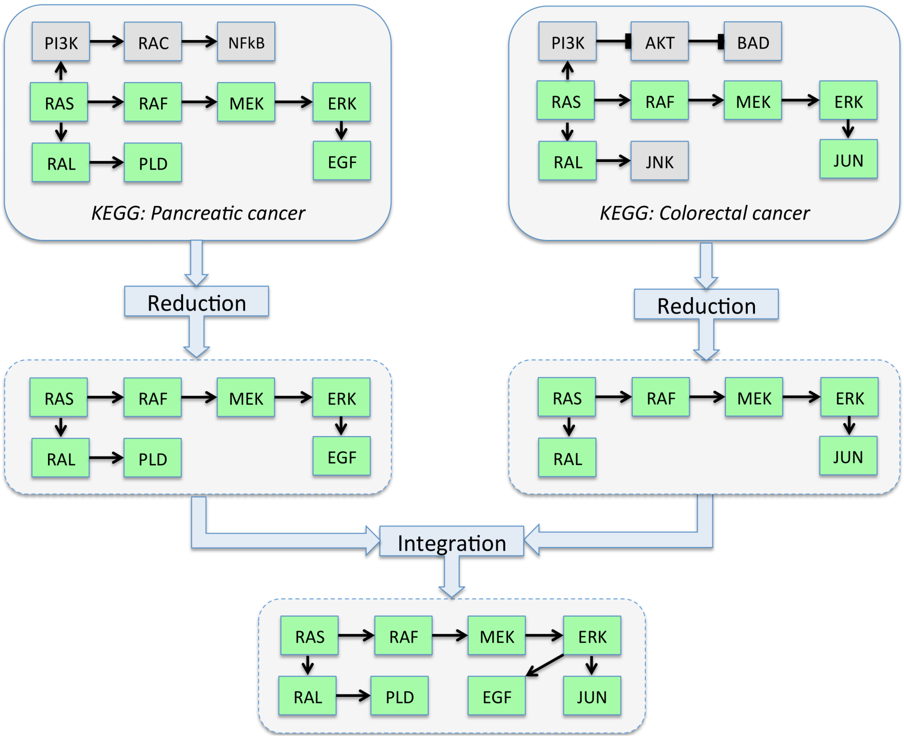


Figure 4.10: Reduction and integration of enriched subnetworks. The goal of the reduction is to remove inactive parts (genes in gray) from the enriched subnetworks to obtain active modules (genes in green). Subsequently, integration across active modules, yielding the union of the observed effects, is applied to obtain an overall picture.

been adapted for the `EnrichmentBrowser` (see again Section 4.2.2, subsection *Set-based enrichment analysis*). Thus, leading edge analysis is also easily derived for other set-based enrichment methods via application of the reduction procedure described for GSEA and SAMGS, and substitution with the respective local and global statistics.

Network-based reduction

Network-based reduction approaches also take the interactions between genes into account. A recent review [29] divides existing approaches for finding active network-modules in three broad categories:

- *Significant area search* methods such as MATISSE [126] typically involve three steps: (1) activity scoring for genes and interactions between them, e.g. based on differential expression, (2) aggregation of activity scores across network regions, (3) detection of high-scoring regions corresponding to active modules.

- *Network propagation* methods such as HotNet [200] and PARADIGM [201] represent a particularly relevant subclass of significant area search methods. These methods employ the hierarchical structure of regulatory networks and test whether signals

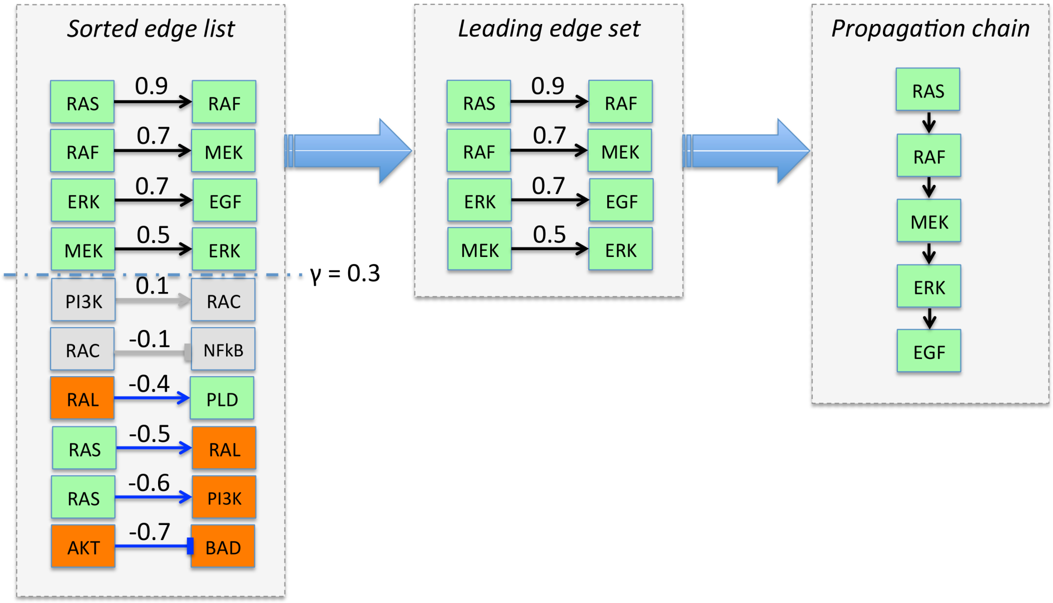


Figure 4.11: Outline of the Sub-GGEA reduction approach. Shown on the left is the edge list of an enriched subnetwork, sorted by consistency (arrow labels). Black and blue arrows depict consistent and inconsistent edges, whereas green and red boxes correspond to up- and down-regulated genes, respectively. Gray color is used to indicate neutral effects on genes and interactions. Application of a consistency threshold $\gamma = 0.3$ yields the leading edge set of highly consistent edges. By incorporation of the known network topology, these high-consistency effects can be arranged along the propagation chain, thereby yielding the active module of the enriched subnetwork.

are propagated along that hierarchy, e.g. by transmitting differential expression from general regulators to specific target genes. Propagation often incorporates concepts related to the consistency concept of GGEA (Section 3.2.2, subsection *Consistency of regulatory interactions*), i.e. whether expression changes are transmitted in agreement with the regulation type (activation/inhibition).

- *Biclustering* methods such as SANDY [202] and cMonkey [203] invoke simultaneous clustering of network interactions and the conditions under which these interactions are active.

Given these considerations, a network-based reduction approach for GGEA is straightforward derived by combining aspects of leading edge and network propagation analysis (Sub-GGEA, Figure 4.11):

1. Sort edges of the induced subnetwork GRN(u) by consistency in decreasing order (Equation 3.9 and 3.21),
2. Reduce the sorted edge list to the leading edge set by cutting the list at a defined consistency threshold γ (suitable choices for γ are discussed in Section 3.3.2, subsection *Consistency threshold*),

3. Arrange the remaining edges in the leading edge set along the propagation chain as implied by the network topology.

Integration

Once the enriched subnetworks have been reduced to active modules, the next step is to integrate findings across them to obtain an overall picture (Figure 4.10). Basic integration is typically carried out by simply taking the union across active modules, thereby removing any remaining redundancy between modules. This is especially beneficial if overlap and crosstalk between enriched subnetworks have not been resolved beforehand as described in Section 4.3.1.

On the other hand, advanced integration approaches also combine findings across different gene set and pathway databases as well as across multiple experiments and omic-platforms in order to identify functionally related themes among them [183,192,204,205]. Depending on the scale of the results, this can, however, substantially increase the size and complexity of the integrated active network. Interpretation is then typically eased by *a priori* grouping of modules into biological functions, e.g. according to gene set categories provided by most databases such as `KEGG` and `MSigDb` (Appendix C.3). For example, `KEGG` divides cellular processes into transport and catabolism, cell motility, cell growth and death, and cellular community. Using these categories to functionally group active modules enables partitioning of the integrated network and immediate identification of significantly altered cell functions.

Chapter 5

Conclusion and Outlook

Studying a phenotypic trait of interest traditionally involves the reduction of a complex system to single genes. Conclusions about what causes the trait under study are then drawn from the detailed analysis of these genes in isolation. However, when analyzed in interaction with the system, additional effects on these genes need to be integrated. Nowadays, genes are measured by the thousands, as analysis capacities have experienced a boost by an unprecedented order of magnitude with the advent of genomic high-throughput assays. Nevertheless, integrating from genome-wide data to understand the system as a whole is inevitably accompanied by a loss in resolution, which can result in an inappropriate simplification of the overall picture. Hence, an important unresolved question in systems biology can be formulated as follows: when scaling up from single genes to systems, how can a detailed understanding of the driving mechanisms be preserved?

Although a suitable trade-off between the reductionistic and the integrative approach is not easily achieved, I have presented several steps towards this goal for the analysis of large-scale gene expression data. Major contributions are the definition of a gene regulatory network model that allows to incorporate additional details on a larger scale (Chapter 2), the development of a method to incorporate such networks in the functional analysis of high-throughput expression data (Chapter 3), and the implementation of a software package that allows the execution of this method in combination with other state-of-the-art methods as well as the in-depth interpretation of the results (Chapter 4).

As I have focused on transcriptomic data measured with microarrays or RNA-seq, future challenges will include to adapt these concepts also for novel platforms and additional sources of genome-wide data. For instance, micro-fluidic lab-on-a-chip devices enable the integration of various small-scale assays such as qPCR on a single chip [206, 207]. Most promisingly, this technology can be applied for gene expression profiling at the resolution of individual cells [208, 209]. On the other hand, recent advances in mass spectrometry have also considerably improved the quality of proteomic data [210, 211]. As proteins are the actual determinants of cellular function, and transcriptomic studies are assumed to only roughly approximate the proteomic landscape, this will eventually result in a more appropriate functional readout of the cell. A comprehensive quantitative characterization of regulatory and metabolic processes can then be achieved by coupling proteomic measurements with complementary information from high-throughput metabolomics [212,213]. Quantifying metabolic reactions catalyzed by enzymatic proteins in conjunction with quantification of the enzymes themselves bridges the gap between genetic potential and functional metabolism [214, 215]. Given these benefits, multi-platform analysis is expected to become routine in the future. As a recent example, The Cancer Genome Atlas project, TCGA [183], has launched a molecular investigation of various cancer types on an unprecedented scale including genotyping as well as transcriptomic and proteomic measurements. However, the issue of losing resolution already observed when analyzing data from single high-throughput platforms becomes even more pronounced for such cross-platform studies. Network-based analysis of such data requires to rewire existing networks across platforms and to define appropriate models of interconnections between the different genomic and metabolic entities [216].

In Chapter 2, I have identified minimum requirements for modeling gene regulatory networks and have outlined, at the example of the diauxic shift, how networks satisfying these requirements can be constructed. Although absent from many existing regulatory networks, I have shown that it is necessary to characterize the conditions under which regulations take place (*context*) and how regulated genes are affected (*effect*). Regulatory entities of key importance in the model are transcription factors. A major limitation arises from the fact that the heterogeneous regulation of these factors themselves is only insufficiently understood. While the expression level is typically taken as an approximation of the activity, there are many examples where transcription factors are activated by other mechanisms such as post-translational phosphorylation [217–219]. Hence, as already slight expression changes on the transcriptional level can have a crucial impact, it is necessary to resolve the dynamics of transcription factor expression over time as accurately as possible, e.g. from the aforementioned single-cell snapshot data [220, 221]. On the other hand, integration of an additional proteomic layer in the model for quantification of phosphorylated transcription factors is essentially required. Although resolving the behavior of transcription factors is already sufficiently complex, future models will also need to incorporate additional layers of transcriptional control. This includes epigenetic control mechanisms which, by modifying the DNA and the histones it is wrapped around, regulate whether promoters of target genes become accessible for transcription factors [222–224]. In addition, important contributions are expected to come from studying the interplay of transcription factors with distal enhancers via 3D chromatin interactions [225–228]. A recent effort into this direction has been carried out in the ENCODE project [83, 229]. Integrating these additional aspects of transcriptional regulation is straightforward, considering the structure of the model suggested for the diauxic shift, composed by a regulatory, a transcriptional, and a metabolic layer (see again Figure 2.6). This also holds for future complementation of transcriptional regulatory networks with post-transcriptional and translational regulatory networks [230–232].

In Chapter 3, I have demonstrated how regulatory networks can be integrated in the enrichment analysis by scoring interactions in functional gene sets with the observed expression data. I have shown that this consistently aligns regulation and expression and allows to explain significant expression changes by regulators within the set, thereby considerably easing interpretation. The concept of consistency is central for GGEA. However, the gap between experimental annotation and observed behavior of regulatory interactions in large-scale assays impairs a precise consistency evaluation. For example, an interaction annotated as activating might never show up as such in transcriptomic data. Reasons for that are plentiful, including erroneous annotation, context dependency, and non-transcriptional features of regulatory networks [233]. Therefore, learning direction and strength of regulatory interactions across experiments is essential to appropriately parameterize interactions of the network, thereby establishing suitable null models the observed behavior can be tested against [234]. On the other hand, future extension of the consistency concept to non-transcriptional features such as the aforementioned metabolic and proteomic layer will allow to build longer chains of causality. For instance, quantification of intra- and extra-cellular metabolites as well as the activity of receptor and kinase proteins can be incorporated in the consistency evaluation of interactions between signal molecules and their cellular perception. Promising initial efforts towards this goal are the recent integration of multi-level omics data into set-based [192, 235] and network-based enrichment analysis [200, 201, 236].

In Chapter 4, I have presented the `EnrichmentBrowser` software package for the combined evaluation of set- and network-based enrichment as well as the detailed exploration of the results. I have demonstrated that this allows to filter out spurious hits of individual methods and increases the confidence in gene sets and pathways accumulating evidence from different methods. Although this is a first important step towards achieving a consensus among the various existing methods, a consistent and objective comparison of the methods is still required [21, 28]. There is broad agreement that gold standards for the evaluation of enrichment methods are needed, yet a conclusive idea how to design such standards is currently missing. Existing solutions of simulating certain scenarios are often biased towards the strengths of individual methods. On the other hand, biological reasoning on real datasets is a fraught procedure, rarely allowing precise conclusions about benefits and disadvantages of the methods. Therefore, construction of suitable benchmark datasets will be a decisive factor for future developments in the area of enrichment analysis [149, 186].

Subsequent biological interpretation of enrichment analysis results is currently centered on `GO` and `KEGG`. However, there are many more existing and presumably upcoming gene set and pathway databases, each of them with unique characteristics (Appendix C.3). Comprehensive integration of these additional resources in the biological interpretation is expected to complement findings obtained with the established resources. On the other hand, using knowledge from pathway databases tends to restrict the interpretation to known mechanisms within pathways. As a consequence, novel hypotheses for hits outside pathways are often under-represented [237]. Leveraging strategies to target those findings will thus considerably add to the understanding of the biological mechanisms resulting in the observed data.

In conclusion, this work has substantially contributed to the field of network-based analysis of gene expression data with respect to regulatory network construction, subnetwork detection, and their biological interpretation. As systems biology in general is a young and rapidly developing research discipline, much remains to be explored and we are just beginning to resolve issues related to model scaling, network annotation, and method evaluation. Nevertheless, ambitious large-scale projects such as ENCODE, TCGA, and the 1000 Genomes Project [238] will continue to release big data waves in increasingly shorter intervals. This will further challenge our analysis capacities and change the way how we are approaching and analyzing the data. On the other hand, this opens new perspectives for genomic research and holds great promise for the future.

Appendix A

Methods of molecular biology

A.1 Quantitative measurement of gene expression

The two major methods for large-scale quantification of transcript abundance in cellular mRNA extracts are based on hybridization (microarrays) or sequencing (RNA-seq). Although microarrays have been established earlier for high-throughput gene expression analysis, RNA-seq is considered superior with respect to reproducibility and data range [11,12]. However, both have in common that they require specific normalization procedures to accurately separate biological from technical effects [13,14].

A.1.1 Microarrays

A microarray consists, in accordance to the definition of `ArrayExpress` [239], of an arrayed series of thousands of microscopic spots of DNA oligonucleotides (*features*), each containing picomoles of a specific DNA sequence (*probes*). This is a short section of a gene, which is used to hybridize a reverse transcribed mRNA sample (*target*) via complementary base pairing between the probe and the target. The target is usually labeled with a fluorophore, which makes the detection and quantification of the hybridization possible. The hybridization degree itself is then taken as an approximation of the expression level of the target.

A.1.2 Next-generation sequencing (RNA-seq)

DNA sequencing, i.e. determination of the precise order of nucleotides of a DNA sequence, has been traditionally carried out using the chain-termination method developed by Sanger *et al.*, 1977 [240]. However, massive parallelization of the sequencing process in next-generation sequencing methods has largely replaced Sanger sequencing [241,242]. RNA-seq exploits next-generation DNA sequencing for the quantification of cDNA that has been reverse transcribed from mRNA samples [243]. This results typically in millions of short *reads*, i.e. DNA sequences of ≈100 nucleotides. When mapping these reads to the reference genome, the number of reads aligned with a gene (*read count*) is then taken as an approximation of its expression level.

A.2 Detection of protein-DNA interactions

Protein-DNA interactions of major importance for the construction of gene regulatory networks are transcription factors binding physically to the promoter of their target genes. Determination of such interactions can be carried out targeting a DNA-binding protein of interest via a specific antibody (ChIP) or untargeted by cleaving the DNA and tracing sections that are protected due to the binding of a protein (DNA footprinting).

A.2.1 Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation [58, ChIP] identifies physical interactions between the DNA and a DNA-binding protein of interest. The experimental procedure is composed of protein-DNA crosslinking, sonication of the DNA to obtain protein-DNA complexes, and immunoprecipitation of the protein of interest using a specific antibody. Subsequent quantification of the bound DNA can then be applied to detect enriched sequences corresponding to specific DNA-binding sites of the protein. The combination of ChIP with the microarray technology [59, ChIP-chip] or next-generation sequencing [244, ChIP-seq] allows the genome-wide identification of the binding sites of a transcription factor under study.

A.2.2 DNA footprinting

DNA footprinting detects regions of the DNA that are protected from enzymatic cleavage by proteins bound to the DNA [56, 245]. The deoxyribonuclease DNase I is typically applied to cut the DNA, and the areas protected by protein can then be sequenced or quantified using various protocols. DNA footprinting can also be combined with ChIP to extract footprints of a specific DNA-binding protein. Recent coupling of DNA footprinting with next-generation sequencing [246, DNase-seq] and subsequent computational motif discovery by digital genomic footprinting [247] has enabled the application of the assay on a genome-wide scale.

Appendix B

Statistical methods

B.1 Hypothesis testing

Hypothesis testing is an essential part of the statistical analysis of gene expression data [248, for an introduction]. For example, it applies to testing of differential expression between sample groups as well as testing the distribution of gene set scores. The basic principle of hypothesis testing is to test a null hypothesis H_0 against an alternative hypothesis H_1. This involves the computation of a test statistic and if the test statistic exceeds a certain value the null hypothesis becomes unlikely and is rejected in favor of the alternative hypothesis. A concept of fundamental importance in hypothesis testing is the p-value. Under the null hypothesis, the p-value corresponds to the probability of observing a value of the test statistic at least as extreme as computed for the data under investigation. Accordingly, a p-value close to zero indicates that it is rather unlikely that the data is in agreement with the null hypothesis. The null hypothesis is rejected if $p < \alpha$, with α being a predefined significance level, typically chosen as 0.05. This is equivalent to the statement that a p-value below α can be considered *statistically significant*, which is often found in the literature.

B.1.1 Student's t-test

Assuming two sample groups to be normally distributed with equal variance σ^2, the two-sample t-test tests whether the means μ_1 and μ_2 of both groups are equal, i.e.

$$H_0: \quad \mu_1 = \mu_2 \tag{B.1}$$

against

$$H_1: \quad \mu_1 \neq \mu_2. \tag{B.2}$$

As the mean of a normal distribution is known to follow Student's t-distribution, the test statistic is chosen accordingly as

$$t = \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{n_1 n_2}}{s\sqrt{n_1 + n_2}} \tag{B.3}$$

with n_1 and n_2 being the number of samples in both groups, and $\bar{x}_1$ and $\bar{x}_2$ the sample means. The assumed equal, but unknown variance σ^2 is estimated via

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}. \tag{B.4}$$

As the assumption of equal variance might not hold in practice, the extension of Welch [249] allows to also apply the t-test with different variances in both sample groups.

With the advent of microarrays, the t-test has become one of the most frequently applied tests for assessing differential gene expression between two sample groups. Notably, Smyth [160] has introduced the moderated t-statistic for microarray data, which prevents large t-scores to occur only due to small gene variances. Although the t-test is suitable for microarray data, where the data distribution is typically roughly normal, this is not the case for RNA-seq data *per se*. Due to the dynamic range of RNA-seq data, often producing extremely large read counts for a number of genes, the variance is typically much larger than the mean, an effect termed *overdispersion*. Overdispersion can be appropriately modeled based on the negative binomial distribution [164, 165]. Recently, Law *et al.* [163] have proposed the `voom`-transformation for RNA-seq data resulting in a mean-variance relationship similar to microarray data, thus enabling the application of the t-test also on `voom`-transformed RNA-seq data.

B.1.2 Fisher's exact test

Fisher's exact test [250, 251] is a test for independence of two categorical variables, each of them having exactly two categories. Observations categorized into one of the categories for both variables are described by a 2 × 2 contingency table as exemplarily depicted in Table 3.1. Here, genes are categorized as *differentially expressed* or *not differentially expressed* as well as being *in gene set* or *not in gene set*.

The corresponding null hypothesis

$$H_0: \quad \text{Differential expression is independent of gene set membership} \tag{B.5}$$

is tested against

$$H_1: \quad \text{Differential expression is associated with gene set membership.} \tag{B.6}$$

The test statistic is the size of the overlap m_{GD}, leading under the null hypothesis to the hypergeometric distribution, i.e. the probability of observing m_{GD} differentially expressed genes in the gene set under investigation corresponds to

$$P(X = m_{GD}) = \frac{\binom{m_G}{m_{GD}}\binom{m_{\bar{G}}}{m_{\bar{G}D}}}{\binom{m}{m_{GD}+m_{\bar{G}D}}}. \tag{B.7}$$

Accordingly, the p-value is computed as the probability of observing an overlap of at least m_{GD} genes as

$$p = 1 - P(X \leq m_{GD}) = 1 - F(m_{GD}) \tag{B.8}$$

with F being the hypergeometric cumulative distribution function.

B.1.3 Permutation testing

Student's t-test and Fisher's exact test are parametric, or distribution-dependent, tests. This means that the distribution of the test statistic under the null hypothesis is obtained from a theoretical probability distribution such as the t-distribution and the hypergeometric distribution. However, in many cases the underlying null distribution is not known or can not be assigned to a known distribution.

Permutation testing is a non-parametric, or distribution-free, alternative procedure. It relies on repeated calculation of the test statistic on randomized data to obtain the null distribution [187]. The p-value for an observed value of the test statistic can then be estimated based on the empirical null distribution. An example of permutation testing is given in Section 3.2.2, subsection *Significance and Ranking*, which assesses the significance of the GGEA score. Although permutation testing is advantageous as it does not depend on a known distribution, it is computationally expensive to obtain a null distribution of suitable granularity, requiring typically $\geq$1,000 permutations [143].

B.1.4 Wilcoxon's rank-sum test

Wilcoxon's rank-sum test [252, 253] is a non-parametric alternative to the t-test that does not require the data to be normally distributed. The test is used to assess whether two samples have been drawn from the same distribution, or from two different distributions X and Y, one being stochastically greater than the other.

The corresponding null hypothesis

$$H_0 : \quad P(X > Y) = P(Y > X) \tag{B.9}$$

is tested against

$$H_1 : \quad P(X > Y) \neq P(Y > X). \tag{B.10}$$

The test statistic is not computed on the observed values themselves, but on the ranks obtained from ordering the observed values. As the name of the test indicates, the test statistic corresponds to the sum of the ranks of either the first or the second sample. For large samples, the null distribution of the rank-sum statistic can be approximated by a normal distribution. However, Wilcoxon's rank-sum test can also be combined with permutation testing, i.e. the observed rank sum is evaluated against values obtained from recomputing the rank-sum statistic in many random permutations. This permutation procedure is e.g. used for estimating the significance of the global gene set statistic from SAFE, which sums up the ranks of the local statistic for each gene [125].

B.1.5 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test [253] is another non-parametric test for equality of two sample distributions. Compared to the t-test (testing whether the means of two distributions are equal) and Wilcoxon's rank-sum test (testing whether one distribution is stochastically

greater), the Kolmogorov-Smirnov test also takes location and shape of the distributions into account, and is therefore a more general test of whether two sample distributions are equal.
The corresponding null hypothesis

$$H_0: \quad \text{The two sample distributions are equal} \tag{B.11}$$

is tested against

$$H_1: \quad \text{The two sample distributions are different.} \tag{B.12}$$

The test statistic is the maximal distance of the empirical cumulative distribution functions of the two distributions to be tested. The Kolmogorov-Smirnov test can also be used to test whether a sample distribution resembles a known distribution. For example, this is used in GSEA to test whether the ranks of the differential expression p-values of a gene set under study represent a sample from a uniform distribution [123].

B.1.6 Multiple testing

As introduced in Section B.1, testing of a hypothesis is done in accordance to a predefined significance level α. The significance level corresponds to the probability of rejecting the null hypothesis although it is true. Hence, when carrying out a test 100 times using the typical significance level $\alpha = 0.05$, one has to expect 5 cases of erroneous rejection of the null hypothesis. In turn, this means that the p-values of the 100 tests have to be expected to false positively indicate statistical significance for 5 of the tests. Thus, multiple testing, i.e. repeatedly carrying out the same test, bears the danger of inflating statistical significance. This is especially important for genomic research, where often thousands of genomic entities are tested simultaneously.
To correct for multiple testing, several procedures have been suggested [166, for an overview]. This includes the stringent Bonferroni correction, multiplying resulting p-values with the number of tests. However, this is only necessary if the tests are independent. This is rarely the case for genomic research where genes are known to influence each other which leads to dependent tests. A less conservative procedure is thus typically preferred, which is frequently chosen to be the correction of Benjamini and Hochberg [254].

Appendix C

Databases

C.1 Genome databases

Genome databases store sequences and functional annotation of complete genomes. For model organisms, there are individual databases such as `SGD` for *S. cerevisiae* [68] or `FlyBase` for *D. melanogaster* [255]. On the other hand, collection of genomes including non-model organisms are available from `Ensembl` [256] or the UCSC Genome Browser [257].

C.2 Gene expression databases

Gene expression databases are archives for large collections of experimental high-throughput data. This includes the Gene Expression Omnibus, `GEO` [147,258], and `ArrayExpress` [239] for microarray data as well as the Sequence Read Archive, `SRA` [259], for RNA-seq data.

C.3 Gene set and pathway databases

Gene sets are simple lists of usually functionally related genes without further specification of relationships between genes. They are frequently used for gene set enrichment analysis, but also as signatures for disease classification [17, 18]. Comprehensive gene set collections are available in the Molecular Signatures Database, `MSigDb` [260,261]. Recently, `Enrichr` [262] has considerably added to this by integrating 35 gene set libraries covering 6 broad categories: transcription, pathways, ontologies, diseases/drugs, cell types, and miscellaneous.

Pathways are specific gene sets, typically representing a group of genes that work together in a biological process. Pathways are commonly divided in metabolic and signaling pathways. Metabolic pathways such as glycolysis represent biochemical substrate conversions by specific enzymes. On the other hand, signaling pathways such as the MAPK signaling pathway describe signal transduction cascades from receptor proteins to transcription factors. Signaling pathways are frequently used for network-based gene set enrichment analysis. A variety of biomolecular pathway databases is available [263, 264]. `Pathguide` [265] lists 547 pathway databases as of November 2015.

C.3.1 Gene Ontology (GO)

The Gene Ontology, `GO` [22, 23], consists of three major sub-ontologies that classify gene products according to molecular function (MF), biological process (BP) and cellular component (CC). Each ontology consists of `GO` terms that define MFs, BPs or CCs to which specific genes are annotated. The terms are organized in a directed acyclic graph (DAG), where edges between the terms represent relationships of different types. They relate the terms according to a parent-child scheme, i.e. parent terms denote more general entities, whereas child terms represent more specific entities. Multiple inheritance is possible, a child term can thus have more than one parent term. Annotation of a gene to a specific term results in the implicit annotation to all of its parents. Annotations are classified with evidence codes of 4 broad categories: experimental, computational, indirect, or unknown.

C.3.2 Kyoto Encyclopedia of Genes and Genomes (KEGG)

The Kyoto Encyclopedia of Genes and Genomes, `KEGG` [24, 25], is a collection of manually drawn pathway maps representing molecular interaction and reaction networks. These pathways cover a wide range of biochemical processes that can be divided in 7 broad categories: metabolism, genetic and environmental information processing, cellular processes, organismal systems, human diseases, and drug development. Metabolism and drug development pathways differ from pathways of the other 5 categories by illustrating reactions between chemical compounds. While drug pathways depict *in vitro* transformations of the chemical structure of related drug components, metabolic pathways display the *in vivo* interconversion of biochemical substrates catalyzed by specific enzymes in defined reaction chains. On the other hand, pathways of the other 5 categories illustrate molecular interactions between genes and gene products. A large fraction of these pathways are signaling pathways displaying the transduction of extra-cellular signals by the interplay of specific receptors, kinases, transcription factors, and target genes (Figure 1.2).

C.4 Regulatory network databases

Identifying appropriate databases for regulatory network information is not trivial. Despite ongoing efforts to establish comprehensive collections of regulatory networks such as the `Network Portal` [266], regulatory knowledge remains widespread distributed over the literature, compendia of experiments and specific databases. Thus, compilation of regulatory networks for specific research questions is cumbersome and typically involves assembling of information across various resources for protein-protein interactions [267] and transcriptional regulatory networks [268]. Although several resources listed in the following attempt to collect regulatory information across species, they often lack the quality of the well-curated networks for individual organisms. Thus, if available, specific resources for the organism under study are typically the better choice.

C.4.1 Cross-species resources

The `iRefIndex` [269] database compiles protein-protein interactions from 13 major databases including `BioGRID` [270] and `IntAct` [271]. Promising initial efforts for establishing a similar database for gene regulatory networks are the `RegTransBase` [272] for transcriptional regulation in prokaryotes and `ORegAnno` [273] for regulatory information from ChIP experiments of 19 species.

C.4.2 Species-specific resources

The availability of species-specific regulatory information is typically restricted to certain model organisms and tightly linked to the complexity of the organism under study. Resources for simple organisms such as the `RegulonDB` [32, 44] for *E. coli*, the `MTB Network Portal` [274] for *M. tuberculosis*, and `Yeastract` [33, 45] for *S. cerevisiae* are among the most comprehensive and detailed resources available for transcriptional regulatory interactions. On the other hand, databases for more complex organisms such as `AGRIS` [275] for *A. thaliana* and `REDFly` [46] for *D. melanogaster* also store more complex regulatory information, e.g. including information on developmental stages, but concentrate predominantly on transcription factor binding sites determined from ChIP experiments. Most importantly, for *H. sapiens* this includes the large series of experiments on tissue-specific transcriptional regulation carried out by ENCODE [229], but also recent computer-assisted curation of the human regulatory network from the literature [276, 277].

Bibliography

[1] Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, and Jackson RB. *Campbell Biology*, chapter 4. The structure and function of large biological molecules. Benjamin Cummings, 10th edition, 2014.

[2] Lodish H, Berk A, Kaiser C, Krieger M, Bretscher A, et al. *Molecular Cell Biology*, chapter 3. Protein structure and function. WH Freeman, 7th edition, 2012.

[3] Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, and Jackson RB. *Campbell Biology*, chapter 17. Gene expression: from gene to protein. Benjamin Cummings, 10th edition, 2014.

[4] Berg JM, Tymoczko JL, Gatto GJ, and Stryer L. *Biochemistry*, chapter 4. DNA, RNA, and the flow of genetic information. WH Freeman, 8th edition, 2015.

[5] Lodish H, Berk A, Kaiser C, Krieger M, Bretscher A, et al. *Molecular Cell Biology*, chapter 7. Transcriptional control of gene expression. WH Freeman, 7th edition, 2012.

[6] Lowe SW and Lin AW. Apoptosis in cancer. *Carcinogenesis*, 21(3):485–95, 2000.

[7] Hanahan D and Weinberg RA. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

[8] Palmblad M, Henkel CV, Dirks RP, Meijer AH, Deelder AM, and Spaink HP. Parallel deep transcriptome and proteome analysis of zebrafish larvae. *BMC Res Notes*, 6:428, 2013.

[9] Ponnala L, Wang Y, Sun Q, and van Wijk KJ. Correlation of mRNA and protein abundance in the developing maize leaf. *Plant J*, 78(3):424–40, 2014.

[10] Bevilacqua A, Ceriani MC, Capaccioli S, and Nicolin A. Post-transcriptional regulation of gene expression by degradation of messenger RNAs. *J Cell Physiol*, 195(3):356–72, 2003.

[11] Malone JH and Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9:34, 2011.

[12] van Dijk EL, Auger H, Jaszczyszyn Y, and Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*, 30(9):418–26, 2014.

[13] Quackenbush J. Microarray data normalization and transformation. *Nat Genet*, 32:496–501, 2002.

[14] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, 14(6):671–83, 2013.

[15] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–54, 2002.

[16] Fang Z, Martin J, and Wang Z. Statistical methods for identifying differentially expressed genes in RNA-seq experiments. *Cell Biosci*, 2(1):26, 2012.

[17] Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, et al. A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst*, 104(4):311–25, 2012.

[18] Fumagalli D, Blanchet-Cohen A, Brown D, Desmedt C, Gacquer D, et al. Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-sequencing technology. *BMC Genomics*, 15:1008, 2014.

[19] Chan SS and Kyba M. What is a master regulator? *J Stem Cell Res Ther*, pii:114, 2013.

[20] Huang da W, Sherman BT, and Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, 2009.

[21] Khatri P, Sirota M, and Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 2012.

[22] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–9, 2000.

[23] Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*, 43(Database issue):D1049–56, 2015.

[24] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 27:29–34, 1999.

[25] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, and Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*, 42(Database issue):D199–205, 2014.

[26] Breitling R, Amtmann A, and Herzyk P. Iterative Group Analysis (iGA): a simple method to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5:34, 2004.

[27] Goeman JJ and Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23:980–987, 2007.

[28] Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, et al. Methods and approaches in the topology-based analysis of biological pathways. *Front Physiol*, 4:278, 2013.

[29] Mitra K, Carvunis AR, Ramesh SK, and Ideker T. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet*, 14(10):719–32, 2013.

[30] Ge H, Liu Z, Church GM, and Vidal M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet*, 29:482–6, 2001.

[31] Jansen R, Greenbaum D, and Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12:37–46, 2002.

[32] Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res*, 41(Database issue):D203–13, 2013.

[33] Teixeira MC, Monteiro PT, Guerreiro JF, Goncalves JP, Mira NP, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Res*, 42(Database issue):D161–6, 2014.

[34] Czarnecki J and Shepherd AJ. Mining biological networks from full-text articles. *Methods Mol Biol*, 1159:135–45, 2014.

[35] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, et al. Wisdom of crowds for robust gene network inference. *Nat Methods*, 9(8):796–804, 2012.

[36] Liu Y, Niculescu-Mizil A, Lozano A, and Lu Y. Temporal graphical models for cross-species gene regulatory network discovery. *J Bioinform Comput Biol*, 9(2):231–50, 2011.

[37] Pesch R and Zimmer R. Complementing the eukaryotic protein interactome. *PLoS One*, 8(6):e66635, 2013.

[38] Bushel PR, Heard NA, Gutman R, Liu L, Peddada SD, et al. Dissecting the fission yeast regulatory network reveals phase-specific control elements of its cell cycle. *BMC Syst Biol*, 3:93, 2009.

[39] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, 2000.

[40] Chen D, Toone WM, Mata J, Lyne R, Burns G, et al. Global transcriptional responses of fission yeast to environmental stress. *Mol Biol Cell*, 14(1):214–29, 2003.

[41] Kanehisa M, Goto S, Sato Y, Furumichi M, and Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–14, 2012.

[42] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–7, 2014.

[43] Pesch R. *Cross-Species Network and Transcript Transfer*. LMU München, München, 2016.

[44] Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*, 39(Database issue):D98–105, 2011.

[45] Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenco AB, et al. YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res*, 39(Database issue):D136–40, 2011.

[46] Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res*, 39(Database issue):D118–23., 2011.

[47] Zaman S, Lippman SI, Zhao X, and Broach JR. How Saccharomyces responds to nutrients. *Annu Rev Genet*, 42:27–81, 2008.

[48] Kel OV, Romaschenko AG, Kel AE, Wingender E, and Kolchanov NA. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res*, 23(20):4097–103, 1995.

[49] Balaji S, Babu MM, Iyer LM, Luscombe NM, and Aravind L. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol*, 360(1):213–27, 2006.

[50] Walhout AJ. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res*, 6(12):1445–54, 2006.

[51] Kim TM and Park PJ. Advances in analysis of transcriptional regulatory networks. *Wiley Interdiscip Rev Syst Biol Med*, 3(1):21–35, 2011.

[52] Schüller HJ. Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Curr Genet*, 43(3):139–60, 2003.

[53] Turcotte B, Liang XB, Robert F, and Soontorngun N. Transcriptional regulation of nonfermentable carbon utilization in budding yeast. *FEMS Yeast Res*, 10(1):2–13, 2010.

[54] Miller JH. Experiments in molecular genetics. *Cold Spring Harbor Laboratory Press*, Cold Spring Harbor:NY, 1972.

[55] Alwine JC, Kemp DJ, and Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA*, 74(12):5350–4, 1977.

[56] Galas D and Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*, 5(9):3157–70, 1978.

[57] Garner MM and Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–60, 1981.

[58] Collas P. The current state of chromatin immunoprecipitation. *Mol Biotechnol*, 45(1):87–100, 2010.

[59] Buck MJ and Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–60, 2004.

[60] Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[61] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[62] Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*, 9(4):326–32, 2008.

[63] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38(Database issue):D105–10, 2010.

[64] Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, et al. Identifying transcription factor functions and targets by phenotypic activation. *Proc Natl Acad Sci USA*, 103(32):12045–50, 2006.

[65] VanGuilder HD, Vrana KE, and Freeman WM. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, 44(5):619–26, 2008.

[66] DeRisi JL, Iyer VR, and Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–6, 1997.

[67] Hu Z, Killion PJ, and Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet*, 39(5):683–7, 2007.

[68] Skrzypek MS and Hirschman J. Using the Saccharomyces Genome Database (SGD) for analysis of genomic information. *Curr Protoc Bioinformatics*, Chapter 1:Unit 1.20.1–23, 2011.

[69] Herrgard MJ, Lee BS, Portnoy V, and Palsson B. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. *Genome Res*, 16(5):627–35, 2006.

[70] Hiltunen JK, Mursula AM, Rottensteiner H, Wierenga RK, Kastaniotis AJ, et al. The biochemistry of peroxisomal beta-oxidation in the yeast Saccharomyces cerevisiae. *FEMS Microbiol Rev*, 27(1):35–64, 2003.

[71] Gurvitz A and Rottensteiner H. The biochemistry of oleate induction: transcriptional upregulation and peroxisome proliferation. *Biochim Biophys Acta*, 1763(12):1392–402, 2006.

[72] Murata T. Petri nets: properties, analysis and applications. *Proc of the IEEE*, 77:541–80, 1989.

[73] Lee DY, Zimmer R, Lee SY, Hanisch D, and Park S. Knowledge representation model for systems-level analysis of signal transduction networks. *Genome Inform*, 15(2):234–43, 2004.

[74] Lee DY, Zimmer R, Lee SY, and Park S. Colored Petri net modeling and simulation of signal transduction pathways. *Metab Eng*, 8(2):112–22, 2005.

[75] Koch I. Chapter 25: Petri nets and GRN models. In Das S, Caragea D, Welch SM, and Hsu WH, editors, *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*, pages 604–37. IGI Global, Hershey, Pennsylvania, 2010.

[76] Koch I, Reisig W, and Schreiber F. *Modeling in Systems Biology: The Petri net approach*. Springer, Berlin, 2010.

[77] Zadeh LA. Fuzzy sets. *Information and Control*, 8:338–53, 1963.

[78] Windhager L, Erhard F, and Zimmer R. Fuzzy modeling. In Koch I, Reisig W, and Schreiber F, editors, *Modeling in Systems Biology: The Petri net approach*, pages 179–204. Springer, Berlin, 2010.

[79] Küffner R, Petri T, Windhager L, and Zimmer R. Petri nets with fuzzy logic (PNFL): reverse engineering and parametrization. *PLoS One*, 5(9):e12807, 2010.

[80] Geistlinger L, Csaba G, Küffner R, Mulder N, and Zimmer R. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, 27(13):i366–73, 2011.

[81] Funahashi A, Tanimura N, Morohashi M, and Kitano H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1:159–62, 2003.

[82] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31, 2003.

[83] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489:91–100, 2012.

[84] Zheng W, Zhao H, Mancera E, Steinmetz LM, and Snyder M. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, 464:1187–91, 2010.

[85] Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, and Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150:1274–86, 2012.

[86] Wu WS, Li WH, and Chen BS. Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data. *BMC Bioinformatics*, 8:188, 2007.

[87] Petricka JJ and Benfey PN. Reconstructing regulatory network transitions. *Trends Cell Biol*, 21:442–51, 2011.

[88] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[89] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5:e8, 2007.

[90] Michoel T, De Smet R, Joshi A, Van de Peer Y, and Marchal K. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst Biol*, 3:49, 2009.

[91] Greenfield A, Hafemeister C, and Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–7, 2013.

[92] Wu M and Chan C. Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data. *Brief Bioinform*, 13:150–61, 2012.

[93] Küffner R, Petri T, Tavakkolkhah P, Windhager L, and Zimmer R. Inferring gene regulatory networks by ANOVA. *Bioinformatics*, 28:1376–82, 2012.

[94] Soranzo N, Bianconi G, and Altafini C. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, 23:1640–7, 2007.

[95] Narendra V, Lytkin NI, Aliferis CF, and Statnikov A. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics*, 97:7–18, 2011.

[96] Naeem H, Zimmer R, Tavakkolkhah P, and Küffner R. Rigorous assessment of gene set enrichment tests. *Bioinformatics*, 28:1480–6, 2012.

[97] Ciofani M, Madar A, Galan C, Sellars M, Mace K, et al. A validated regulatory network for Th17 cell specification. *Cell*, 151(2):289–303, 2012.

[98] Qian J, Lin J, Luscombe NM, Yu H, and Gerstein M. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19:1917–26, 2003.

[99] Mordelet F and Vert JP. SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24:i76–82, 2008.

[100] De Smet R and Marchal K. Advantages and limitations of current network inference methods. *Nat Rev Microbiol*, 8:717–29, 2010.

[101] Simpson EH. The interpretation of interaction in contingency tables. *J R Stat Soc (Series B)*, 13:238–41, 1951.

[102] Pearl J. *Causality.* Cambridge University Press, 2nd edition, 2009.

[103] Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res*, 36:D866–70, 2008.

[104] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39:D1005–10, 2011.

[105] MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, and Fraenkel E. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:113, 2006.

[106] Dorogovtsev SN and Mendes JF. *Evolution of networks: from biological nets to the internet and WWW*. Oxford University Press, 2003.

[107] Pavlidis P and Gillis J. Progress and challenges in the computational prediction of gene function using networks: 2012-2013 update. *F1000Res*, 2:230, 2013.

[108] Gillis J and Pavlidis P. The impact of multifunctional genes on "guilt by association" analysis. *PLoS One*, 6:e17258, 2011.

[109] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34:166–76, 2003.

[110] Mordelet F and Vert JP. A bagging SVM to learn from positive and unlabeled examples. Technical report, 2010.

[111] Jason Ernst, Beg QK, Kay KA, Balazsi G, Oltvai ZN, and Bar-Joseph Z. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Comput Biol*, 4:e1000044, 2008.

[112] Yip KY, Kim PM, McDermott D, and Gerstein M. Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels. *BMC Bioinformatics*, 10:241, 2009.

[113] Holloway DT, Kon M, and DeLisi C. Classifying transcription factor targets and discovering relevant biological features. *Biol Direct*, 3:22, 2008.

[114] The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, 38:D331–5, 2010.

[115] Ozcan S and Johnston M. Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose. *Mol Cell Biol*, 15:1564–72, 1995.

[116] François J and Parrou JL. Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae. *FEMS Microbiol Rev*, 25:125–45, 2001.

[117] Lorenz MC and Heitman J. Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains. *Genetics*, 150:1443–57, 1998.

[118] Morano KA, Grant CM, and Moye-Rowley WS. The response to heat shock and oxidative stress in Saccharomyces cerevisiae. *Genetics*, 190:1157–95, 2012.

[119] Myers CL, Barrett DR, Hibbs MA, Huttenhower C, and Troyanskaya OG. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7:187, 2006.

[120] Ambroise J, Robert A, Macq B, and Gala JL. Transcriptional network inference from functional similarity and expression data: a global supervised approach. *Stat Appl Genet Mol Biol*, 11(1):Article 2, 2012.

[121] Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95:14863–8, 1998.

[122] Khatri P and Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21:3587–3595, 2005.

[123] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102:15545–15550, 2005.

[124] Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, et al. Gene-set analysis and reduction. *Brief Bioinform*, 10:24–34, 2009.

[125] Barry WT, Nobel AB, and Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21:1943–9, 2005.

[126] Ulitsky I and Shamir R. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol*, 1:8, 2007.

[127] Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, et al. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet*, 3:e96, 2007.

[128] Lee HK, Hsu AK, Sajdak J, Qin J, and Pavlidis P. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14:1085–94, 2004.

[129] Küffner R, Zimmer R, and Lengauer T. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16:825–36, 2000.

[130] Gatti DM, Barry WT, Nobel AB, Rusyn I, and Wright FA. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11:574, 2010.

[131] Ihaka R and Gentleman R. R: A language for data analysis and graphics. *J Comp Graph Stat*, 5:299–314, 1996.

[132] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5:80, 2004.

[133] Goecks J, Nekrutenko A, Taylor J, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11:86, 2010.

[134] Keller A, Backes C, Gerasch A, Kaufmann M, Kohlbacher O, et al. A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics*, 25:2787–94, 2009.

[135] Miller FD and Kaplan DR. Neurotrophin signalling pathways regulating neuronal apoptosis. *Cell Mol Life Sci*, 58:1045–53, 2001.

[136] Schramm G, Wiesberg S, Diessl N, Kranz AL, Sagulenko V, et al. PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics*, 26:1225–1231, 2010.

[137] Ohmichi M, Decker S J, Pang L, and Saltiel A R. Nerve growth factor binds to the 140 kd trk proto-oncogene product and stimulates its association with the src homology domain of phospholipase C gamma 1. *Biochem Biophys Res Commun*, 179:217–23, 1991.

[138] Borrello MG, Pelicci G, Arighi E, De Filippis L, Greco A, et al. The oncogenic versions of the Ret and Trk tyrosine kinases bind Shc and Grb2 adaptor proteins. *Oncogene*, 9:1661–8, 1994.

[139] Evangelopoulos ME, Weis J, and Kruttgen A. Neurotrophin effects on neuroblastoma cells: correlation with trk and p75NTR expression and influence of Trk receptor bodies. *J Neurooncol*, 66:101–10, 2004.

[140] Walker SR, Ogagan PD, DeAlmeida D, Aboka AM, and Barksdale EM Jr. Neuroblastoma impairs chemokine-mediated dendritic cell migration in vitro. *J Pediatr Surg*, 41:260–5, 2006.

[141] Meier R, Mühlethaler-Mottet A, Flahaut M, Coulon A, Fusco C, et al. The chemokine receptor cxcr4 strongly promotes neuroblastoma primary tumour and metastatic growth, but not invasion. *PLoS One*, 2:e1016, 2007.

[142] Darios F and Davletov B. Omega-3 and omega-6 fatty acids stimulate cell membrane expansion by acting on syntaxin 3. *Nature*, 440:813–7., 2006.

[143] Larson JL and Owen A. Moment based gene set tests. *BMC Bioinformatics*, 16:132, 2015.

[144] Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, et al. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–8, 2004.

[145] Maurer LM, Yohannes E, Bondurant SS, Radmacher M, and Slonczewski JL. pH regulates genes for flagellar motility, catabolism, and oxidative stress in Escherichia coli K-12. *J Bacteriol*, 187(1):304–19, 2005.

[146] Lin L, Wagner MC, Cocklin R, Kuzma A, Harrington M, et al. The antibiotic gentamicin inhibits specific protein trafficking functions of the Arf1/2 family of GTPases. *Antimicrob Agents Chemother*, 55(1):246–54, 2011.

[147] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue):D991–5, 2013.

[148] Li X. *ALL: A data package.* R package version 1.7.0.

[149] Tarca AL, Draghici S, Bhatti G, and Romero R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136, 2012.

[150] Bhatti G and Tarca AL. *KEGGdzPathwaysGEO: KEGG Disease Datasets from GEO.* R package version 1.3.1.

[151] Villaveces JM, Koti P, and Habermann BH. Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv Appl Bioinform Chem*, 8:11–22, 2015.

[152] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

[153] Luo W and Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1, 2013.

[154] Eden E, Navon R, Steinfeld I, Lipson D, and Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, 2009.

[155] Wu C, Zhu C, and Jegga AG. Integrative literature and data mining to rank disease candidate genes. *Methods Mol Biol*, 1159:207–26, 2014.

[156] Donato M, Xu Z, Tomoiaga A, Granneman JG, Mackenzie RG, et al. Analysis and correction of crosstalk effects in pathway analysis. *Genome Res*, 23(11):1885–93, 2013.

[157] Ganju J and Ma GJ. The potential for increased power from combining p-values testing the same hypothesis. *Stat Methods Med Res*, 2014.

[158] Han Y and Garcia BA. Combining genomic and proteomic approaches for epigenetics research. *Epigenomics*, 5(4):439–52, 2013.

[159] R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL: `http://www.R-project.org`.

[160] Smyth GK. Linear models and empirical Bayes for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):A1, 2004.

[161] Risso D, Schwartz K, Sherlock G, and Dudoit S. GC-content normalization for RNA-seq data. *BMC Bioinformatics*, 2:480, 2011.

[162] Carlson M. *hgu95av2.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95av2).* R package version 2.14.0.

[163] Law CW, Chen Y, Shi W, and Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, 15:R29, 2014.

[164] Robinson MD, McCarthy DJ, and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–40, 2010.

[165] Love MI, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15:550, 2014.

[166] Shaffer JP. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–576, 1995.

[167] Gene set file formats [online]. URL: `http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#Gene_Set_Database_Formats`.

[168] Pathway xml format [online]. URL: `http://www.kegg.jp/kegg/xml`.

[169] Alexa A and Rahnenführer J. *topGO: Enrichment analysis for Gene Ontology.* R package version 2.20.0.

[170] Tenenbaum D. *KEGGREST: Client-side REST access to KEGG.* R package version 1.5.2.

[171] Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8:242, 2007.

[172] GSEA [online]. URL: `http://www.broadinstitute.org/gsea`.

[173] SAMGS [online]. URL: `https://www.ualberta.ca/yyasui/SAM-GS`.

[174] Geistlinger L, Csaba G, Dirmeier S, Küffner R, and Zimmer R. A comprehensive gene regulatory network for the diauxic shift in Saccharomyces cerevisiae. *Nucleic Acids Res*, 41(18):8452–63, 2013.

[175] Zhang JD and Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. *Bioinformatics*, 25(11):1470–71, 2009.

[176] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009.

[177] Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, 13:226, 2012.

[178] Dutta B, Wallqvist A, and Reifman J. PathNet: a tool for pathway analysis using topological information. *Source Code Biol Med*, 7(1):10, 2012.

[179] Huntley MA, Larson JL, Chaivorapol C, Becker G, Lawrence M, et al. ReportingTools: an automated results processing and presentation toolkit for high throughput genomic analyses. *Bioinformatics*, 29(24):3220–1, 2013.

[180] Long L, Gentleman R, and Hahne F. *biocGraph: Graph examples and use cases in Bioinformatics.* R package version 1.30.0.

[181] Podpecan V, Lavrac N, Mozetic I, Novak PK, Trajkovski I, et al. Segmine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, 12:416, 2011.

[182] Sales G, Calura E, Martini P, and Romualdi C. Graphite web: Web tool for gene set analysis exploiting pathway topology. *Nucleic Acids Res*, 41(Web Server issue):W89–97, 2013.

[183] The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45(10):1113–20, 2013.

[184] The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

[185] Rahman M, Jackson LK, Johnson WE, Li DY, Bild AH, and Piccolo SR. Alternative preprocessing of RNA-sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics*, pii:btv377, 2015.

[186] Tarca AL, Bhatti G, and Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One*, 8(11):e79217, 2013.

[187] Phipson B and Smyth GK. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*, 9:A39, 2010.

[188] Merico D, Isserlin R, Stueker O, Emili A, and Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, 5(11):e13984, 2010.

[189] Rhee SY, Wood V, Dolinski K, and Draghici S. Use and misuse of the gene ontology annotations. *Nat Rev Genet*, 9(7):509–15, 2008.

[190] Newton MA and Wang Z. Multiset statistics for gene set analysis. *Annu Rev Stat Appl*, 2:95–111, 2015.

[191] Barriot R, Sherman DJ, and Dutour I. How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. *BMC Bioinformatics*, 8:332, 2007.

[192] Sass S, Buettner F, Mueller NS, and Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res*, 41(21):9622–33, 2013.

[193] Cao J and Zhang S. A bayesian extension of the hypergeometric test for functional enrichment analysis. *Biometrics*, 70(1):84–94, 2014.

[194] Bauer S, Gagneur J, and Robinson PN. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res*, 38(11):3523–32, 2010.

[195] Alexa A, Rahnenführer J, and Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–7, 2006.

[196] Grossmann S, Bauer S, Robinson PN, and Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, 23(22):3024–31, 2007.

[197] Ma J, Sartor MA, and Jagadish HV. Appearance frequency modulated gene set enrichment testing. *BMC Bioinformatics*, 12:81, 2011.

[198] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–73, 2003.

[199] Yan X and Sun F. Testing gene set enrichment for subset of genes: Sub-GSE. *BMC Bioinformatics*, 9:362, 2008.

[200] Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*, 47(2):106–14, 2015.

[201] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–45, 2010.

[202] Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, and Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–12, 2004.

[203] Reiss DJ, Baliga NS, and Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7:280, 2006.

[204] Kim TM, Chung YJ, Rhyu MG, and Jung MH. Inferring biological functions and associated transcriptional regulators using gene set expression coherence analysis. *BMC Bioinformatics*, 8:453, 2007.

[205] Parikh JR, Xia Y, and Marto JA. Multi-edge gene set networks reveal novel insights into global relationships between biological themes. *PLoS One*, 7(9):e45211, 2012.

[206] Mark D, Haeberle S, Roth G, von Stetten F, and Zengerle R. Microfluidic lab-on-a-chip platforms: requirements, characteristics and applications. *Chem Soc Rev*, 39(3):1153–82, 2010.

[207] Rothbauer M, Wartmann D, Charwat V, and Ertl P. Recent advances and future applications of microfluidic live-cell microarrays. *Biotechnol Adv*, 33(6):948–61, 2015.

[208] Citri A, Pang ZP, Südhof TC, Wernig M, and Malenka RC. Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat Protoc*, 7(1):118–27, 2011.

[209] Stahlberg A and Kubista M. The workflow of single-cell expression profiling using quantitative real-time PCR. *Expert Rev Mol Diagn*, 14(3):323–31, 2014.

[210] Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. A draft map of the human proteome. *Nature*, 509(7502):575–81, 2014.

[211] Wilhelm M, Schlegl J, Hahne H, Moghaddas GA, Lieberenz M, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–7, 2014.

[212] Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, 4(11):e1000282, 2008.

[213] Bartel J, Krumsiek J, and Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J*, 4:e201301009, 2013.

[214] Abraham PE, Giannone RJ, Xiong W, and Hettich RL. Metaproteomics: extracting and mining proteome information to characterize metabolic activities in microbial communities. *Curr Protoc Bioinformatics*, 46:13.26.1–14, 2014.

[215] Chavez-Dozal A, Gorman C, and Nishiguchi MK. Proteomic and metabolomic profiles demonstrate variation among free-living and symbiotic vibrio fischeri biofilms. *BMC Microbiol*, 15(1):226, 2015.

[216] Imam S, Schäuble S, Brooks AN, Baliga NS, and Price ND. Data-driven integration of genome-scale regulatory and metabolic network models. *Front Microbiol*, 6:409, 2015.

[217] Steen HC and Gamero AM. STAT2 phosphorylation and signaling. *JAKSTAT*, 2(4):e25790, 2013.

[218] van Loosdregt J and Coffer PJ. Post-translational modification networks regulating FOXP3 function. *Trends Immunol*, 35(8):368–78, 2014.

[219] Lu T and Stark GR. NF-kB: Regulation by methylation. *Cancer Res*, 75(18):3692–5, 2015.

[220] Ocone A, Haghverdi L, Mueller NS, and Theis FJ. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–96, 2015.

[221] Li Y, Varala K, and Coruzzi GM. From milliseconds to lifetimes: tracking the dynamic behavior of transcription factors in gene networks. *Trends Genet*, 31(9):509–15, 2015.

[222] Vaissiere T, Sawan C, and Herceg Z. Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutat Res*, 659(1):40–8, 2008.

[223] Breiling A and Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin*, 8:24, 2015.

[224] Annalisa I and Robert S. The role of linker histone H1 modifications in the regulation of gene expression and chromatin dynamics. *Biochim Biophys Acta*, S1874(15):189–93, 2015.

[225] Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98, 2012.

[226] Sanyal A, Lajoie BR, Jain G, and Dekker J. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–13, 2012.

[227] Shlyueva D, Stampfel G, and Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, 15(4):272–86, 2014.

[228] Levine M, Cattoglio C, and Tjian R. Looping back to leap forward: transcription enters a new era. *Cell*, 157(1):13–25, 2014.

[229] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[230] Dassi E and Quattrone A. Tuning the engine: an introduction to resources on post-transcriptional regulation of gene expression. *RNA Biol*, 9(10):1224–32, 2012.

[231] Beckwith EJ and Yanovsky MJ. Circadian regulation of gene expression: at the crossroads of transcriptional and post-transcriptional regulatory networks. *Curr Opin Genet Dev*, 27:35–42, 2014.

[232] Meyer SU, Stoecker K, Sass S, Theis FJ, and Pfaffl MW. Posttranscriptional regulatory networks: from expression profiling to integrative analysis of mRNA and microRNA data. *Methods Mol Biol*, 1160:165–88, 2014.

[233] Markowetz F, Bloch J, and Spang R. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21(21):4026–32, 2005.

[234] Sedgewick AJ, Benz SC, Rabizadeh S, Soon-Shiong P, and Vaske CJ. Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM. *Bioinformatics*, 29(13):i62–70, 2013.

[235] Sass S, Buettner F, Mueller NS, and Theis FJ. RAMONA: a web application for gene set analysis on multilevel omics data. *Bioinformatics*, 31(1):128–30, 2015.

[236] Krämer A, Green J, Pollard JJ, and Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–30, 2014.

[237] Huang SS and Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal*, 2(81):ra40, 2009.

[238] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:6874, 2015.

[239] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, et al. ArrayExpress update–simplifying data submissions. *Nucleic Acids Res*, 43(Database issue):D1113–6, 2015.

[240] Sanger F, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. *PNAS*, 74(12):5463–7, 1977.

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

[242] Pettersson E, Lundeberg J, and Ahmadian A. Generations of sequencing technologies. *Genomics*, 93(2):105–11, 2009.

[243] Wang Z, Gerstein M, and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009.

[244] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–80, 2009.

[245] Brenowitz M, Senear DF, and Kingston RE. DNase I footprint analysis of protein-DNA binding. *Curr Protoc Mol Biol*, Chapter 12:Unit 12.4, 2001.

[246] Song L and Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2, 2010.

[247] Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4):283–9, 2009.

[248] Ewens WJ and Grant RG. *Statistical Methods in Bioinformatics: An Introduction.* Springer, Berlin, 2005.

[249] Welch BL. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1):28–35, 1947.

[250] Agresti A. A survey of exact inference for contingency tables. *Stat Science*, 7(1):131–153, 1992.

[251] Corder GW and Foreman DI. *Nonparametric Statistics: A Step-by-Step Approach*, chapter 8. Tests for nominal scale data: chi-square and Fisher exact tests. Wiley, 2nd edition, 2014.

[252] Wilcoxon F. Individual comparisons by ranking methods. *Biometrics*, 1(6):803, 1945.

[253] Corder GW and Foreman DI. *Nonparametric Statistics: A Step-by-Step Approach*, chapter 4. Comparing two unrelated samples: the Mann-Whitney U-test and the Kolmogorov-Smirnov two-sample test. Wiley, 2nd edition, 2014.

[254] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc*, 57(1):289–300, 1995.

[255] FlyBase Consortium. FlyBase : a database for the Drosophila research community. *Methods Mol Biol*, 420:45–59, 2008.

[256] Cunningham F, Amode MR, Barrell D, Beal K, Billis K, et al. Ensembl 2015. *Nucleic Acids Res*, 43(Database issue):D662–9, 2015.

[257] Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, et al. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*, 43(Database issue):D670–81, 2015.

[258] Edgar R, Domrachev M, and Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, 2002.

[259] Leinonen R, Sugawara H, and Shumway M. The sequence read archive. *Nucleic Acids Res*, 39(Database issue):D1921, 2011.

[260] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, and Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–40, 2011.

[261] Liberzon A. A description of the Molecular Signatures Database (MSigDB) web site. *Methods Mol Biol*, 150:153–60, 2014.

[262] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, et al. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14:128, 2013.

[263] Bauer-Mehren A, Furlong LI, and Sanz F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Syst Biol*, 5:290, 2009.

[264] Ooi HS, Schneider G, Lim TT, Chan YL, Eisenhaber B, and Eisenhaber F. Biomolecular pathway databases. *Methods Mol Biol*, 609:129–44, 2010.

[265] Bader GD, Cary MP, and Sander C. Pathguide: a pathway resource list. *Nucleic Acids Res*, 34(Database issue):D504–6, 2006.

[266] Turkarslan S, Wurtmann EJ, Wu WJ, Jiang N, Bare JC, et al. Network portal: a database for storage, analysis and visualization of biological networks. *Nucleic Acids Res*, 42(Database issue):D184–90, 2014.

[267] Ooi HS, Schneider G, Chan YL, Lim TT, Eisenhaber B, and Eisenhaber F. Databases of protein-protein interactions and complexes. *Methods Mol Biol*, 609:145–59, 2010.

[268] Bacha J, Brodie JS, and Loose MW. myGRN: a database and visualisation system for the storage and analysis of developmental genetic regulatory networks. *BMC Dev Biol*, 9:33, 2009.

[269] Razick S, Magklaras G, and Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9:405, 2008.

[270] Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res*, 43(Database issue):D470–8, 2015.

[271] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42(Database issue):D358–63, 2014.

[272] Cipriano MJ, Novichkov PN, Kazakov AE, Rodionov DA, Arkin AP, et al. RegTransBase–a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics*, 14:213, 2013.

[273] Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*, 36(Database issue):D107–13, 2008.

[274] Turkarslan S, Peterson EJ, Rustad TR, Minch KJ, Reiss DJ, et al. A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis. *Sci Data*, 2:150010, 2015.

[275] Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, and Grotewold E. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res*, 39(Database issue):D1118–22, 2011.

[276] Thomas P, Durek P, Solt I, Klinger B, Witzel F, et al. Computer-assisted curation of a human regulatory core network from the biological literature. *Bioinformatics*, 31(8):1258–66, 2015.

[277] Han H, Shim H, Shin D, Shim JE, Ko Y, et al. Trrust: a reference database of human transcriptional regulatory interactions. *Sci Rep*, 5:11432, 2015.

Acknowledgements

I would like to thank all contributors of the publications integrated in this work, namely Simon Dirmeier, Stefan Altman, Tobias Petri, Nicola Mulder, Gergely Csaba, Robert Küffner, and Ralf Zimmer, for the good cooperation and their courtesy to include the shared work in here.

I am especially grateful to Gergely Csaba, for all his technical advice and support; to Robert Küffner, for being a humorous but merciless reviewer of most of my papers; and to Ralf Zimmer, for engaging in my academic development since early in my undergraduate studies, being a mentor during my diploma thesis as well as my PhD thesis, and for being encouraging as well as challenging up to the present day.

My deepest thanks are going to my parents, Ute Rosenbaum and Helmut Geistlinger. Thank you for showing me my path in life. This work is dedicated to you.

Curriculum Vitae

Ludwig Geistlinger, born in Leipzig, Germany, on December 10th, 1984.

Education

1995 - 2003	Abitur (A-levels Equiv.), Wilhelm-Ostwald-Gymnasium, Leipzig, Germany
2004 - 2010	Dipl.-Bioinf. (M.Sc. Equiv.), LMU & TU München, München, Germany
2011 - 2016	PhD, LMU München, München, Germany

Experience

2006 - 2010	Student assistant, Institute of Genetic Epidemiology, Helmholtz-Center Munich, Neuherberg, Germany
2008	Student assistant, Institute of Biostatistics, LMU München, München, Germany
04-07/2009	Research stay, Shanghai Center for Bioinformation Technology, Shanghai, China
04-10/2010	Research stay, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa
07/2011	Research visit, Research and Training Center on Bioinformatics, Lomonosov Moscow State University, Moscow, Russia
10/2015	Research visit, Escola de Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, São Paulo, Brazil

Publications

Silva VH, Regitano LCA, Geistlinger L, Pertille F, Giachetto PF, Brassaloti RA, Zimmer R, Coutinho LL. Genome-wide detection of CNVs and their association with meat tenderness in Nelore cattle. *PLoS One*, under review, 2016.

Geistlinger L, Csaba G, Zimmer R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics*, 17:45, 2016.

Petri T, Altmann S, Geistlinger L, Zimmer R, Küffner R. Addressing false discoveries in network inference. *Bioinformatics*. 31(17):2836-43, 2015.

Geistlinger L, Csaba G, Dirmeier S, Küffner R, Zimmer R. A comprehensive gene regulatory network for the diauxic shift in Saccharomyces cerevisiae. *Nucleic Acids Res*, 41(18):8452-63, 2013.

Kurome M, Geistlinger L, Kessler B, Zakhartchenko V, Klymiuk N, Wuensch A, Richter A, Baehr A, Kraehe K, Burkhardt K, Flisikowski K, Flisikowska T, Merkl C, Landmann M, Durkovic M, Tschukes A, Kraner S, Schindelhauer D, Petri T, Kind A, Nagashima H, Schnieke A, Zimmer R, Wolf E. Factors influencing the efficiency of generating genetically engineered pigs by nuclear transfer: multi-factorial analysis of a large data set. *BMC Biotechnol*, 13:43, 2013.

Geistlinger L, Csaba G, Küffner R, Mulder N, Zimmer R. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, 27(13):i366-73, 2011.

Wang Z, Ding G, Geistlinger L, Li H, Liu L, Zeng R, Tateno Y, Li Y. Evolution of protein phosphorylation for distinct functional modules in vertebrate genomes. *Mol Biol Evol*, 28(3):1131-40, 2011.

Marzi C, Albrecht E, Hysi PG, Lagou V, Waldenberger M, Tönjes A, Prokopenko I, Heim K, Blackburn H, Ried JS, Kleber ME, Mangino M, Thorand B, Peters A, Hammond CJ, Grallert H, Boehm BO, Kovacs P, Geistlinger L, Prokisch H, Winkelmann BR, Spector TD, Wichmann HE, Stumvoll M, Soranzo N, März W, Koenig W, Illig T, Gieger C. Genome-wide association study identifies two novel regions at 11p15.5-p13 and 1p31 with major impact on acute-phase serum amyloid A. *PLoS Genet*, 6(11):e1001213, 2010.

Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J, Illig T, Suhre K. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, 4(11):e1000282, 2008.

Software

Geistlinger L, Csaba G, Zimmer R. EnrichmentBrowser: seamless navigation through combined results of set-based and network-based enrichment analysis. *Bioconductor*, R package version 2.0.0, 2014.

Risso D, Geistlinger L, Dudoit S. EDASeq: exploratory data analysis and normalization for RNA-seq data. *Bioconductor*, R package version 2.4.0, 2015.