Expression Data Analysis and Regulatory Network Inference by Means of Correlation Patterns

Tobias Hartmut Petri

München 2017

Expression Data Analysis and Regulatory Network Inference by Means of Correlation Patterns

Tobias Hartmut Petri

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

vorgelegt von
Tobias Hartmut Petri
aus München

München, den 26. April 2017

Erstgutachter: Prof. Dr. Ralf Zimmer
Zweitgutachter: Prof. Dr. Karsten Borgwardt
Tag der mündlichen Prüfung: 28. Juli 2017

Eidesstattliche Versicherung

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

Hiermit erkläre ich, Tobias Petri, an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Ort, Datum	Doktorand

Contents

List of Abbreviations **xi**

List of Figures **xv**

List of Tables **xvii**

German Abstract **xix**

English Abstract **xxi**

1 Introduction **1**

1.1 From Complex Systems to Correlated Data 2

1.1.1 Indirect Observations 2

1.1.2 Regulatory Systems 3

1.1.3 Formation of Correlation Patterns 4

1.1.4 Detection of Correlation Patterns 6

1.2 Assessment of Biological Relevance 7

1.2.1 Statistical Testing 7

1.2.2 Sources of Variation and Noise 9

1.3 Modelling Correlation Patterns 11

1.3.1 Model Regularization 11

1.3.2 Coping with Outliers 11

1.3.3 Multiple Hypothesis Testing under Dependency 12

1.3.4 Influence of Subgroups on Model Evaluation 13

1.4 This Work 15

2 Correlated Mutations **19**

2.1 Overview 20

2.2 Contact Promotes Conservation 20

2.2.1 A Co-evolution Anecdote 20

2.2.2 Co-Evolution among Proteins 21

2.2.3 Co-Evolution within Individual Proteins 21

2.2.4 A Real-World Problem of Antibiotic Resistance 22
2.3 Related Work 23
2.4 Methods 24
2.4.1 Data Acquisition 24
2.4.2 Alignment Correlation 25
2.5 Results 26
2.6 Discussion 27

3 Artifact Detection 29
3.1 Experimental Artifacts 30
3.2 Existing Databases contain Artifacts 30
3.3 Noise Detection on Probe Level Data 31
3.4 Related Work 32
3.5 Outline and Experimental Setup 32
3.6 Microarray Measurements 34
3.6.1 Data 34
3.6.2 Summarization and Normalization 34
3.7 Quality Assessment 35
3.7.1 Probe Noise Score 35
3.7.2 Probe Noise Plot 36
3.7.3 Replicate Scatter Plot 36
3.7.4 Introducing Artificial Noise 36
3.8 Artifact Detection 37
3.8.1 ϵ-criterion 37
3.8.2 Window Criterion 37
3.9 Correction of Corrupted Probes 38
3.10 Evaluation of Artifact Detection 38
3.11 Probe Noise Plots 38
3.12 Insufficient Correction by Summarization 39
3.13 Sensitivity of Summarization to Noise 42
3.14 Identification and Correction of Corrupted Probes 42
3.15 Accuracy of Artifact Identification 44
3.16 Compared Methods 44
3.17 Evaluation Results 45
3.18 Conclusions 47

4 Differential Analysis 49
4.1 Differential Analysis 50
4.2 Experiment Specificity 50
4.3 Overview of *Padesco* 52
4.3.1 Patterns 52

4.3.2 Deviations . . . 53
4.3.3 Evaluation . . . 53
4.3.4 Scoring . . . 53
4.4 Related Work . . . 53
4.5 Data Sets . . . 54
4.6 Basic Protocol . . . 55
4.6.1 Differential Analysis . . . 56
4.6.2 Support Vector Regression . . . 56
4.6.3 Model Building . . . 58
4.6.4 Fold-change Matrix . . . 58
4.6.5 Selection of Parameters during Training . . . 61
4.6.6 Scoring Performance Deviation . . . 61
4.6.7 Measure of Expectation . . . 61
4.6.8 Permutation Test for the Evaluation of *Padesco* . . . 62
4.7 Results . . . 62
4.7.1 Evaluation of Expression Fold-Change Predictions . . . 62
4.7.2 Prediction of Fold-changes for Individual Genes . . . 64
4.7.3 Permutation Test Based Evaluation . . . 64
4.7.4 Specific Candidates . . . 66
4.8 Discussion . . . 68

5 Confidence Recalibration 71
5.1 Genome-scale Models . . . 72
5.2 The Need for Computational Approaches . . . 73
5.3 Integration of Known Topologies . . . 74
5.4 Network Inference Schemes . . . 76
5.4.1 Basic setup . . . 76
5.4.2 Expression-Based Approaches . . . 77
5.4.3 Supervised Approaches with Topology Prior . . . 78
5.4.4 Applied Inference Approaches . . . 81
5.4.5 Supervised Function Prediction . . . 84
5.4.6 Shared Issues . . . 85
5.5 Material and Methods . . . 87
5.5.1 Overview of Training Data . . . 87
5.5.2 Estimating the Size of the Yeast Regulatory Network . . . 88
5.5.3 A Network-Only Model . . . 89
5.5.4 Evaluation Metrics for Network Prediction . . . 90
5.5.5 Prediction Setup and Validation . . . 91
5.5.6 Confidence Recalibration (*CoRe*) . . . 91
5.5.7 Analysis of Interactions in Network Motifs . . . 93
5.5.8 Functional Coherence . . . 95

5.5.9 Derivation of Modules from the Predicted Network 97
5.6 Results 99
5.6.1 Network Predictions without Expression Data 99
5.6.2 Simpson's Paradox 99
5.6.3 Implications of Simpson's Paradox 101
5.6.4 Correction through Score Recalibration 102
5.6.5 Application of *CoRe* to Network Inference 105
5.6.6 A Comprehensive Yeast Network 107
5.6.7 Novel predictions 108
5.6.8 Performance based on Binding Sites 110
5.6.9 Robustness of *CoRe* 111
5.6.10 Dependency of Distribution Parameters on TF Out-Degree . 112
5.6.11 Out-Degree and the Number of Predictions 112
5.6.12 Score Distributions Based on Probability Estimates 114
5.6.13 Improving Regulator-Specific Predictions 115
5.6.14 Numerical Values of Performance Estimates 116
5.7 Discussion 120

6 Conclusion 123
6.1 Complex Systems 124
6.1.1 Co-Evolution and Correlated Mutations 124
6.1.2 Deviation Patterns 124
6.1.3 Removal of Noise and Artifacts 125
6.1.4 Network Evaluation 125
6.2 Outlook 126

7 Appendix 127

Acknowledgements 137

List of Abbreviations

ABC	ATP-binding cassette
ALL	Acute Lymphocytic Leukemia
AML	Acute Myeloid Leukemia
APT	Affymetrix Power Tools
ATP	Adenosin Triphosphate
AUC	Area under the ROC
AUPR	Area under the PR curve
AUROC	Area under the ROC curve (→AUC, used to contrast AUPR)
CC	Correlation Coefficient
CGS	Hypothetically Complete Gold Standard
ChIP	Chromatin Immunoprecipitation
CV	Cross Validation
DAVID	Database for Annotation, Visualization and Integrated Discovery
DE	Differential Expression
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
DREAM	Dialogue for Reverse Engineering Assessments and Methods

ENCODE	Encyclopedia of DNA Elements
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FWER	Family-wise Error-Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRN	Gene Regulatory Network
GUI	Guided User Interface
HUVEC	Human Umbilical Vein Endothelial Cells
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
LOO	Leave-one-experiment-out
LOOCV	Leave-one-out CV
MBR	Microarray Blob Removal
MIAME	Minimum Information About a Microarray Experiment
miRNA	Micro Ribonucleic Acid
MLL	Mixed Lineage Leukemia
MMM DB	Many Microbe Microarray database
mRNA	messenger RNA
n-CV	n-fold Cross-Validation
NCBI	National Center for Biotechnology Information
ODE	Ordinary Differential Equation
P50	Precision-50

Padesco	Pattern Deviation Scoring
PDS	Positive Dependence through Stochastic Ordering
PR Curve	Precision-Recall Curve
RMA	Robust Microarray Averaging
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
ROC	Receiver Operator Characteristic
SEREND	SEmi-supervised REgulatory Network Discoverer
SIRENE	Supervised Inference of REgulatory NEtworks
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
TCS	Two-component System
TF	Transcription Factor
TN	True Negatives
TP	True Positives
TPR	True Positive Rate

List of Figures

1.1 A Simple Regulatory Petri Net 5
1.2 Imbalanced Label Distributions 14
1.3 Correlation Types 16

2.1 Co-Evolution Contact Graph 27
2.2 Co-Evolution Sampling 28

3.1 Measurement Artifacts 33
3.2 Replicate Scatter Plots 40
3.3 Simulation of Fold-changes 41
3.4 Spiked Gene ST Arrays 43
3.5 Precision-Recall Curves 46

4.1 Differential Expression Setup 51
4.2 Schematic View of the Fold-Change (fc) Matrix $\mathcal{F}$. 59
4.3 Outline of *Padesco* 60
4.4 Scatter Plot of Predicted *vs.* Measured Fold-changes 63
4.5 Precision and Number of Detected Genes 65
4.6 Common Regulation and Experiment Deviation 69

5.1 Properties of Yeast Interactions 75
5.2 Supervised Inference, Regulator-Specific 80
5.3 Recalibration Outline 92
5.4 Motif Prediction Preferences 96
5.5 Functional Coherence Measure 98
5.6 Score Recalibration in Network Predictions 104
5.7 Assessment of Predicted Interactions 106
5.8 Module Representation of the Complete Result Network 109
5.9 Robustness of Score Recalibration 112
5.10 Degree Dependency of Location Parameters 113
5.11 Out-Degree and the Number of Predicted Targets 114
5.12 Degree Dependency of Probability Estimates 115

7.1 DG75-eGFP Exon Arrays 128
7.2 DG75-10/12 Exon Arrays 129
7.3 Probe Noise Plots for DG75-10/12 130
7.4 Replicate Scatter Plot for Newly Transcribed RNA 131
7.5 Scatter Plots of ϵ-criterion for Meta-Probesets 132
7.6 Scatter Plots of ϵ-criterion using Fold-changes 133
7.7 Scatter Plots of ϵ-criterion using Complementarity 134
7.8 Probe Noise Plots for Spiked Noise 135

List of Tables

1.1 Contingency Table for Few Individuals 8

2.1 Matrix Correlations 26

4.1 Classification Assignment for the Evaluation 62

5.1 Classification of Inference Approaches 81
5.2 Evaluation of Motif Classes 93
5.3 Performance of Integrative Approaches 110
5.4 Regulator Performance using only Promotor Binding Information . 111
5.5 Regulator Performance after Calibration 111
5.6 Evaluation of Expression Features 118
5.7 Evaluation of Expression Features using Randomized Topologies . . 118
5.8 Evaluation of Expression and Connectivity Features 119
5.9 Connectivity Features using Random Topologies 119

Zusammenfassung

Mit der Weiterentwicklung von Hochdurchsatztechniken steigt die Anzahl verfügbarer Daten im Bereich der Molekularbiologie rapide an. Es ist heute möglich, genomweite Aspekte eines ganzen biologischen Systems komplett zu erfassen. Korrelationen, die aufgrund der internen Abhängigkeits-Strukturen dieser Systeme enstehen, führen zu charakteristischen Mustern in gemessenen Daten. Die Extraktion dieser Muster ist zum integralen Bestandteil der Bioinformatik geworden. Durch geplante Eingriffe in das System ist es möglich Muster-Änderungen auszulösen, die helfen, die Abhängigkeits-Strukturen des Systems abzuleiten. Speziell differentielle Expressions-Experimente können Muster-Wechsel bedingen, die wir verwenden können, um uns dem tatsächlichen Wechselspiel von regulatorischen Proteinen und genetischen Elementen anzunähern, also dem regulatorischen Netzwerk einer Zelle.

In der vorliegenden Arbeit beschäftigen wir uns mit der Erkennung von Korrelations-Mustern in molekularbiologischen Daten und schätzen ihre praktische Nutzbarkeit ab, speziell im Kontext der Kontakt-Vorhersage von Proteinen, der Entfernung von experimentellen Artefakten, der Aufdeckung unerwarteter Expressions-Muster und der genomweiten Vorhersage regulatorischer Netzwerke.

Korrelations-Muster sind nicht auf Expressions-Daten beschränkt. Ihre Analyse im Kontext konservierter Schnittstellen zwischen Proteinen liefert nützliche Hinweise auf deren Ko-Evolution. Muster die auf korrelierte Mutationen hinweisen, würden in diesem Fall auch in den entsprechenden Proteinsequenzen auftauchen. Wir nutzen eine einfache Sampling-Strategie, um zu entscheiden, ob zwei Elemente eines Pathways eine gemeinsame Schnittstelle teilen, berechnen also die Wahrscheinlichkeit für deren physikalischen Kontakt. Wir wenden unsere Methode mit Erfolg auf ein System von ABC-Transportern und Zwei-Komponenten-Systemen aus dem *Firmicutes* Bakterien-Stamm an.

Für räumlich aufgelöste Expressions-Daten wie Microarrays enspricht die Detektion von Artefakten der Extraktion lokal begrenzter Muster. Im Gegensatz zur Erkennung von Rauschen stellen diese innerhalb einer definierten Region Ausreißer dar. Wir entwickeln eine Methodik, um mit Hilfe eines Sliding-Window-Verfahrens, solche Artefakte zu erkennen und zu entfernen. Das Verfahren erkennt diese sehr zuverlässig. Zudem kann es auf Daten diverser Plattformen, wie Custom-Arrays, eingesetzt werden.

Als weitere Möglichkeit unerwartete Korrelations-Muster aufzudecken, entwickeln wir *Padesco*. Wir extrahieren häufige und wiederkehrende Muster, die über Experimente hinweg konserviert sind. Für ein bestimmtes Experiment sagen wir vorher, ob ein Gen von seinem erwarteten Verhalten abweicht. Wir zeigen, dass *Padesco* ein effektives Vorgehen ist, um vielversprechende Kandidaten eines differentiellen Expressions-Experiments auszuwählen.

Wir konzentrieren uns in Kapitel 5 auf die Vorhersage genomweiter regulatorischer Netzwerke aus Expressions-Daten. Hierbei haben sich Korrelations-Muster als nützlich für die datenbasierte Abschätzung regulatorischer Interaktionen erwiesen. Wir zeigen, dass für die Inferenz eukaryotischer Systeme eine Integration zuvor bekannter Regulationen essentiell ist. Unsere Ergebnisse ergeben, dass diese Integration zur Überschätzung netzwerkübergreifender Qualitätsmaße führt und wir schlagen eine Prozedur – *CoRe* – zur Verbesserung vor, um diesen Effekt auszugleichen. *CoRe* verbessert die False Discovery Rate der ursprünglich vorhergesagten Netzwerke drastisch. Weiterhin schlagen wir einen Konsensus-Ansatz in Kombination mit einem erweiterten Satz topologischer Features vor, um eine präzisere Vorhersage für das eukaryotische Hefe-Netzwerk zu erhalten.

Im Rahmen dieser Arbeit zeigen wir, wie Korrelations-Muster erkannt und wie sie auf verschiedene Problemstellungen der Bioinformatik angewandt werden können. Wir entwickeln und diskutieren Ansätze zur Vorhersage von Proteinkontakten, Behebung von Artefakten, differentiellen Analyse von Expressionsdaten und zur Vorhersage von Netzwerken und zeigen ihre Eignung im praktischen Einsatz.

Abstract

With the advance of high-throughput techniques, the amount of available data in the bio-molecular field is rapidly growing. It is now possible to measure genome-wide aspects of an entire biological system as a whole. Correlations that emerge due to internal dependency structures of these systems entail the formation of characteristic patterns in the corresponding data. The extraction of these patterns has become an integral part of computational biology. By triggering perturbations and interventions it is possible to induce an alteration of patterns, which may help to derive the dependency structures present in the system. In particular, differential expression experiments may yield alternate patterns that we can use to approximate the actual interplay of regulatory proteins and genetic elements, namely, the regulatory network of a cell.

In this work, we examine the detection of correlation patterns from bio-molecular data and we evaluate their applicability in terms of protein contact prediction, experimental artifact removal, the discovery of unexpected expression patterns and genome-scale inference of regulatory networks.

Correlation patterns are not limited to expression data. Their analysis in the context of conserved interfaces among proteins is useful to estimate whether these may have co-evolved. Patterns that hint on correlated mutations would then occur in the associated protein sequences as well. We employ a conceptually simple sampling strategy to decide whether or not two pathway elements share a conserved interface and are thus likely to be in physical contact. We successfully apply our method to a system of ABC-transporters and two-component systems from the phylum of *Firmicute* bacteria.

For spatially resolved gene expression data like microarrays, the detection of artifacts, as opposed to noise, corresponds to the extraction of localized patterns that resemble outliers in a given region. We develop a method to detect and remove such artifacts using a sliding-window approach. Our method is very accurate and it is shown to adapt to other platforms like custom arrays as well.

Further, we developed *Padesco* as a way to reveal unexpected expression patterns. We extract frequent and recurring patterns that are conserved across many experiments. For a specific experiment, we predict whether a gene deviates from its expected behaviour. We show that *Padesco* is an effective approach for selecting promising candidates from differential expression experiments.

In Chapter 5, we then focus on the inference of genome-scale regulatory networks from expression data. Here, correlation patterns have proven useful for the data-driven estimation of regulatory interactions. We show that, for reliable eukaryotic network inference, the integration of prior networks is essential. We reveal that this integration leads to an over-estimate of network-wide quality estimates and suggest a corrective procedure, *CoRe*, to counterbalance this effect. *CoRe* drastically improves the false discovery rate of the originally predicted networks. We further suggest a consensus approach in combination with an extended set of topological features to obtain a more accurate estimate of the eukaryotic regulatory network for yeast.

In the course of this work we show how correlation patterns can be detected and how they can be applied for various problem settings in computational molecular biology. We develop and discuss competitive approaches for the prediction of protein contacts, artifact repair, differential expression analysis, and network inference and show their applicability in practical setups.

Chapter 1

Introduction

1.1 From Complex Systems to Correlated Data

This work is concerned with the detection of correlation patterns in bio-molecular data and their practical application in terms of contact prediction (Chapter 2), artifact detection (Chapter 3), unexpected patterns (Chapter 4), and network inference (Chapter 5). In the following we provide an overview of how correlation patterns may arise from complex systems and discuss their usage in the course of estimating the biological relevance of observations.

1.1.1 Indirect Observations

Throughout this work and in most areas of bioinformatics we examine biological processes. These processes occur within organisms, usually on the molecular scale. All data we currently collect provides snapshots of system read-outs. We cannot observe the system connections directly, but we must infer them. This strongly impacts the exploration of the underlying molecular mechanisms.

Of course, this problem is not new. The founder of one of the most important theories of the past centuries, “the theory of evolution”, Charles Darwin [48], could not provide a plausible mechanism for his observations. While he could observe phenotypes, he was not aware of the rules of inheritance, let alone today’s knowledge of the molecular level. Manifold discoveries since then supported the theory’s validity and relatively recent technical developments like sequencing have opened up a molecular point of view as well.

With the advance of systematic sequencing of full genomes (3,981 eukaryotes, and 88,449 prokaryotes in January 2017, following NCBI statistics [177]) our understanding of the underlying processes has greatly improved. Similarities and differences between organisms can be captured in detail.

Large-scale biochemical methods allow us to capture the regulation for not only few but millions of molecules [14, 67, 107, 249]. Still, we do not observe the mechanisms, but effects.

For instance, cancer is likely caused by an imbalance among gene products, metabolites and a failure of DNA repair mechanisms. The phenotypical observations are typically limited to a tumor’s biochemical properties and the host’s environment like the absence or presence of compounds, proteins, and genes. These properties exhibit strong heterogeneity both intra- and inter-tumoral [2, 10]. This may potentially imply diverse causes for tumor growth and thus, varying experimental data across patients. Effectively, the detection of mechanisms remains challenging despite more and more data being available.

The more data experiments capture on an increasingly detailed level, the more experimental layers exists. The degree of indirection is increasing and the underlying system is more difficult to control. For example, degradation processes may skew the observed outputs before the read-out takes place [128, 183].

Given experimental data, We can extract and derive distinctive features. In case of molecular processes, this is usually a characteristic change in the concentration of a molecule. A collection of features is often referred to as a pattern or a molecular fingerprint. Patterns are useful to summarize observations and provide a snapshot of the analyzed system.

The discovery of characteristic patterns is a major challenge in molecular biology and the comparison of patterns is crucial for a wide range of applications [131, 194, 199, 218, 221]. Chapter 4 and 5 discuss problems associated with pattern extraction from expression data.

The more complex the observed system is, the more difficult it is to observe direct effects. Most biological systems are robust and redundant, such that an enforced intervention may be buffered and spread into a variety of small changes [138].

In order to achieve reproducible and interpretable results, it is essential to acquire sensible controls. A robust complex system is more difficult to control and controlling entities of interest will likely affect linked outcomes or produce side effects that may not be observed due to the indirect nature of our data. With this in mind, it is extremely important to examine experimental data with great care to reduce the chance of obtaining false positive results. In Chapter 5 we are particularly concerned with bias emerging from complex network structures and how to counteract it.

1.1.2 Regulatory Systems

A basic regulatory network is a dependency model of two basic molecular components: genes and transcription factors (TFs). TFs are Deoxyribonucleic Acid (DNA) or Ribonucleic Acid Polymerase (RNAP) binding gene products. They may bind other TFs, gene products and small molecules as well. TFs trigger the activity of genes by binding their associated regulatory regions, either promoter regions (DNA-binding elements) or RNAP, and change their target gene's expression level. The regulated gene products may themselves be TFs. The entirety of all TFs and regulated or target genes can be represented as a network.

This network is very complex and any associated model is necessarily simplified. Many other reactants like micro RNA (miRNA), co-factors, inorganic molecules and chemical reaction environment are usually neglected. Yet, if there is a causal relationship among two targets of the same TF these targets could exhibit a similar measurable pattern. The two gene products share a common regulatory mechanism. They are expected to be abundant simultaneously, at least in general. The two signals are *correlated.* The inversion of this argument is that given two signals are correlated, they may interact or share a common regulator. This observation is used to infer regulatory networks in practice [11, 12, 34, 41, 45, 51, 59, 50, 66, 105].

In this context, we have successfully applied regulatory network inference via correlation pattens [144]. The key idea is to guess an initial executable model and simulate its expected data. The emerging patterns can be compared to actual observations and judged by their similarity. By iterative model refinement and consecutive simulation, we approach the underlying circuitry of the network.

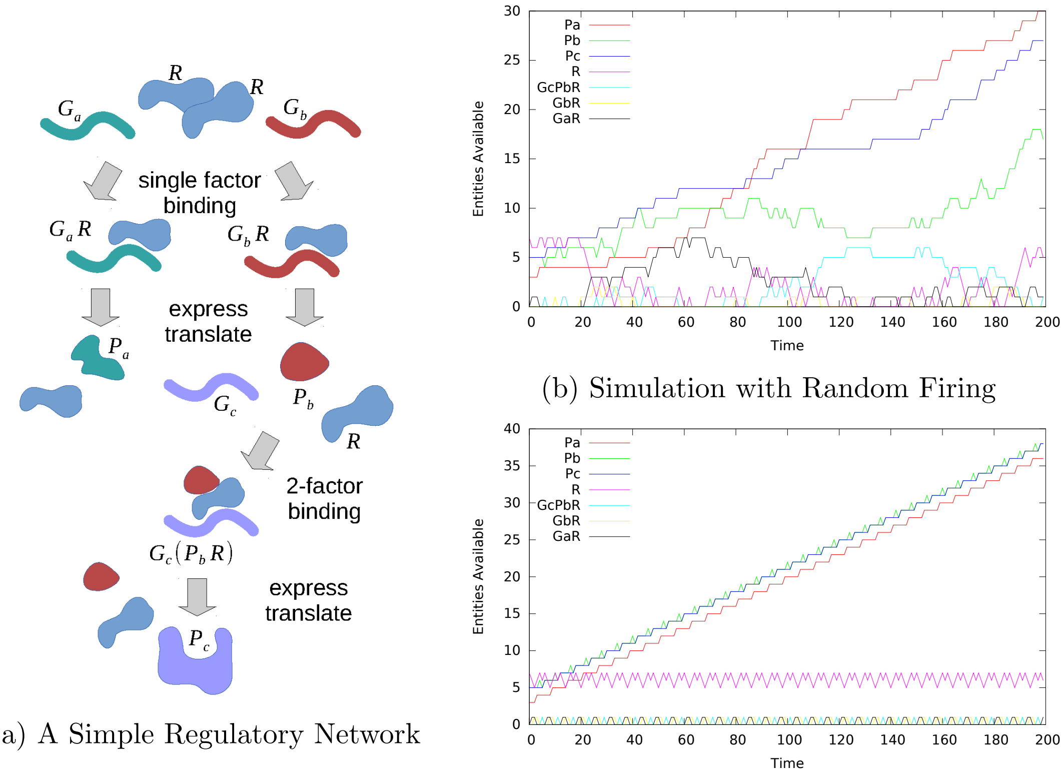
Unfortunately, this procedure breaks down for large networks. The number of networks capable to generate the observed data is virtually unlimited. Additionally, in complex networks large effect sizes may not exist at all. Buffering increases the robustness of the system [138]. Disruptions are scattered across the network and lead to numerous weak downstream effects. Furthermore, many unrelated signals will also exhibit similar patterns simply by chance. They are correlated but they neither share a common causal mechanism nor are they related otherwise. Especially within complex systems, correlation can often be expected by chance alone. Exploiting previous knowledge, the integration of more data, and faster modes of computation become necessary. In fact, we could show that the integration of experimental annotation into correlation approaches could improve regulatory network prediction quality [143].

1.1.3 Formation of Correlation Patterns

In the following, we demonstrate why correlations occur in the context of molecular systems. We simulate data emerging from a network and plot a time-course of the observed data. We apply Petri nets [173] as a conceptually simple way to model and simulate this prototypic regulatory system.

Petri nets consist of places and transitions. Each transition is linked to input places via incoming arcs and is linked to output places via outgoing arcs. Firing a transition resembles the process flow of the respective reaction. It consumes tokens from its input places and produces tokens on its output places. Each token may represent a single molecule, while for more realistic modeling, the tokens may represent molecular concentrations [144].

We implemented a Petri net resembling a prototypic regulatory network shown in Figure 1.1a and simulated the number of available entities in the network at a given time. Each arc carried a weight of one, such that each firing of a transition would consume one entity from each input place and produce one entity to each output place. For Figure 1.1b the firing transition was chosen randomly among the set of transitions ready to fire, while for Figure 1.1c we used the canonical left-to-right (reading) order of Figure 1.1a. While this simplified model may not be very realistic, the stochastic order of fired transitions mimics the simultaneous nature of real biological systems given limited resources (gene copies, regulator



(a) A Simple Regulatory Network

(b) Simulation with Random Firing

(c) Simulation with Ordered Firing

Figure 1.1: Petri net topology for a prototypic regulatory network. A single transcriptional activator R mediates the production of proteins via intermediate complexes. For simplicity, the transcription step is not modeled separately. The regulator R is crucial to produce protein P_a from gene G_a and P_b from G_b. Further, R and P_b mediate the production of P_c via binding G_c. The intermediate complexes G_aR, G_bR and (G_c, P_b, R) model the physical binding of regulators to target genes.

molecules).

We observe that an ordered firing rule produces a predictable pattern of intermediate complexes and a monotonously growing amount of proteins. The regulator R exposes a cyclic pattern, after its release from the genomic regulatory site. The same holds true for intermediate complexes. P_a, P_b, and P_c are correlated due to their common regulator R. In the case of P_c this regulation is mediated via the production of P_b.

For the stochastic firing order of transitions, the correlation is obfuscated. The emerging pattern is quite complex. We can observe a temporary drop in P_b due to its incorporation into the intermediate complex (G_c, P_b, R), as well as cyclic patterns of R. These are far less obvious as the observed signals incorporate much more noise in general. Notably, our model does not include differences in protein half lives, binding affinities, or other factors that may lead to the formation of more complex patterns [159].

Additionally, in practice, the actual experimental setup is unlikely to capture 200 incremental measurements. Technical limitations, time restrictions, and budget restrictions lead to lower-resolution patterns captured at few specific time points only. Furthermore, measurements are commonly taken from distinct samples that may differ in terms of their initial molecular concentrations or even the circuitry of the underlying system. Awareness of these limitations is essential for the analysis of correlation patterns.

We analyzed the inference of regulatory networks from correlation patterns on multiple scales, ranging from small artificial regulatory systems to real-word systems on the genome scale. We developed various approaches to overcome existing limitations in this regard [143, 144, 192] (see Chapter 5).

1.1.4 Detection of Correlation Patterns

We will now take a closer look at how correlation patterns can be detected and give a short overview of applications in the fields of expression data analysis and network inference.

Numerous definitions exist to describe the correlation among measurements, outcomes or control values [79]. Still, the most widely used correlation is Pearson's Correlation [79, 251]:

$$\rho_{XY} = Corr[XY] = \frac{Cov[XY]}{\sigma[X]\sigma[Y]} \tag{1.1}$$

Here, Cov denotes the covariance of the random variables X and Y and σ is their respective variance. Pearson's ρ is relatively easy to apply, implement and its properties are well-understood. The coefficient $\rho(X, Y)$ is zero if the two random variables X and Y are stochastically independent. If Y is the result of a linear function of X, then $\rho(X, Y)$ equals 1, which is suggesting a perfect linear dependency of both variables. Since ρ is considering linear dependencies only, one can construct complex non-linear patterns that will effectively result in a correlation of zero, the most prominent of which being Anscombe's dataset [6].

Often a slope is shown for a scatter plot of X and Y. This slope integrates noise in Y only, as the linear fit follows the linear relation $X = Y + \epsilon$. Yet, in most biological setups both reference values X and experiment values Y are prone to noise. The use of Deming's Regression [52] as a special case of total least squares [95] may be more appropriate for the estimation of a linear fit, and more intuitive in cases where symmetrical slopes for (X, Y) and (Y, X) are expected.

The applications of correlation analysis in bioinformatics are manifold. In the context of expression data analysis the applications range from clustering [64, 110, 142, 261], network inference [51, 61, 73, 144], tissue classification [43, 242], module detection [12, 129] to function prediction [21, 156], to name a few.

The covariance is closely related to the correlation, as it can be seen as an unscaled version of the correlation (see Equation 1.1). The co-variance matrix derived from all pairwise covariances of variables (like regulators and targets) in a system is of particular interest for network inference, as its inverse corresponds to the conditional independence among random variables [75, 163, 213]. A more detailed discussion on how network inference approaches exploit correlation patterns can be found in Chapter 5, Section 5.4.

Throughout this work we apply inference methods that heavily rely on either correlation estimates or measures of similarity. In particular, kernel functions are an integral part of Support Vector Regression (SVR) and Support Vector Machines (SVM) in general [182, 229] (see Chapters 4 and 5). Notably, kernel functions have an interpretation as covariance functions [205]. Thus, their application is a way to detect non-trivial correlation patterns in the input data. When non-linear kernels are employed, this enables the detection of non-linear correlation patterns as well. We discuss further advantages of kernel-based pattern detection and the sensitivity of correlation estimates to outliers in Section 1.3.

In Chapter 2, we discuss how correlated mutations can be used to distinguish physically contacting proteins from other members of the same pathway. In this context, the correlation definition is modified to estimate sequence mutations [94].

1.2 Assessment of Biological Relevance

In both expression data analysis and network inference, we are dealing with scientific hypotheses that describe traceable changes or behaviour. Results obtained under reproducible conditions and from standardized experiments help to verify the validity of a hypothesis. A key question that arises in this context is whether some observed outcome is also biologically relevant. In the following, we will provide several problem descriptions and approaches for the estimation of biological relevance.

1.2.1 Statistical Testing

Proving relevance requires sensible reasoning and embedding of novel evidence into existing knowledge. An important tool in this regard is the testing of statistical hypotheses (see Gravetter and Wallnau [99], p.223ff.). Aspects of a scientific hypothesis can be contrasted by a corresponding null-hypothesis. The null-hypothesis represents the assumption that, for given a perturbation or treatment, no effect can be observed in the system or the population, respectively. If it is unlikely that the data obtained by an experimental perturbation is observed given the null-hypothesis is true, we may reject it. We would instead opt for the alternative hypothesis, a hypothesis in accordance with the scientific hypothesis, stating that the induced system changes lead to an observable effect. The statistical test-result could then serve as additional evidence and underpin biological relevance.

If the specific pre-conditions of a statistical hypothesis test are met, its outcome is a test-statistic. From the statistic together with knowledge on how it is distributed, an associated p-value can be obtained. It reflects how likely a statistic of this or higher magnitude is observed. A p-value is deemed significant if an α-level of significance is met. Defining α before the actual test is necessary to achieve a valid significance test. This experiment-specific threshold is commonly chosen as 0.05, and a significant p-value leads to the rejection of the null-hypothesis. Insignificant values are often erroneously seen as counter evidence to biological relevance [181]. p-values have been controversially discussed, mostly in terms of their misinterpretation and an increase in publishing bias [102, 178, 181].

The achievable statistical significance in a test is dependent on the underlying sample size. As an example, we assume 144 individual entities (say organisms or cell cultures) that share some observable feature. We modify some of these entities and observe whether the modification leads to a certain effect. A 2×2 contingency table is then compiled (see Table 1.1). The scientific hypothesis here is that the modification leads to the observed effect.

system state	normal	modified
no effect	83	30
effect	17	14

Table 1.1: An example system encompassing 144 individuals. 44 individuals are modified. All individuals are scattered by some measurable effect. We are interested in whether the modification is linked to the observable phenotype.

A null hypothesis associated with this 2-variable (effect, modification) setup is: the observable effect is independent of the applied modification. A statistical test suited for this analysis is the Chi-squared test (see Gravetter and Wallnau [99], p.559ff.). We choose an α-level of 0.05 and obtain a p-value of 0.08. The null-hypothesis is not rejected, and we consequently gain no evidence that the modification influences the phenotypic effect. Yet, we cannot reject the initial scientific hypothesis. Half of all modified entities show an effect while only some 20% do in the normal state. The observation is not significant, but it is in agreement with our hypothesis. Relying on p-values alone for the testing of scientific hypotheses is not sufficient.

In general, a reduction of statistical testing to p-values is problematic. It may lead to a premature rejection of scientific hypotheses, such that insignificant p-values are interpreted as an estimate for the probability that the underlying scientific hypothesis is wrong. Censoring of experiments by statistical significance alone would eventually lead to the reporting of artificially inflated effect sizes, as for equal sample sizes more extreme observations are necessary to obtain significant results [178]. In the course of the detection and modelling of correlation patterns multiple testing issues may arise, which are discussed in Section 1.3.3.

Statistical significance tests are an integral part of scientific hypothesis testing. Carefully applied and in combination with contextual knowledge they provide a well-defined tool to back empirical observations and help to evaluate the biological significance of our results.

1.2.2 Sources of Variation and Noise

Robustness against Perturbation. Most biological systems are robust enough to work under changing conditions and maintain homeostatic conditions. Even major fluctuations may often be buffered, both on the organism level and the molecular level. In particular, regulatory systems have established robust mechanisms to overcome changes [138].

The robustness of complex systems limits our ability to to control them during experiments. We may thus modify a system, but cannot obtain measurable differences when compared to a given control state. Similarly, a perturbation that is strong enough to result in a signal would likely entail side-effects on other parts of the system. Therefore, experiments like gene-knockdowns and knockout are not expected to yield simple cause-effect chains, in particular for eukaryotic and higher organisms. For regulatory networks, we have shown that this lack of traceable signaling results in the breakdown of correlation-based inference [143].

Technical Variation. Technical variation refers to the observed variation of experiments that occurs when the same biological entity is measured several times as a technical replicate. Commonly, this variation is treated as noise, as no causal biological explanation should exists. In fact, bio-molecular analysis must reliably control cellular mechanisms. Unintended, yet systematic side-effects can arise and lead to biases. Nonetheless, for many modern techniques the magnitude of technical variation is small and biases are well-documented [164, 253, 264].

Biological Variation. Biological variation is the observable difference among individuals measured by biological replication. Given that both sufficient technical replicates and biological replicates are available, we can use hypothesis testing (see Section 1.2.1) to detect significant changes across different experimental conditions. Biological replicates are usually more costly than technical replicates as they require the complete experiment to be repeated. The magnitude of noise across biological re-sampling is typically much larger than that expected from technical replication [22]. For complex systems in particular, a system's initial state during an experiment is difficult to both estimate and control. As the source of variation is on the level of individuals (cell-lines, samples, organisms), many observed entities are expected to be de-regulated by naturally occurring processes.

Repeated sample collection may be imprecise, in particular when analyzing mammals or living organisms. Tissue samples likely contain impurities from different tissues. Even carefully selected tissues contain multiple cell types that have distinct functions and thus, biochemical properties. *In-silico* dissection methods have been applied to tackle these problems [1, 96, 221, 243]. Commonly, we would treat biological variation as noise, similar to experimental variation, unless we have a mechanistic explanation for observed deviations.

Experimental Limitations. The experimental techniques available determine the degree of observation that is possible. Current experimental techniques often rely on copy numbers as a proxy for the activity of molecules or proteins, often complemented by data on methylation, phosphorylation or binding affinities [15, 67, 107]. Yet, the actual biological activity is not directly coupled to the copy number. A well-studied example is the correlation of protein-level and expression-level, where large discrepancies have been reported and attributed to numerous factors like protein half-life and translation efficiency [100, 106, 159]. When estimating copy numbers, we must further be aware that the dynamic range of experiments is narrower than the actual molecular abundance in a cell. Experimental outcome at the extreme ends of a method's dynamic range cannot be reliably observed.

Reliability of Published Data. Published data may contain undocumented errors and biases [87, 145, 152]. The ENCODE Consortium [67] provided the scientific community with a complex, large-scale data collection for numerous tissues. A recent re-analysis by Gilad *et al.* [87] revealed that part of the data seems to suffer from a flawed experiment design and batch-effects following an initial discussion by Lin *et al.* [152]. Tissues usually feature highly conserved structures, in particular within the class of mammals. In terms of gene expression data, tissues from mouse and human are expected to be more similar than different tissues of the same organism. Surprisingly, Lin *et al.* [152] reported that the ENCODE data indicates that species samples would in fact form more similar clusters. For publicly available data, an appropriate analysis for undocumented effects must be undertaken to prevent avoidable downstream effects. In Chapter 3 we discuss the presence of artifacts in GEO [14], their detection and their removal.

1.3 Modelling Correlation Patterns

1.3.1 Model Regularization

Bioinformatics has established a strong system's viewpoint within the field of molecular biology [138]. In principle, it is possible to design a highly complex model of gene regulation capturing all known regulatory factors and targets. Unfortunately, an unambiguous parameter fitting of this model would require more than the available data. To overcome this problem it is necessary to introduce simpler models with fewer parameters or implicit coefficient shrinkage [63, 93, 236]. Often model smoothness or a restriction of model complexity is achieved via penalizing model complexity [266]. This is also referred to as regularization (see Tsuda *et al.* [215], p.42ff. and Hastie *et al.* [113], p.34/144ff.). This may lead to fewer parameters and thus, more general models. Memorizing individual patterns (rote learning) is prevented and the prediction of previously unobserved events may be improved. Simplified models likely interpret some true signals as noise. Yet, the general trade-off among erroneously removing true positive results and over-fitting a model cannot be bypassed. In Chapter 4 and Chapter 5 we discuss the application of regularized SVR models for the detection of robust patterns.

1.3.2 Coping with Outliers

Outliers are extreme values in the data. They can have an explanation on the molecular level, and a process of generation we have not yet described. In general, they are generated by a different mechanism than the rest of the data [265]. Thus, the perception of whether a data-point is treated as an outlier depends on the underlying model. The outlier degree can be expressed numerically [26, 141].

Under linear model assumptions, any non-linear effect may be interpreted as an outlier. Simultaneously, linear correlation estimates are highly sensitive when a dataset contains outliers (see Gravetter and Wallnau [99], p.499ff.), leading to wrong conclusions about the degree of correlation present in the data. Robust estimation techniques (Press *et al.* [198], p.818ff.) and modeling via support vectors (like the epsilon-insensitive tube or the maximum-margin property) would help to discard outlier data-points from the model [215, 229] and detect robust correlation patterns in the presence of outliers. The interpretation of model residuals as local outliers and subsequent filtering may lead to robust estimates of co-expression patterns by suppressing false positive correlations [81].

For general and molecular biology applications, the definition, the detection and the treatment of outliers has been studied intensively, in particular with regard to high-dimensional problems [26, 70, 141, 265]. Simulations have shown, that the expected value range of correlation-like measures as the cosine similarity narrows with growing dimensionality, affecting the capability to distinguish data-points via their similarity or via their respective distance [202, 265]. In Chapters 4, we implement an empirical novelty detection strategy reporting unexpected patterns of individual genes. The artifact detection approach described in Chapter 2 can be interpreted as a local outlier detection strategy.

1.3.3 Multiple Hypothesis Testing under Dependency

When coping with whole genome data the testing of hypotheses usually involves repeated evaluations of a model on the same data. In this case significant outcomes are expected by chance. Under the null-hypothesis p-values are expected to be equally distributed [20] and by the definition of the α significance value, we expect $n * \alpha$ significant results after n tests. Multiple testing correction has been suggested to overcome this problem [92, 121]. In particular, methods controlling the False Discovery Rate (FDR, the number of false positives among all significant tests) have gained popularity [17, 232], by restricting the view on significant tests rather than a family-wise error-rate (FWER, probability of at least one false positive). These approaches assume independent test statistics. Yet, bio-molecular data is characterized by strong functional associations [91]. The resulting test statistics are likely dependent. As a consequence, multiple strategies to test and cope with multiple hypotheses from correlated data have been developed [18, 92, 108, 133, 150, 210]. Following the classification used by Goeman and Solari [92], we can distinguish three major classes of dependency modelling. The first class allows for arbitrary dependencies by conservative corrections, *e.g.*, the Bonferroni correction. The second class relies on the positive dependence through stochastic ordering (PDS) condition, resulting in less conservative results for some test distributions. The third class estimates the dependency structure via permutation tests. While the latter case allows for adaption of the observed dependency structure of p-values, we have to pay particular attention to the setup to avoid invalid interpretations [91].

In Chapter 5 we develop the so-called κ-transformation, which is motivated by controlling the FDR in a network inference setting by repeated randomization of the original gold-standard network.

1.3.4 Influence of Subgroups on Model Evaluation

In the following we will describe a particular aspect of evaluation of these models via global measures like the Receiver Operator Characteristic (ROC, see Chapter 5, Section 5.5.4). This analysis resembles the problems discussed in Chapter 5.

For the evaluation of predictive methods, it is common to relate predictions and observed labels via measures of correlation. These are often closely related to statistical dependency tests. For example, a re-scaled Area Under the Receiver Operator Characteristic Curve (AUC/AUROC) would follow a Mann-Whitney U-distribution as used by a Wilcoxon Rank Sum Test [162]. An ROC plot therefore closely relates to a rank-based evaluation, as it is based on a set of sortable predictions (a ranking criterion or confidence) together with a known binary labeling. It provides a both visual and, in case of the AUC, numeric way to judge how well the predictions match the dichotomy given by the labels.

The computation of an ROC involves the incremental shifting of an internal threshold, plotting the sensitivity (True Positive Rate, TPR) versus 1 minus specificity (False Positive Rate, FPR). This curve provides a visual measure of the correlation among the predictions and the actual labeling. The visual inspection provides a powerful tool to compare predictive algorithms.

The analysis of ROCs can be misleading though [49, 89, 109]. Further, predictions may be clustered into sub-groups with specific features. A procedure that integrates these features, intended or unintended, via some correlated property (availability of data, measurement magnitude or scale) can become unspecific on the group level. The training data may be divided into sub-groups of distinct label distributions (see Chapter 5). Predictive models that neglect these groups focus on an increased overall performance. Yet, if the group-wise label distributions differ, some groups are more likely to increase the overall performance and training algorithms may detect and reward such groups. Effectively, an algorithm could rank groups by their label distribution. An ROC analysis will then correctly report reasonable performance, despite the ranking within each group may be random (see Figure 1.2). From a biological perspective, the prediction must be seen critical, in particular if conclusions for individual groups are drawn. This setup resembles the key problems discussed in Chapter 5 on the prediction of regulatory targets using topology information.

The analysis in Figure 1.2 is concerned with how strongly the ROC analysis is affected from the presence of sub-groups, given that the training algorithm may access this piece of information. In practice, this may occur by combining group-wise predictions in retrospect (see Chapter 5).

As shown in Figure 1.2, the predictions within each group may be random. For balanced label-distributions the area under the ROC is 0.5. Yet, reducing the number of positive instances per sub-group by 1% each (decrease by $5, 10, 15, \dots$)

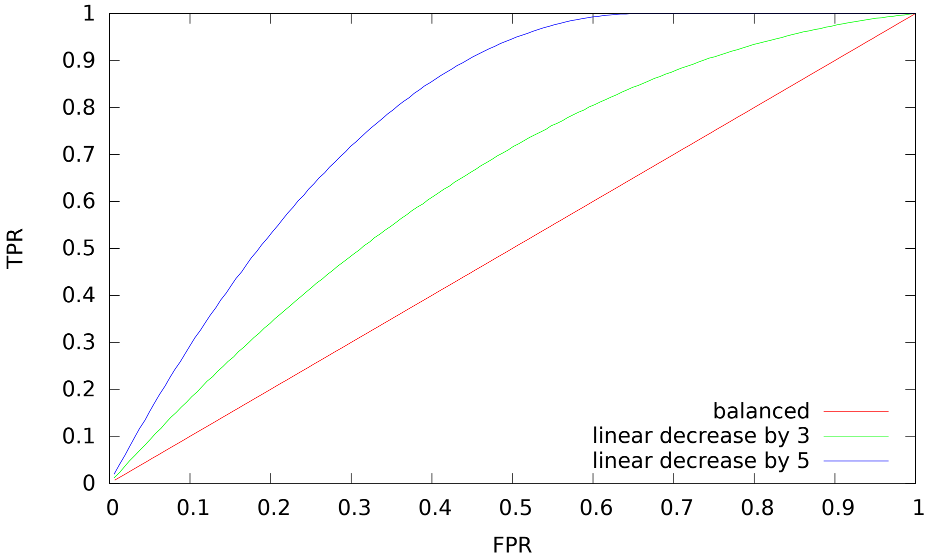


Figure 1.2: We analysed a set of 150 groups of 5,000 predictions each. The predictions within each group cannot be distinguished and receive equal confidence values. We show how a ranking of groups would affect an ROC analysis. The predictions in each subgroup all receive a score proportional to the number of positive labels in this group. Three different label distributions are shown: in the 'balanced' case all subgroups contain 500 positive instances. As all scores are equal, the ROC resembles a random predictor. In the 'linear decrease by k' case, the n-th subgroup contains $500 - n * k$ positive instances. The Area Under the Curve is 0.64 for a linear decrease by $k = 3$ and 0.79 for a linear decrease by $k = 5$.

yields an ROC of 0.78. Still, for each individual sub-group, the prediction is random. The effect of unbalanced groups may be strongly misleading for the interpretation of results on the sub-group level.

For problems that contain relatively few positive labels (like network inference) the FPR axis of ROCs is governed by the amount of true negatives (TN). When group-ordering occurs as discussed above, this means that the FPR is damped by the large amount of negatives and most positive labels can be discovered (TPR) for relatively low FPR values (in Figure 1.2 this occurs for around 30% of all negatives). This may imply seemingly accurate predictions. Yet, when comparing several algorithms, the subgroup order may be a confounder that needs to be corrected.

The ROC correctly states that the overall prediction is superior to random guessing. While this is correct, the prediction may not be useful in practice. The most confident predictions are randomly ordered in their respective groups -

starting with the first one.

This result is somewhat counter-intuitive and highlights the importance of multiple, orthogonal, evaluation measures. Complex systems would likely lead to correlated patterns and thus the formation of subgroups in the data. In these cases specific aspects of model building must be considered during evaluation. We discuss this problem in the context of pattern-based network inference in Chapter 5.

1.4 This Work

This work is concerned with the detection of correlation patterns in bio-molecular data and their practical application in terms of contact prediction (Chapter 2), artifact detection (Chapter 3), pattern discovery (Chapter 4) and network inference (Chapter 5).

In the course of this thesis we successfully took part in two rounds of DREAM challenges [160] for the inference of both dynamic small-scale [144] and static large-scale [143] network inference challenges on real-world data. Therefore, some of the most important results of this work are centered around the reconstruction of regulatory networks as well [192]. A conceptual overview of all chapters is given in Figure 1.3.

In Chapter 2 we focus on the protein level rather than the expression level and focus on a definition of correlation anchored in evolutionary biology that quantifies the coupled evolution of proteins via their alignments. Here, we seek to predict contacting protein families among two-component-systems in a *bactitracin* resistance context. We combine an established set of methods in combination with a bootstrapping strategy to screen protein families for potential mutually conserved regions. Randomly occurring mutations are contrasted with those caused by physical constraints. A potential source of bias is the prevalence of certain organisms, as was recently discussed [126], where bacteria are considered the most diverse branch in the tree of life by far.

In Chapter 3 we examine noise stemming from technical artifacts on microarrays and provide means to visually inspect and filter or replace them. The noise type discussed here can be observed directly. Visual inspection of data should be an integrated part of any analysis. Visual artifacts and features may hint to biases, sources of noise or hidden correlations that might skew black-box analysis and derived statistics.

In Chapter 4 we use predictive models to decide whether the gene fold-changes in an experimental context are predictable from the changes in other genes, *i.e.*, expected. We use Support Vector Regression (SVR) to train gene-wise models capable to predict fold-changes for a particular gene. This model can be seen as a weighted neighborhood of a gene – a functional network. Notably, at the core of the

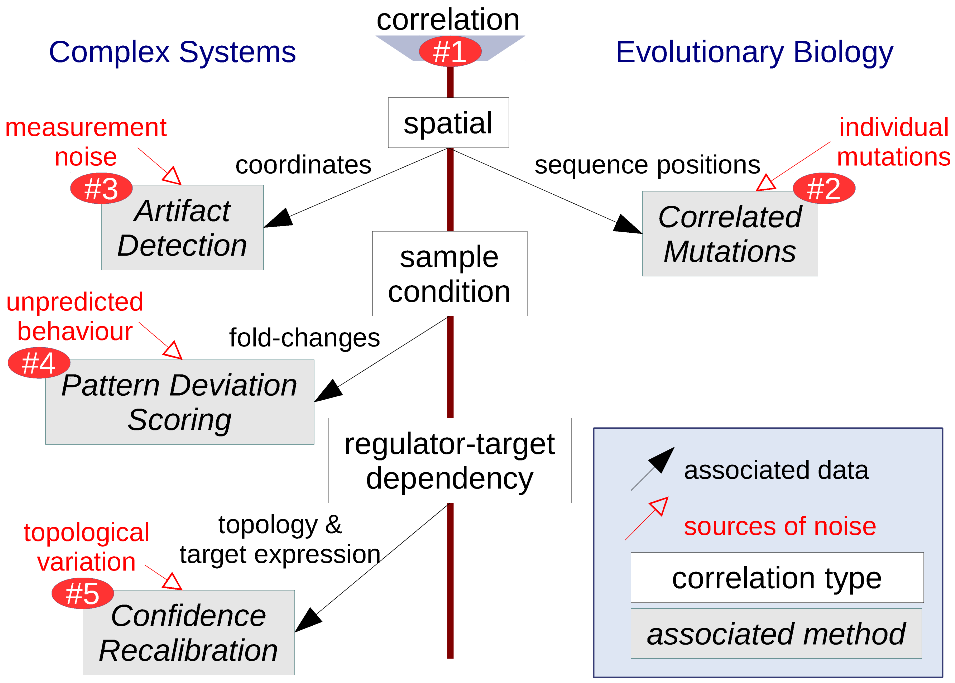


Figure 1.3: In this work, we examine correlation structures throughout four specific bioinformatics problem settings (chapter numbers in red). We analyze correlated sequence mutations at pairwise sequence positions in Chapter 2. A spatial definition of correlation is applied in Chapter 3 to detect and correct for technical artifacts. Chapter 4 deals with predictable fold-change patterns of individual genes, yet network dependencies are not modeled. In Chapter 5 correlations among regulators and their targets are used to predict novel targets and rectify some key problems of eukaryotic network inference. The major sources of noise range from silent mutations (1), measurement noise (2), which is virtually present across all experiments, unpredicted system behaviour (3) and topological variability (4), *i.e.*, dependencies that are not consistently present in a biological system.

SVR we apply a linear kernel as a measure of similarity among experiments that is closely related to the correlation measures discussed above. We apply robust techniques to overcome the phenotypic noise in these experiments.

In Chapter 5 we examine state-of-the-art approaches to data-driven regulatory network inference. Our analysis revealed that existing evaluation methods rely on assumptions that are critical for a sensible estimation of network-quality and completeness. First, we confirm that a related problem described for functional prediction [89] is present in our setting as well: the prevalence of highly connected nodes (hubs). We show that the topology of the network directly affects supervised measures of network-quality like ROC curves.

In most cases the confidence values predicted for a regulation are regulator-specific and are not directly suited for a network-wide candidate selection or as an estimate of method performance. The regulator's degree correlates with the location parameters of its confidence value distribution. This simple connection impacts both evaluation and network composition. Regulators with few known interactions would not receive any novel prediction at all while novel predictions would cluster for larger regulators, increasing false positives levels. We developed a strategy, *Confidence Recalibration* (*CoRe* [192]), to achieve a balanced prediction across regulators and a network-wide decrease in false discoveries. The selection of newly predicted interactions is driven by method-specific confidence values.

Chapter 2

Correlated Mutations of Protein Contacts

Background

This chapter is based on Dintner *et al.* [55]. The results have been achieved in joint work with colleagues from the Department of Microbiology at the Ludwig-Maximilians-Universität, München. At this time Evi Berchtold was my student assistant. This chapter provides a bioinformatics point of view. For mechanistic and microbiological details of the two-component systems and ABC transporters we confer to the original publication [55]. We point out that all wet-lab experiments as well as the manual curation of input alignments specific to the *Firmicute* bacterial strains was performed by our colleagues from microbiology (see below). For this work we designed and implemented a bootstrapping framework for alignments, mapping of protein contact regions and data visualization and we provided statistical routines.

Contributions

The research setup was designed by Susanne Gehard (SG) and Thorsten Mascher (TM) with bioinformatics support from Tobias Petri (TP) and Ralf Zimmer (RZ). Experiments, selection of ortholog families and manual alignment curation was carried out by Sebastian Dintner (SD), Anna Staroń (AS) and SG. Bioinformatics routines have been conceived and implemented by TP. Additional bioinformatics experiments have been performed by Evi Berchtold (EB).

2.1 Overview

In the following we will describe a way to use evolutionary conservation as a measure of correlation that can be used to distinguish contacting from non-contacting proteins. This chapter is a singleton in the context of this work in terms of the underlying data, as we rely on sequences rater than expression measurements. Yet, in terms of spatial correlations and the computation of sensible backgrounds it resembles key ideas applied throughout this work (see Figure 1.3)

We rely on a seminal method [94] to quantify contacts co-evolving across bacterial strains and estimate the degree of co-evolution among two proteins. A simple method is then described to distinguish contacting from non-contacting proteins with only incomplete data on ortholog genes. The method is successfully applied to confirm a direct regulatory interaction for two-component systems and ABC transporters in *Firmicute* bacteria.

2.2 Contact Promotes Conservation

2.2.1 A Co-evolution Anecdote

The notion of protein co-evolution is borrowed from the classical notion of co-evolution on the species level. Both animal behavior and features are often coupled across species. A well-known example is the symbiosis of the clown-fish subfamily *Amphiprioninae* with certain anemones. The poisonous host serves as shelter for the clown-fish whereas the anemone inflicts serious wounds for other fish. Similarly, the clown-fish would reduce parasites and predator threats for the host.

Obviously high-level adaption has taken place and a beneficial interface has developed. In case that a change in one of the two organisms would break this interface the consequences could be drastic: the fish may loose its immunity or the anemone would be prone to parasites. The same holds true for external influences, say, an increased acidification of oceans that affects the composition of the protective mucous layer of the clown-fish skin. To adopt to these external forces a simultaneous change among the symbiotic partners is necessary to counteract negative effects. This is referred to as co-adaption.

Evolution is usually a smooth process of continuous adaption in a heterogeneous environment that requests for a robust balance among variation and selection. Its consequences are hidden to individuals and must be observed across generations and on the population level. No individual fish can adapt to environmental consequences. Two populations with sufficient variation may be capable to buffer disruptions and maintain the beneficial interface.

Features that are not crucial for the interaction stability may change though.

Clown-fish and anemones differ in both poison and resistance type and actually many sub-families do. There exists a strong host-specificity. Yet, while the key interface remains coupled, other features like size or color may change without affecting the mutual co-operation.

2.2.2 Co-Evolution among Proteins

Protein families are usually characterized by the conservation of some functional domain or characteristic binding sites. Similar to clown-fish and anemones, co-evolution of distinct families may also occur on the molecular level and in fact, several concepts that hold on the macroscopic level can be transfered to the proteins level.

A proteins environment is determined by its host. Related bacteria can therefore host similar but not identical proteins referred to as protein families. Two proteins that enable some specific function may interact physically. Although these pairs may not be identical across families it is important to note that the essential functional regions are often conserved. A change to the interface of one interaction partner would require a co-adaption of the second protein.

Commonly, interfaces are located on the surface of the protein. Their structure is necessary to maintain the function of an interaction and is crucial to the underlying system. Most functional contacts are transient. Any malfunction would seriously hamper signal transduction and thereby the metabolism of the host.

Two contacting regions with a joint function are likely subject to simultaneous or co-evolution. Whenever an amino acid is altered by a non-silent mutation in protein A then the ability to bind protein B may be disrupted leading to deleterious effects for the host. On the population level, there may exists an additional alteration of B such that a functional interaction among the altered proteins A' and B' is maintained. On the sequence level, a simultaneous protein exchange is observed when the pairs A-B and A'-B' are compared. The function is maintained despite of differences in the interface of the contacting protein regions. We refer to such systems as ortholog.

The multiple alignment of ortholog amino acid sequences can be used to visualize and characterize the conserved regions of coupled sequence changes. Given two protein families, a degree of conservation has been suggested to estimate their mutual dependency [94].

2.2.3 Co-Evolution within Individual Proteins

The simultaneous change of residues that are in contact can as well be observed within a protein. Intra-protein contacts are crucial for the tertiary structure and the adaption of contacting residues provides a mechanism to maintain protein function. Obviously, intra-molecular changes may affect surface interfaces and thus, promote co-evolution among families as well. In practice, both types of co-evolution exist simultaneously. Co-adaption promotes co-evolution, yet there is a chance that co-evolution would happen by chance. A causation can therefore not be concluded from the information encoded in protein families.

2.2.4 A Real-World Problem of Antibiotic Resistance

More than twenty years ago, in 1992, Harold Neu (†1998) discussed mechanisms and the impact of antibiotics resistance in pathogens [180]. He concludes:

> The need for new antibiotics will continue because bacteria have a remarkable ability to overcome each new agent synthesized.

He describes that selection of mutants due to inappropriate (or even necessary) application of antibiotics would lead to further crisis in the future. Today, we know that this prediction was perfectly true and still, the problem is even more pressing than it was twenty years ago [187].

Therefore, the understanding of the underlying mechanisms of antibiotic resistance remains a crucial topic. The complex interactions among cell-wall proteins like transporters, signaling proteins and small molecules allow a rapid adaption of resistance types and require for detailed models capturing the respective binding sites and their mutual influence. While the amino acid sequence of most key players can be determined, even for larger families of orthologs, their 3D-structures remain unknown. Here, computational predictions of proteins that interact would lead to a deeper understanding of the underlying pathways.

ATP-binding cassette (ABC) transporters define a membrane-bound superfamily of proteins that feature both trans-membrane domains and a specific nucleotide-binding domain. They are crucial for the active transport of substrates across membranes and constitute an important building block in the detection of antimicrobial peptides in *Firmicutes* bacteria. Their gene expression is regulated by so-called two-component systems (TCS) with kinase functionality. For these systems, no sensor domain is known. The ABC transporter permease however, features a large extracellular region. Therefore, we suggest a direct regulatory interaction between the ABC transporters and the TCS [55]. We point out that for example in *Bacillits subtilis* there is an "absolute requirement" for the presence of both units to detect antimicrobial peptides. In particular, the extracellular domain of the ABC could constitute the detection domain, and hypothesize that a regulatory interaction among both subunits is in fact common in *Firmicutes.*

In this chapter, we describe how correlation measures capturing co-evolutionary evidence may help to support this hypothesis. We follow the approach of Goh *et* *al.* [94] (see Section 2.3) to distinguish functional, potentially direct interactions from non-interactions and provide some additional ways to visualize and guide further experiments.

2.3 Related Work

In this section we provide a brief overview of the field of protein co-evolution. For the problem of bacterial ABC transporters we found that a modified variant of the seminal method of Goh *et al.* [94] provided the fastest and most practical way to come up with a sensible estimate of co-evolution.

Co-evolution on the protein level is motivated by the observation of mutual influence among species as described above. Pazos and Valencia [190] provide a detailed historical outline and also discuss the critical difference between co-adaption and co-evolution. One of the earliest theoretical discussions on the matter can be found in Lapedes *et al.* [146].

Goh *et al.* [94] introduced a way to quantify the distance among two phylogenetic trees in terms of their co-evolution. First, for two protein or protein domains sequences are collected from related organisms or bacterial strains. These proteins are usually subject to divergent evolution. Their sequences may differ strongly, except for regions that are crucial to the protein function or an interaction. A multiple alignment is constructed for each protein family and a matrix of all induced pairwise alignment distances is compiled. A linear correlation coefficient is computed among these matrices, referred to as matrix correlation (CC, see Section 2.4.2).

Pazos and Valencia [189] extend the approach of Goh *et al.* and applied the basic method on a larger and more heterogeneous dataset. Basically, the approach termed *mirror tree* resembles that of Goh *et al.*. The name is following the idea of measuring the similarity of family-wise phylogenetic trees. The actual method does not rely on the deviation of a phylogenetic tree though but relies on the comparison of the matrices derived from the respective family (see Section 2.4.2). In principle, as for Goh *et al.* all information of a phylogenetic tree is present in the matrices.

Ramani and Marcotte [204] applied the basic idea to identify interacting proteins among two sets of homologous sequences. While for ortholog sequences the interacting pairs are given by the species, this piece of information is missing for general homology. Ramani and Marcote provide an iterative optimization algorithm that swaps the pairwise protein-protein interaction assignments until the best possible overall assignment is found.

On the residue level, contacts are often modeled as weighted bipartite graph. Each node represents a single residue (the corresponding aligned position). The edge weight is proportional to the distance of the two positions in contact. The problem of contact prediction can then be seen as the reconstruction of a residue contact graph. Gloor *et al.* [90] use mutual information of multiple alignment positions to decide on single contacting positions. Similar approaches introduce row and column weighting [97], use specific substitution correlation [60] or apply molecular dynamics simulations to refine the initial mutual information based results [216].

Jones *et al.* [132] discuss that consideration of indirect effects is necessary to overcome low accuracies in residue contact prediction (typically 20-40% correctly assigned contacts). They argue that this is caused by the introduction of many false positives due to so-called "chains of covariance". The effect can be observed due to two reasons: (1) phylogenetic bias and (2) indirect coupling effects.

Two non-contacting residues may seemingly co-evolve although two direct coevolution events exist in-between. This problem has been addressed via the analysis of higher-order contacts using triplets [111], by message passing algorithms in a Bayesian context [30, 250, 31] or by matrix decomposition approaches [75]. Ekeberg *et al.* [65] show that the main setting is equivalent to the general problem of *inverse statistical mechanics.*

2.4 Methods

2.4.1 Data Acquisition

We obtained six distinct sequence families as described in Dintner *et al.* [55]: BceR, BceS, BceA, BceB, YycG and OppB. The membrane-bound BceS and intracellular BceR constitute the two-component systems. They were derived by querying and filtering homologs of the TCS in *B. subtilis.* The membrane-bound BceB and BceA constitute the transporter and ATPase system. YycG and OppB were obtained to serve as negative controls known not to interact with any of the other families. The so-called 'core' set of othologs contained 26 sequences with proteins available for all six families. It spans highly related bacterial strains. The Bce homologs could be obtained across 180 distinct organisms, yielding the 'extended' set.

For each data set, a multiple sequence alignment of all sequences ('complete') was generated using ClustalW2, using a Gonnet280 substitution matrix, gap open 10, gap extension 0.2 and gap distance penalty 5. Additionally, 50 randomly chosen subsets of size 20 were sampled from each core and extended set and realigned using aboves parameters.

For some of the family pairs known physical interactions had been previously reported, namely BceA/BceB and BceR/BceS. Similarly, all pairwise interactions among YycG and OppB do not feature physical interaction.

2.4.2 Alignment Correlation

The method of Goh *et al.* takes two sets of orthologous multiple sequence alignments as an input. For each set of orthologs a global multiple alignment of n sequences is computed.

We then transform each alignment comprising the n sequences of a familiy k to a matrix of n^2 sequence-pair similarities $M_k := (m_{ij}^k) \in \mathbb{R}^{n \times n}$, where $i, j \in \{1, \ldots, n\}$ is a sequence pair of the respective multiple alignment. Since we are dealing with highly related sequences, we rely on the induced pairwise average sequence identity (number of identical positions in the alignment divided by the average length of both sequences) rather than a distance measure. For the purpose of computing a correlation these measures yield very similar results.

For two sequence families, we obtain two matrices $M_1, M_2 \in \mathbb{R}^{n \times n}$ that express all pairwise sequence identities. Following Goh *et al.* [94] a linear correlation coefficient CC is then computed as

$$CC(M_1, M_2) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (m_{ij}^1 - \overline{m^1})(m_{ij}^2 - \overline{m^2})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (m_{ij}^1 - \overline{m^1})^2} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (m_{ij}^2 - \overline{m^2})^2}} \tag{2.1}$$

where $\overline{m^k}$ is the mean of all matrix entries in M_k.

Goh *et al.* suggest to compute a p-value by re-assignment of ligands to other receptors an repeatedly compute the random of this novel mapping. In practice, the rows of M_1 are shuffled. This should reflect the probability that a correlation coefficient is obtained by chance, yet the original work does not discuss the nature of this p-value in more detail. The permutation determines the nature of this p-value. The procedure relies on a permutation of the pairwise association of orthologs. Thus, the resulting p-value can be interpreted as the probability that a new phylogenetic assignment would yield the observed correlation or above. This p-value is therefore not a measure of the likelihood that the two protein are functionally related or not. Instead, we obtain a set of non-interacting proteins from the same set of organisms and compute a negative background CC.

While the negative background is an effective tool to judge the magnitude of an observed CC, the effects of less conserved strains remain. Outliers to the family in terms of sequence conservation may disrupt measures of co-evolution. Family members that are less related to the majority of sequences in the family may enforce non-characteristic shifts within the corresponding alignments. These may have a strong impact on measures of coupled mutations as they promote alignment entropy. Consequently, estimates of co-evolution are likely influenced by these outlier sequences. We tackle this problem by a simple sub-sampling approach. From n sequences we randomly select $2 * n$ subsets containing 75% of organisms and construct novel multiple alignments for each such subset. A sampled variant of the CC is computed. This results in a distribution of CCs that reflects the impact of individual sequences for two protein families. To obtain comparable results, we restrict the sampling size of larger families to that of the smaller family.

2.5 Results

Our aim is to apply a measure of phylogenetic co-evolution to gather evidence for a functional interplay of the two protein families BceS and BceB, a membrane-bound histidine kinase and a ABC transporter with an extracellular sensor for antimicrobial peptides. We applied a measure of co-evolution based on multiple alignments of protein families to each pair of sequence sets (BceS, BceR, BceB, BceA, OppB and YycG). The resulting correlation coefficients are shown in Table 2.1.

As discussed above the CC (see Equation 2.1) has no direct interpretation. We cannot judge from its magnitude whether it provides any evidence on co-evolution among the two protein families. In fact, the topology of the selected sequences could lead to a strong correlation, even for unrelated proteins.

Family	OppB	YycG	BceB	BceA	BceS
BceR	0.37	0.52	0.94 (0.78)	0.91 (0.91)	*0.92* (0.84)
BceS	0.41	0.51	**0.97** (0.82)	0.9 (0.72)	
BceA	0.43	0.62	*0.93* (0.79)		
BceB	0.44	0.55			
YycG	0.64				

Table 2.1: Matrix correlations among all six families of proteins of *Firmicutes* analyzed. The values are CC values (see Equation 2.1) for the *core* set of 26 proteins. The bracketed values reflect the *extended* set of 180 sequences, where available. BceA-BceB and BceR-BceS (italic) constitute known control interactions. BceS-BceB (bold) is the hypothetical interaction among the two-component system and the transporter permease. All correlations with OppB and YycG serve as negative controls.

Therefore we have introduced positive and negative controls as a frame of reference. We observe that the negative controls are far less conserved than the positive controls, *i.e.*, the known interactions. Strikingly, the postulated interaction of the two membrane-bound components BceS and BceB exposes an even higher co-evolution coefficient than the positive controls.

Furthermore, we computed a simple contact graph among all families (see Figure 2.1). This graph features two major components. The first component entails highly correlated and interacting families that show correlated mutations exceeding a threshold of 0.7 in their co-evolution coefficient. The second component consists of non-interacting entities that are on par with the background correlation. A clear distinction can be made between the contacting families of the Bce-homologs and the known negatives of OppB and YycG.

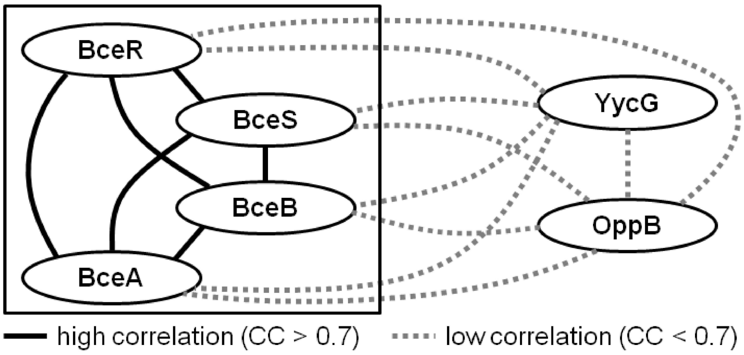


Figure 2.1: A contact graph among protein families based on a correlation measure among pairwise sequence identities. Coefficients are calculated from the complete datasets of 26 sequences, using all combinations of protein family pairs. The graph visualizes the correlation among all six analyzed protein families. Grey dotted lines encode lower correlation levels ($CC < 0.7$) whereas solid lines encode higher correlation levels above ($CC \geq 0.7$). The components of Bce-like modules are grouped by a box. A clear-cut pattern of intra-family co-evolution is obtained suggesting a potentially direct regulatory interaction of the ABC transporter and two-component systems.

To estimate the influence of outlier sequences we sampled 50 subsamples of size 20 ($2 * n$, 75% of sequences) from the collected families and re-computed the multiple alignments for each subset. We then computed the CC for each sampling. The distribution of CCs for all pairs is shown in Figure 2.2).

We observe that the BceR-BceS distribution clearly exceeds the negative controls on average. Few samples exist where a sub-sample would actually obtain a worse correlation coefficient than the highest observed negative controls. By contrast, the observed correlation is on par with positive the positive controls.

2.6 Discussion

In this chapter we have provided an overview of the field of co-evolution that can be used to discover protein interactions. We focussed on the application of an approach by Goh *et al.* [94] to a real-world problem with a direct connection to

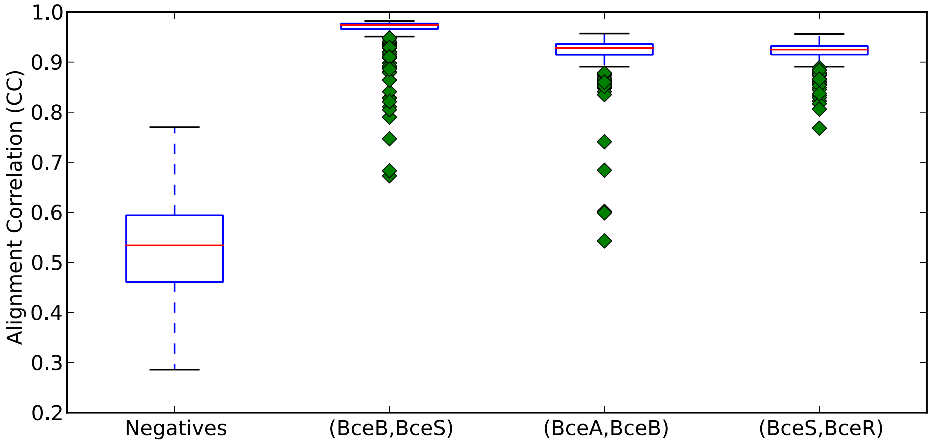


Figure 2.2: Correlation coefficients from random sub-samples of 20 sequences (among 26) were chosen for negatives and 180 sequences among all other pairs. Each box shows the 25%, 50% and 75% quantiles; the whiskers encode the 1.5-fold Inter-Quartile-Range; green diamonds are outliers exceeding this range. The set of "Negatives" consists of all pairs involving at least one YycG- or OppB-homolog, known not to directly interact with the other families.

antibiotics resistance in *Firmicute* bacteria. Therefore, we designed and implemented a prediction pipeline to determine the functional dependency among two families of ortholog proteins from their multiple alignments. Since in our case, both families are membrane-bound, they provide an inherently difficult target for protein structure prediction, such that structural evidence for a contact is difficult. A sequence-only method was therefore promising. With the advance of high-throughput sequencing these methods will gain further importance.

We used a measure of co-evolution in combination with carefully selected positive and negative controls to gather supporting evidence for the contacting and simultaneous conservation of both proteins. We extended the original method by a conceptually simple sub-sampling and obtained sensible contrasts between known negative and positive regulations well-suited for further visualisation.

Together with the results presented in the paper by Dintner *et al.* [55], the alignment correlation could successfully provide evidence for a completely new mechanism of regulation that is widely conserved among *Firmicutes*.

Chapter 3

Detection of Spatial Artifacts

Background

This chapter is based on Petri *et al.* [193]. We present the results of joint work of Tobias Petri (TP), Evi Berchtold (EB) and Caroline Friedel (TP). We analyzed a novel type of exon arrays that was capable to distinguish newly synthesized RNA. During the initial analysis we detected strong artifacts on the analyzed exon arrays. The best practice then was to repeat the affected (if not all) experiments. Yet, for reasons like limited funding or time restrictions, repetition is often impracticable. Due to missing sensible off-the-shelves solutions for the replacement of probe level data, we decided to further investigate on the subject. We came up with a conceptually simple and generally applicable outlier detection and probe replacement approach. In contrast to existing multi-step algorithms that provided an overwhelming amount of parameters, we focussed on a simple visually guided selection.

Contributions

The experimental setup was designed by TP and CF. Data preparation was done by EB and TP. TP implemented the spatial smoothing filter. EB implemented the probe replacement routines. TP and EB ran experiments and evaluation runs. EB adapted the routines for interactive use. The interpretation and discussion of results was done by TP and CF.

3.1 Experimental Artifacts

Noise within experiments can be harmful for downstream analysis. Noise and artifacts present within raw data are usually simple to detect. Unfortunately, they will likely corrupt further analysis and therefore leading to biased results. In this chapter, we discuss the presence of artifacts in microarray databases and their impact on standard routines that are widely applied in the context of microarrays.

3.2 Existing Databases contain Artifacts

Hybridization-based DNA microarrays are a key technology for high-throughput quantification of expression levels for thousands of genes [153, 220]. State-of-the-art microarrays now allow the genome-wide analysis of transcript abundance not only for entire genes but also for individual exons, for alternatively spliced transcripts and even for a large fraction of non-coding genomic regions [19, 82]. Thus, despite the increasing prevalence of alternative methods such as RNA-seq [249], RNA microarrays remain important for the analysis of many biological processes such as miRNA-based regulation [74], alternative splicing patterns across human tissues [98] or the role of alternative splicing in stem cell differentiation [212] and cancer [147].

Recently, Langdon *et al.* [145] reported that *all* human Affymetrix microarrays available in the Gene Expression Omnibus (GEO) [14] contain spatial defects to some degree. Thus, quality control for microarrays remains a major issue.

Although many methods and software tools have been developed for quality assessment of microarrays [24, 77, 134, 255], detection of spatial artifacts is not yet routinely applied. Furthermore, it is usually not clear how to proceed once such artifacts have been detected. The two alternatives are (1) to either completely exclude or (2) to include the corresponding arrays for any subsequent analysis. In the first case, the corresponding measurements are not available for gene expression profiling and may even have to be repeated if they are crucial to the analysis. This can be cost-intensive, for instance if corresponding samples have been used up. In the second case, one has to assume that normalization and summarization methods can correct for the measurement errors.

The latter assumption is based on the construction of microarrays where probes of the same probeset are not contiguous on the array. Thus, smaller artifacts due to uneven hybridization or other experimental problems may only affect a subset of probes for a probeset. The general assumption is that summarization methods – which combine the values for individual probes to a probeset value, such as *RMA* [130] – can estimate the probeset value correctly despite measurement errors for some probes.

We illustrate that this assumption is often invalid by showing that even small artifacts on the array can have a significant effect on the overall expression values of many probesets, not only the ones affected by the artifact. Furthermore, we introduce two simple but effective approaches for the identification of corrupted probes: (1) a threshold-based approach and (2) an extension of this approach that takes into account the neighborhood of a probe, *i.e.*, spatial information of the array. We show that the use of spatial information improves the identification of defective probes as well as the reproducibility of probeset intensities after summarization. Finally, we propose two strategies to either correct probe values using probeset information or to filter corrupted probes, both of which improve summarization accuracy as well as reproducibility between replicates. In this way, we can recover even arrays with large artifacts for downstream analysis that otherwise would have to be discarded.

3.3 Noise Detection on Probe Level Data

As outlined above, the GEO exhibits spatial defects in a substantial fraction of microarrays. Nevertheless, in contrast to quality assessment, artifact detection is not widely used in standard gene expression analysis pipelines. Furthermore, although approaches have been proposed to detect diverse types of spatial noise on arrays, the correction of these artifacts is mostly left to either summarization methods or the corresponding arrays are completely discarded.

We show that state-of-the-art robust summarization procedures are vulnerable to artifacts on arrays and cannot appropriately correct for these. To address this problem, we present a simple approach to detect artifacts with high recall and precision, which we further improve by taking into account the spatial layout of arrays. Finally, we propose two correction methods for these artifacts that either substitute values of defective probes using probeset information or filter corrupted probes. We show that our approach can identify and correct defective probe measurements appropriately and outperforms existing tools.

While summarization is insufficient to correct for defective probes, this problem can be addressed in a straightforward way by the methods we present for identification and correction of defective probes. As these methods output CEL files with corrected probe values that serve as input to standard normalization and summarization procedures, they can be easily integrated into existing microarray analysis pipelines as an additional pre-processing step.

3.4 Related Work

Although not commonly included in standard microarray analysis pipelines, a number of methods have been previously proposed for visualization of microarray artifacts as well as identification and/or correction of corrupted probe measurements (see also Arteaga-Salas *et al.* [8] for an overview of methods published before 2008). One of the most frequently used approach is *Harshlight* [233], which identifies and masks local artifacts based on statistical and image processing methods. Artifacts are grouped into three classes based on the variation from the median array: compact defects affecting only a few probes, diffuse defects affecting larger areas and extended defects that are even larger and may thus invalidate the whole array. Probes within defects can either be excluded from the analysis or be replaced by the median intensity across replicates.

An alternative method for identifying artifacts from raw intensity values is *Microarray blob remover (MBR)* [230], which operates in two steps. First, broad areas are determined in which more than half of the probes are above the kth percentile of probe intensities, where k may be in the range of 60 to 100. These candidate areas are then further refined and probes flagged to be within artifacts are added to the 'outlier entries' section in CEL files.

In addition, several other methods have been proposed based on comparisons to reference arrays [9, 127, 167, 206]. Reimers and Weinstein [206] calculate the log fold-change of the probe value compared to the trimmed median of reference arrays to visualize spatial artifacts but do not aim to identify individual defective probes. Arteaga-Salas *et al.* [9] use a modification of Upton and Lloyd's approach [239] to identify areas in which the largest fold-changes compared to the median of all arrays all stem from the same array. Having identified arrays with defects, they then try to correct the original values using the values for the probe on the other arrays. A similar approach is also pursued by Hulsman *et al.* [127] as part of their normalization pipeline.

The most recent approach, *caCORRECT2* [167], uses a z-score-like statistic (h-score) to estimate whether a probe value on a given array is consistent with the observed distribution for all other arrays. Corrupted probes are then flagged if they have high h-scores and are contained in regions of high h-scores. Corrected values for these probes are then estimated both from the other probes in the same probeset as well as the other arrays using singular value decomposition.

3.5 Outline and Experimental Setup

Our analysis is structured into two parts. First, we illustrate that state-of-the-art robust methods for summarization of probeset values from the individual probe

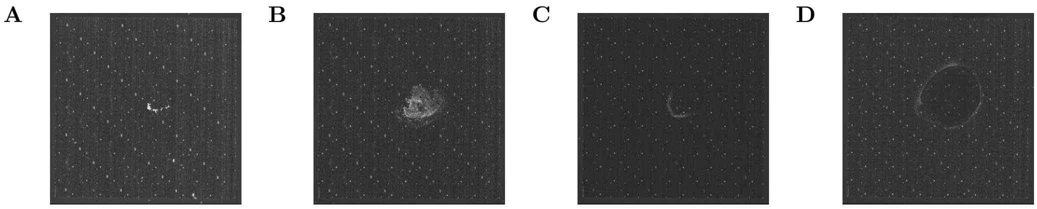


Figure 3.1: Measurement artifacts observed on different arrays of our dataset: total RNA for replicates 1 (A) and 3 (B) in DG75-10/12 cells; total RNA for replicate 2 (C) in DG75-eGFP cells; newly transcribed RNA for replicate 1 (D) in DG75-eGFP cells.

values cannot appropriately correct for measurement artifacts. Second, we present methods for the identification of probes affected by measurement artifacts and the correction of artifacts by replacing the values of affected probes or modifying probeset definitions.

To evaluate the performance in correcting for artifacts, we used 18 exon array measurements of DG75 and DG75-10/12 B-cell lines (see Section 3.6) for which distinctive measurement artifacts were observed in some samples (see Figure 3.1). These measurements included three replicates each of total RNA, newly transcribed RNA labeled for 60 min with 4-thiouridine [57, 135] and the complementary unlabeled pre-existing RNA. As newly transcribed and pre-existing RNA should sum up to total RNA, these experiments provide a true biological control for the assessment of quality problems and their correction.

In this study, we focused mostly on the measurements of the DG75-10/12 cells. In this case, 2 out of 3 total RNA measurements showed substantial spatial artifacts in the images of the arrays but the corresponding measurements of newly transcribed and pre-existing RNA were free of defects or showed only very small or weak artifacts allowing us to use these as biological control (see Figure 7.3). The largest artifact affecting a sizable amount of probes was observed in replicate 3 and a smaller one in replicate 1. Replicate 2 was artifact-free in total RNA, although slight defects were observed in pre-existing and newly transcribed RNA. As we only required the total RNA sample of replicate 2 as a control for the other two replicates, this was not a problem.

3.6 Microarray Measurements

3.6.1 Data

We used RNA measurements for two cell lines using both Affymetrix GeneChip Human Gene 1.0 ST and Exon 1.0 ST arrays: 1) the B-cell line DG75 transduced to express 10 out of 12 miRNAs encoded by the Kaposi's sarcoma-associated herpesvirus (KSHV) (DG75-10/12) and 2) DG75 transduced to express eGFP (DG75-eGFP) as control [247]. For each cell line and array type, total RNA was quantified. In addition, RNA synthesis and decay was measured using a recently developed method for labeling of newly transcribed RNA using 4-thiouridine (4sU) [135, 57]. This allows the separation of total cellular RNA (T) into the labeled newly transcribed RNA (N) and the unlabeled pre-existing RNA (P) as well as quantification of de novo transcription and decay in a single experimental setting.

For each cell line and each RNA fraction three replicates were measured resulting in a total of 18 arrays for each microarray platform. The Gene 1.0 ST measurements were recently published [56]. Exon 1.0 ST measurements were performed in the same way. However, in this case considerable experimental artifacts were observed for several of the 18 arrays resulting in distinctive stains visible in the array images (see Figure 3.1, 7.1 and 7.2). These artifacts were probably a consequence of a drying out of the central part of the array during the hybridization step resulting in artificially high values for the corresponding probes.

3.6.2 Summarization and Normalization

Two steps that are generally performed first in a microarray experiment are normalization and summarization. Normalization is applied to allow the comparison of results from different replicates and conditions. Summarization estimates overall expression values for each probeset from the individual probe measurements.

Normalization. In this study, we used quantile normalization, which is commonly used in combination with *RMA* summarization. If newly transcribed (N) as well as pre-existing (P) RNA have been quantified in addition to total cellular RNA (T), an additional normalization step has to be applied to account for the different amounts of RNA between the fractions [57]. Since $T = N + P$ has to hold approximately for all probes, the linear model $T = \lambda_1 N + \lambda_2 P$ minimizes the sum of residuals for $\lambda_i \in R^+$, $i \in \{1, 2\}$. The corresponding λ_i can be found by linear regression [57], which can be applied both on the summarized probeset values as well as the individual probe values themselves. If the fold-change between replicates is used to calculate the probe noise score a loess normalization is additionally applied before fold-change calculation.

Summarization. One of the most widely used summarization techniques is Robust Microarray Averaging *RMA*, which estimates both an overall expression value for each probeset and the probe-specific measurement error by fitting a linear model to the probe values [130]. Thus, this method implicitly estimates the noise level for each probe and effectively subtracts the estimated noise from the probe when calculating the overall probeset values. This explains why it is commonly assumed that this method can correct for measurement artifacts [154]. For our purposes, the Affymetrix Power Tools (APT) were used for summarization (`http://www.affymetrix.com`). An alternative implementation is provided by the *affyPLM* [24] library, which also provides access to the estimated residuals. However, due to its considerable memory usage when working on exon arrays the *affyPLM* library could not be applied to all exon array measurements together.

3.7 Quality Assessment

3.7.1 Probe Noise Score

To assess the level of noise for individual probes, different criteria can be applied. If measurement errors are explicitly modeled as in the *RMA* approach, residuals can be used to assess reliability of the corresponding approach [24]. The higher the absolute values of the residuals, the stronger the effect of measurement errors on this probe. The global residual level (calculated by the APT subroutine *qcc* [154]) can be used to indicate which arrays are suited as control and which are likely to contain artifacts.

A general probe-level noise score for probe j can be calculated as the fold-change compared to a control:

$$s_j = \left| \log_2 \frac{v_j + c}{v'_j + c} \right| \tag{3.1}$$

Here, v_j is the intensity for the probe on the corrupted array and v'_j is their value on the control. The pseudocount c corresponds to the estimated detection limit (in our case $c = 16$). Both v_j and v'_j can be measured directly or can be derived values, *e.g.*, using measurements of newly transcribed and pre-existing RNA as described in the normalization section. In the latter case, the normalized sum of N and P serves as a control for the measurement of T, *i.e.*, $v_j = T$ and $v'_j = \lambda_1 N + \lambda_2 P$. Alternatively, replicates may serve as a control. If it is not possible to determine a suitable control, the fold-change against the median probe intensities of all replicates can be used. This corresponds to the error image used by *Harshlight* [233].

3.7.2 Probe Noise Plot

As each probe has a defined location on the array, the noise level of individual probes can be visualized by plotting the noise score of the probe against this location. For a more intuitive visualization, the noise score is color-coded and the location represented by the x- and y-axis. If residuals from the *RMA* model are plotted, this corresponds to the *residual plot* proposed by Bolstad et al. [24]. To calculate residuals for noise plots, we used the *affyPLM* implementation of *RMA* which provides access to these residuals. In this case, residual estimation for a specific array was based on the three replicates for the corresponding condition.

3.7.3 Replicate Scatter Plot

To evaluate correction of artifacts, probeset values for the affected arrays are plotted against the control values. If no artifacts are observed, summarized probeset values should be highly reproducible between the replicates. Instead of replicates for the same condition and RNA fraction, the complementarity of the total, newly transcribed and pre-existing RNA fractions can be exploited.

3.7.4 Introducing Artificial Noise

Measurement errors were introduced artificially (*spiked*) in exon array measurements by selecting a noise level δ and spiking each probe according to this probability. The raw measured values of spiked probes were then replaced by an artificial level drawn from a log-normal distribution with mean μ and standard deviation σ (in our case $\mu = \log_2(850)$ and $\sigma = 1$ were inferred from the intensities within the real artifacts). Only probes corresponding to core probesets defined by Affymetrix were spiked and included in the summarization. Simulations for each selected value of δ were repeated 100 times.

Furthermore, real-life stains were projected onto the Gene ST arrays to create realistically shaped artifact patterns. For this purpose, our artifact detection approach (see Section 3.8) was applied to the exon array measurements with artifacts to detect the location of the corrupted probes. The exon array artifacts were then scaled down to the dimensions of the gene arrays and transferred to the artifact-free gene arrays. For this purpose, 2×2 rectangles of probes on the exon arrays were mapped to one probe on the gene arrays and the maximum value of any of the probes in this rectangle was used for the spiked probe. To account for the overall larger intensities on the gene arrays, the resulting value was multiplied by the ratio of the 75 percentile of the intensity distribution on the gene arrays relative to the corresponding 75 percentile for the exon arrays.

3.8 Artifact Detection

We propose two alternative approaches to identify probes which are affected by significant measurement errors. The first method is based on a simple threshold criterion, the second approach extends this method by including the neighborhood information on the array.

3.8.1 ϵ-criterion

The ϵ-criterion is based on the noise score defined in equation 3.1 and simply applies a threshold t on this score. Thus,

$$\epsilon(s_j) = \begin{cases} \text{true} & s_j > t \\ \text{false} & \text{otherwise} \end{cases} \tag{3.2}$$

If $\epsilon(s_j)$ is true, probe j is flagged as corrupted. Thresholds can be adjusted manually by analyzing both probe noise and replicate scatter plots.

3.8.2 Window Criterion

As measurement artifacts usually affect a specific region on the array and, accordingly, a set of probes closely located to each other, we propose a method which takes into account the neighborhood information. Note that noisy regions, *i.e.*, artifacts on the chip do not necessarily yield a higher mean intensity. An experimenter looking at a defective chip will spatially group noisy spots. Our window criterion follows this intuition, in terms of a spatial correlation among spots. Thus, for estimating the reliability of a specific probe we take into account the values of the probes in a window around this probe. For our purposes, we used a 2D window of dimension $(2k + 1) \times (2k + 1)$ with the probe considered in the center of the window (here $k = 25$ was used).

We calculate a weighted average of the probe noise scores in this window:

$$sw_j = \frac{\sum_{p \in P} s_p \cdot w(p, j)}{\sum_{p \in P} w(p, j)} \tag{3.3}$$

where P is the set of probes in the window, s_j the noise score of the probe p and $w(p, j)$ is the weight of probe p in the window for j. The weight is calculated as $1/d(p, j)$ if $p \neq j$ where d is the distance between probes. In this study, we used the euclidean distance on the probe coordinates but alternative distances can be used. If $p = j$, the weight is set to 2. If residuals from *RMA*-like methods are used as noise scores, s_p is set to the absolute value of the residuals. For probes close to the borders of the array, the window will be cut off at the respective sides. Subsequently, the ϵ-*criterion* is applied to the window-based noise scores.

3.9 Correction of Corrupted Probes

For correction of corrupted probe values we use two alternative approaches. In the first case, we replace the intensities of the corrupted probes by the mean intensity of the remaining probes of the corresponding probeset in the CEL file. This correction only takes into account probe values measured with the same array, thus, differences in intensity distributions between arrays do not have to be considered. If all probes of a probeset are corrupted it is not possible to infer a meaningful probeset intensity. Thus, we set all probe intensities to 0 resulting in a probeset intensity of 0. These probesets should be excluded from further analysis.

The alternative method consists in removing corrupted probes from the probeset definition by modifying the PGF annotation file provided by Affymetrix. It should be noted that it is also possible to completely exclude affected probes from the summarization procedure using the "--kill-list" option of the Affymetrix Power Tools. Yet, downstream tools may request for the filtered probe values, thus, direct probe value correction is far more robust than complete removal.

3.10 Evaluation of Artifact Detection

To evaluate the performance of artifact detection, Gene ST arrays were spiked as described above. For each threshold applied, we then calculated *true positives* (spiked probes that are filtered, TP), *false positives* (probes not spiked but filtered, FP) as well as *true negatives* (probes neither spiked nor filtered, TN) and *false negatives* (spiked probes not filtered, FN). To evaluate different approaches over all possible thresholds, we used Precision-Recall curves for which

$$precision = TP/(TP + FP) \tag{3.4}$$

is plotted on the y-axis against

$$recall = TP/(TP + FN) \tag{3.5}$$

on the x-axis for all possible thresholds.

3.11 Probe Noise Plots

Although the array images already gave a first clue to the artifacts observed in our example, this was only due to the high intensity values of the affected probes and not all defects can be identified so easily. Thus, instead of intensities, we visualize *probe noise scores* that quantify the deviation from a control or a linear model as used by *RMA* for example. As a control, we use additional replicates that are artifact-free in the relevant region or the biological control from RNA labeling experiments (see Section 3.7). If residuals from the *RMA* models are used, the probe noise plots correspond to the *residual plot* proposed by Bolstad *et al.* [24].

Probe noise plots and residual plots for the arrays analyzed are shown in Figure 7.3. Here, we used as controls for total RNA either the artifact-free replicate 2 or the normalized sum of newly transcribed and pre-existing RNA of the corresponding sample. While different noise scores pick up the artifacts similarly well for the defective replicates 1 and 3, a striking observation was made for replicate 2 in the *RMA* residual plots. Here, an additional stain showed up in the center of the array, which is not observed in the original image. Most likely, the *RMA* model, which is based on several replicates (in this case all three total RNA measurements), was biased by the stains on the other two arrays at this location leading to large residuals for replicate 2. This provides a first indication that summarization suffers from these artifacts.

3.12 Insufficient Correction by Summarization

To evaluate whether the final probeset values can nevertheless be estimated correctly by summarization, we used *replicate scatter plots* that compare probeset levels between the affected array and a control (see Figure 3.2 A,B and Figure 7.4). Here, the same controls as for the probe noise plots were used and probesets were colored according to the fraction of probes that were flagged as corrupted by our simple thresholding approach on the probe noise scores (the *ϵ-criterion*, see Section 3.8.1).

As expected, the deviation to the control is substantial for probesets with all probes affected as no reasonable estimation is possible. In contrast, if only 75% or less (0-3 probes for most probesets) probes were affected, we did not see a correlation between the number of defective probes and the deviation from the diagonal. Instead, all probeset levels were affected to some degree. Strikingly, the deviation for replicate 1 with the small stain was stronger than for replicate 3 with the largest stain. Furthermore, this deviation was most pronounced for probesets with high expression values, which were not even affected by the stain.

One possible explanation is that this is an effect of the normalization – in this case quantile normalization – that has to be performed before summarization. It might compensate for the extremely high values for some of the probes by reducing the levels of the remaining probes. When omitting quantile normalization, the strong deviation for highly expressed genes in replicate 1 is reduced (Figure 3.2 C,D). For replicate 1 there is a bias even for the uncorrupted probesets (A) that can be reduced by omitting quantile normalization (C). If probe correction is applied

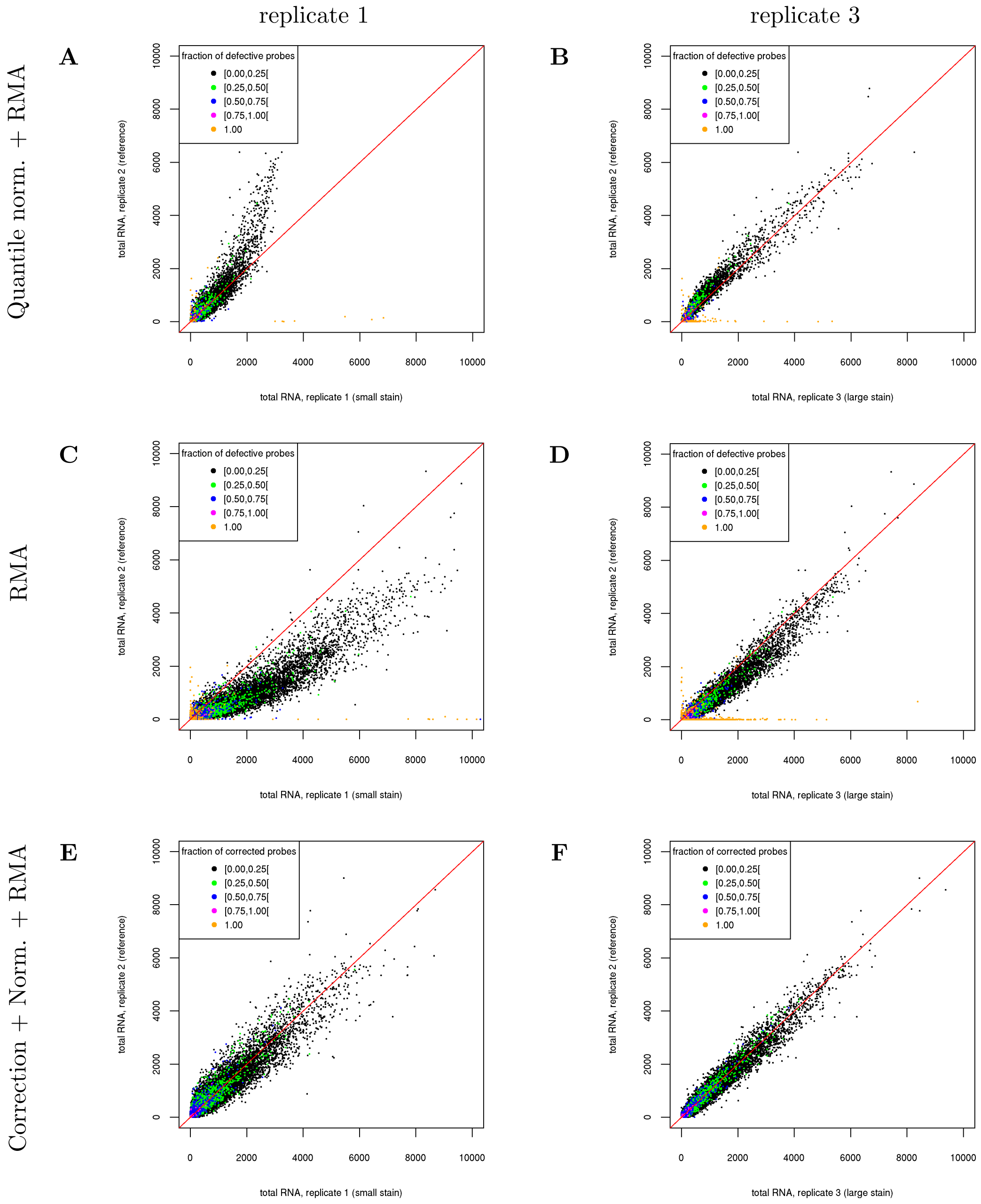


Figure 3.2: Replicate scatter plots comparing total RNA for replicates 1 (A, C, E) and 3 (B, D, F) against the artifact-free replicate 2 for the exon array measurement in DG75-10/12 cells. The subfigures A and B show the results using both *RMA* and quantile normalization, C and D using only *RMA* without quantile normalization and E and F after probe correction. Probesets are colored according to the percentage of their probes that are flagged as corrupted by the ϵ-*criterion* (noise scores using newly transcribed and pre-existing RNA as control).

prior to normalization and summarization (E,F), this bias is removed. The results are shown for the correction method which replaces the probe value by the mean of the unaffected probes in the same probe set. In this case, the intensity of probesets for which all probes are corrupted are set to zero. Results for the filtering approach in which affected probes are removed from the probeset definition are very similar. Nevertheless, even without normalization, the nonlinear behavior for both replicates in comparison to replicate 2 is still observed.

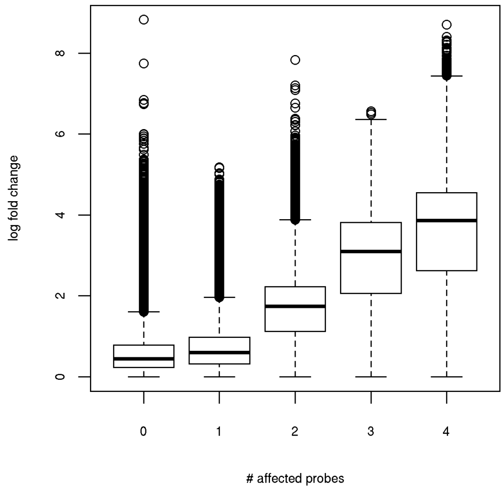


Figure 3.3: A boxplot of the $\log_2$ fold changes for probesets with $0, 1, 2, 3$ or 4 spiked probes in the simulation in which 5% of all probes were spiked in total ($\delta = 0.05$). Here, probesets with the same number of spiked probes were pooled across all simulation results. For the case of 0 spiked probes, probesets were selected randomly from the pooled set as there were too many probesets for loading into R. In this case, each probeset was selected with a probability of 0.01. The more probes of a probeset are spiked, the higher is the fold-change between replicates. We observe a very strong correlation between the number of affected probes and fold-change biases on probeset level, which may seriously harm downstream analyses.

Exon arrays also offer the possibility to summarize probe values to meta-probesets that correspond to genes. As there are more probes per meta-probeset the effect of the artifacts should be smaller. Nevertheless, we still observed a systematic shift from the diagonal in the corresponding replicate scatter plot, although for replicate 3 the deviation was much smaller (see Figure 7.5). In contrast to probeset level summarization, however, omitting quantile normalization did not reduce this deviation.

3.13 Sensitivity of Summarization to Noise

To systematically analyze the influence of measurement artifacts on summarization, we performed the following experiment using three replicates of total RNA measured with exon arrays for the DG75-eGFP cell lines. These measurements were basically artifact-free with only a very small stain in one replicate, which could be easily corrected using our ϵ-*criterion* (see Figure 7.7). Here, only 6380 probes (out of >5.5 million features on the array) were identified as corrupted and 6335 probesets had 1 corrupted probe, 21 had 2 and only one had 3. This is a much smaller number than observed for the substantial artifacts on the DG75-1012 arrays.

We then introduced artificial measurement artifacts into the corrected DG75-eGFP arrays (*spiking*, see Section 3.6). Depending on a noise level δ, probes to be spiked were chosen randomly with probability δ and their intensity values were drawn randomly from a log-normal distribution (with parameters $\mu = \log_2(850)$ and $\sigma = 1$). Mean intensity values were taken from corrupted probes identified by the ϵ-*criterion* on the DG75-10/12 total RNA measurements (mean intensity values ~850) to provide a realistic level of noise. Spiking was performed for only one of the arrays and the remaining arrays were used as control. After spiking the raw values on the array, we performed summarization and normalization. To assess the effect on the resulting probeset levels, we evaluated the average $\log_2$ fold-changes in probeset levels between each pair of spiked array and spike-less control for noise levels in the range of 0.01 to 0.1. For each noise level, random spiking was repeated 100 times.

Comparing the $\log_2$ fold-change against the number of spiked probes for each probeset (Figure 3.3), we found a very clear trend: if only one probe is affected, the median fold-change is slightly higher than for probesets not affected by spiking. However, if more than one probe is spiked, the fold-changes increase substantially. Thus, variance of the probeset levels are increased considerably even if only few probes are affected. This larger variance can lead to low or no statistical significance for differentially expressed genes and as a consequence reduce the sensitivity of gene expression profiling.

3.14 Identification and Correction of Corrupted Probes

To address the problem of measurement artifacts for summarization, we propose a two-step approach in which we first identify corrupted probes and then correct for these corrupted probes in one of two ways. The first correction method consists

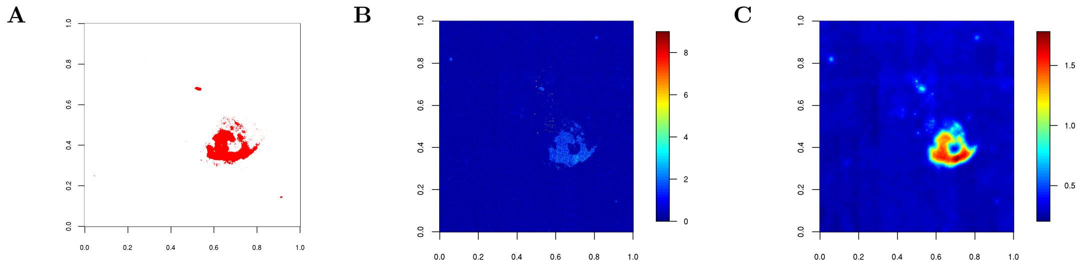


Figure 3.4: Illustration of the results on the spiked Gene ST arrays. Both shape of the artifact and intensities of the spiked probes were transfered from exon arrays containing artifacts. A) shows the spiked probes in red and B) and C) the probe scores based on fold changes between replicates using only the probe information itself (B) or also its neighborhood (C). For both B and C the overall shape of the spiked stain can easily be identified, but only when using the *window-criterion* (C) all probes within this area are identified. Furthermore, in B there are more probes with high noise scores that were not spiked (false positives).

in replacing the probe value by the mean of the remaining unaffected probes for the given probeset. The second alternative consists in removing the probe from the analysis, for instance by re-defining probeset definitions to exclude the affected probes. As several analysis tools including the APT suite cannot handle missing values appropriately and even the *de facto* standard of present and absent flags is often ignored by downstream tools, the first method is more robust than the second.

To identify the corrupted probes we use a simple threshold criterion based on probe noise scores calculated either from fold-changes to a control or *RMA*-derived residuals. Here, we developed two approaches that calculate the probe noise score either for each probe alone (ϵ-*criterion*) or as a distance-weighted mean of the noise scores within a 2D-window around the probe (*window-criterion*, see Section 3.8.2). The latter approach is based on the observation that measurement artifacts, *e.g.*, due to uneven hybridization, usually affect several closely located probes and not only individual probes. Probes with a noise score above a certain threshold are then flagged as corrupted.

To correct the DGF75-10/12 measurements and to evaluate the performance of correction appropriately, we pursued the following procedure to avoid over-fitting. If we compared the corrected and summarized probeset values between replicates (Figure 3.2 E-F), detection of corrupted probes was based on the ratio of total RNA and the normalized sum of newly transcribed and pre-existing RNA and vice versa (see Figure 7.6).

These results show a significant improvement after probe value correction. Both with and without quantile normalization, the distinctive deviation for large expression values seen before in replicate 1 is no longer observed. Instead, for both defective replicates 1 and 3, variance is symmetrical on both sides of the diagonal. This was true both for the correction using mean values of unaffected probes of the same probeset (Figure 3.2 E-F, Figure 7.6) as well as for the filtering approach in which the affected probes were removed from the probeset definition (not shown). Here, the mean absolute deviation from the diagonal decreased from 12.2 in the original data to 7.34 and 4.7 for the first and second correction method, respectively. Thus, even the simple ϵ-*criterion* could successfully identify the defective probes and probeset values could be corrected appropriately, with slightly better results obtained by removing affected probes instead of using values from unaffected probes.

3.15 Accuracy of Artifact Identification

To perform a systematic analysis of the performance in detecting measurement artifacts, we used Gene ST array measurements of the same samples that were measured with the exon arrays. The Gene ST measurements were free of artifacts and have been published recently [56]. Artificial stains were spiked into these artifact-free Gene ST measurements by projecting the artifact observed in total RNA of replicate 3 for the DG75-10/12 cells from the exon arrays to one sample of total RNA measured with gene arrays as described in the methods section (Figure 3.4 A). We used the pattern of the stain on a real-life example instead of random selection or some other spatial pattern to perform a realistic simulation of noise and fair comparison of the approaches.

3.16 Compared Methods

We compared the ϵ-*criterion* and *window-criterion* using probe noise scores based on

1. fold-changes between replicates, calculated from all 3 replicates of total RNA for the DG75-eGFP cells including the spiked replicate.

2. fold-changes between total RNA and normalized sum of newly transcribed and pre-existing RNA corresponding to the spiked replicate.

3. *RMA* residuals calculated based on all 18 replicates using the *affyPLM* library.

These approaches were additionally compared against *Harshlight* [233] and *MBR* [230], which were applied to the 6 array measurements of total RNA. *Harshlight* does not provide noise scores per se but relies on downstream algorithms to decide on affected probes. To compute precision and recall values, probes were ranked by their fold-change to the corresponding median probe value across all arrays. This score is used by *Harshlight* in its initial step. For our purposes, it was additionally incremented by a constant value for probes flagged as defects by *Harshlight* such that all flagged probes ranked higher than any other probe. As *MBR* is only available as a GUI, we investigated only a small number of values for the parameter k (60-80 in increments of 5; values larger than 80 were found to have only very small recall).

Additionally, we planned to evaluate performance of *caCORRECT2* [167] as well as the method by Reimers and Weinstein [206], which both are available as web-servers. However, as none of the two programs had yielded a result 24 hours after uploading the data to the web-servers, we aborted the evaluation. Thus, it appears that these methods did not scale well to the size of the Gene ST arrays used in this study, which are substantially larger than older Affymetrix arrays but still much smaller than the exon arrays. Alternatively, in particular for the Reimers and Weinstein method, the web-servers might no longer be maintained. The method by Hulsman *et al.* for identifying location artifacts is only available as an intermediate step within their normalization pipeline and could not be evaluated on its own.

3.17 Evaluation Results

Figure 3.4 illustrates the spiked artifact as well as the probe noise scores calculated using either only the probe information alone or including also the probe neighborhood using the window-based approach. Here, the probe scores were calculated from the fold-changes between replicates. The window approach results in a much smoother change of scores and high noise scores within the complete spiked area. If scores are calculated on each probe alone, we observe large variations in the spiked area with not all probes having high scores. Similar results are observed for the other types of noise scores (see Figure 7.8), indicating that the window approach results in higher sensitivity in identifying defective probes.

To compare the different approaches, Precision-Recall curves were calculated (Figure 3.5). For this purpose, precision in identifying defective probes is plotted on the y-axis against recall on the x-axis for decreasing thresholds for flagging a probe corrupted. Here, several interesting observations can be made. First, the noise scores based on fold-changes to either replicate or newly transcribed plus pre-existing RNA samples perform almost identically using the ϵ-*criterion*.

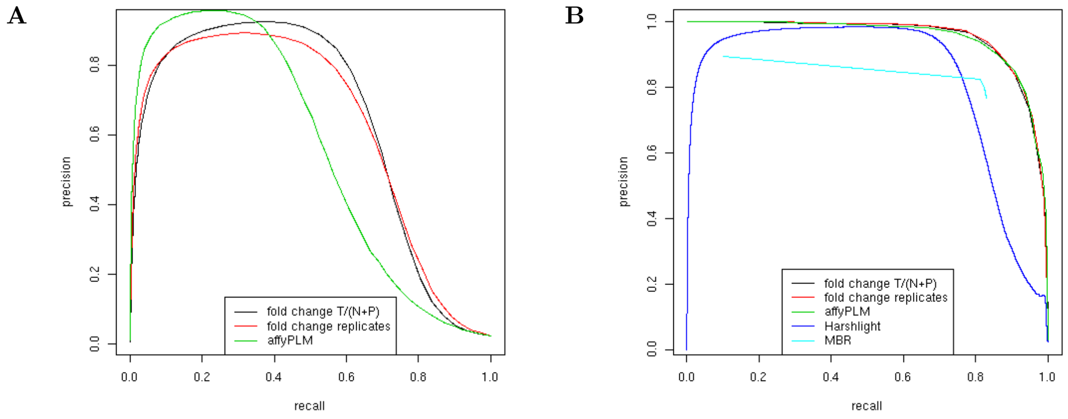


Figure 3.5: Precision-Recall curves for spiked Gene ST measurements. Here, artifacts were projected from the exon array measurements onto the gene arrays to produce realistic noise patterns. Three different scoring approaches were compared both for the simple threshold approach, the *ϵ-criterion* (A), and its cumulative variant, the *window-criterion* (B), which takes into account the probe neighborhood information. The scoring approaches compared are: (i) absolute log fold change between total RNA and normalized sum of newly transcribed and pre-existing RNA (*fold change $(T/N + P)$*, see Section 3.7.1); (ii) absolute log fold change between replicates (*fold change replicates*); (iii) residuals determined with the *RMA* summarization approach using the *affyPLM* model (*affyPLM*). These results show that the window-based approach improves the performance of all used methods, resulting in almost identical performance for all of them, which is superior to the performance of both *Harshlight* and *MBR*.

In contrast, the scores based on the *RMA* residuals show a higher precision for low recall values but this precision deteriorates more rapidly for increasing recall values.

Second, performance of all probe scores improves considerably when we apply the *window-criterion*. By taking the local information of a probe's neighborhood into account, recall can be increased significantly while the number of probes mistakenly flagged as corrupted is reduced. Furthermore, when using the *window-criterion* the differences between the scoring approaches disappear and all scoring methods show a very similar performance. Here, the reason for the poor performance of the *ϵ-criterion* at low recall are a few isolated probes with high noise scores on the arrays that were not spiked and thus, are counted as false positives. While these outliers might also be interesting, they do not indicate a systematic artifact. Accordingly, smoothing over the scores in the neighborhood of these probes reduces their noise level. This enables us to find an appropriate threshold between spiked and spike-less probes independent of the scoring method used.

Finally, the performance of the different *window-criterion* variants was compared to *Harshlight* and *MBR*. While *Harshlight* performs similarly well for intermediate recall values, precision is very low when trying to reach full recall. At a recall of 85% of spiked probes, the fraction of correctly flagged probes is only less than 50%, whereas for the *window-criterion* more than 90% of the flagged probes had been spiked. Thus, it appears that *Harshlight* uses too strict requirements on probe quality and, accordingly, tends to flag too many probes as defective. Additionally, modern platforms like gene and exon arrays appear to cause problems to *Harshlight* due to either calibration or technical issues. Using default settings large diffuse defects are detected even for artifact-free arrays and spike-in probes used for calibration are detected as compact defects.

MBR also performed worse than all *window-criterion* variants at all recall values but outperformed *Harshlight* in a small range. It should be noted that the parameter k used by *MBR* allowed only very little tuning of performance. For k=80 (the largest value investigated), recall was as low as 0.1, then increased dramatically to 0.81 for k=75 and then only increased moderately up to 0.83 for the smallest allowed value of k=60. At the same time, precision varied only between 0.90 for k=80 and 0.77 for k=60.

3.18 Conclusions

In this chapter, we discussed that frequently applied normalization and summarization procedures may be vulnerable even to small spatial defects. We illustrated the necessity of integrating artifact detection and correction into standard gene expression analysis pipelines.

We proposed a general and simple approach for the identification and correction of these artifacts that relies on control measurements, and technical or biological replicates. Furthermore, we have shown that, if available, newly synthesized, total and pre-existing RNA fractions may guide this process. By additionally taking the probe neighborhood into account, we can furthermore improve the detection accuracy compared to more complex multi-step approaches. Thus, even if a substantial amount of probes is defective on an array, the remaining measurements can still be sensibly analyzed.

In a later project we could successfully apply our corrective procedure to the analysis and correction of custom arrays in a veterinary medicine context [140].

Chapter 4

Scoring the Deviation of Expression Patterns from Known Behavior

Contributions

This chapter is based on Petri *et al.* [195]. The experiments described in this chapter were carried out in collaboration with Robert Küffner (RK). The experimental setup was discussed and designed by RK and Tobias Petri (TP). Machine learning routines have been implemented by TP. Leave-one-out experiments have been designed and run by TP. Additional predictive routines have been implemented by RK. The evaluation was implemented by TP and RK. The interpretation of results was done by TP and RK. Ralf Zimmer (RZ) critically reviewed results and provided substantial feedback in the course of this work.

4.1 Differential Analysis

In the last chapter we discussed the problem of microarray artifacts as an example of experimental noise that is relatively easy to detect. In this chapter, we focus on a specific interplay of correlation and noise in differential analysis. To understand this, we first need to sketch the problem setting and the way it is commonly approached.

To pin down the key differences between a diseased and a healthy condition, we usually apply some kind of differential analysis, *i.e.*, both conditions are measured, compared and screened for the most striking differences. We would usually like to detect genes that are differentially regulated and show a statistically significant change in terms of their mRNA abundance level as measured by chips or next-generation sequencing. By comparing the variance within replicates and the inter-condition changes, a list of significant and strongly regulated genes can be compiled. Figure 4.1 shows a prototypic setup that is commonly used.

In this setup candidate genes may be significant and regulated but less interesting in the context of the conditions analyzed. In particular, genes that follow an expected behaviour given the state of other genes. Among the processes that are frequently regulated are inflammatory responses – often involving members of the interleukin family. The underlying mechanisms must maintain their function across a wide range of diseases. They may constitute large fractions of the observed differential behaviour. Similarly, multi-functional genes likely show up in several experiments as well. For the well-known tumor-suppressor gene $P53$ for example it is nearly impossible to name a single context that would best describe its mode of action: according to the *GeneCards* resource (*http://www.genecards.org/*) the gene is associated with some 30 Gene Ontology molecular function terms and more than a hundred pathways.

4.2 Experiment Specificity

As described above, the differential analysis of genes comparing several experimental conditions or treatments routinely estimates which genes change significantly. Multi-functionality and context-free behaviour therefore lead to false positive candidates. Few genes are regulated individually and the commonly observed behavior may thus be a consequence of changes in other genes. Existing approaches like co-expression analysis aim at resolving such patterns. The knowledge of such a background set of experiments can be used to compute expected gene behavior based on known links. It is particularly interesting to detect previously unseen specific effects in other experiments.

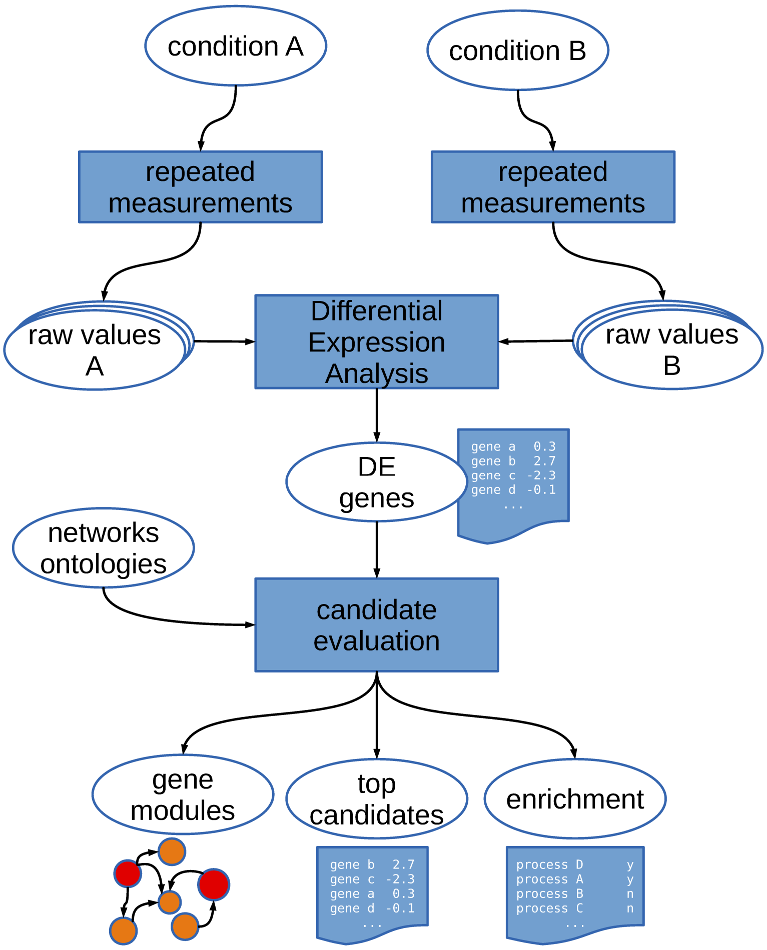


Figure 4.1: A common experimental setup for differential expression analysis. Two conditions are monitored using the same measurement technique and the resulting raw values (possibly with replicated experiments) are used to compute differentially expressed (DE) genes. These are then fed into a candidate evaluation which usually integrates further resources like networks or ontologies. Commonly, this results in the detection of gene modules that are jointly regulated as well as top candidates suited to discriminate the input conditions. Processes that are enriched for differential genes may hint at novel modes of action.

In this chapter we describe a novel method to spot genes deviating from expected behavior (PAttern DEviation SCOring – *Padesco*). It uses linear regression models learned from a background set of differential experiments to arrive at gene specific prediction accuracy distributions. We use these distributions to decide whether a gene is predicted worse (or better) than expected. This provides a novel way to estimate the experiment specificity of each gene.

There is no generic procedure to assess whether a candidate classified as specific is correctly identified. To provide a useful estimate of our procedure we therefore resort to a simulation that introduces specific genes into existing experiments. The resulting validation procedure provides an estimate of the detection rate for these candidates. We show that *Padesco* can identify the experiment specific behaviour of a gene with an average accuracy of about 85 percent.

4.3 Overview of *Padesco*

Candidate gene lists in a differential setting may contain several hundreds of genes. Detailed biological downstream studies are usually not feasible for all of these genes. Further filtering towards more promising candidates is therefore necessary. Moreover, most candidates are likely indirect targets of initially affected genes or, more generally, they follow a pattern which can be observed similarly in other experiments. Such genes may not be of immediate interest. In return, striking differences to known behavior indicate specificity for a certain experiment and such genes are suited for further analysis. We will now introduce how *Padesco* models common patterns and in which way differences to known patterns are obtained.

4.3.1 Patterns

Hirsch et al. [119] noted that disease specific effects eventually trigger core biological pathways and frequently lead to “a transcriptional signature that is common to a diverse set of human diseases”. Such signatures can be learned and used for experiment specific predictions. *Padesco* detects how well the behavior of a gene can be derived from other genes. It allows to detect genes which show both differential and unexpected – and thereby interesting – behavior. The target gene patterns we learn are derived through Support Vector Regression (SVR) and basically constitute linear models describing its dependencies to other genes. Since we aim at unexpected *changes* rather than *states* we use fold-changes, not expression values to describe gene behavior.

4.3.2 Deviations

Not all differential genes detected by differential expression analysis are specific. A background set of experiments must be heterogeneous to assess this. *Padesco*'s key idea is that we can decide whether the behavior of a gene can be predicted worse than expected and is a specific gene. It is important to note that there may well be genes which are difficult or easy to predict in general. Our scoring scheme is designed to account for this individual prediction complexity by estimating an empirical distribution of deviations in a cross-validation (CV) setting.

4.3.3 Evaluation

Evaluation of differential expression results is difficult. Simulations may show methodological strengths and weaknesses, but biological evaluation is only possible through comparison to published knowledge or downstream experiments. We discuss *Padesco*'s performance both by means of an exhaustive simulation experiment and a detailed discussion of literature supporting genes found to be interesting by our approach. The simulation shows that genes deviating from their common behavior would be neglected due to differential expression analysis alone, since they often show only moderate differential expression.

4.3.4 Scoring

Padesco is trained on a background set of experiments consisting of 1,437 microarrays from 25 experiments sharing 4,117 genes. A leave-one-out cross-validation (LOOCV) across all experiments is done yielding predictions for a genes fold-change for all pairs of arrays within the omitted experiment. We estimate how well a gene can be predicted by deriving the empirical distribution of its deviations from the measured fold-changes. We then devise a score based on the median absolute deviation to score a gene in an unseen experiment. We assume that a gene in a differential setting is interesting if it exhibits a change in its gene expression. We therefore use differentially expressed genes. Furthermore, a gene may be predicted better than expected, which points at stronger presence of a trained gene-gene dependency within this experiment. Although the problem is related we do not focus it in this work. If a gene is predicted worse than expected this suggests changes in a known dependency structure.

4.4 Related Work

Padesco is a natural extension to co-expression approaches [64, 149, 208, 238] as well as residual scoring schemes [139, 199]. Co-expression aims to construct gene sets by clustering or the construction of co-regulated gene sets across many samples. Our trained patterns are similar, yet not identical to these previously derived co-expression patterns. Mentionable differences are the use of fold-changes rather than raw measurements and our predictability criterion. Predicting a gene's expression from other genes has been previously applied to estimate condition-specific deviations, referred to as residual scoring [139, 199]. These approaches do not derive predictive models in terms of the prediction of novel experiments. To achieve meaningful residual scores they are applied on homogeneous training data like certain disease subtypes. These scoring schemes fail on heterogeneous experiment data. Instead, *Padesco* omits complete experiments from training and captures actual predictability rather than model residuals. It therefore aims to bridge the gap between the detection of predicted co-regulations and residual scoring. For heterogeneous training sets a background sensitive view on differential experiment results is provided.

In general, identified sets of differentially expressed genes should maximize sample discrimination. The markers, or gene signatures, would then provide promising targets for further analysis. Previously, gene expression profiling was used to identify transcription signatures for breast cancer prognosis classification [241] and gene expression profiles have been used to reveal pathway deregulation in prostate cancer [208]. We rely on comprehensive resources like GEO [62]) that enable the analysis of co-expression across many experiments. The comparison can for instance be quantified by measures of reproducibility in-between experiments [83].

Padesco's linear models are related to those used for imputation of missing values (IMV). Here, either data from single [114] or multiple experiments [123] is used. SVRs have been used to tackle the problem of IMV as well [137], yet to our knowledge neither fold-change predictions nor the imputation of known values has been examined in detail.

4.5 Data Sets

Padesco is trained on a set of experiments compiled by Lee *et al.* [149]. It consists of 3,924 microarrays from 60 human data sets. These sets comprise 62.2 million expression measurements. They consist of 10 to 255 samples. Genes are filtered for a minimum amount of variance across samples. We restrict the data set to Affymetrix array platforms (HG-U95A, HG-U95Av2, HU6800, HuGeneFl, HG-U133A and HG-U133comb). Although absolute expression levels of single experiments and platforms are usually not comparable, the fold-changes used by *Padesco* are suited for further analysis spanning multiple experiments.

We restrict our analysis to a subset G of 4,117 genes that occur in more than 50% of all experiments and contain at least 75% non-missing measurements. 25

experiments with a total of $p = 1437$ microarrays fulfill this constraint. We obtain a matrix of fold-changes $\mathcal{F}$ by sampling p pairs of arrays (Figure 4.2 and Section 4.6.3) from the space of all possible pairs within experiments.

For 13 selected experiments we combine arrays into sample groups, *e.g.*, tumor *vs.* normal samples. This is analogous to usual differential expression analysis (see Section 4.1) to compare the expression of genes between different sample groups. Overall, 62 sample groups have been analyzed. Comparisons are performed as one sample group versus the others from the same experiment, *i.e.*, one comparison is performed for experiments with $n = 2$ sample groups and n comparisons otherwise. We conducted 59 comparisons in total.

4.6 Basic Protocol

Padesco uses a two step approach for the selection of candidate genes from expression measurements. Prior to its application, expression patterns must be trained on a background set of experiments (Section 4.5). We apply Support Vector Regression (see Sections 4.6.2 and 4.6.3) to train one model for each gene given this set. The predicted labels are within-experiment fold-changes of this gene. Training features are the fold-changes of all other genes. For a new experiment (not contained in bs), *Padesco* selects genes by two consecutive filter steps. First, the measured genes are analyzed for differential expression (see Section 4.1) based on Wilcoxon's rank sum test [254]. In general, any differential expression approach can be applied as *Padesco* does not rely on a particular method.

The novel second filter step is based on an analysis of the trained regression models. We discard genes that conform to the patterns learned from the background experiments. We analyze the gene prediction errors using **residual scoring** (see Section 4.6.6) by comparing its predicted against the observed fold-changes. We then assess its pattern conformance in terms of the distribution of its residuals across the LOOCV as described below. Our basic work flow is shown in Figure 4.3.

To arrive at a background distribution of errors for each gene, we perform a leave-one-out cross-validation omitting each experiment once. Each fold induces $|G|$ models. The prediction performance can therefore be evaluated independently for each gene in each experiment. The background-training set for an (experiment e, gene g)-pair contains all but this experiment using the gene as dependent variable. Once trained, the application of *Padesco* involves only one prediction per gene using the model trained on all known experiments.

4.6.1 Differential Analysis

We apply a differential microarray analysis setup as primary filter step (see Section 4.6.3). Within the compendium of experiments (see Section 4.5) sample groups can be defined.

Each experimental sample groups is compared using the Wilcoxon rank sum statistic [254]. For a given gene, a p-value is computed for each sample group comparison as measure of significance of differential expression. All p-values are corrected for multiple testing using the procedure of Benjamini-Hochberg [17]. We assume genes as being differentially expressed if they exhibit a significance α equal or below 0.01.

4.6.2 Support Vector Regression

Padesco is based on the training of predictive regression models using ν-Support Vector Regression (SVR [228, 240], see Section 4.6.3). Support Vector Machines (SVMs) have acquired general acceptance for microarray applications and have been used for a wide range of tasks including experiment and tissue classification [40, 80]. They have been shown to yield very competitive results for applications in molecular biology [215]. In the following we borrow notations from Smola and Schölkopf, 2003 [229] to provide an overview of SVR training.

SVMs have originally been designed for classification tasks. They build upon the idea that two distinct classes of instances can be discriminated by a separating hyperplane. To achieve an unique and optimal solution the hyperplane spans a maximal margin between the classes. Modifications to the maximum margin hyperplane requirement allow for the adoption to other problems like outlier detection, clustering and regression.

To represent missing data throughout this chapter we define the set of real-valued numbers together with a missing value μ as

$$\mathbb{R}_\mu = \mathbb{R} \cup \{\mu\}. \tag{4.1}$$

All (predictive) SVM and SVR formulations can handle data of the form

$$\mathcal{D} = \{(x_i, y_i) \,|\, i = 1 \dots l\} \subseteq \mathcal{X} \times \mathcal{L}. \tag{4.2}$$

$\mathcal{X}$ is a user-defined input space of instance data x_i and $\mathcal{L}$ is the set of associated labels y_i. In our case, $\mathcal{X} = \mathbb{R}_\mu^d$, *i.e.*, a d-dimensional real-valued vector and $\mathcal{L} = \mathbb{R}_\mu$. A canonical labeling function λ is given by

$$\begin{aligned} \lambda : \mathcal{X} \mapsto \mathcal{L},\, \lambda(x_i) = y_i \\ i = 1 \dots l \end{aligned} \tag{4.3}$$

For the evaluation of model accuracy, it is common to partition $\mathcal{D}$ into a training set $\mathcal{T}$ and an induced test-set $\mathcal{P}$.

Given a dot-product defined in $\mathcal{T}$ the training of an SVR searches for $w \in \mathbb{R}^d$ parameterizing the linear function

$$f(x) = \langle w, x \rangle + b \quad x \in \mathcal{T},\, b \in \mathbb{R}. \tag{4.4}$$

Two observations about f are crucial to enable sensible predictions of the instances in $\mathcal{P}$. First, the norm of w should be minimal and secondly, the function f may not exist for all possible pairs $(x_i, y_i) \in \mathcal{T}$. By introducing slack variables $\xi_i, \xi_i^\star$, these pairs may enter a feasible solution and the following convex optimization problem can be constructed [229]

$$\begin{aligned}&\text{minimize}\,\frac{1}{2}\|w\| + C\sum_{i=1}^{l}(\xi_i + \xi_i^\star)\\ &\text{subject to}\begin{cases} y_i - \langle w, x_i\rangle - b & \leq \epsilon + \xi_i \\ -y_i + \langle w, x_i\rangle + b & \leq \epsilon + \xi_i^\star \end{cases} \quad \forall i = 1 \ldots l\end{aligned} \tag{4.5}$$

We speak of an ϵ-insensitive loss function $|.|_\epsilon$ that penalizes the prediction of labels $f(x)$ deviating from the actual label y as

$$|y - f(x)|_\epsilon = \max\{0, |y - f(x)| - \epsilon\}. \tag{4.6}$$

No penalty is applied for instances that have less than ϵ deviation from their label. By comparing the model predictions of $\mathcal{P}$ with the actually observed labels more accurate estimates of model performance are possible. We exploit this fact for scoring of deviations in *Padesco*.

Equation (4.5) may be reformulated into a dual objective function that avoids a direct estimation of w. Instead, the solution relies solely on dot-products among input instances, whereas only some of the instances, the so-called support vectors, enter the model. The algorithm is expressible in terms of dot-products and allows for the application of kernel functions. Depending on the kernel chosen, the regression model then resembles non-linear solutions as well. The discussion by Smola and Schölkopf [229] provides in-depth details on the use of kernels and solutions to the dual formulation.

An extension to the ϵ-SVR is the so-called ν-SVR, which embeds ϵ itself as variable during training. Instead, the ν parameter controls the number of support vectors and the expected fraction of errors by allowing the ϵ-insensitive tube to have a flexible width. The parameter C controls how strongly deviations from the optimal model are penalized. In our experiments we apply the libSVM implementation of ν-SVR [36].

4.6.3 Model Building

Raw Data. We derive SVR models to predict gene expression fold-changes. A model that predicts a gene's fold-change is trained on fold-changes of the other $(n-1)$ genes. Additionally, to obtain experiment-specificity we skip one experiment at a time. Below, we introduce some notation to clarify our approach.

The raw data matrix R contains data from p individual arrays $A = \{a_1, \ldots, a_p\}$. Each array captures expression values for n genes $G = \{g_1, \ldots, g_n\}$. For p arrays that capture n genes each we obtain a matrix of expression values $(r_{ij}) = R \in \mathbb{R}_{\mu}^{p \times n}$. Let $E = \{e_1, \ldots e_m\}$ the set of m experiments and $S_e = \{s_e^1, \ldots, s_e^{m_e}\}$ the induced sample groups of an experiment $e \in E$. For convenience of notation the experiment that subsumes $a \in A$ is referred to as e^a and its corresponding sample group as s_e^a.

Missing Values. Missing values are frequently encountered in micro-array experiments. Throughout this chapter, missing values are treated implicitly by the application of a linear kernel. Dimensions that contain missing values are consequently skipped and do not contribute to the scalar product. In addition we examined an orthogonal coding scheme [248], zero-imputation and average-imputation. We found that neither did increase the average performance of *Padesco*. Missing data that is encountered as an instance's label is not considered sensible and skipped during further analysis (both training and prediction).

4.6.4 Fold-change Matrix

Fold-changes capture relative effects within an experiment and are less prone to experiment specific bias than raw values. We therefore transform the initial matrix R to a matrix of fold-changes F. Fold-changes are computed among array pairs $a, b \in A$, $a \neq b$ that share the same experiment $e^a = e^b$ but belong to different conditions $s_e^a \neq s_e^b$. We refer to these pairs as sensible. The fold-change of a gene g contrasting array a and b is computed as

$$f_{ab}^g = \begin{cases} \mu & \text{if } (r_{ag} = \mu) \text{ or } (r_{bg} = \mu) \\ log_2(\frac{r_{ag}+c}{r_{bg}+c}) & \text{otherwise.} \end{cases} \tag{4.7}$$

By adding a constant $c \in \mathbb{R}$ fold-changes derived from very low gene expression is buffered and prevents artificially high values. From the space of all sensible array pairs we randomly sample p pairs and obtain $(f_{ij}) = F \in \mathbb{R}_{\mu}^{p \times n}$ for each array pair i and gene j (see Figure 4.2).

Training Experiment Specific Models. Each SVR model is trained to predict fold-changes of a gene g in an experiment e. Next, we illustrate how the training of

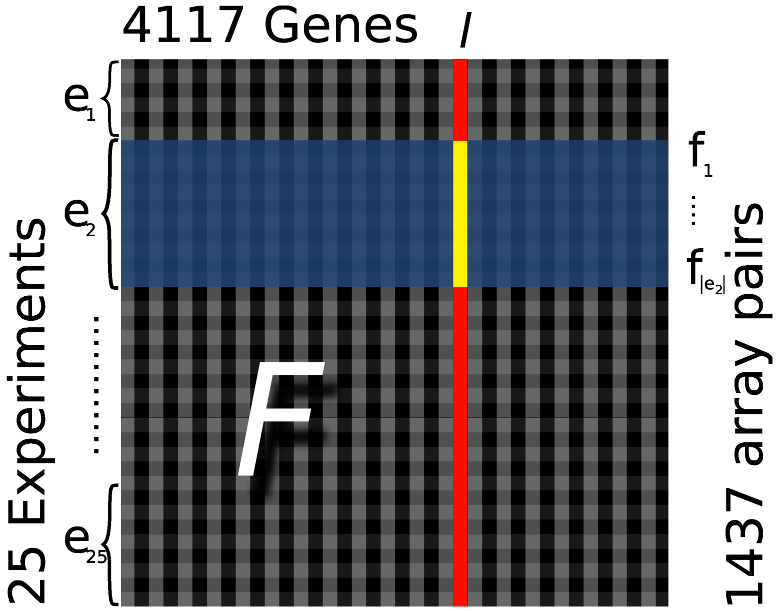


Figure 4.2: **Schematic view of the fold-change (fc) matrix $\mathcal{F}$. rows**: fc vectors of array pairs; **columns**: genes; for an experiment e, we sample $|e|$ pairs of arrays that belong to this experiment but differ in their condition or sample group. **Training**: an SVR models is build to predict a gene l in experiment e (yellow). Training data is given by the fold-changes of other experiments (gray), whereas training labels are the fold-changes of l (red). **Prediction**: fcs in the test-set (blue) are used as features to predict the test labels of gene l (yellow). They result in predicted fcs that are compared to measured fcs (residual scoring). **Deviation**: based on a leave-one-out validation each gene is assigned a background distribution of known deviations. A median absolute deviation based score (*padscore*) is derived to estimate whether a prediction is better or worse than expected.

a single (g, e)-specific model relates to the general SVR training setup as discussed in Section 4.6.2.

Each row of F corresponds to a single instance. Let $\mathbf{f}_t = (f_{tk})$, $k \in G \setminus \{g\}$ the fold-changes of an array pair $t \in A \times A$. Then $(\mathbf{f}_t, f_{tg}) \in \mathcal{X}$ represents an instance with its associated label. The training data $\mathcal{T} \subset \mathcal{X}$ encompasses all array pairs (rows) of F that are not associated with experiment e. The fold-changes for g in all other experiments $E \setminus \{e\}$ are the training labels for $\mathcal{T}$ (see Figure 4.2, red). Note that $\mathcal{T}$ depends on the gene and experiment that are examined. The measured labels for experiment e and gene g are $\mathbf{f}_e^g$ (see Figure 4.2, yellow). We emphasize that the target experiment e is omitted completely during training, *i.e.*, all its sample group comparisons. This setting avoids over-fitting to similar conditions and enables the estimation of an actual prediction performance. Repeating the hold-out for each experiment results in $|E|$ models for each gene thus, $|G| * |E|$ models are trained overall. The factor $|E|$ is relevant for the initial training only. The prediction step is reduced to the evaluation of labels for a single experiment of interest, *i.e.*, the trained models are used to predict experiment-specific fold-change pairs (see Figure 4.3).

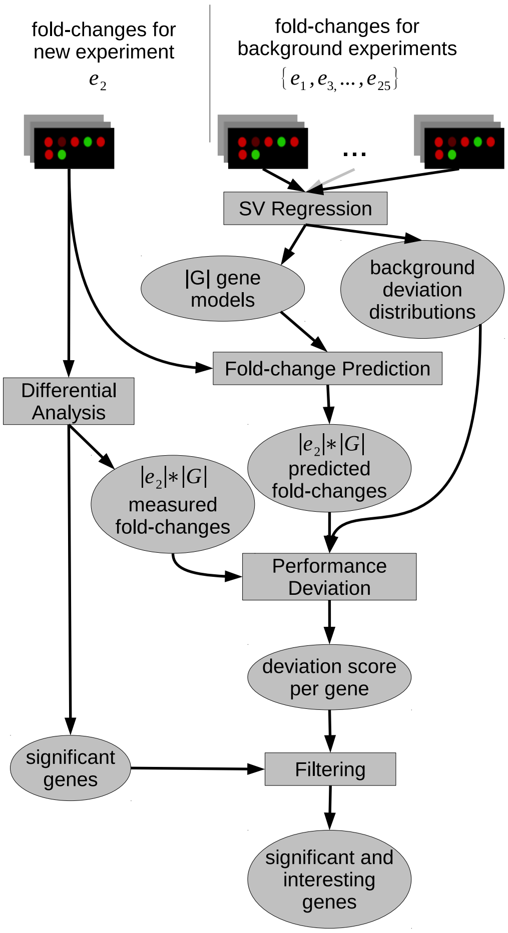

fold-changes for new experiment
e_2
fold-changes for background experiments
$\{e_1, e_3, \dots, e_{25}\}$
...
SV Regression
|G| gene models
background deviation distributions
Fold-change Prediction
Differential Analysis
$|e_2| * |G|$ predicted fold-changes
$|e_2| * |G|$ measured fold-changes
Performance Deviation
deviation score per gene
significant genes
Filtering
significant and interesting genes

Figure 4.3: **Outline of *Padesco*.** A matrix $\mathcal{F}$ (see Figure 4.2) of gene expression fold-changes (fc) is computed. A LOOCV is performed for each gene and omitting each experiment once resulting in $|G| * |E|$ models. For a new experiment e, the fcs of g are predicted using the model for (g,e) and are then compared to the measured fcs (*residual scoring*). Based on the LOOCV a deviation score (*padscore*) is derived. It enables deviation filtering of significant genes.

4.6.5 Selection of Parameters during Training

We examined a linear kernel as well as the radial basis kernel. The final parameters have been chosen as $(C, \nu) = (100, 0.2)$ using an exhaustive grid search combined with leave-one-experiment out (LOO) for hyper-parameters $C \in \{10^i | i = -2, -1, \dots, 4\}$ and $\nu \in \{0.2, \dots, 0.8\}$. On average the linear kernel provides similar performance to the radial basis kernel (screening $\gamma \in \{10^i | i = -4, -3, \dots, 1\}$) and a stable prediction performance across a wide range of parameters.

4.6.6 Scoring Performance Deviation

We now assume that the experiment e has not been observed beforehand. We like to predict a gene of interest g. First, we calculate the fold-changes for array pairs in this experiment. This corresponds to a sub-matrix of $\mathcal{F}$ (see Figure 4.2; sub-matrix: blue, corresponding labels: yellow). We are especially interested in the actual fold-change values $\mathbf{f}_e^g$ and use the (g, e)-specific model (unaware of e) to obtain $\mathbf{f}_e'^g$, the predicted vector of an experiment's fold-changes.

To obtain a measure of correspondence between $\mathbf{f}_e^g$ and $\mathbf{f}_e'^g$, we compute an un-centered Pearson correlation $\rho_{g,e}$. Given the two n-dimensional vectors $\mathbf{f}_e^g = (f_{e,i}^g)$, $i = 1 \dots n$ and $\mathbf{f}_e'^g = (f_{e,i}'^g)$, $i = 1 \dots n$ it is calculated as

$$\begin{aligned} \rho_{g,e} = & \quad \frac{\mathbf{f}_e^g \cdot \mathbf{f}_e'^g}{\|\mathbf{f}_e^g\| \cdot \|\mathbf{f}_e'^g\|} \quad = \\ = & \quad \frac{\left(\sum_{i=1}^n f_{e,i}^g \cdot f_{e,i}'^g\right)}{\sqrt{\sum_{i=1}^n \left(f_{e,i}^g\right)^2 \sum_{i=1}^n \left(f_{e,i}'^g\right)^2}}. \end{aligned} \tag{4.8}$$

We also compute a discretized version of $\mathbf{f}_e^g$. We set all fold-changes above a threshold t_f to 1 and all below to 0. We compute an Area Under Curve (AUC) value by varying t_f for $\mathbf{f}_e'^g$. The threshold for $\mathbf{f}^g$ is fixed at 2.

4.6.7 Measure of Expectation

Given and AUC and a measure of correlation we have an estimate to tell how well a gene performs in a single experiment. Yet, we cannot say whether this is more or less than we expected. To arrive at a deviation of known patterns we compute the empirical distribution D^g from $(\rho_{g,x})$, $x \in E$. The deviation for an experiment e is expressed in units of median absolute deviations (*padscore*) with respect to this distribution. Given the median med^g of D^g and the corresponding median absolute deviation mad^g the *padscore* is given by:

$$(med^g - \rho_{g,e})/mad^g \tag{4.9}$$

In order to be retained by *Padesco* genes must simultaneously satisfy the significance level of differential expression in a specific sample group given by a certain α-level as well as a minimum *padscore*. These genes are referred to as *specific*. Remaining *unspecific* genes are significantly regulated, but could be predicted by the SVR models.

4.6.8 Permutation Test for the Evaluation of *Padesco*

Gold standards on the experiment-specific expression of genes are not available. We use a permutation test to simulate genes deviating from their common behavior. By copying (*spiking*) the expression values from a significant gene g^+ to an insignificant gene g^- within an experiment e we force genes to violate their common behavior and trivially become significant. We have two choices in parameters here. First, we can choose a *z-score* level t_z for significance. Second, we can choose a *padscore* level t_p to select interesting genes. The following permutation test selects these thresholds and estimates the associated performance for spike-in controls. We sample from the significant genes, and spike into the insignificant genes. We then recompute the *padscore* for the previously insignificant gene. This process is repeated s times where s is the number of significant genes. $SPIKE$ denotes the set of spiked genes. After all repeats have been computed we obtain sensitivity=$tp/(tp + fn)$ and precision=$tp/(tp + fp)$ and repeat the evaluation for all possible thresholds t_z and t_p in the experiment. A gene g's recomputed *padscore* p determines whether it is *tp*, fn or fp based on these thresholds (see Table 4.1).

Type	Abbreviation	Condition
true positive	tp	$g \in SPIKE \wedge p \geq t_p$
false positive	fp	$g \notin SPIKE \wedge p \geq t_p$
false negative	fn	$g \in SPIKE \wedge p < t_p$

Table 4.1: **Classification assignment for the evaluation**. A threshold on *z-score* (t_z) and *padscore* (t_p) is chosen. After spiking (see Section 4.6.8) a gene g's recomputed *padscore* p determines its type.

4.7 Results

4.7.1 Evaluation of Expression Fold-Change Predictions

We use the uncentered correlation to measure how well the regression models can predict expression fold-changes of gene g in experiment e. The prediction

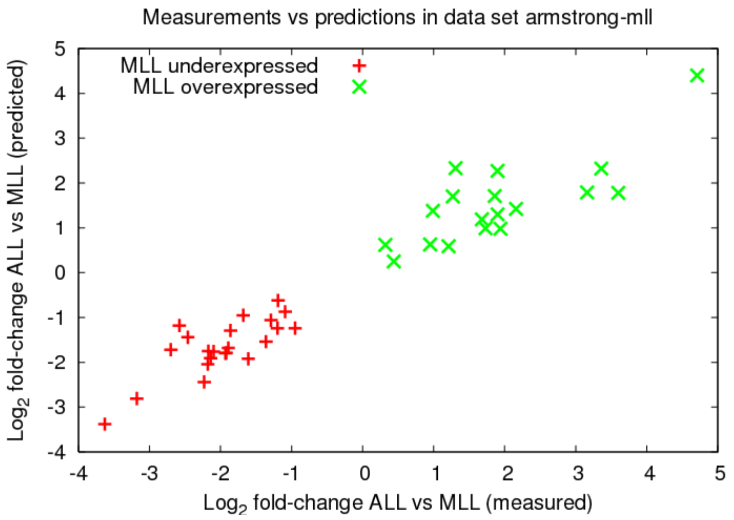


Figure 4.4: **Scatter plot of predicted *vs.* measured fold-changes.** Microarray studies frequently derive gene signatures, *i.e.*, sets of differentially expressed genes discriminating between experimental conditions. Gene signatures can be predicted well by our SVR models, as shown here for a gene signature distinguishing ALL and MLL leukemic genotypes published along with the data set of armstrong-mll [7]. For every gene, our SVR models predict expression changes between Acute Lymphocytic Leukemia (ALL) and Mixed Lineage Leukemia (MLL) correctly, although the precise values of the measured fold-changes are not reproduced exactly. A gene is depicted as a single point that corresponds to the average of all fold-changes of this gene across array-pairs comparing the conditions ALL and MLL (see Section 4.6.3).

performance is significantly better than random for the majority of genes: the uncentered correlation achieved an average value of $\rho_u = 0.7$ in our experiment. 91% of the predictions exhibit a $\rho_u > 0$. After discretization we achieve an AUC of 81% on 15.7% cases with a fold-change of more than two. We argue that the remaining specific candidates that cannot be predicted well are particularly interesting because they exhibit an experiment-specific expression that could not be learned from the training data.

4.7.2 Prediction of Fold-changes for Individual Genes

Gene signatures, *i.e.*, sets of genes that are differentially regulated between different cellular states are frequently published along with microarray experiments. Such signatures are expected to yield diagnostic markers that could help to differentiate between healthy and sick individuals. Here, we examine a gene signature that has been compiled by [7] to distinguish a particular chromosomal translocation involving the MLL (mixed-lineage leukemia) gene from the regular ALL (acute lymphoblastic leukemia) genotype. The MLL translocation is significant as it frequently leads to an early relapse after chemotherapy. In Figure 4.4 we compare our predictions against the experimentally determined expression fold-changes for this gene signature. For all genes, the direction of differential expression can be correctly derived from our predictions, although the values of the measured fold-changes are not reproduced exactly. Similar results have been obtained for other published signatures.

4.7.3 Permutation Test Based Evaluation

Padesco filters genes based on a standard differential expression *z-score* (Wilcoxon test, see Section 4.6.1), and a novel *padscore* (see Equation 4.1) indicating experiment-specific expression. This second score indicates whether genes can be predicted less well than expected from the training (background) set of experiments. They are selected due to their *padscore* since they do not conform to their trained patterns. Cutoffs on the two scores are required for the selection of specific candidate genes that exhibit differential as well as experiment specific expression. The permutation test introduced in Section 4.6.8 generates artificial pattern deviations through spiked genes. As shown in figure 4.5 at a *padscore* cutoff of 2.0 specific candidates are accurately detected (85% precision). The precision increases for differentially expressed genes ($z > 3$). Based on the evaluation we picked a *padscore* threshold of 2.0 and a *z-score* threshold of 3.0 to receive a moderate number of candidates exhibiting a high precision. Thereby we selected some 250 specific candidates.

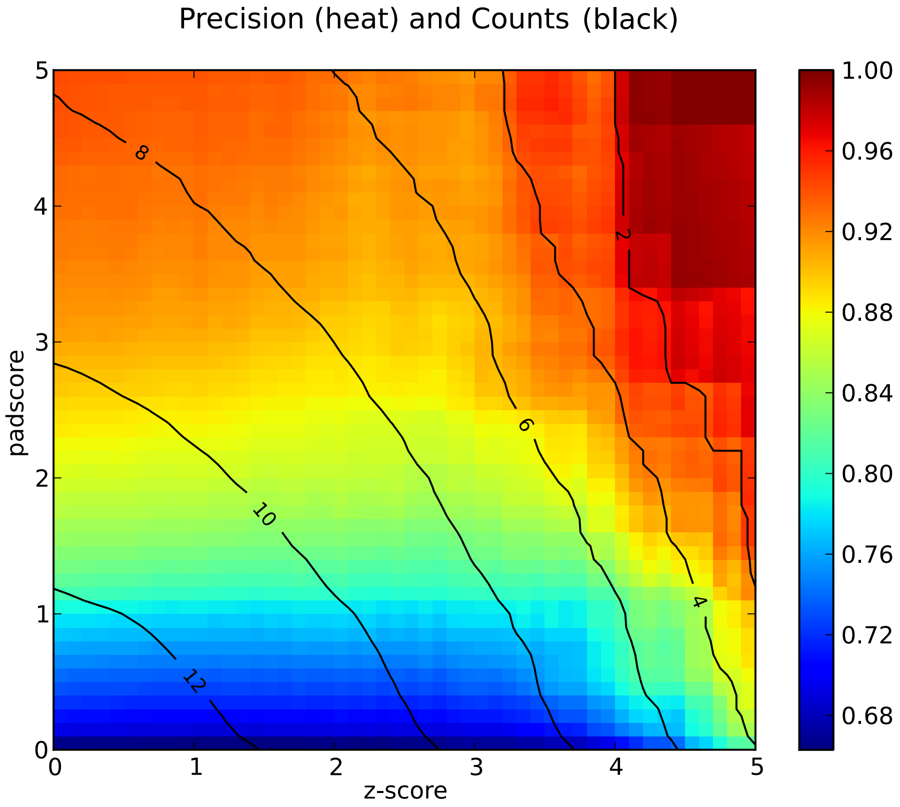


Figure 4.5: Precision (percent, colormap) and the number of detected genes (log base 2, contour) as a function of padscore *padscore* and z-score. As gold standards are not available, we estimate the performance of specific candidate gene detection by a permutation test. This test evaluates how well known spike-in controls can be recovered by *Padesco* for arbitrary z-score (differential expression) and *padscore* (experiment specific expression) thresholds. Specific candidate genes can be reliably identified (85% precision) using a *padscore* above 1.5 even if they exhibit only moderate levels of differential expression (*z-score* < 4). By combined *z-score* and *padscore* thresholds candidate gene lists can efficiently be reduced for follow-up studies. Analyzed here are 59 condition comparisons from 13 experiments. At the chosen *padscore* (2.0) and z-score (3.0) thresholds, some 250 specific candidates (contour) are detected by *Padesco*.

4.7.4 Specific Candidates

In this section we discuss sample results in two experiments that examine prostate cancer [225, 252] and one study on leukemia [7]. As a further example we provide results from a *Toxoplasma gondii* infection study by Chaussabel *et al.* [38]. *Padesco* does not aim to whole relevant pathways but allows to focus on a small subset of interesting genes for further analysis. Vogelstein and Kinzler [246] discuss key pathways which are likely to be disturbed to promote cancer in almost any cell type. Development of cancer is strongly coupled to the perturbation of one or more such pathways. The interleukin 2 pathway has been shown to be deregulated in many cancer types and was also described to be involved in prostate cancer. IL2RB dimerization with the α-subunit leads to a higher affinity towards interleukin-2. IL-2 treatment was previously shown to lead to reduced prostate tumor growth in rats [115, 169]. Another cancer therapy using IL-2 has been developed by Otter *et al.* [185].
Eicosanoids are known to interact with immune messengers like interleukins. In the first experiment by Welsh *et al.* [252] tumor samples were compared against normal and HUVEC (Human Umbilical Vein Endothelial Cells) samples, where 27 genes have been detected as differentially expressed. For an initial screening of functional significance we apply a gene ontology over-representation analysis (DAVID, [53, 125]) on the differentially expressed genes, *i.e.*, without *padscore* filter. As for [252] a screening for significance (Benjamini-Hochberg corrected scores, $\alpha = 0.01$) shows no significant enrichment, yet 3 genes (IPR, EPR3R and CYT450J) are found to belong to the eicosanoid metabolism (p=0.08). With *padscore* filter Interleukin 2 receptor β (IL2RB) is the only gene found to be interesting in this experiment. *Padesco* reported an unusual expression of IL2RB in the tumor samples, which could explain the decoupling of eicosanoid pathway members from IL2RB regulation.

The second examined experiment on prostate cancer is described in Singh *et al.* [225]. Here, 60 differentially expressed genes were identified. DAVID analysis shows no significant over-representation. After *Padesco*-filtering, 4 genes remain that we discuss in the following. HCK (*padscore* = 7.2), an src related tyrosin kinase is most interesting in terms of the *padscore*. Smith *et al.* [227] describe its association with gpl130 and the formation of a complex with IL-6R which promotes high affinity binding of IL-6. In prostate cancer, IL-6 is a key protein. It has been suggested to contribute to prostate cancer progression towards an androgen-independent state. We observe MGC17330 (PIK3IP1, *padscore* = 2.6) to be the second *padscore* relevant gene. It is a negative regulator of PI3K. Src kinases are upstream mediators for the PI3K signaling pathway with important roles in proliferation, migration and survival. It is described to be a tumor suppressor in heptacellular carcinomas [71]. It shows only a weak positive fold-change in this experiment which may explain why it fails as a suppressor here. Mutations within the PI3K pathway have been described by Vogelstein and Kinzler [246] to be involved in a number of tumor types. The third gene found is ATP2B1 (PMCA1, *padscore* $= 2.6$). It is a Ca^{2+} ATPase subunit. An unusual reduction in gene expression can be observed in our data. This reduction is also discussed by Roderick and Cook [209]. Ca^{2+} pumps are likely to provide good therapeutic targets for anticancer drug development as suggested by Monteith *et al.* [168] and Roderick and Cook [209]. They emphasize the role of Ca^{2+} (intracellular calcium) in the life-and-death decisions of the cell such that disturbed control of Ca^{2+} may lead to an inappropriate cell fate. The fourth candidate – C2orf3 (*padscore* $= 2.03$) – has also been identified by an approach by Hong *et al.* [122]. They analysed three prostate cancer sets but since more than a hundred genes are identified and C2orf3 is no top-ranking gene this candidate has not been subject to further discussion. The transcription repressor binds GC-rich sequences of the epidermal growth factor receptor, beta-actin and calcium-dependent protease promoters.

Slightly below the *padscore* threshold STK38 (*padscore* $= 1.84$) has been described to exhibit cancer specific alternative splicing variants [148]. In Rozanov *et al.* [211] it is suggested as a part of the downstream network of MT1-MMP, a key regulator linked to tumorgenesis and metastasis. Similar, CPD (*padscore* $= 1.47$) is a metallo carboxypeptidase family enzyme. It is described to have shown lower levels of gene expression in colon carcinomas. We observe a similar downregulation in our data. The well-known PSMA (or PSA) also exposes carboxypeptidase activity associated with increased invasiveness of prostate cancer [86]. Similarily, GTF2B/TFIIB (*padscore* $= 1.46$), the general transcription factor 2B plays a major role in the transcription of eukaryotic genes. Minucci and Pelicci [166] suggested Histone deacetylases (HDACs) as promising targets in cancer therapy, partly due to their DNA binding capabilities. GTF2B/TFIIB exhibits auto-acetyltransferase function regulated by acetylation while acetylation also impairs activities of enzymes involved in DNA metabolism and repair. DNA repair is a key mechanism which has to be bypassed to allow for tumor development due to Vogelstein and Kinzler [246]. With *padscore* of 1.11 KLF6 is predicted only slightly worse than usually, yet with a much smaller offset than HCK. The Krueppel-like factor 6 is a well known tumor suppressor gene. On the other end it is necessary to track for the negative end of the results *i.e.*, the genes which are predicted better than usual. As an example branched amino acid transferase BCAT2 (*padscore* $= -3.30$) is among the differentially expressed genes yet could be predicted *better than expected* in this experiment. To our knowledge the enzyme has not been described in the context of prostate cancer. The third data set [7] compares Acute Lymphoblastic Leukemia (ALL) to Myeloid Lineage Leukemia (MLL). We can recover most specific genes originally published (some are filtered due to our criteria among all integrated experiments). An interesting exeption is LGALS1 which shows considerable expression in both the MLL and the AML samples, but not in ALL. Using a clustering procedure Armstrong *et al.* [7] describe it to be MLL specific.

Our top candidate here does not meet the *padscore* = 2 requirement yet RhoA (*padscore* = 1.99) is involved in cytokinesis and has been described as both pathogen target as well as regulator of oncogenensis [32]. Ordonez-Moran *et al.* [184] associates RhoA with colon carcinoma (which is not part of our background set). A second gene, ETS1 (*padscore* = 1.86) belongs to the familiy of ETS transcription factors and has been shown to regulate telomerase activity at gene transcription level. It constitutes a well known oncogene with potential effects on telomer stability. In Yeoh *et al.* [260] BCR-ABL samples 2 genes exhibit an extreme *padscore* of 4.8 and 5.9: PHLDA2 and PSMA6. While PSMA6 has recently been suggested as a new prognostic marker in acute monocytic leukemia [39], PHLDA2 (pleckstrin homolog-like domain family member 2) is located in a region considered to be an important tumor supressor gene region. The top gene among the identified candidates in the prostate cancer subtype of the Butte *et al.*dataset [33] is SLC30A3 (*padscore* = 6.81), a zinc transfer transporter from the solute carrier family (Prostate tissue in general shows an increased zinc content - around 10-fold higher than other tissues). SLC30A3 has been suggested to be a member of the apoptotic pathway by Ackland *et al.* [4].

Commonly SLC30A3 lowers intracellular zinc concentrations by mediating zinc efflux. Mouse TRAMP models suggest that both too high and too low zinc uptake have drastic effects on prostate tumor sizes, consequentially this candidate's dysregulation could be crucial for tumor tissue. Prasad *et al.* [197] argue that an optimal zinc level is crucial as a protective instance against cancer development.

DOOST (*padscore* = 8.53) is involved in the metabolism of glycoproteins and has been suggested to mediate processes associated with cell-adhesion or invasion [112] and has been frequently described in the context of gastric cancer which was not yet included in our background set of experiments.

Chaussabel *et al.* [38] analyze diverse parasite infections on human macrophages and dendritic cells. We find CCR1 to be the most prominent gene (*padscore* = 9.29) in the Toxoplasma infection subgroup. Mice lacking this chemokine receptor CCR1 have shown dramatically increased mortality after Toxomplasma gondii infection [136].

4.8 Discussion

Genes are not regulated individually [149, 238]. Frequently, patterns of co- or anti-regulation can be observed such that the up-regulation of a gene A is a good hint that another gene B will also be up-regulated while a third gene C

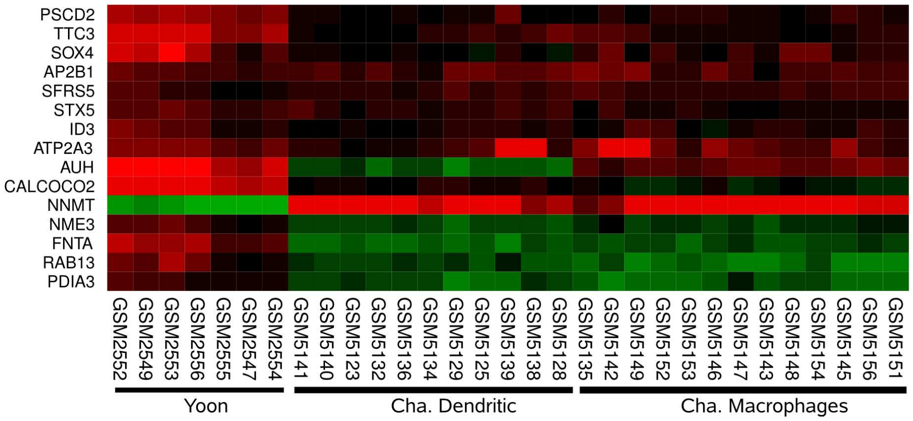


Figure 4.6: **Common Regulation and Experiment Deviation.** The heat map shows a cluster of genes which are usually correlated (pairwise Pearson's Correlation above 0.8, 'yoon-p53 (Yoon)' [262] is given as reference). Here, AUH with a *padscore* of 5.09 is detected as interesting in 'chaussabel-parasite (Cha)' [38].

will rather be down-regulated. The disruption of such patterns pinpoints genes with experiment-specific expressions. We call such genes *specific* candidates in contrast to the remaining *unspecific* candidates that exhibit only generic expression patterns. *Padesco* detects *specific* candidates by analyzing fold-change based co-expression patterns with Support Vector Regression models trained on a background set of microarray experiments. After training, we select *specific* candidate genes via a two stage filter. The first filter step is a routine analysis of differential expression (significant genes). A novel second filter selects genes that show deviations from generic expression patterns predictable by linear models (interesting genes).

In order to avoid the predictions of false *specific* candidates *Padesco* depends on a good prediction performance of the underlying SVR models. The prediction performance can be evaluated rigorously as the prediction target experiment is excluded from training in a leave one out cross-validation setting where all conditions of particular experiments are left out. This not only excludes condition specific but also experiment specific biases. We examined 4,117 genes across 25 experiments consisting of 1,437 individual microarrays. Predictions by *Padesco* are better than expected by chance in 91% of the cases. Segal *et al.* criticized that gene signatures rarely help to identify the involved biological processes or the causal regulatory mechanisms. Hirsch *et al.* [119] further argued that a gene signatures frequently do not represent specific attributes of the measured biological conditions. We analyzed gene signatures published together with the corresponding microarray experiments. These signatures were selected by the authors of the corresponding studies to discriminate between experimental conditions (sample groups). We found that expression changes for genes in signatures are predicted well by our SVR models trained on other, unrelated experiments. An example is the signature distinguishing ALL and MLL [7]. Although the ALL/AML signature certainly provides discriminating marker genes, it does not capture experiment specific expression patterns according to *Padesco*.

The extent of differential expression alone does not indicate experiment *specific* involvement of genes. Based on the prediction performance we identified specific candidates genes that exhibit experiment *specific* expression, *i.e.*, expression changes that cannot be explained (predicted) by our models. This analysis is related to co-expression studies and complements differential expression analysis. It enables to focus on concise candidate lists for follow-up studies that consist of experiment-specific candidates only. We screened for filter thresholds and estimated *Padesco*'s performance from permutation tests as comprehensive gold standards for the experiment specific expression of genes are not available. This newly devised simulation approach suggests that *specific* candidates are identified reliably by *Padesco* ($> 85\%$ precision at *padscore* > 1.5) even if they show only marginal levels of differential expression. On the other hand, more than 90% of the genes selected by differential expression alone exhibit only generic expression patterns and could be excluded from further studies. *Specific* candidates are likely to represent characteristic features of the corresponding experimental conditions.

We evaluated *Padesco* selected genes for two data sets on prostate cancer. Besides interesting new candidates, we found several genes with a known involvement in the disease. Some of them, such as IL-2RB, have already been reported as promising drug targets. We demonstrated that such examples are more difficult to detect by differential expression analysis alone. Instead, differential expression tends to pick up genes that act similarly in other, biologically unrelated experiments. Thus, in combination with differential expression analysis, *Padesco* is a promising protocol for the detection and analysis of particularly distinctive features of microarray experiments.

Chapter 5

Reduction of Network Bias via Confidence Recalibration

Background

This chapter is based on Petri *et al.* [192]. Supervised network reconstruction has been around for a couple of years. When we successfully participated in the DREAM3 and DREAM4 challenges [144, 143], we observed that the results for a prediction of eukaryotic networks were not satisfactory. Tackling these problems would eventually boost the existing network inference performance [192]. Valuable input came from Ludwig Geistlinger, who by then had annotated the regulatory network of the diauxic shift in yeast and provided us with more detailed annotations to our result clusters and further valuable input data.

Contributions

Tobias Petri (TP), Stefan Altmann (SA), Ludwig Geistlinger (LG) and Robert Küffner (RK) compiled the data and conducted experiments. TP and RK developed the correction method, evaluation routines, functional coherence scoring and the 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, RK and LG wrote the Bioinformatics paper with suggestions from Ralf Zimmer (RZ). RK and RZ supervised the project.

5.1 Genome-scale Models

Gene Regulatory Networks (GRN) provide an important means to capture the interplay among regulatory factors like proteins and their genetic target regions. Currently, our knowledge of these networks is limited. In particular, large-scale models are restricted to topological features. Neither the change of reactions over time nor reaction contexts are modeled. These features would be tremendously useful to obtain a more accurate representation of the interplay among proteins, genes, and other factors. Yet, even large-scale data cannot provide these models with the necessary level of detail to obtain a sensible parameterization. Notably, the amount of data required to accurately capture as few as 10 entities is demanding [144] from both data availability and computational aspects. For larger systems, detailed modeling in terms of simulations grows infeasible. Most approaches would then resort to measures of correlation to estimate interactions [143].

Similarly, the experimental setup may restrict the set of observable interactions. Other limitations, such as the availability of specific antibodies, affect context specificity as well. The timing of an experiment is crucial, as the majority of regulations is transient rather than constitutively active. For these reasons, most large-scale models capture key aspects of a regulatory system, simplified, and independent of specific contexts. Nonetheless they provide helpful abstractions and provide elemental building blocks for further, more detailed modeling and prediction tasks.

In-silico predictions may complement existing GRNs and integrate a wider range of experimental conditions. The use of large-scale expression compendia may enable accurate predictions of novel regulations and complement other sources of information.

However, expression data alone is often not sufficient to infer interactions in eukaryotes such as yeast [143, 160]. Supervised inference methods have been proposed to support the inference process by known interactions in addition to expression data. We find that methods exploiting known targets show an unexpectedly high rate of false discoveries. Many interactions suggested with a high confidence are random.

In terms of network-wide prediction quality, naïve baseline methods (see Section 5.5.3, random target assignments) seem to be on par even with recent, sophisticated methods. This result may come as a surprise, but it can be explained by a key property of GRNs: their topology.

We show that the origin of the observed discrepancy is the assumption that individual regulator predictions can be integrated to obtain a complete network. Yet, network topology strongly impacts the regulator-wise properties of predicted confidences and it must be accounted for. Otherwise, these networks suffer from what we term *High Degree Preference* (*HDP*).

Larger regulators or hubs acquire more predicted interactions and they can be predicted with higher confidences in general. Unfortunately, these issues are hidden from the most frequently used evaluation and cross-validation setups. This leads to overly optimistic performance estimates for existing approaches. In this regard, supervised network inference resembles the well-known Simpson's Paradox as explained in Section 5.6.2.

To tackle this problem, we devise a corrective procedure (Confidence Score Recalibration, *CoRe*, Section 5.5.6) to obtain globally consistent results. We then benchmark supervised and other reference approaches and at the example of yeast we show that a reliable inference of interactions in eukaryotes is feasible. After recalibration, the detected interactions exhibit a better functional consistency and are capable of explaining the formation of expression patterns across many biological processes.

As similar inference techniques are also employed for function prediction, knowledge transfer or hypothesis generation, our recalibration is likely extensible to other predictions that feature network-based predictions as well.

CoRe considerably improves the results of network inference methods that exploit known targets. Predictions then display the biological process specificity of regulators more accurately and enable the inference of genome-wide regulatory networks in eukaryotes. For yeast, we propose a network with more than 22,000 confident interactions.

5.2 The Need for Computational Approaches

Regulators such as transcription factors physically bind to specific nucleotide sequences to regulate the expression of target genes. The binding sites of regulators have been determined by experimental protocols such as Chromatin Immuno-Precipitation (ChIP [263]) or deoxyribonuclease (DNase) footprinting [179]. Similarly, binding studies helped elucidating network architecture in the ENCODE project [84], but they also report interactions that are not associated with changes in target expression [258]. By contrast, the expression profiling of TF knockout mutants [42] detects interactions that exhibit changes in target expression, but this technique is prone to indirect or spurious effects [124]. Therefore, binding studies and expression data should be analyzed in combination as both types of experiments complement each other.

Although the number of conducted TF-binding and TF-knockout studies is growing [196] the discovery of novel regulations detected with each additional study decreases. A combination of experimental results and computational inference approaches is likely to provide more comprehensive networks.

The Yeastract database [3] compiles yeast interactions from two types of experiments. The first type detects physical binding of TFs to promoter regions of target genes. The second one tests whether a perturbation (knockout, silencing, activation) of a TF leads to changes in the expression of putative targets. We speak of *active* interactions if they are observed in both types of studies, *i.e.*, if the TF both binds to and effects transcriptional changes in a corresponding target gene. Just 9% of all interactions detected by binding studies are confirmed (Figure 5.1a). It is important to note that interactions are unevenly distributed among the TFs (TF out-degrees = number of known targets per TF, Figure 5.1b).

To demonstrate that network inference is necessary, we estimated the size of the complete yeast regulatory network (see Section 5.5.2). We treated the number of interactions as a function of the available binding studies and found that a hypothetically complete network would contain 3.5 times the number of interactions in Yeastract ($3.5 * 29398 \approx 105000$ interactions) given an estimated limit value of binding studies (Figure 5.1c).

Based on this estimation, we reason that 2.5 times the number of currently available binding studies would be required to obtain half of the completed network ($2.5 * 356 \approx 900$ studies). Furthermore, our results suggest that 50% of all "active" interactions are currently known. However, the low confirmation rate of 9% impedes their identification and separation from the inactive ones. Inference methods are potentially able to close that gap.

5.3 Integration of Known Topologies

Many inference methods use expression data exclusively. An interaction is predicted if a TF and its putative target are coexpressed. Such *expression-based* approaches were successfully applied to infer prokaryotic networks [161, 73, 165, 101]. For eukaryotes, useful results have been achieved for restricted gene sets like respiratory genes, yet, they perform hardly better than random in general [124, 165, 257, 143, 160, 231, 176]. Interactions in eukaryotes are difficult to infer as observable dependencies between the expression of regulator and target are weaker and context-dependent [257, 143, 160]. One reason is the increased level of complexity and the combinatorial nature of the eukaryotic regulation of transcription [179].

In this context, *a priori* known interactions are referred to as *topological priors*. The integration of such prior information yielded promising results previously [171, 101]. For the prediction of novel targets of a specific TF the restriction to a *local* context for this factor is crucial. Otherwise the resulting model may be too unspecific, resulting in a performance drop. TF specific signals detected within the expression patterns of known targets may enable more reliable predictions. Such *supervised* methods are now widely employed. Their use is not limited

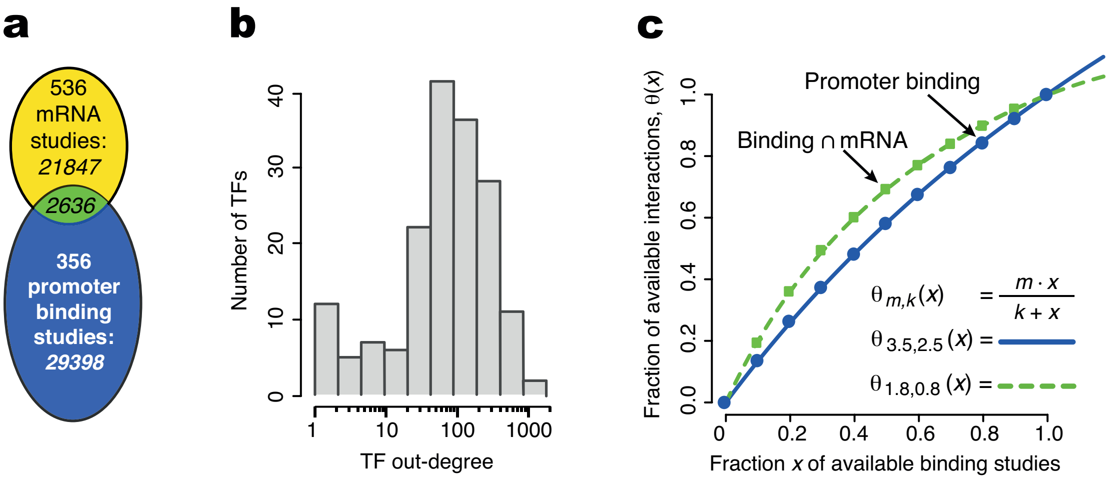


Figure 5.1: **Properties of yeast interactions. (a)** The Venn diagram depicts the number of interactions (italics) in Yeastract obtained from 536 mRNA expression studies (yellow), 356 promoter binding studies (blue), or both (green). **(b)** shows the distribution of TF out-degrees in Yeastract binding studies. **(c)** plots the fraction of interactions contained in random subsets of binding studies as a function of subset size ($x = 1.0 \mathrel{\hat{=}} 356$ studies). Fractions are plotted for interactions from binding studies (blue circles, ordinate: $1.0 \mathrel{\hat{=}} 29398$ interactions) and for interactions detected in both study types (green squares, $1.0 \mathrel{\hat{=}} 2636$ interactions). We fit first order Hill functions $\theta(x)$, shown as lines, to estimate the ratio of expected to known interactions m. Thus, an infinite number of promoter binding studies (blue line) would detect $m = 3.5$ times the currently known 29,398 interactions. The second parameter k indicates that $k = 2.5$ times the currently available 356 studies are required for detecting half the expected interactions.

to the detection of gene regulatory interactions [200, 171, 51], but they are used for the prediction of protein-protein interactions [259, 245], drug-target interactions [23], and gene functions [21] as well. Furthermore, subsets of active TFs may be determined by the expression of known targets [175, 44], and increase the reliability of predicted targets [200, 171, 51]. *Integrative* methods incorporate further types of information such as TF binding sites (SEREND, [68]) or chromatin profiles [69]. Here, we investigate whether eukaryotic networks are accurately inferred by methods exploiting topology priors. We find that the composition of most inferred networks is skewed towards hubs and that most existing network evaluations cannot detect this effect.

They do not adequately integrate (local) topology aspects and, thereby, overestimate the overall network quality substantially. This effect resembles Simpson's Paradox, well-known in statistics, and causal theory in particular [224, 191]. We develop a conceptually simple recalibration strategy and demonstrate how it can be applied for the inference of a confident genome-scale regulatory network in yeast.

5.4 Network Inference Schemes

In this section we provide a schematic overview and classification of existing approaches to network inference and introduce necessary terminology.

5.4.1 Basic setup

A regulatory network is modeled as a directed graph $N = (G, I)$, where the vertice set G is the set of all genes, and the set of edges I represents regulator-gene interactions. The set of genes that regulate other genes is $R \subseteq G$. The set of target genes is $T \subseteq G$. In general, regulators may be targeted by regulators as well, leading to more complex network patterns, *i.e.*, motifs (see Section 5.5.7). We denote known targets of r within the network N as $T_N(r) \subseteq G$. The vertex out-degree of r in a network N is $|r|_N^{out}$, the in-degree is $|r|_N^{in}$.

A pair $(r, t) \in R \times G$ is called an *experimentally supported* regulation if there exist TF-binding studies that detected it. Otherwise (r, t) is a potential regulation. To express the degree of experimental support (r, t) is associated to a weight $w_{rt} \in \mathbb{N}$. It is chosen as the number of TF-binding studies that confirm (r, t). The matrix of all weights is $W \in \mathbb{R}^{|R|\times|G|}$. Some algorithms require an induced binary label, which is given by:

$$l_{rt} = \min(w_{rt}, 1) \tag{5.1}$$

The majority of approaches relies on expression data. We define an experiment as an array of fold-changes for all genes. Thus, the set of experiments E consists of p array pairs. Each experiment provides fold-changes for a specific experimental condition or replicate. For p experiments and n genes we obtain a matrix of fold-changes:

$$M = (m_{eg}) \in \mathbb{R}^{p \times n} \tag{5.2}$$

Submatrices of M are written as $M_{E'}^{G'}$, with experiment rows $E' \subseteq E$ and gene columns $G' \subseteq G$, respectively. We denote the row vector that corresponds to e by $\mathbf{m}_e$ and the column vector associated to g by $\mathbf{m}^g$.

Most approaches estimate real-valued confidences for both supported and potential regulations. For all pairs $(r, t) \in R \times G$ this results in a matrix of confidences

$$\hat{C} = (\hat{c}_{rt}) \in \mathbb{R}^{|R| \times |G|} \tag{5.3}$$

Ranking the pairwise confidences may then reveal promising candidates for novel regulations or identify regulations that are not sufficiently supported by expression data.

5.4.2 Expression-Based Approaches

The class of expression-based approaches relies on expression data omitting network topology. Most of them are *unsupervised* and often resemble *lazy* learners. The majority of expression-based approaches either relies on some predefined *information theoretic* measure of dependency or the extraction of estimates from *linear model* fits.

Apart from topology information, additional annotation A may be available. For an experiment $e \in E$ the corresponding annotation a_e represents a complex annotation type that may include information on knocked out genes, the particular siRNA applied, or genes being over-expressed. This may help to reduce otherwise unexplained variation. If these pieces of information are used, we speak of approaches that are *annotation-aware*. They have shown promising results [143, 101, 116]. These models assume that experiment perturbations can be traced within the data and attributed to network structures. In practice though, and for expression data in particular, observable effects of individual perturbations may be weak and dispersed across the network. Furthermore, the annotation of perturbations and conditions must be available for all experiments and structured uniformly. Despite of standards like MIAME [25], few existing resources like the Many Microbe Microarray database [72] provide the necessary level of detail. Yet, without substantial and ongoing effort such resources quickly become obsolete.

Information-theoretic dependency. The branch of information-theoretic approaches estimates regulator-target confidences $\hat{c}_{rt}$ by comparing the expression data of a regulator r and a potential target t across experiments. These methods rely on some predefined dependency function:

$$\begin{aligned} d &: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}, \\ d\left(\mathbf{m}^r, \mathbf{m}^t\right) &= \hat{c}_{rt} \end{aligned} \tag{5.4}$$

Common choices for d are correlation and mutual information based measures [5, 73, 161]. The result matrix $\hat{C}$ is post-processed and analyzed to obtain the most likely candidates for regulations.

A well-known problem affecting these approaches is that the estimated confidences are symmetrical. Several extensions have been introduced to estimate the network of immediate effects effects [16, 75, 50].

Linear models. Unlike information-theoretic approaches linear modeling approaches [101, 104, 116, 222] do not rely on a given dependency function. Instead, they model the expression levels $\mathbf{m}^t$ of a potential target as a (linear) function d_t of other fold-changes using an influence vector β^t:

$$\begin{aligned} d_t &: \mathbb{R}^{p\times(n-1)} \to \mathbb{R}^p, \\ d_t\left(M^{G\setminus\{t\}}\right) &= M^{G\setminus\{t\}}\beta^t = \mathbf{m}^t \end{aligned} \tag{5.5}$$

Equation 5.5 represents a general structure of the equation system associated with target-centric approaches. Individual approaches strongly differ in the way the parameter vector β^t is estimated. Some approaches transform the input data to reflect changes in concentration over time resembling an ordinary differential equation system (ODE [101]). Due to the number of experiments available Equation 5.5 has no unique solution. Thus, the computation of β^t resorts to regularization strategies to obtain unique solutions [101, 158, 160]. A crucial step is then to extract regulator-target confidences $\hat{\mathbf{c}}_t$ from the influence vector β^t. Even if the underlying system may not exhibit linear behaviour regularized linear regression models [235, 266] seem to provide reasonable approximations.

By using the network knowledge W as an initial parameter estimate, β^t can be used to integrate prior knowledge, effectively resulting in *supervised* variants [101] (see Section 5.4.3).

5.4.3 Supervised Approaches with Topology Prior

The class of *supervised* approaches integrates prior network knowledge (see Figure 5.2). This is also referred to as *topological prior.* We speak of *pattern-centric* approaches, as their underlying models are designed to detect expression patterns with predictive power for the inference of novel regulations. Depending on the way these models are build, pattern-centric approaches are classified as *global* or *local.*

Global pattern-centric. These models analyze simultaneous changes that occur in both regulator and target, also referred to as covariance patterns. To distinguish patterns of regulator-target interplay and independent pairs they rely on topology information and thus, resemble a *supervised* variant of information-theoretic approaches. The underlying models are not regulator-specific and referred to as *global.* Unlike unsupervised approaches the dependency function is parameterized in an *eager* way in a separate training step that incorporates the topology information W.

This results in a W-specific global model d_W:

$$\begin{aligned} &d_W : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}, \\ &d_W \left(\mathbf{m}^r, \mathbf{m}^t\right) = \hat{c}_{rt} \end{aligned} \tag{5.6}$$

These algorithms are trained using the training labels l_{rt} for the binary case or w_{rt} in a regression setup (see Section 5.4.1). The key idea is to train a model d_W capable to distinguish patterns of regulatory interactions from independent patterns given experiment fold-changes and thus, infer novel regulations [28, 200].

Notably, as the topology information enters the training of these models, measures have to be taken to avoid over-fitting. Commonly, regularization or restricted training by cross-validation is applied (see Section 5.4.4).

Local regulator-centric. These methods [171, 244, 172] are closely related to global approaches. Yet, the trained models are bound to specific regulators (see Figure 5.2). They distinguish targets from non-targets for one regulator. These models predict a regulator-specific confidence via the dependency function $d_{W,r}$ (see Equation 5.7, Section 5.4.4 provides a naïve baseline approach). During the training all available expression data is analyzed for each regulator, exploiting available information on targets and non-targets. The resulting model predicts the confidence $\hat{c}_{rt}$ using the target's expression pattern across experiments:

$$\begin{aligned} &d_{W,r} : \mathbb{R}^p \to \mathbb{R}, \\ &d_{W,r} \left(\mathbf{m}^t\right) = \hat{c}_{rt} \end{aligned} \tag{5.7}$$

One-class. It has been argued that false negative interactions may mislead supervised methods. Consequently, approaches using only confirmed interactions have been developed to separate real regulations from false positive ones [37, 35, 85, 172]. We refer to this class of approaches as *one-class*, resembling the idea of wrong regulations being outliers to the single true class of regulations.

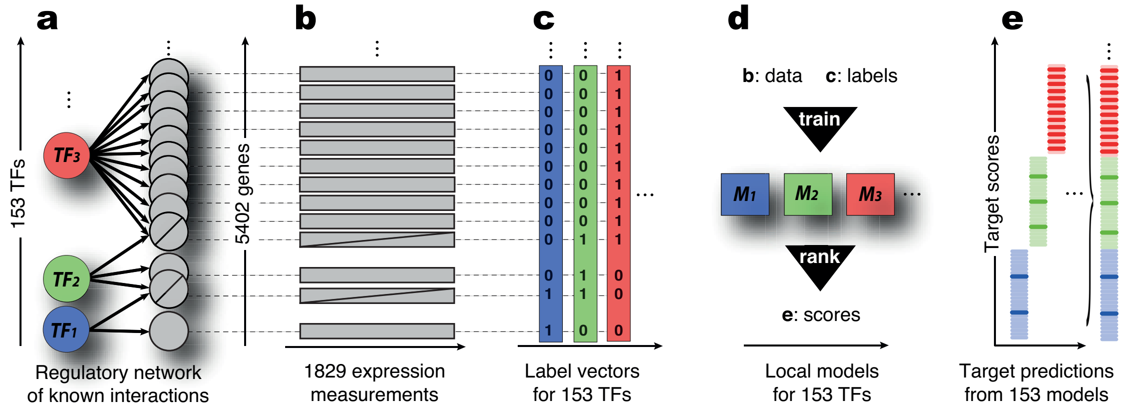

a
b
c
d
e
153 TFs
TF3
TF2
TF1
5402 genes
Regulatory network of known interactions
1829 expression measurements
Label vectors for 153 TFs
b: data c: labels
train
M1
M2
M3
rank
e: scores
Local models for 153 TFs
Target scores
Target predictions from 153 models

Figure 5.2: **Supervised regulator specifc inference.** Supervised inference methods can utilize **(a)** known interactions as well as **(b)** an expression data matrix. **(c)** Known interactions are transformed into regulator-specific label vectors of length 5,042: each gene is labeled 1 if it is targeted by the regulator and 0 otherwise. **(d)** A model M_i is trained for regulator i. Each model consists of n sub-models, where n cross-validation splits are used to avoid over-fitting (not shown). The model incorporates the structure prior (a+c) and target expression (b) to distinguish known from non-target genes. **(e)** All potential regulations are predicted by each model and the respective targets are ranked by the predicted confidence scores. A simplified example is shown whereas known targets (saturated) are indistinguishable from non-targets (pastel). Yet, even if all models produce random confidences common evaluation routines would assess the union of all models' predictions as accurate. This effect can be attributed to the fact that large regulators (red, high out-degree) systematically achieve higher confidences than smaller ones (green/blue, low out-degree).

Integrative methods. The explicit use of additional data sources has been analyzed, ranging from sequence binding motifs [68] to the semi-automated integration of experimental outcomes in an iterative fashion [44]. SEREND [68] is a state-of-the-art integrative method for GRN prediction. It utilizes TF binding site information and expression data. Three logistic regression classifiers are trained: (i) using expression data, (ii) using binding sites and (iii) using the two initial predictions via a meta classifier. The approach is local as classifiers are trained separately for each TF.

Table 5.1 provides an overview of the most important dichotomies that arise in the context of network inference.

Table 5.1: Characteristics and types of inference approaches by algorithmic aspect

inference aspect	type	characteristics
network utilization	expression-based	no topology integration
	supervised	with topology integration
handling of interactions	one-class	missing regulations considered unknown
	two-class	missing regulations considered negative
model building	lazy	no trained model
	eager	predictive model
data handling	integrative	further data sources integrated
	non-integrative	expression and topology data
modeling strategy	global	one model for all regulations
	local	one model for each regulator

5.4.4 Applied Inference Approaches

In this section, we provide details on the approaches that are compared in this chapter, in particular the supervised prediction schemes (see Section 5.4.3) that are essential to the so-called supervised inference of regulatory networks (SIRENE) protocol [172].

Predictive Correlation. A simple way to come up with a predictive supervised dependency function d_r for a potential target $x \in G$ is to compare $\mathbf{m}^x$ (the experiment fold-change values for gene x) to all known targets $t \in T_N(r)$. We use an average of all Pearson's correlations of each $\mathbf{m}^t$, $t \in T_N(r)$ and $\mathbf{m}^x$:

$$d_r(x) := \frac{\sum_{t \in T_N(r)} \rho(\mathbf{m}^x, \mathbf{m}^t)}{|T_N(r)|} \tag{5.8}$$

This dependency only uses previously known regulations. It provides a baseline comparative approach in [85] as well (see Figure 5.7, page 106, method 3).

Decision Trees. Decision trees are decision structures which classify genes with regard to the values of $\mathbf{m}^x$. We applied decision trees to train local models. In particular, a TF-specific decision tree imposes an order for experiment examination. Nodes in a tree represent the expression measurements (columns of M) and the corresponding threshold to optimally distinguish between targets and non-targets of the given TF. For each putative target, the prediction procedure starts at the root node and decides for each level which of the possible decision branches is chosen. The choice is based on the node-specific threshold and expression level of the examined target. Leaves assign predictions on whether or not the tested target is regulated by the given TF. For training and prediction of decision trees we rely on C4.5 [256] via probabilistic thresholds.

A single decision tree is error-prone wherefore usually many trees on subsets of data are build and integrated via meta-learning techniques like boosting or bagging. Here, we employ bagging [201] to arrive at a dependency function by computing the empirical confidence values for each prediction. In each cross-validation fold (see Section 5.5.5), we trained 20 trees each using 80% of the positive and 20% of the negative examples in the training fold. Each possible interaction therefore received a confidence score averaged from 20 trees.

Random Forests. An extension to decision trees are random forests, which sacrifice the ability of model interpretation in favor of predictive power. This tree learner builds a set of predictive decision trees on experimental subsets and uses a majority voting procedure across all trees to arrive at a decision. Decision values returned are a matrix of class probabilities (one column for each class and one row for each input). Probabilities are calculated from the votes of each generated tree. For random forests (R-package *randomForest* [151]) we used default parameters as selected by the corresponding cited software packages.

Two-class SVM classification. Support vector classification (SVC [46]) provides a robust learning technique based on the optimal separation of two-class high-dimensional input vectors. SVMs solve two- or multi-label classification problems with high accuracy and enforce a regularized solution [214]. In practice, SVMs are expected to generalize well to previously unseen regulatory predictions.

The use of SVM models for regulator-centric pattern detection has been suggested following global [28, 200] and local [171] prediction schemes, whereas the latter have shown superior performance. In all cases the separation of regulatory from non-regulatory interactions is enforced.

All pairwise similarities of potential regulator target measurements $\mathbf{m}^i$ and $\mathbf{m}^j$ for $i, j \in T$ are used to derive a maximum-margin hyper-plane separating targets from non-targets. Training set members that lie on this margin are called support vectors. The similarity measure is a positive semi-definite kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, *e.g.*, a linear kernel or a radial basis function (RBF). A parameter C controls the amount of misclassification allowed during model building. In case of the RBF kernel the bandwidth γ controls how far two instances may be apart to be considered similar. The all-against-all pairwise kernel evaluations are then transformed into a convex optimization problem.

For network inference, we define a regulator-specific dependency function $d_{W,r}$ as the distance of a potential target to the hyperplane. We used the implementation of libSVM [37] either directly or via R-wrappers [54].

Supervised one-class SVM. It has been argued that information on non-targets may be unreliable and thus, merely known positive targets should be used to derive regulatory interactions. One-class SVMs build predictive local models based on only positive examples and provide a statistical outlier-detection for targets to be predicted [37].

Graphical Lasso and Penalized Regression. The graphical LASSO (Least Absolute Shrinkage and Selection Operator) method has been proposed by Tibshirani for the estimation of linear models [235]. Lasso fits a generalized linear model via penalized maximum likelihood. This method uses L_1 penalties and hence provides automatic feature selection. The L_1 penalty causes a subset of the solution coefficients to become zero [113]. This corresponds to a feature selection and results in a sparse model with regard to gene coefficients. The approach has been adapted using the R package *glmnet* [78, 223].

Elastic Net. The Elastic Net combines Lasso and ridge regression by a simultaneous optimization of both L_1 and L_2 penalties. The ridge penalty (L_2) shrinks the coefficients of correlated variables towards each other. The elastic net penalty can be used for regression or classification [113]. The elastic net algorithm has first been proposed by Zou and Hastie [266] for the analysis of microarray data and construction of classification rules. It has been used for various studies with different extensions and settings: the inference of expression values of yeast genes during the DREAM3 challenges where it performed best [104]. The elastic net is used for gene selection in the gene expression analysis framework [13]. Previous work by Shimura *et al.* [222] applied the elastic net with an extension of the vector autoregressive (VAR) model to infer gene networks from microarray experiments. As in the case of Lasso, the R package *glmnet* [78, 223] is used.

Direct Integration of Network Topology. Methods that integrate prior knowledge of topology usually rely on the data induced by known regulator-target interactions. They do not explicitly integrate the adjacency matrix of the underlying graph.

Say we derive a model for $r \in R$. Then, in order to integrate the knowledge of known targets in a training set, we could extend the vector of expression data of each potential target $t \in G$ by the information on other known regulators. In particular, the (fold-change) column vector $\mathbf{m}^t$ is concatenated to the vector $\mathbf{w}^t := (w_{tj})$, $j \in \{R \backslash \{r\}\}$. For the predictive model $d_{W,r}$ the regulator information for r is excluded to avoid over-fitting. In a cross-validation setting all interactions in the current test are treated as non-existing with zero weight.

Consensus predictions across methods. To compute a consensus across multiple methods we apply a rank merging procedure [160]. For regulators $r \in R$ and targets $t \in T$ each method m provides a confidence value $\hat{c}^m_{rt}$. We use the average rank across all m as a simple consensus score.

5.4.5 Supervised Function Prediction

Many types of relationships including gene regulatory interactions are subsumed using the generic term *functional associations*. Different kinds of functional associations between proteins or other biological entities further encompass protein-protein interactions, drug-target interactions as well as protein annotations. The latter are associations that link proteins to biological processes or protein functions.

A related important concept is that of a *gene set*. In case of interactions, a TF is associated with all genes contained in a corresponding gene set that comprises all target genes of this TF. A biological process is assumed to be described by a gene set containing all genes known to be relevant for that process. Essentially, a gene set is specific to a given entity such as a TF or a biological process and covers a given type of functional association. Supervised inference can thus be applied to infer functional associations of interest if (i) suitable datasets are available and can be structured as a data matrix and (ii) prior knowledge can be provided as a (partial) set of genes in the form of a label vector. Again, in case of interaction inference, (i) could be large scale gene expression data and (ii) would be (part of) the genes regulated by a given TF.

In this context, the prediction of functional associations or simply function prediction and the prediction of regulatory interactions may be viewed as special cases where function is a common property among a set of genes such as the targets of a single transcription factor. Function prediction is of interest in many biological use cases as predictive models complement prior knowledge and thereby enable a deeper understanding of both novel and existing associations. In the process of prediction, associations across different TFs or biological processes are prioritized based on prediction confidence to enable the selection, evaluation or experimental follow-up of promising candidates.

5.4.6 Shared Issues

The prediction of different kinds of functional associations shares important properties and issues. For instance, Myers *et al.* [174] focus on predicting gene functions, *i.e.*, pathway-gene associations that are predicted separately for each pathway of interest, and thus, on predictions derived via local models. They argue that the evaluation of functional annotation may be influenced by the uneven size and different properties of certain biological processes. Inclusion or exclusion of the ribosome pathway (among 98 other KEGG pathways) makes the difference between co-expression data being the most or least, respectively, informative dataset. Myers *et al.* conclude that each process should be evaluated in isolation to overcome *HDP*. However, this might not be a practical solution if functional associations must be obtained across biological processes or if transcriptional networks must be obtained across TFs.

It is further important to note that, *vice versa*, proteins spanning a broad range of functions are much more likely to be confirmed as correct members of an arbitrary functional class in comparison to specific proteins with a narrow range of functions. For protein-protein interaction networks, Gillis and Pavlidis [88] discuss that the number of functions a protein exposes is coupled to its node degree. For an arbitrary functional category being predicted a ranking based on the network node degree will perform better than expected by chance and it can unintentionally skew quality estimates. This is referred to as *multi-functionality bias.* They conclude that there are no suitable techniques available that substantially reduce it without undesired side-effects. In particular, Gillis and Pavlidis argue that an entirely different problem structure may arise, such that it is often unclear whether a fix is preferred or not.

Preferences in the selection of predicted interactions have been observed [51, 5]. They describe that TFs with many known targets receive disproportionately many predictions while hardly any predictions are assigned to TFs with few known targets. De Smet and Marchal [51] conclude that this is due to the fact that less information is available for the TFs with few known targets and that, based on this observation, supervised approaches should not be applied to infer interactions for TFs with few known interactions. While we can confirm this observation, we find that the problem's origin is different (see Section 5.5.6) and we argue that it is linked to the algorithmic approach. In this chapter, we provide evidence that independent confidence distributions lead to a skewed overall integration. We demonstrate that *HDP* can be tackled by an appropriate recalibration (*CoRe*, see Section 5.5.6).

While for larger regulators some sensitivity in detecting true novel regulations may be lost, the amount of false predictions is drastically reduced. For most smaller regulators *CoRe* boosts sensitivity and the true positive rate and enables the prediction of novel targets, even in case of low-degree TFs. For network inference, supervised inference in combination with a recalibration like *CoRe* is preferred to introduces a regulator-wise empirical false discovery estimate and compute sensible overall networks.

5.5 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 the two main 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 5.3).

5.5.1 Overview of Training Data

We obtained five yeast expression compendia from:

1. The DREAM5 Network 4 (DN4) expression data set [160] comprises 536 expression measurements of 5950 yeast genes compiled from 59 publications. We computed 369 $\log_2$ fold change vectors from this expression compendium. A wide range of experimental conditions, including gene, drug and environmental perturbations, partially conducted in time courses is covered.

2. The compendium containing 904 chips of 6777 yeast genes was obtained from the Many Microbe Microarray Database (M^{3D} [72]). The data set was built from 62 experiments. After conversion to fold change values the data set contained 727 vectors of length 6777.

3. Hue *et al.* [124] performed a comprehensive study of TF knockout experiments. The GEO accession is number is `GSE4654`. It contains expression measurements of 263 transcription factor knockout strains under different experimental conditions. The data set was transformed into 269 $\log_2$ fold change values each measuring 6429 genes. This and the following two compendia focus on steady-state TF deletion and over-expression measurements that we obtained as $\log_2$ fold change values from the GEO database.

4. The study of Chua *et al.* [42]: in contrast to the previous compendium, the data set with GEO accession number `GSE5499` consists of knockout but also over-expression experiments for 55 TFs. The data set contains 270 $\log_2$ experiment fold change values for 6307 yeast genes.

5. From various manually selected GEO data sets [14], we obtained additional 194 independent gene knockout measurements for 6307 genes.

Case-control pairs were selected from 2,442 yeast microarrays as described by Küffner *et al.* [143] 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 [3], augmented by a study of MacIsaac *et al.* [157] featuring combined genome-wide chromatin immunoprecipitation (ChIP) data in combination with two conservation-based motif discovery algorithms, PhyloCon and Converge [157].

We filtered genes that were not contained in the expression data. We excluded TFs regulating less than 6 known targets to enable sensible training and cross-validation. The resulting reference standard contains 153 TFs, 4,870 target genes and 24,462 interactions derived from 356 TF-target binding assays.

5.5.2 Estimating the Size of the Yeast Regulatory Network

We aim to estimate what fraction of regulatory interaction are currently known in yeast. In summary, we compiled 29,398 interactions from 356 TF-to-promoter binding studies as well as 21,847 interactions from 536 gene expression studies. In the latter case, interactions are assumed between a regulator and a target if the target expression changes in regulator deletion or over-expression mutants. Since expression studies would introduce potentially indirect interactions we restrict the gold standard to interactions determined by binding studies. However, these expression studies play an important role in the estimation of the yeast network as described in the following.

Each published study would contribute a small fraction of regulations to the complete network. Measurement bias and study overlap likely introduce saturation effects in the discovery of novel interactions. Thus, we like to estimate the completeness of the yeast regulatory network by empirical limit analysis. An important assumption here is that increasing the number of studies would converge towards a hypothetically completed gold standard (GGS).

We repeatedly sample (10,000 times) a fraction of x from the set of all studies that make up the gold standard. This subset induces a partial regulatory interaction network. The (average) fraction of regulatory interactions detected for x parts of all studies is denoted by $\Theta(x)$. $\Theta(x)$ is not expected to depend linearly on x, but should follow a saturation curve and be convergent towards the CGS. We therefore decided to model the expected dependency in terms of a Hill coefficient [117]:

$$\Theta(x) = \frac{m * x}{k + x} \tag{5.9}$$

The two parameters of this equation have a direct interpretation in terms of the network completeness. First, the parameter m is the fraction of interactions in the CGS relative to all currently known regulations, *i.e.*, $m = 1.0$ would imply the currently known gold standard is complete.

Secondly, k is the fraction of available studies when half of all completed gold standard interactions are detected. The coefficients have been estimated using the sample mean of the interaction count for a given x such that the root mean squared deviation was minimized.

Assuming that not all possible interactions are yet known, m will be greater than 1. Thus, scaling the number of currently known interactions by m would approximate the total number of interactions in the CGS.

We use the approach to contrast the convergence of regulations derived from (i) binding studies (ii) the intersection of binding studies and regulator perturbation-based expression profiling. We sampled from all binding studies in both cases, but in (ii) the population of sampled interactions was limited to the intersected set. As a consequence, $m = 1$ corresponds to the number of interactions supported by both promoter binding as well as TF perturbation studies. The tuple (k, m) was estimated separately for both scenarios.

5.5.3 A Network-Only Model

It has been shown previously that the node degree can seriously impact predictive performance estimates [88]. To estimate the predictive power of a network N's topology we define a naïve regulator-specific confidence mapping

$$d_{W,r}(t) = |r|_N^{out} \tag{5.10}$$

where $|r|_N^{out}$ is simply the number of known targets of a regulator in network N, *i.e.*, its out-degree. Consequently, all targets t of r receive the same score, namely the out-degree of r. Obviously, this dependency function cannot distinguish real from random targets: larger regulators affect more targets and trivially obtain higher scores (see Figure 5.2, page 80). The calculation of the predictive quality then reflects the baseline expected by random guessing.

We assume that the likelihood for any novel target to be regulated by a larger factor is higher as well. Any evaluation that computes a factor-wise performance measure would observe that the predictions are indeed random and no real target can be distinguished from random targets.

Any compilation of all individual predictions into a single list will hide this effect. In fact, ranking regulations among large regulators and their possible targets higher than smaller regulator's interactions is likely superior to any random prediction. Arguably, the result network predicting all possible interactions for say, the 5% largest regulators and nothing else is superior to a complete random solution. It is important to observe that common measures like ROC and PR curves share this global viewpoint and would score degree-sorting better than random predictions.

5.5.4 Evaluation Metrics for Network Prediction

In general, we compared predicted interactions to experimentally confirmed interactions, *i.e.*, the gold-standard. True positives (TP) are predicted interactions that can be confirmed by the gold standard. True negatives (TN) are neither predicted nor in the gold standard. False negatives (FN) are not predicted but present in the gold-standard while false positives (FP) are regulatory interactions that are predicted but are not confirmed. Canonical measures are the precision $pr = TP/(TP + FP)$, the sensitivity $sn = TP/(TP + FN)$ as well as specificity $sp = TN/(TN + FP)$.

Each predictive method results in a list of confidence values $\hat{c}_{rt}$ covering all potential regulatory interactions (r, t) among regulator $r \in R$ and target $t \in G$. A ranked list of regulations is obtained via sorting by confidence. All methods below inherently deal with ties present in these lists by averaging results in intervals of equal confidence.

We computed three performance metrics commonly used to estimate the quality of predictive methods:

1. The Precision-50 (P50) is the maximal number of predictions that exceed or equal a precision of 50% when lowering a confidence threshold on the predicted scores. The higher the number, the more interactions may be actually predicted with sufficient reliability in practice.

2. The precision recall curve (PR) is the precision pr as a function of sensitivity sn. To vary sensitivity all possible thresholds for predictions within the ranked list are screened. The AUPR is the area under the PR.

3. By contrast, the AUC is the area under the receiver operator characteristics curve (ROC). The ROC is the sensitivity sn as a function of (inverse) specificity $1 - sp$. Similar to the PR all possible confidence thresholds are screened and plotted accordingly.

Random predictions are expected to receive an AUC of 0.5. *Vice versa*, an AUC of above 0.5 would imply a non-random covariance of the prediction scores and the gold-standard. The best possible AUC value is 1.0 if predictions and gold-standard perfectly agree. We point out that no perfect, complete gold-standard exists. Therefore further assessment and quality estimates are mandatory.

5.5.5 Prediction Setup and Validation

The trained models (Figure 5.2d and Figure 5.3d) assign a confidence score to each possible regulation (r, t). Ranking all putative interactions results in a list of $|G|$ confidence scores for each regulator. This list is compared to a gold standard N_{gold} that contains known or experimentally confirmed interactions (see Section 5.5.1). For each $r \in R$ we set up a 3-fold cross-validation (3-CV). The set of all network nodes G of N_{gold} is split into n stratified sets. For local models, a scoring model $d_{W,r}$ is built on the $n-1$ splits and the n-th set is predicted. Additionally, a single *global* model d_W is trained (see Equation 5.6) for all regulators using combined feature vectors, *i.e.*, feature vectors of regulator and target represent an interaction. For global models we split the set of nodes G into k stratified folds (w.r.t. the number of regulations). Overall, we train $|R|$ local models $d_{W,r}$ using the network topology W. Each model is capable to predict $|G|$ confidence estimates $\hat{c}_{rt}$ specific to a regulator r (see Equation 5.7 and Figure 5.3b+c). For each split the CV is repeated k times. A corresponding stratified n-fold split is set up across all regulators to train global models.

To estimate the quality of local or global methods we combine all predictions across all regulators (which is not necessary for global methods) and sort them by their confidences. As previously suggested [171], we apply so-called micro-averaging, *i.e.*, the complete list of interactions ranked by their confidences is compared to the corresponding gold-standard annotation. By contrast, macro-averaging would combine regulator-wise performance metrics instead. Macro-averaging is relatively complex to interpret and far less frequently applied. The assessment compares the predictions to a reference standard of *a priori* known interactions, for instance by the area under the receiver operator characteristics curve (AUC). Such a cross-validated AUC analysis is a standard approach for the assessment of inference methods [171]. We calculate several quality estimates for each method. For a detailed definition of all applied evaluation metrics see Section 5.5.4.

To estimate the functional consistency of a prediction we compute the expected biological function overlap of novel predicted targets to known targets. A detailed description of this approach is given in the Section 5.5.8.

5.5.6 Confidence Recalibration (*CoRe*)

Randomized topologies are generated to share key statistics with the reference standard of known interactions (Figure 5.3a+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 [58], 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 5.3). Let $D_r^{(i)}$ be 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 5.3f).

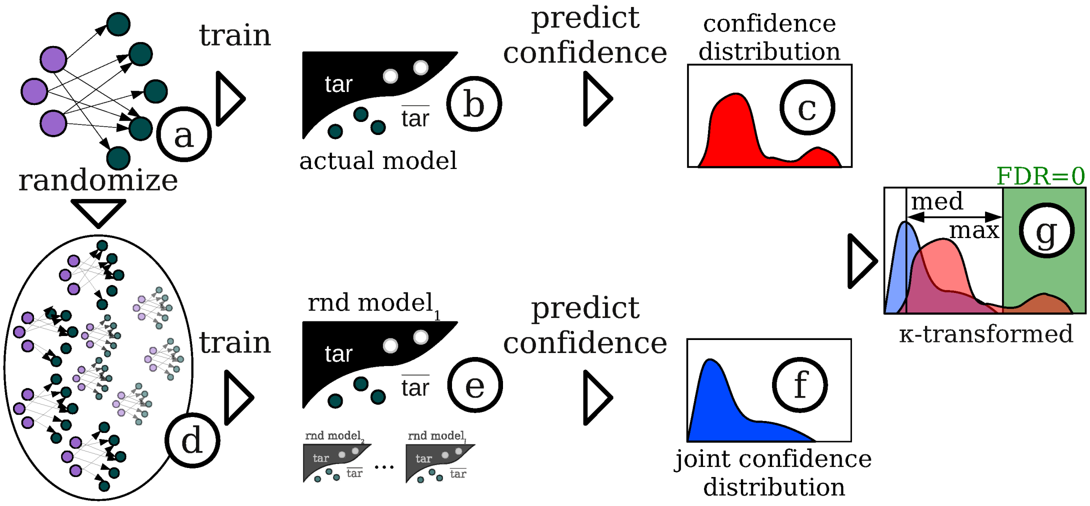


Figure 5.3: **Outline of the recalibration approach.** Based on the known network **(a)**, a regulator-specific model **(b)** is trained to predict potential targets for this regulator. This results in a confidence score distribution for each regulator **(c)**. Additionally, we generate random networks **(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 (green area in **(g)**) that exceed values observed for random networks.

D'_r denotes the *randomized complement* of D_r. By comparing these two distributions we 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 5.3c+g):

$$\kappa_{rt} = \frac{\hat{c}_{rt} - med(D'_r)}{\max(D'_r) - med(D'_r)}. \qquad (5.11)$$

Scores are 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.

5.5.7 Analysis of Interactions in Network Motifs

It is desirable to estimate the predictive power of an approach in the context of known motif contexts. In the following we describe how we measure motif dependency in the context of these motifs as present in a gold-standard. In particular, we contrast two motif types at a time to obtain sensible positive and negative classes to classify each prediction (a regulation exists or not) as true positive (TP), false positive (FP), true negative (TN) or false negative (FN). Given this definition common performance values like AUROC can be computed.

Table 5.2: Assignment of True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN) in a gold standard motif context. The regulatory interaction (r,t) between a regulator $r \in R$ and its target $t \in T$ is predicted if the regulatory interaction confidence $\hat{c}_{rt}$ exceeds a given cutoff h. The in- and out-degree of gene $g \in G$ in the network N is $|g|_N^{in}$ and $|g|_N^{out}$, respectively. Screening the cutoff allows the computation of ROC and PR curves. Each motif-contrast (the comparison of two distinct motif classes) is separated by a horizontal line and evaluated individually. All interactions that match neither class are discarded for this contrast. Some motifs require the presence or absence of an additional regulator $r' \in R$. This table resembles the classes in Figure 5.4.

Motif Class	Gold-Standard Context	$\hat{c}_{rt} \leq h$	$\hat{c}_{rt} > h$
regulation	$(r,t) \in N_{gold}$	FN	TP
no regulation	$(r,t) \notin N_{gold}$	TN	FP
regulation, low t in-degree	$(r,t) \in N_{gold},\ \lvert t\rvert_{gold}^{in} < d_t$	FN	TP
no regulation, low t in-degree	$(r,t) \notin N_{gold},\ \lvert t\rvert_{gold}^{in} < d_t$	TN	FP
regulation, high t in-degree	$(r,t) \in N_{gold},\ \lvert t\rvert_{gold}^{in} \geq d_t$	FN	TP
no regulation, high t in-degree	$(r,t) \notin N_{gold},\ \lvert t\rvert_{gold}^{in} \geq d_t$	TN	FP
regulation, low r out-degree	$(r,t) \in N_{gold},\ \lvert r\rvert_{gold}^{out} < d_r$	FN	TP
no regulation, low r out-degree	$(r,t) \notin N_{gold},\ \lvert r\rvert_{gold}^{out} < d_r$	TN	FP
regulation, high r out-degree	$(r,t) \in N_{gold},\ \lvert r\rvert_{gold}^{out} \geq d_r$	FN	TP
no regulation, high r out-degree	$(r,t) \notin N_{gold},\ \lvert r\rvert_{gold}^{out} \geq d_r$	TN	FP
auto-regulation	$(r,t) \in N_{gold},\ (r = t)$	FN	TP
no auto-regulation	$(r,t) \notin N_{gold},\ (r = t)$	TN	FP
directed regulation	$(r,t) \in N_{gold},\ (t,r) \notin N_{gold}$	FN	TP
reverse regulation	$(r,t) \notin N_{gold},\ (t,r) \in N_{gold}$	TN	FP
feed-forward	$(r,t) \in N_{gold},\ \exists r' \in R : (r',t),(r,r') \in N_{gold}$	FN	TP
cascade	$(r,t) \notin N_{gold},\ \exists r' \in R : (r',t),(r,r') \in N_{gold}$	TN	FP
direct regulation	$(r,t) \in N_{gold},\ \nexists r' \in R : (r',t),(r,r') \in N_{gold}$	FN	TP
cascade	$(r,t) \notin N_{gold},\ \exists r' \in R : (r',t),(r,r') \in N_{gold}$	TN	FP

Simple regulations. In principle, simple regulations are no motifs. Thus, it is straightforward to decide whether a predicted regulation is present in the gold-standard (TP) or not (FP). Similarly, a gold-standard regulation that is missed by the prediction is FN while a TN is reported by neither prediction nor gold-standard. To get a more specific idea of the influence of node degree we restrict the set of regulations that are considered for AUROC analysis (see Figure 5.4).

Auto-regulation. It is useful to decide how well predictions can resolve auto-regulatory loops. Then two classes do exist in the gold standard: (1) auto-regulation and (2) non-auto-regulation. For each regulator-target pair we check whether a predicted regulation exists in the gold-standard (TP) or not (FP). It is also correct to predict no regulation if no regulation is present in the gold-standard (TN), yet would imply a FN otherwise.

Directed interactions. The simplest motif involving two distinct entities of the network is a directed interaction. If no reverse regulation is present in the gold-standard, then a predicted regulation is considered TP and FP if the gold-standard features a reverse regulation. By contrast, it is considered FN not to predict a regulation if the reverse regulation is present in the gold-standard and TN if is not.

Feed-forward loops and cascades. In case of regulations embedded within feed-forward loops the definition of classes is slightly more complicated for the set of non-feed-forward loops is too general. Instead, we restrict the analysis to feed-forward-loops and cascades in this case. All other motifs are neglected. For each regulator-target pair we check whether a regulation is predicted and if that is the case if the gold-standard context of the regulation is a feed-forward loop (TP) or a cascade (FP). The prediction of no regulation is considered a FN if a gold-standard feed-forward context is present. In case of a cascade motif it is correct not to predict any regulation (TN).

Direct regulation and cascades. Similarly, for cascade motifs, the contrasting classes are regulations without existing bypass on the one hand and on the other hand cascades. Thus, the prediction of a direct regulation while only a bypass is actually present in the gold standard is considered FN. Consequently, it is correct not to predict an interaction (TN). For the positive class, the prediction of a direct regulation is correct (TP) since no cascade is present. If we miss the direct regulation despite there is no existing bypass in the data we consider the missing regulation a FP.

In general, we classify different types of regulatory interactions according to the network patterns surrounding them. Each interaction defined in the gold-standard N_{gold} is assigned to one or more types (see Figure 5.4) and predicted confidences are evaluated in this context (see Table 5.2). Given a prediction method we evaluate the specific advantages or disadvantages for each interaction type.

For a given method we analyze the list of confidences for all possible $|R| * |T|$ regulatory interactions. The types are defined by the gold-standard network. The list of confidence values is restricted to include only one type of interaction at a time (see Figure 5.4). Then, for the remaining interactions, AUC values are computed as guided following the assignments defined in Table 5.2. The resulting AUC values are motif-specific and may be compared across several methods.

For a given threshold h an interaction (r, t) is predicted if $\hat{c}_{rt} > h$. The interaction is considered correct in the motif context if it is supported by the gold standard. Each type induces a subset of both gold-standard regulations and non-regulations. This is necessary to arrive at sensible contexts, *e.g.*, the restriction to high out-degree regulators.

The filtered set of interactions is then relevant for the motif of interest. Regulations that do not match any class are discarded for this type. Motifs of up to three nodes $(r, r', t) \in (R \times R \times T)$ are analyzed. We define degree cutoffs d_r and d_t to distinguish low from high node degrees.

5.5.8 Functional Coherence

Network inference methods suggest additional interactions that are not yet contained in the gold standard of experimentally supported interactions. We defined a *functional coherence* score to determine whether biological functions [234] – annotated by gene ontology (GO) processes to the known, experimentally supported targets of a given regulator r – match the functions of newly predicted target genes (see Figure 5.5).

A functional profile for r was defined based on the known targets t in the gold standard network. The profile is represented by a vector $ont_R(r) \in \mathbb{R}^K$, where K is the number of functional categories, such that functions associated to many targets of the given TF receive higher weights. The functional coherence of newly predicted targets was then evaluated by comparing the profile vector to according profiles $ont_G(t)$ of each predicted target. The d-th component of $ont_G(t)$ is 1 if t is associated to the d-th functional category, and 0 otherwise. It reflects how well novel target predictions correspond to the functional annotations of targets in the gold standard.

The functional coherence measure depends on the functional representation of r as a vector of K GO biological processes $ont_R(r) \in \mathbb{R}^K$. Each dimension $d = 1 \ldots K$ is the statistical significance of an intersection set, *i.e.*, of genes that

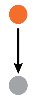
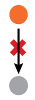
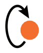
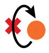
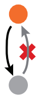
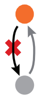
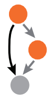
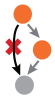
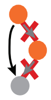
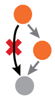
class 1	class 0	Methods -specific preference for network motifs
		**All interactions** (i.e. no restriction), restriction with respect to **TF out-degree, or TG in-degree** (5 motifs, see text).
		**Auto-regulatory loops**.
		**Direction of interactions**.
		**Feed-forward loops**, i.e. prediction of interactions in the presence of additional indirect paths.
		**Cascade motifs**, i.e. prediction of interactions in the absence of additional indirect paths vs. the incorrect prediction in the presence of indirection.

Figure 5.4: **Motif prediction preferences.** We analyzed method-specific preferences that depend on whether predicted interactions (orange=transcription factor or TF, grey=target gene or TG) take part in 9 different network motifs. Our analysis evaluated, in terms of AUC, how well correct and incorrect predictions (black interaction = class 1 and black crossed-out interaction = class 0, respectively) can be distinguished. The motif context was defined by the presence or absence of further edges in the gold standard (gray interactions). The first row yielded 5 motifs based on additional restrictions on the black interactions: (i) no restriction, (ii) low target in-degree (≤ 2 TFs), (iii) high target in-degree (> 2 TFs), (iv) low TF out-degree (≤ 25 targets) and (v) high TF out-degree (> 25 targets).

are both known targets of a given regulator r as well as associated with the d-th biological process. The significance of the overlap was calculated as functional enrichment score of the targets $T_{N_{gold}}(r)$. For a given functional category, it was computed as a hypergeometric z-score $h_z(x, N, n, k)$ given the number of genes k in the category, the number of genes n known to be regulated by r, the number N of all possible targets in the gold standard and the number of genes x in the intersection (see Figure 5.5). Similarly, each $t \in T_{N_{pred}}(r)$ was then assigned to a vector $ont_G(t) \in \{0, 1\}^K$ encoding the membership of t in each process. For a regulatory interaction (r, t), we then computed the functional coherence as the normalized scalar product $cons_{rt} := \langle ont_R(r), ont_G(t)\rangle$.

We then selected a set of regulatory interactions $\{(r, t) \,|\, c_{low} \leq s_r(r, t) < c_{high}\}$ for each interval of prediction scores $c = \langle c_{low}, c_{high}\rangle$. Each interval is associated with a row in a two-dimensional density map that displays a histogram across equally sized bins of coherence scores.

5.5.9 Derivation of Modules from the Predicted Network

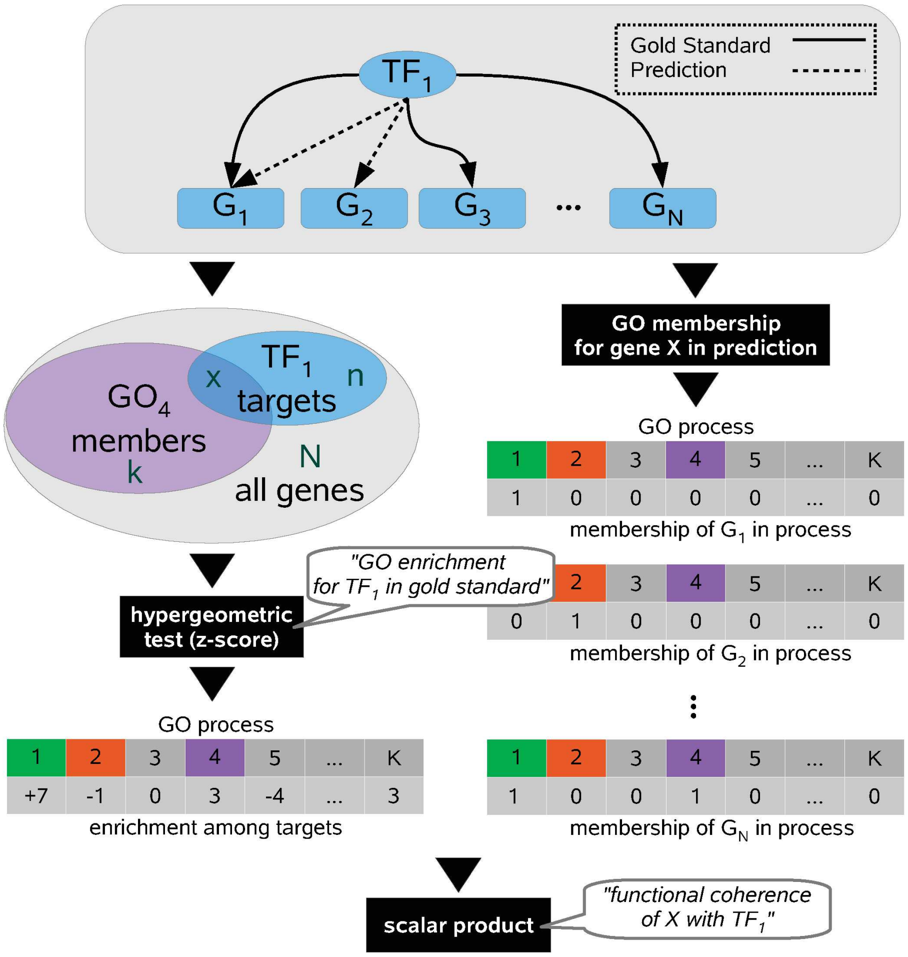
We applied a k-means clustering approach using an euclidean distance metric on the predicted network N_{pred}. We represent each TF as a binary vector of all targets $t \in G$ (the set of all genes, see above). An interaction was encoded as 1, non-interactions as 0. The representation resulted in a matrix M_N with 153 rows (TFs) and 3,747 columns (targets). Clustering was performed in two dimensions: (1) clustering of TFs and (2) clustering of targets. For both clusterings, k is screened randomly 100 times in the range of 8 to 15. Overall, 10,000 bi-clusterings were prepared. We filtered the result to retain only biclusters with a minimum density of 40% predicted interactions. Subsequently, bi-clusterings were ranked based on the retained bi-clusters using

- the number of bi-clusters in the bi-clustering n_b
- the number of interactions n_i covered by the bi-clustering
- the number of TF clusters n_t and
- the number of target clusters n_g

by the empirical ranking criterion

$$n_i - (n_b * n_t * n_g). \tag{5.12}$$

The criterion is designed to cover as many interactions as possible within a minimal number of clusters. The key result of this procedure, the set of highest scoring bi-clusters, is Figure 5.8. Here, TF clusters are connected to target clusters they regulate. Interaction clusters then represent the bi-clusters derived by this procedure. A detailed discussion of TF and target clusters is given in Section 5.6.6.


Gold Standard
Prediction
TF1
G1
G2
G3
...
GN
GO4 members
k
x
TF1 targets
n
N
all genes
hypergeometric test (z-score)
"GO enrichment for TF1 in gold standard"
GO membership for gene X in prediction
scalar product
"functional coherence of X with TF1"

GO process

1	2	3	4	5	...	K
+7	-1	0	3	-4	...	3

enrichment among targets

GO process

1	2	3	4	5	...	K
1	0	0	0	0	...	0

membership of G1 in process

	2	3	4	5	...	K
0	1	0	0	0	...	0

membership of G2 in process

⋮

1	2	3	4	5	...	K
1	0	0	1	0	...	0

membership of GN in process

Figure 5.5: **Functional Coherence Measure.** For each TF (top), a measure of functional coherence is derived by assessing the overlap of functional annotations of (1) its experimentally supported targets and (2) newly predicted putative targets among G. In a first step (left side), we apply the hypergeometric test to analyze the enrichment of functional annotations among the experimentally supported targets. The enrichment score is computed with respect to observing an overlap of x or more genes among targets of TF_1 and those genes annotated with the hypothetical GO category GO_4. The table 'enrichment among targets' denotes this enrichment as z-scores for all $1..K$ GO processes in the second row. Positive or negative z-scores denote process annotations that are enriched or depleted, respectively, among the targets of TF_1. Each table of newly predicted targets of TF_1 (right side) refers to a single gene, which might either be part of GO_4 or not, hence assigning 0 or 1, respectively (column 4 of the second row). Finally, functional coherence is computed as a scalar product among an enrichment vector (left table) and a gene-specific vector (right table). Note that if the coherence for a known TF_1-target such as G_N is calculated, it is removed from the calculation of the enrichment vector in a leave-one-out setup.

5.6 Results

5.6.1 Network Predictions without Expression Data

Expression data is the principal source of information exploited to infer interactions. However, by disregarding expression data in a network-only approach, basic issues of regulator-specific methods can be illustrated. An analogous approach was suggested previously for function prediction [88]. For the network-only approach, we assigned confidence scores based on the out-degree of regulators such that scores for targets of a regulator A are always higher than scores for targets of a regulator B if A has the higher out-degree. In contrast, scores among the candidate targets of a single regulator are distributed uniformly so that true and false targets of a given regulator are indistinguishable (Figure 5.2e and Section 5.5.3).

Accordingly, we calculated a cross-validated AUC for a single network combining all regulator-specific confidences as suggested [171]. In addition, we determined the AUC for all regulators separately. The latter indeed resulted for each regulator in an AUC of 0.5 expected for random predictions. However, the integration of the same predictions across regulators into a joint confidence score distribution resulted in an AUC of 0.798, seemingly indicating a substantial performance. Thus, despite the fact that individual predictions were random, an integrated network can exhibit a substantial enrichment of true TF targets at higher scores (Figure 5.2e).

5.6.2 Simpson's Paradox

A working example. We described that taking a regulator-wise or network-wide viewpoint for the evaluation of inference approaches may result in strikingly divergent outcomes. We therefore start with an example of Simpson's Paradox to clarify the conditions that lead to the (seemingly) paradox situation. For the evaluation of network inference we aim to compare two methods A and B. Each methods provides us with a confidence for each potential regulatory interaction. There are two common approaches to evaluate the result network. (1) We sort all predicted regulations based on their assigned confidence values and compute some canonical network-wide quality measure (like an AUC). This is known as micro-evaluation. (2) We sort the predicted regulations *per regulator* and compute local quality measures. This is often referred to as macro-evaluation. In practice, a situation may occur where micro-evaluation suggests that B is superior to or on par with A and, simultaneously, most or all macro-evaluations would prefer A. This seems to be paradox, because we intuitively think that a method that is better for all sub-problems (or subsets) should perform better for the complete set as well. This reversal given two points of view (complete and subsets) is often referred to as Simpson's Paradox [191, 224].

Observations on real-world inference. Mapped to our setting, method A is a regulator-specific machine learning model that can be used to predict novel targets from known regulator target patterns, *e.g.*, a random forest approach. Method B randomly re-assigns the known regulations to random targets and then uses method A on the shuffled network and data. Method C is a baseline method that works free of data would predict the number of known targets for each regulator as a confidence value for all its targets (see Section 5.5.3).

For a network-wide estimate of quality (like an AUROC) both A and B seem to be on par. For example a random forest model achieves a micro-evaluation AUC of 79.6. The same model being trained on a shuffled topology achieves 72.9 (see Table 5.7, page 118).

In practice, both models would be considered to yield useful results given their overall performance. Yet, for an averaged macro-evaluation we observe 63.1 for standard random forests and 49.4 using randomized topologies. Notably, a model that provides random predictions for almost all regulators obtains a global quality of more than 70 percent. The model quality is also evident in the Precision-50: for shuffled random forest predictions the P50 is 0, whereas 6,996 regulations can be predicted at 50% precision otherwise.

While both networks are of similar overall quality with respect to the micro-evaluation AUC, the regulator performance is crucial. It seems inconsistent that the AUC fails to recognize this shortcoming as it provides a network-wide point-of-view.

Simpson's Paradox motivates confidence recalibration. The Simpson's Paradox refers to the counter-intuitive interpretation of observed results. In fact, both the micro-evaluation AUC and the average macro-evaluation are correct. The common perception is that a network cannot be correct globally, but random for each regulator. This view neglects an important aspect: both methods A and B have access to the degree of a regulator. This prior information may override the predictions that individual, regulator-specific models provide. In fact, we observed a strong degree-dependency for predicted confidences in all models, and the micro-evaluation AUC would benefit from ranking larger regulators first, while macro-evaluations do not rely on this ranking.

We can by now tell that the Simpson's Paradox is induced by the integration of topology information. Strikingly, method C yields an AUC of 79.8, a score that is superior to methods that integrate data. Since the AUC itself is a reasonable quality measure one may argue to choose this globally best model. This argument is easily disproved: The regulator-wise quality is essential for almost any kind of application, and method C cannot rank the predictions for individual regulators – neither can method B.

To resolve the Simpson's Paradox would then mean to select a network-wide set of regulations with reasonable performance whereas individual regulators should maintain the quality that state-of-the-art predictive methods can provide. To tackle this problem, and bridge the gap that leads to Simpson's Paradox, we suggest to capture the regulator-specific nature of B as a random background and use it to contrast the results of the corresponding method A. We refer to this as confidence calibration (*CoRe*). This is the motivation behind the κ-transformation procedure (see Section 5.5.6) as key element of *CoRe*.

Obviously, while we aim to uncover regulator-wise information, the topology information should not be cancelled out completely: it is implicitly reflected by an increased κ-value, *i.e.*, the degree-specific contrast among random and non-random confidence values.

As expected the Simpson's Paradox and the *HDP* disappears upon recalibration. While the macro-evaluation AUC stays the same, the semi-global P50 estimate for these networks slightly drops. Yet, by design, the estimated network-wide false discovery rate is drastically reduced.

5.6.3 Implications of Simpson's Paradox

We followed the SIRENE approach [171] and trained local models based on Support Vector Machines 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 5.5.1) 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 (Section 5.5.3). Strikingly, an evaluation across all regulators yielded an AUC of 0.798, a score above the AUC achieved by SIRENE.

As discussed, 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 [224, 191]: each of the regulator-specific distributions achieves an AUC of 0.5, while the AUC of the joint distribution suggests non-random performance (see Section 5.6.2).

This results from the fact that predicted confidence score distributions are heterogeneous across regulators and are characterized by different scale and location parameters (Figure 5.6a, gray boxes, 104). In particular, score distributions for regulators with many known targets (high out-degree) such as *ste12* are wider and systematically above average following the *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 remain unspecific while biologically more specific signals are likely being missed [188]. 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 [88].

To examine whether the paradox is an artifact of SVMs we trained further model classes (among others 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 *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 (see Section 5.5.2), 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 5.6a, red and gray boxes). 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 5.6f), 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 5.6a). 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 [219]. We therefore tested whether known and predicted targets of the same TF exhibit substantial functional overlaps (Section 5.5.8). We observed that the high proportion of random scores (like *ste12*) concealed most of the signal as interactions with higher scores hardly showed an increased functional coherence (Figure 5.6b).

5.6.4 Correction through Score Recalibration

We introduce a confidence recalibration (*CoRe*) as a wrapper for existing methods (Section 5.5.6). 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 5.6c). 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 above 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 5.6g), 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 (Section 5.5.8). Regulatory patterns were more coherent for the corrected network (Figure 5.6e).

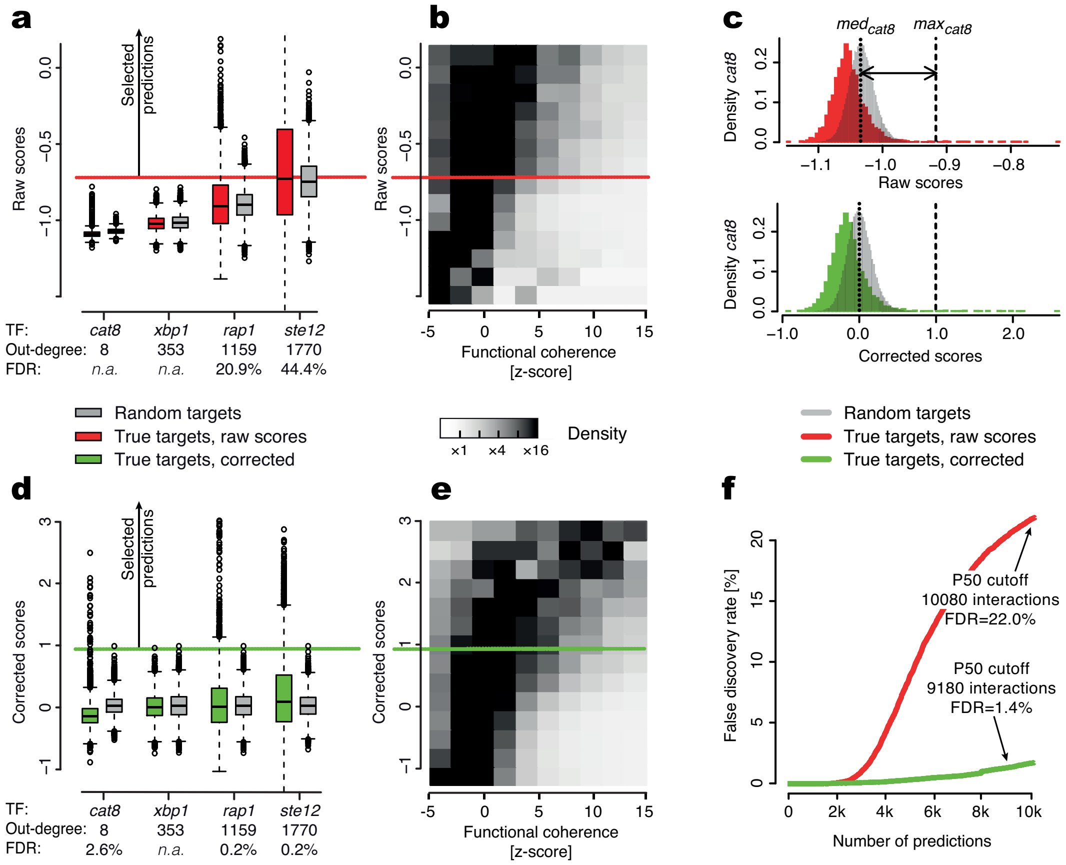


Figure 5.6: **Score recalibration in network predictions. (a)** We trained support vector machines (SVMs) for each TF (see Figure 5.2). Putative target genes were selected by a threshold (red line) on the resulting TF-specific scores (red boxplots). Additional SVMs were trained on random networks (gray boxplots) and false discovery rates (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 (red) were recalibrated via randomized distributions (gray): for each TF, the median *med* (dotted line) and maximum *max* (dashed line) are mapped to 0.0 and 1.0, respectively. **(d+e)** show boxplots, prediction threshold (green line), 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).

5.6.5 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 (see Section 5.4.4).

SIRENE [171] 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 (Section 5.5.7) to capture method- and topology-specific preferences (Figure 5.7b). 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 [73]. These methods are unable to detect auto-regulation as in this case both expression profiles would be identical. Confirming previous findings [160], 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 5.7c). We observed that integrative methods like SEREND [68] suffered from false positive predictions. This is likely due to the low specificity of positional weight matrices (PWMs [120]) predicting targets only for 6.5% of all regulators (see Section 5.6.8). These methods were not further analyzed. Of the remaining five methods (methods 5-12 in Figure 5.7), the best results were obtained from regulator-specific SVMs and decision trees trained on bootstrap samples (bagging). In Section 5.6.13 we discuss methodological extensions such as the integration of multiple predictors to perform a consensus prediction.

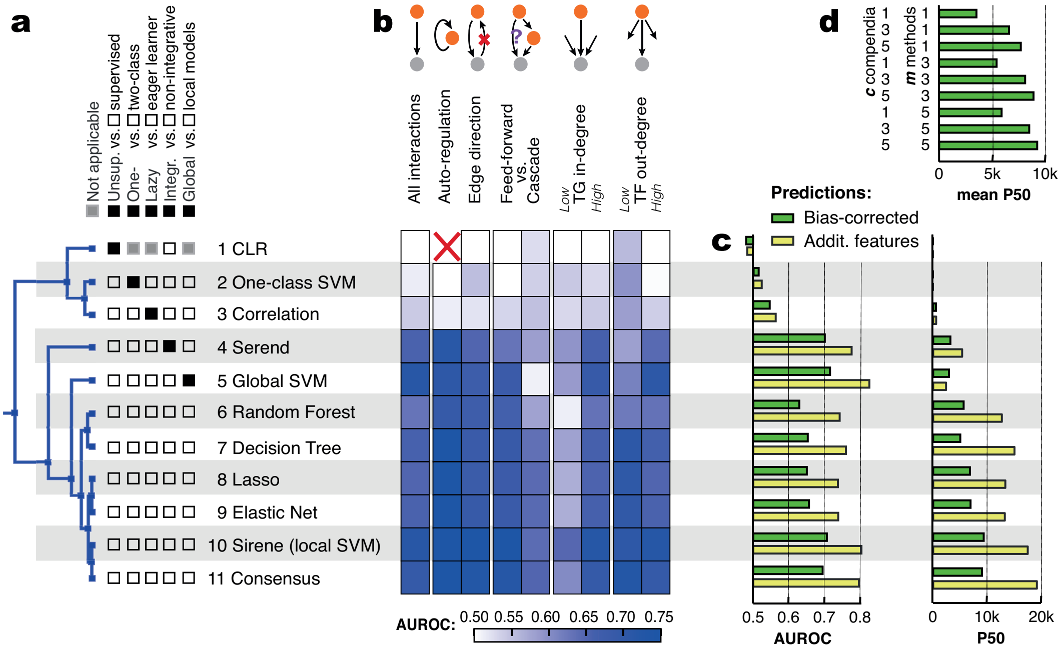


Figure 5.7: We analyzed the predictions of 11 different inference methods across five yeast gene expression compendia. **(a)** The dendrogram groups methods according to the similarity of their predictions. Properties that discriminate between different classes of methods are indicated by the check boxes. **(b)** shows if interactions in particular network motifs are easier (blue) or harder (white) 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) (green). Furthermore, we encoded experimentally determined targets of TFs into additional features (yellow, see Section 5.4.4 and 5.6.13) 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.

5.6.6 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 11,593 $(24462 - 12869)$ reference standard interactions (see Figure 5.1a) lacked an observable effect on expression and were 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 5.8 and Section 5.5.9).

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. Representative genes were 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 5.8). According to GO [234], most of the 190 genes of the *hxt2* cluster belong either to sugar transport (*hxt* genes) or glycogen metabolic process (*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 [186], whereas under glucose starvation glycogen storage is catabolized to produce glucose preferably for fermentation [76].

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 [155].

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 [170], initiated by the *hxt2* cluster via increased glucose uptake.

5.6.7 Novel predictions

In the following, we briefly describe examples (i) for novel predictions missing in current gold standards (the activation of *cat2* and *tes1* by *pip2/oaf1* and *adr1*) as well as (ii) for an interaction contained in the gold standard not supported by our predictions (the regulation of *hap4* by *cat8*). This latter interaction may be an example for a 'quiet' interaction not associated with expression changes of the target.

Genes involved in peroxisomal beta-oxidation in *S. cerevisiae* are repressed in the presence of glucose, de-repressed on non-fermentable carbon sources such as ethanol, and further induced by more than ten-fold in the presence of oleate [103]. Examples of gene products involved in the breakdown of fatty acids include *pot1*, *pox1*, *fox2*, *sps19*, and *cta1*. The transcriptional up-regulation of these genes is driven by the *pip2/oaf1* transcription factor, binding to the oleate response element (ORE), and by *adr1*, binding to another upstream activating site, UAS1 [118]. *Cat2*, a carnitine O-acetyltransferase, and *tes1*, an acyl-CoA thioesterase are also enzymes involved in fatty acid breakdown, currently postulated to be regulated by *pip2/oaf1* [118]. We predicted that the transcription of *cat2* and *tes1* is also activated by *adr1*, which has not been reported before (or only indirectly as for *tes1* [226]) but seems plausible given the known regulation of beta-oxidation genes by *pip2/oaf1* and *adr1*.

Cat8 and *hap4* are major transcriptional regulators of the diauxic shift [217]. *cat8* especially activates the transcription of gluconeogenic genes via binding to a carbon source responsive element (CSRE) in their promoter. *Cat8* itself is transcriptionally regulated in dependence on the carbon source, where positive regulation on non-fermentable carbon sources is carried out by the *hap2/3/4/5* complex [237]. *Hap4* is the activator subunit of the *hap2/3/4/5 complex*, especially driving the expression of genes involved in respiration and the TCA-cycle. *hap4* is also the regulatory subunit of the complex, as it is the only one whose level is regulated by the carbon source itself. Interestingly, it seems that *hap4* and *cat8* are mutually activating each other, as *hap4* transcription has been shown to be *cat8*-dependent [27]. In our network, the regulation of *hap4* by *cat8* was not predicted. This is in agreement with current studies, which assign the carbon source dependent regulation of *hap4* rather to *rds2* [237].

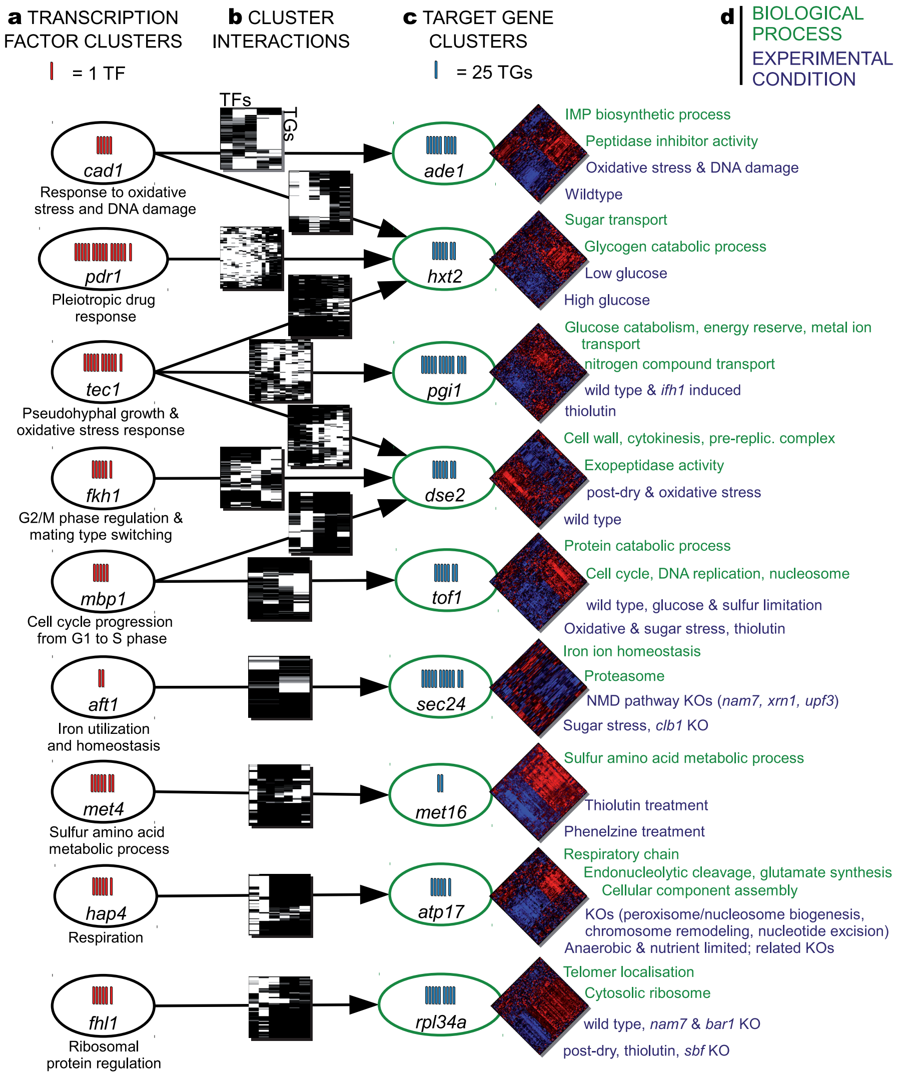

a TRANSCRIPTION FACTOR CLUSTERS
= 1 TF
b CLUSTER INTERACTIONS
c TARGET GENE CLUSTERS
= 25 TGs
d BIOLOGICAL PROCESS
EXPERIMENTAL CONDITION
TFs
TGs
cad1
Response to oxidative stress and DNA damage
pdr1
Pleiotropic drug response
tec1
Pseudohyphal growth & oxidative stress response
fkh1
G2/M phase regulation & mating type switching
mbp1
Cell cycle progression from G1 to S phase
aft1
Iron utilization and homeostasis
met4
Sulfur amino acid metabolic process
hap4
Respiration
fhl1
Ribosomal protein regulation
ade1
hxt2
pgi1
dse2
tof1
sec24
met16
atp17
rpl34a
IMP biosynthetic process
Peptidase inhibitor activity
Oxidative stress & DNA damage
Wildtype
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
Iron ion homeostasis
Proteasome
NMD pathway KOs (nam7, xrn1, upf3)
Sugar stress, clb1 KO
Sulfur amino acid metabolic process
Thiolutin treatment
Phenelzine treatment
Respiratory chain
Endonucleolytic cleavage, glutamate synthesis
Cellular component assembly
KOs (peroxisome/nucleosome biogenesis, chromosome remodeling, nucleotide excision)
Anaerobic & nutrient limited; related KOs
Telomer localisation
Cytosolic ribosome
wild type, nam7 & bar1 KO
post-dry, thiolutin, sbf KO

Figure 5.8: **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. A representative gene is displayed for each TF and target cluster. Cluster interaction maps (black=interaction, white=no interaction) comprise a total of 11232 (50.5%) interactions. **(d)** Thus, depicted TF modules are likely to trigger expression responses (heatmaps: red=up-, blue=down-regulation) in respective target modules and associated biological processes (green annotation). The heatmaps display the differential expression of these target modules under the indicated knockout (KO) and other experimental conditions (blue annotation).

5.6.8 Performance based on Binding Sites

SEREND trains classifiers for each TF individually and is based on local models for prediction. For the application to yeast, we used positional weight matrices (PWMs) obtained from the JASPAR database [29] and derived PWM promoter matching scores via CUREOS [203]. As detailed in Section 5.4.3, SEREND separately trains two logistic regression classifiers to predict GRIs from expression data and TF promoter binding sites, respectively. A third classifier is employed to combine the predictions from the other two classifiers.

Table 5.3: Performance (AUC) of SEREND across TFs

Data	Micro	Macro	Corrected
Motif	79.6	56.9	59.7
Expression	79.3	66.2	61.1
Combined	80.4	61.2	65.5

SEREND's confidence scores for putative GRIs are reported for each of the three classifiers, which enabled us to separately evaluate the performance. Large difference in performance between regulator-wise and network-wide quality measures (Table 5.6) suggest that SEREND would preferentially attach novel regulations to larger regulators. Table 5.3 indicates that each of the individual scores is susceptible, as shown by inflated micro-averaged AUC values.

We next analyzed the TF-specific performance achieved using only the information on binding sites. Table 5.4 demonstrates the strong shift towards new targets for high-degree TFs. The two TFs ($ste12$, $rap1$) with the highest out-degrees exhibit the lowest AUC performance but account for 80% of the predictions. This shows that the networks estimated by SEREND may profit from a reduction in False Discoveries by score recalibration. Table 5.5 depicts the results after recalibrating SEREND's sequence binding scores using *CoRe*. After the recalibration, the predictions are balanced with respect to TF out-degree (compare Figure 5.11). No significant predictions were obtained for $ste12$, indicating that predictions achieved before were independent of the regulator-specific model and entirely due to *HDP*. However, even after recalibration, suitable numbers of targets were predicted (empirically, we required that the number of targets predicted for a given TF should be $>$ 10% of its out-degree) for only 10 out of 153 (6.5%) TFs while no or very few (as in case of $fkh1$) targets were predicted for the majority of TFs.

Table 5.4: Examples for some regulator-specific performance using only promoter binding information

TF orf	Gene	Outdegree	Predicted	AUC
YHR084w	*ste*12	1770	1609	53.4
YNL216w	*rap*1	1159	736	67.7
YJR060w	*cbf*1	313	157	86.7
YBR049c	*reb*1	502	151	87.7
YKL112w	*abf*1	459	94	86.0
YDR207c	*ume*6	166	70	83.3
YEL009c	*gcn*4	284	46	79.9
YOL028c	*yap*7	174	18	84.8

Table 5.5: TF-specific performance of promoter binding after recalibration

TF orf	Gene	Out-degree	Predicted	AUC
YKL112w	*abf*1	459	532	86.0
YJR060w	*cbf*1	313	361	86.7
YEL009c	*gcn*4	284	237	79.9
YBR049c	*reb*1	502	198	87.7
YOL028c	*yap*7	174	129	84.8
YGL131c	*snt*2	23	79	93.2
YDR207c	*ume*6	166	75	88.3
YMR043w	*mcm*1	238	60	69.8
YBL005w	*pdr*3	107	24	75.5
YKL109w	*hap*4	159	15	65.8
YIL131c	*fkh*1	207	13	71.8

5.6.9 Robustness of *CoRe*

As described in Section 5.5.6, the proposed recalibration, *CoRe*, is based on random transcriptional networks. Figure 5.9 shows how the number of used random networks influences the resulting evaluation metrics. Shown are the results obtained from all possible subsets of the ten generated random networks used in this study. Using a larger number of random networks boosts the scores and decreases their variance. The differences in averaged performance estimates decrease as more random networks are used, indicating ten networks enable a sufficiently accurate recalibration. In addition, the results from evaluation metrics without recalibration are shown, demonstrating the substantial over-estimation of performance.

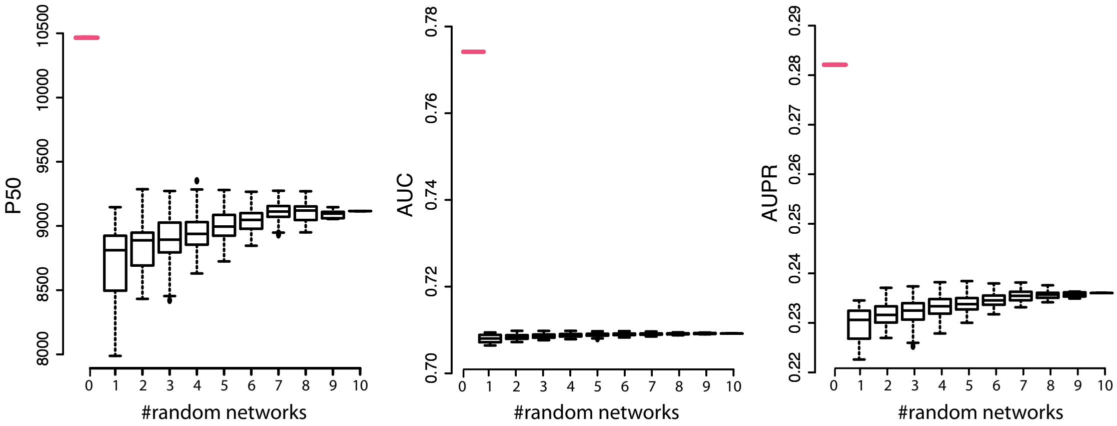


Figure 5.9: **Robustness of score recalibration** The boxplots depict how the number of random transcriptional networks (abscissa) used for the recalibration influences the results from various evaluation metrics (ordinate) including the AUC, the AUPR and the P50 measures. For comparison, we also show the results obtained without score recalibration (red bar, based on no, *i.e.*, 0, random networks).

5.6.10 Dependency of Distribution Parameters on TF Out-Degree

In this section we further examine the properties of score distributions and their dependence on the TF out-degree. Figure 5.10 depicts the dependency of distribution parameters on the TF out-degree. Both median and maximum of the score distributions exhibit a strong positive correlation with respect to the TF out-degree. Across the range of TF out-degrees, the maximum shows a higher slope than the median. This indicates that the score distribution is not only shifted but also scaled in dependence on the out-degree. As shown in the next section, a threshold on the non-recalibrated scores will therefore select more targets for TFs for which many targets are assigned by the gold standard.

5.6.11 Relationship between TF Out-Degree and the Number of Predicted Targets

GRIs are typically selected by applying a precision-based threshold (P50 for a precision of 50%) on a global list of predictions ranked by confidence scores [171]. In case of non-recalibrated scores, Figure 5.11 shows that thereby, an excessive number of predictions are selected for high-degree TFs while all predictions may be rejected in case of low-degree TFs. The P50 threshold can also be applied to each TF individually (corresponding to a macro-evaluation), but this leads to similar results. In contrast, a global P50 criterion applied to calibrated confidence

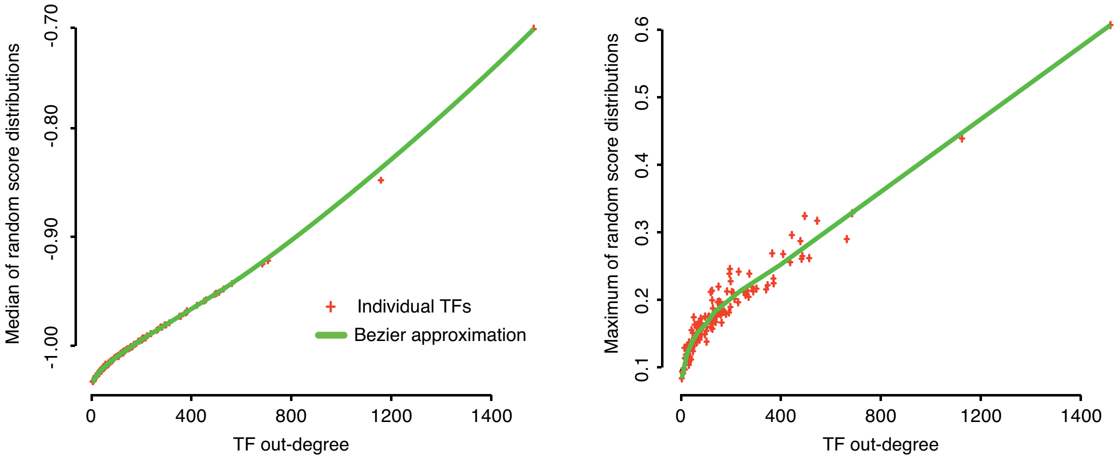


Figure 5.10: **Degree dependency of location parameters.** For each TF, simple location parameters are estimated such as the median (left panel) and maximum (right panel) from score distributions derived from random gold standards. The plot depicts the dependencies of these parameters (ordinate) on the TF out-degree (abscissa). Shown as red crosses are the parameters as estimated from individual TFs and their approximation via Bezier curves (green line). Score distributions were derived from local SVM models obtained from the Sirene approach.

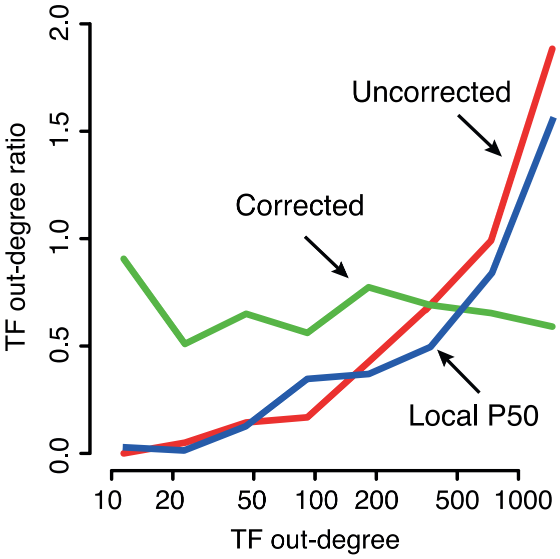


Figure 5.11: **Influence of TF out-degree on the number of predictions.** The ratio of predicted to gold-standard targets (ordinate) is depicted across the range of TF out-degrees (=number of gold-standard targets, abscissa) before (red) and after (green) recalibration with *CoRe*. Using raw confidence scores, TFs with many targets in the gold standard would receive an overly large number of newly predicted targets. Here, we select the highest scoring targets across all TFs such that a precision of 50% (P50 criterion) is obtained. As an alternative that corresponds to macro-evaluation, the P50 criterion is applied to each TF individually (local P50, blue).

scores results in a balanced ratio of predicted to known TF targets, *i.e.*, data points in Figure 5.11 are parallel to the abscissa.

5.6.12 Score Distributions Based on Probability Estimates

As an alternative to the raw confidence scores employed by methods such as SIRENE [171], Platt scores have been proposed [120]. Platt scores transform the raw confidence scores into probability estimates that scale between 0 and 1. As shown in Figure 5.12, Platt scores derived from randomized gold standards exhibit similar degree dependencies as the raw confidence scores depicted in Figure 5.6a (page 104). Thus, the transformation into Platt scores alone is not sufficient to correct for the *HDP* effect.

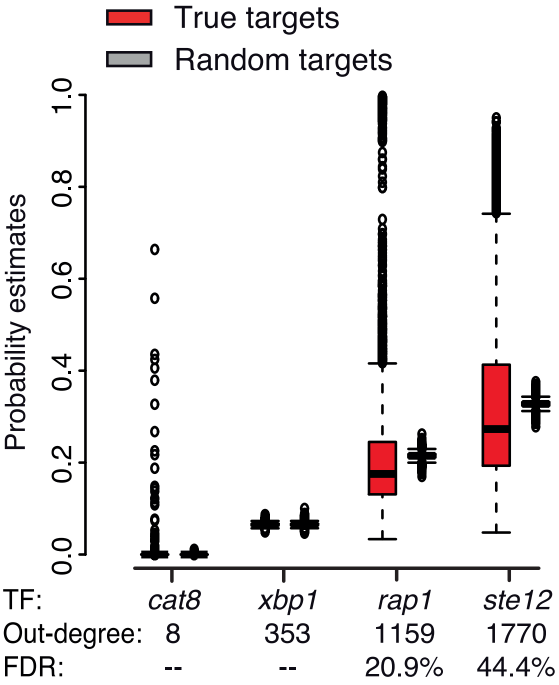


Figure 5.12: **Degree dependency of probability estimates.** Support vector machines were trained and applied as described, but resulting confidence scores were transformed to probability estimates (also referred to as Platt scores) via the *libsvm* SVM library [37]. The probability estimates exhibit increased means and variances in case of confidence score distributions derived for high out-degree TFs. For each TF, the distribution of confidence scores is displayed in a left boxplot for true targets (red) and in a right boxplot for random targets (grey, not visible due to small variance).

5.6.13 Improving Regulator-Specific Predictions

In order to increase the number of correctly predicted interactions, we implemented three improvements. First, we integrated the five methods selected in the previous section into a consensus to obtain a single network (see Section 5.4.4). This integration is potentially beneficial to exploit complementary advantages of different methods [160]. We re-ranked interactions according to the average calibrated score across all methods and selected the top-ranking interactions with a precision of 50% or better. This consensus spanned a network of 8,726 predicted interactions (see Figure 5.7c, page 106). To examine compendia-specific effects, we built consensus networks from predictions derived from subsets of expression compendia and subsets of methods (see Figure 5.7d). The integration of further methods or further compendia generally led to an increased performance.

The second improvement is motivated by the fact that genes are frequently regulated by more than one TF and that several (often functionally related) TFs regulate overlapping sets of targets [207]. Local methods predict targets for a single TF at a time and cannot take such combinatorial regulation into account. We therefore encoded the set of known regulators of a gene as additional training data (see Section 5.4.4). The true targets of each modeled regulator are excluded from the training to avoid over-fitting. We observed that the explicit encoding of known regulations roughly doubled the number of P50 interactions, yielding 18,724 interactions (see Figure 5.7c). This corresponded to a threshold on the κ confidences of 0.92.

Finally, we aimed to include gold-standard interactions predicted with moderate confidence. We therefore extended the predicted network by gold-standard regulations that met a relaxed confidence threshold of 0.46 (compare P50=0.92, see Figure 5.6c). The fact that gold standard interactions have been determined experimentally provides an increased confidence, justifying the relaxation of the threshold. This further increased the size of our final yeast network to 22,231 interactions containing 153 TFs and 3,747 target genes. Of all predicted regulations, 12,869 are contained in the gold standard while 9,362 are novel predictions. Among the 29,398 gold standard interactions (Figure 5.1a), even by the already relaxed threshold, more than half (56.2%) were not confirmed by our approach. These 'quiet' interactions apparently have no regulatory effect visible in our data.

5.6.14 Numerical Values of Performance Estimates

As a reference, the Tables 5.6 (gold standard, recalibrated) and 5.7 (randomized networks, recalibrated) provide exact values of the recalibrated network performance measures as depicted by the green bars in Figure 5.7c. When topology features are included explicitly we obtain the values shown in Tables 5.8 and 5.9, respectively. The associated values shown are:

- auc := 'area under the receiver operator characteristics curve'
- aupr := 'area under the precision recall curve'
- fmb := 'optimal f-measure for variable threshold'
- p50 := 'number of predictions for a precision of 50%'

In addition, these tables summarize various evaluation approaches and contain further performance estimates such as the F-measure. We compare several evaluation setups, including micro- *vs.*macro-averaging, the influence of additional training features encoding known regulators, raw *vs.*recalibrated confidence scores as well as experimentally derived gold standard *vs.*random networks. See Section 5.5.4 for details on the scores and their computation.

Due to the long run-time, the random networks were not processed via CLR and thus, calibrated scores were not computed. Note that the consensus is constructed from the five approaches employing local models, namely Random forest, Decision tree, Lasso, Elastic net and local SVM, which corresponds to the SIRENE approach.

Even without recalibration, macro-evaluation takes a regulator-wise viewpoint and enables sensible local performance estimation that may complement the network-wide point of view. Detailed results are shown in Tables 5.6 and 5.8. However, macro-evaluation does not provide a mechanism to select interactions from a wide range of degrees. Due to the degree dependency of confidence scores the resulting networks will preferentially consist of larger regulators (compare Section 5.6.11).

Table 5.6: Evaluation for fold-change expression features (measure definition, see 5.6.14).

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmb	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
CLR	47.9	2.8	5.7	14	52.0	3.7	7.8	-	-	-	-	-	-	-
One-class SVM	50.3	6.5	12.5	0	50.9	4.0	8.3	51.4	3.6	6.4	0	50.9	4.0	8.3
Correlation	54.5	5.1	7.0	514	54.9	5.6	10.4	54.5	5.1	7.1	456	54.9	5.6	10.4
SEREND	81.5	22.4	28.6	3684	60.8	12.5	18.6	70.4	21.1	28.5	3332	61.2	12.9	19.2
Global SVM	81.8	21.2	29.1	2692	67.6	11.4	17.3	71.6	15.8	21.3	3054	67.6	11.4	17.3
Random forest	79.6	24.0	28.2	6996	63.1	11.4	17.8	62.9	17.9	23.1	5554	63.1	11.4	17.8
Decision tree	79.3	19.5	23.8	4110	66.1	13.3	20.1	65.3	17.0	23.1	4426	66.1	13.3	20.1
Lasso	81.4	25.7	29.9	8000	67.5	13.5	20.0	65.0	19.7	26.5	6490	67.4	13.5	20.1
Elastic net	81.6	25.9	29.9	8090	67.8	13.7	20.3	65.5	20.1	26.7	6576	67.8	13.7	20.3
Local SVM	78.4	28.2	33.5	10466	68.3	16.2	23.2	70.9	23.6	29.8	9130	68.3	16.2	23.2
Consensus	82.2	28.7	32.3	9918	68.5	15.2	21.8	69.3	23.5	29.3	8726	68.5	15.8	22.8

Table 5.7: Evaluation for random networks on fold-change features.

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmb	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
Random forest	72.9	9.9	16.1	0	49.4	3.1	6.4	47.8	3.5	7.5	0	49.4	3.1	6.4
Decision tree	69.9	5.2	10.1	0	49.6	3.1	6.3	48.4	3.2	6.4	0	49.6	3.1	6.3
Lasso	71.4	7.8	14.2	0	49.1	3.1	6.4	47.7	3.9	8.9	0	49.1	3.1	6.4
Elastic net	71.6	7.9	14.5	0	49.0	3.0	6.2	48.0	4.1	9.4	0	49.0	3.0	6.2
Local SVM	67.2	9.4	16.0	20	49.8	3.1	6.4	50.2	3.1	5.8	0	49.8	3.1	6.4

Table 5.8: Evaluation results for features extended by topology information.

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmb	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
CLR	46.9	2.8	5.7	6	38.0	3.2	6.7	-	-	-	-	-	-	-
One-class SVM	51.0	6.8	12.9	14	52.2	4.4	9.0	52.5	3.8	6.8	8	52.2	4.4	9.0
Correlation	56.7	5.5	7.7	504	59.4	9.3	14.6	56.7	5.5	7.7	480	59.4	9.3	10.4
SEREND	81.5	22.4	28.6	3684	60.8	12.5	18.6	78.5	24.3	31.6	5364	67.3	18.5	25.6
Global SVM	87.7	27.4	36.3	2630	79.1	17.8	23.3	82.7	20.5	26.2	2692	79.1	17.8	23.3
Random forest	85.4	34.3	37.8	13374	74.2	20.3	27.4	74.2	30.5	36.2	12990	74.2	20.3	27.4
Decision tree	86.1	36.4	39.3	15188	76.2	26.4	33.0	75.7	34.3	39.3	15586	76.2	26.4	33.0
Lasso	86.8	37.1	39.5	15030	77.6	23.2	30.5	73.7	30.6	37.1	13382	77.6	23.2	30.5
Elastic net	86.8	37.0	39.4	15046	77.8	23.5	30.9	73.8	30.6	37.1	13160	77.8	23.5	30.9
Local SVM	85.0	38.8	42.4	17520	77.9	28.3	35.5	80.3	36.8	41.6	17194	77.9	28.3	35.5
Consensus	88.1	42.0	43.3	18472	79.5	27.6	34.1	79.3	39.9	43.5	18724	79.5	29.8	36.5

Table 5.9: Evaluation results for features on random networks extended by topology information.

	Micro				Macro			Micro, calibrated				Macro, calibrated		
Method	auc	aupr	fmf	p50	auc	aupr	fmb	auc	aupr	fmb	p50	auc	aupr	fmb
Random forest	73.0	9.9	16.1	0	49.6	3.1	6.4	48.2	3.6	7.7	0	49.6	3.1	6.4
Decision tree	70.4	5.3	10.3	0	49.9	3.1	6.3	48.7	3.3	6.6	0	49.9	3.1	6.3
Lasso	71.6	8.0	14.5	2	49.1	3.0	6.2	47.9	4.0	9.0	0	49.1	3.0	6.2
Elastic net	71.8	8.1	14.8	0	49.1	3.1	6.2	48.1	4.1	9.3	0	49.1	3.1	6.2
Local SVM	67.8	9.5	16.2	26	50.0	3.1	6.4	49.8	3.1	5.8	0	50.0	3.1	6.4

5.7 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 of these approaches, but also demonstrated that data-driven inference is both necessary and feasible in eukaryotes.

Even in well-studied eukaryotes such as yeast, where roughly 900 publications on experimental TF-binding studies are available, current networks are far from complete and they 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 for both the detection of novel active regulations and the pruning of inactive regulations.

We reported three crucial findings based on the analysis of a wide spectrum of data-driven inference methods [51, 174]. 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 [160, 176]. We also showed that topology priors lead to Simpson's paradox [191, 224] distorting the 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 [51, 5], and it has been doubted whether a correction is possible or sensible [174, 88]. 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 we identified methods suitable to derive a corrected, accurate yeast regulatory network of active regulations. We describe disadvantages of several methods, which were omitted in downstream experiments due to their poor 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 any 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*.

In this chapter we discussed the use of target correlations by means of pattern detection and supervised inference to predict novel regulations from expression data. Existing topological primers are inherently uncertain, due to a lack of context specificity and experimental noise. *CoRe* in combination with local models allows to adapt a given network to specific data. This is achieved either via filtering regulations that are not supported by the data at hand or by augmentation of missing regulations, leading to highly specific and noise-reduced networks.

Chapter 6

Conclusion and Outlook

6.1 Complex Systems

In this work, we developed and applied several methods to detect and to make use of complex correlation patterns in complex systems. We discuss their formation from complex bio-molecular systems and procedural setups that lead to potential biases. Our particular focus was to overcome the degree of false positive discoveries in these contexts and provide sensible evaluation routines. Finally, we have developed and discussed competitive approaches for the prediction of protein contacts, artifact repair, differential expression analysis, and network inference and show their applicability in practical setups.reliable predictions in a real-world setting.

6.1.1 Co-Evolution and Correlated Mutations

Knowledge of protein contacts is essential to analyze signal transduction pathways. Within pathways, important interfaces among two proteins are often conserved and thus, subject to correlated mutations that maintain them. In Chapter 2, we used an alignment-based similarity definition to compute pairwise correlations from multiple-sequence alignments covering signal-transduction entities of bacterial strains. Coupled mutations can be exploited if they are conserved across these strains or if they happen simultaneously. Other mutations are treated as noise. They are unexplained in this context and counteract directed evolution. We found that our approach could successfully separate contacting from non-contacting proteins and provide evidence for an important bio-molecuar pathway of *bacitracin* resistance [55].

6.1.2 Deviation Patterns

The detection of genes that are differential across two or more conditions is an important step to identify the key players of diseases. Yet, many candidate genes are unspecific to the condition that is examined. Their regulation may correlate with inflammatory responses, cell maintenance or the experimentally induced synchronization of specimen. Therefore, the resulting gene lists may be diluted by false positive candidates and subsequent analyses like gene set enrichment may be misleading. We developed *Padesco* to detect frequent correlations and to select candidate genes that are specific for an experiment (see Chapter 4). The method is capable of predicting deviations from expected behaviour for individual candidate genes. The expected measurements are used to obtain experiment-specific deviation scores via a robust score transformation. A key feature of *Padesco* is the possibility to encode correlation patterns as predictive models. We applied robust regression models to deal with the noise inherent to expression data. Overall, *Padesco* provides effective means to reduce false positive candidates when dealing with differential experiments.

6.1.3 Removal of Noise and Artifacts

Microarrays have become a standard and integral part of any wet lab. They are cheap and therefore often complement other experimental setups. Yet, recent studies have shown that a large amount of published microarrays contains errors like stains or blotches (see Chapter 3, [193]). This may lead to false positive results. Many artifacts affect distinct array regions. Affected measurement values would therefore correlate in the affected array region. In Chapter 3, we exploited this observation and developed an approach to detect artifacts using a sliding-window approach. We provided an effective imputation procedure to replace corrupted probes. The replacement is promising for setups where otherwise the experiment design would become invalid, in particular if few specimen are available and they cannot be replaced. We could show the practical use of our method in a veterinary medicine context studying the impact of conjugated trans-fatty acid nutrition in dairy cattle [47, 140]. Slaughtering experiments of live stock are time-consuming and difficult to repeat, wherefore our imputation approach enabled an integral analysis of all available arrays [140].

6.1.4 Network Evaluation

As part on the work on this thesis, we successfully took part in two rounds of DREAM challenges [160] for the inference of both dynamic small-scale [144] and static large-scale networks on real-world data [143]. Actually, some of the most important results in this work are centered around the reconstruction of regulatory networks [192]. Chapter 5 summarizes the most essential lessons we learned from our earlier work and highlights important pitfalls of network inference in the presence of correlation structures.

In Chapter 5 we discussed how sub-group effects influence and confound network inference. We showed that combining individual sub-network predictions is non-trivial and may lead to false positives if the specific properties of predicted regions are neglected. We further discuss that the evaluation of combined networks results in over-estimates of network quality (see Section 1.3.4). We used numerous independent evaluation schemes and appropriate visualizations to highlight these problems. We therefore devised *CoRe*, a conceptually simple method to reduce false positive predictions for the sensible combination of predicted sub-networks. To aid the biological interpretation of networks in inference contexts, we introduced a measure of functional coherence based on pre-defined ontologies (see Section 5.5.8) and a novel way of modular network visualization (see Section 5.6.6).

It is clear that there exists neither a single model that is correct nor a perfect network that is suitable for all contexts. Yet, we could improve the quality of previous predictions via the removal of hidden correlations and by employing evaluation metrics sensitive to False Negatives. *CoRe* can be used to sensibly integrate sub-network predictions that provide a confidence score for each regulation. This holds true for most current predictive models. As *CoRe* wraps these models the prediction algorithm remains interchangeable. Regardless of the applied model, we could show a strong positive effect on both the True Positive Rate and the False Negative Rate. We further improved the result network by computing a consensus network. Here, regulators with few targets known beforehand could achieve remarkably more and accurate target predictions.

6.2 Outlook

In the future, it seems promising to use *CoRe* as a network filter as well. Through targeted training experiment sets – similar to those used for *Padesco* (see Chapter 4) – regulations that are unexpected for a trained network context could be removed. This may lead to context-specific networks that are semi-automatically derived by the selection of certain ontologies and their associated experiments.

In this work, we have developed promising ways to reduce False Positive results for expression data analysis and network inference. We showed that correlation patterns provide powerful contexts and that they may help to reduce misleading outcomes. With the advance of ever more automated analysis pipelines, the thorough evaluation of intermediate and final results is more essential than ever to prevent error propagation from becoming virtually intractable. The reduction of error, noise and bias will still demand for tailored solutions and the sensible contextualization of experimental data.

Chapter 7

Appendix

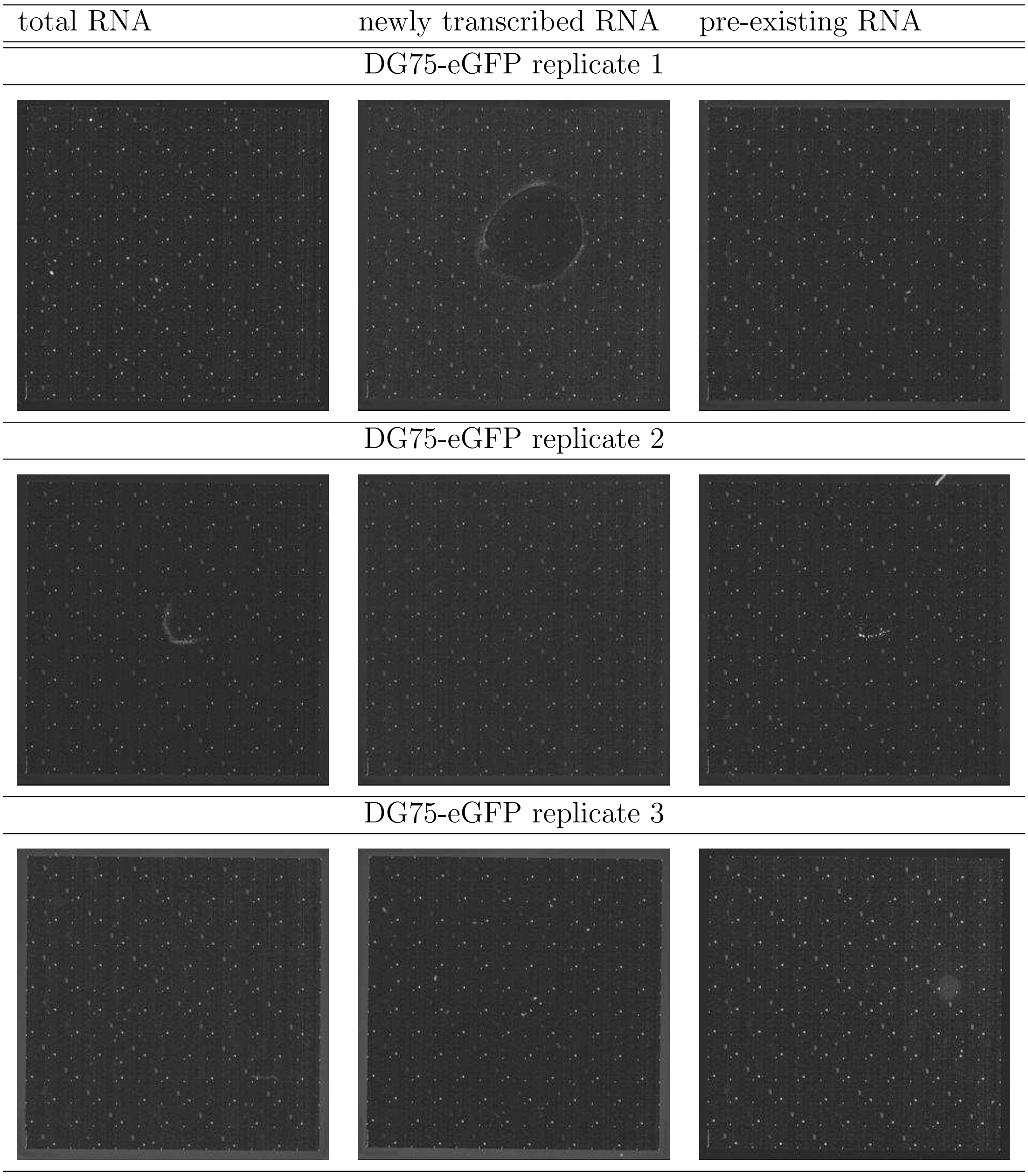


Figure 7.1: **GFP-labeled replicates**.Three fractions are measures, wheras the sum of pre-existing and newly transcribed RNA should sum to the total amount of RNA.

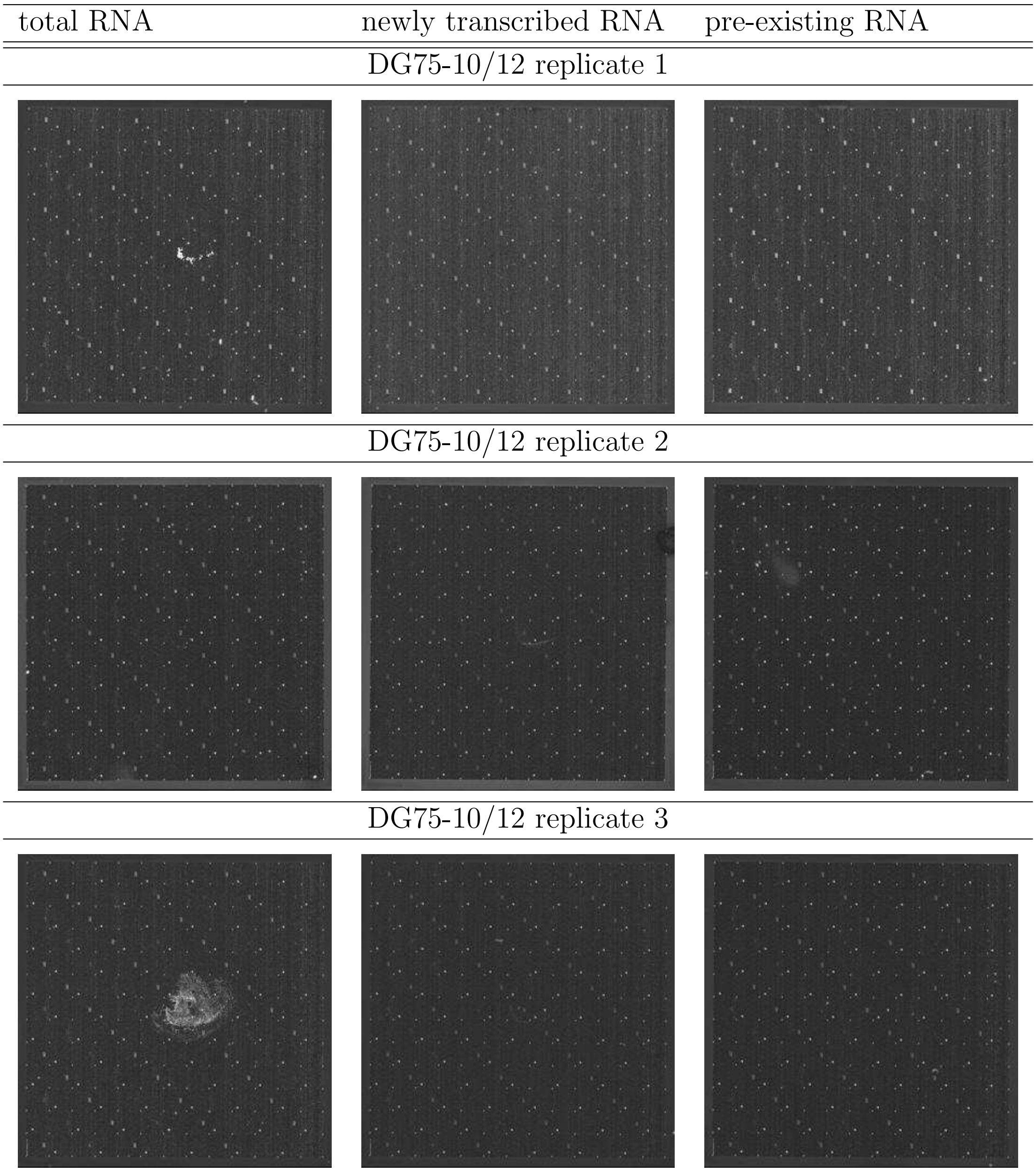
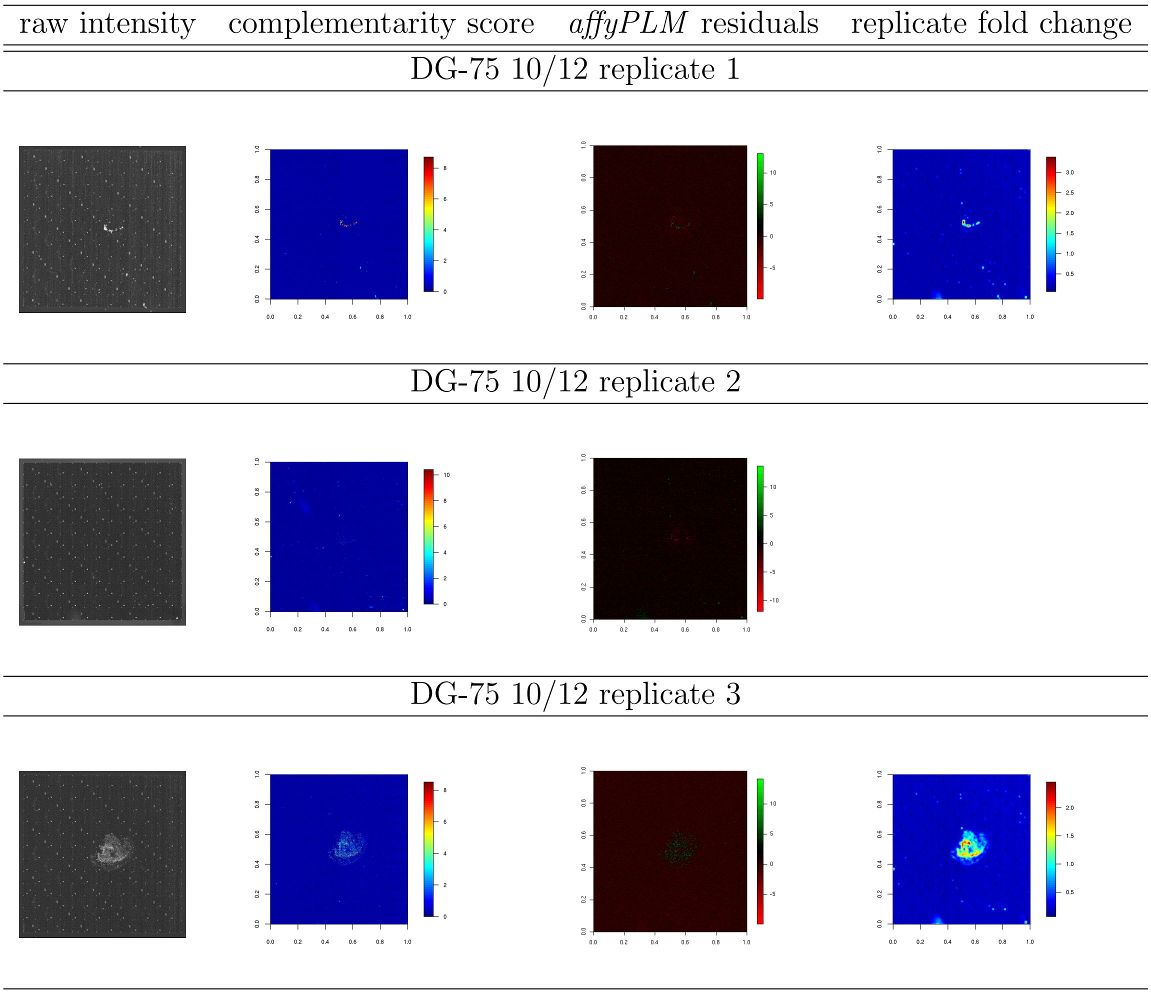


Figure 7.2: **10/12 replicates**. Three fractions are measures, wheras the sum of pre-existing and newly transcribed RNA should sum to the total amount of RNA.



raw intensity | complementarity score | *affyPLM* residuals | replicate fold change

DG-75 10/12 replicate 1

DG-75 10/12 replicate 2

DG-75 10/12 replicate 3

Figure 7.3: Probe noise scores for the three replicates of total RNA in DG75-10/12 cells measured with exon arrays. From left to right the columns are (1) raw intensities, (2) the noise score based on the fold-changes of total RNA to the sum of newly transcribed and pre-existing RNA (complementary score), (3) the *affyPLM* residuals and (4) noise scores based on the fold-change between replicates. Replicate 2 was used as control for case (4), therefore no probe noise plot could be created based on its replicate noise scores.

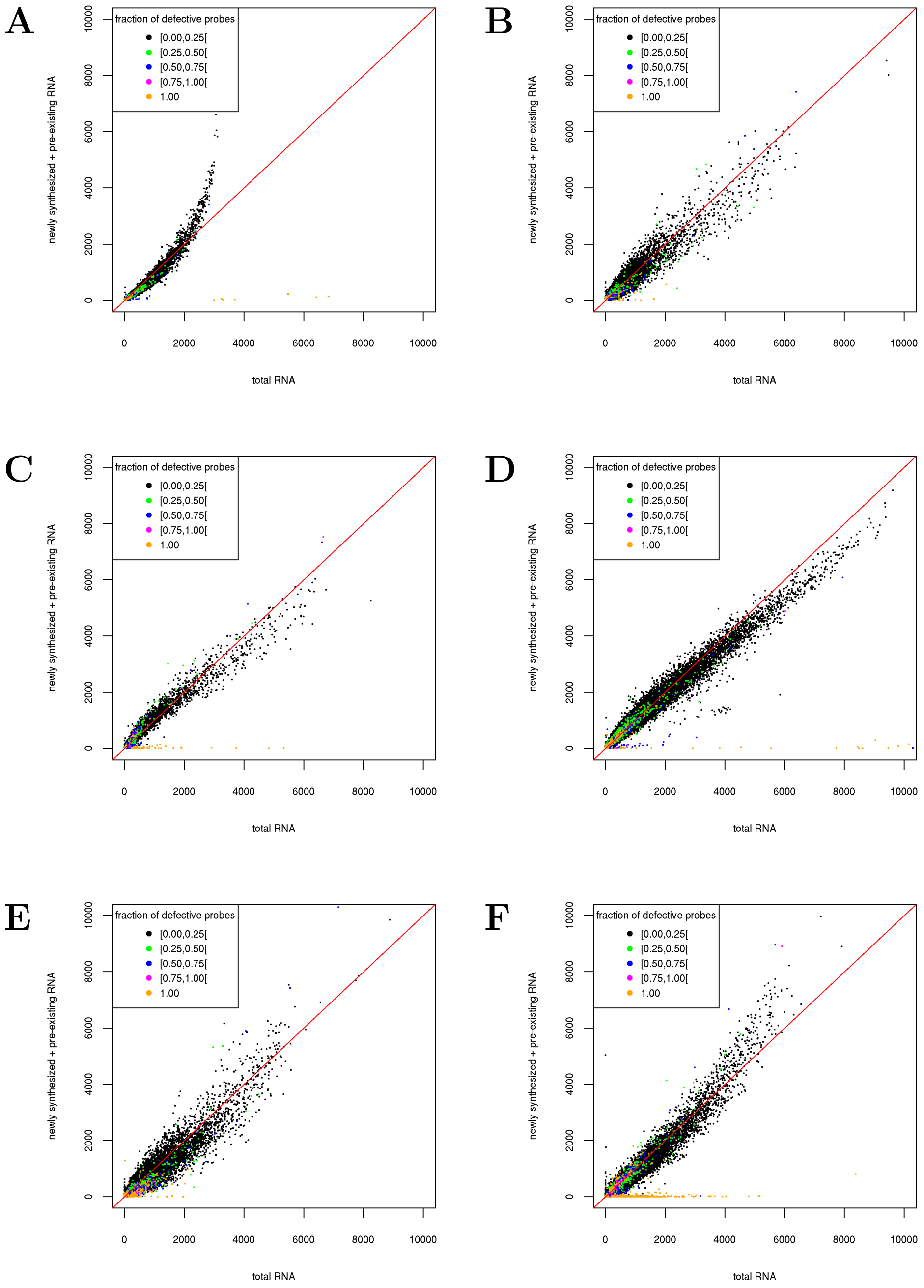


Figure 7.4: Replicate scatter plots comparing total RNA against normalized sum of newly-transcribed and pre-existing RNA for replicates 1 (A, D), 2 (B, E) and 3 (C,F) for the DG75-10/12 exon array measurements. Subfigures A-C show the results using both RMA and quantile normalization, D-F using only RMA without quantile normalization. Here, the y-axis shows the control for the total RNA measurements. Deviations are observed for the two replicates 1 and 3 containing artifacts.

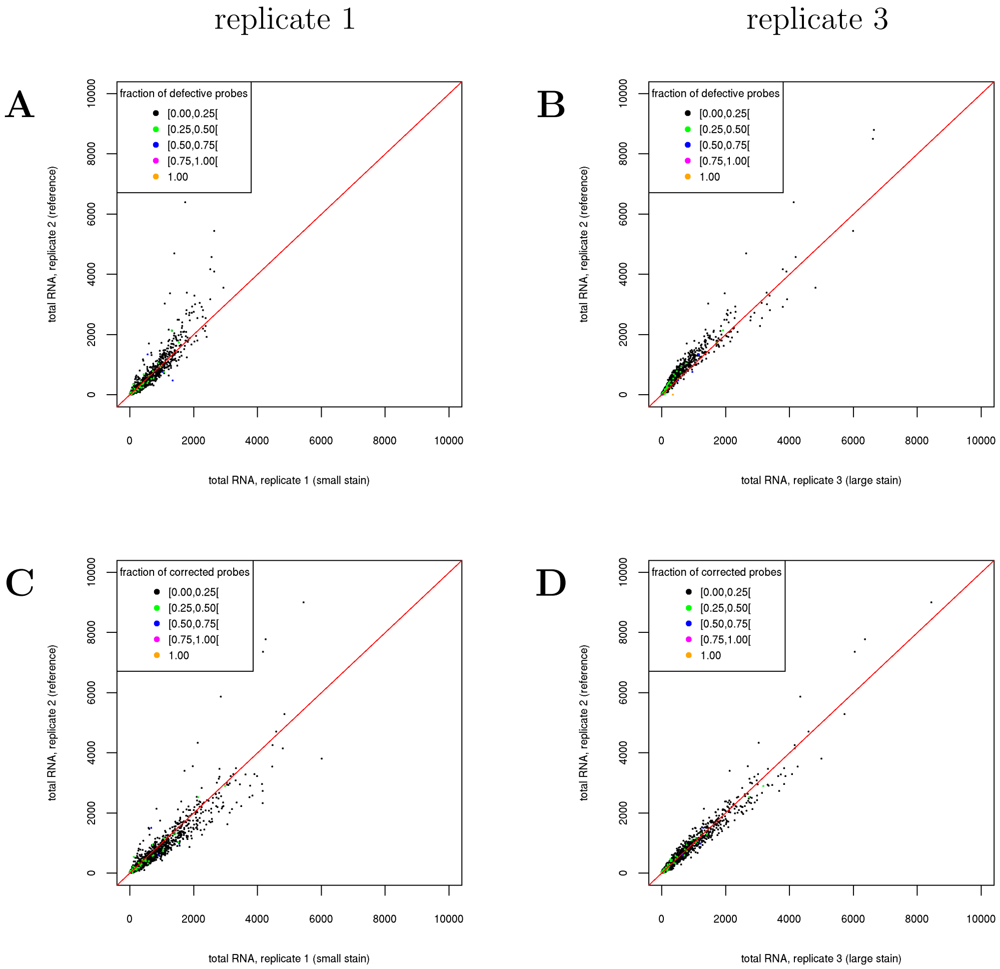


Figure 7.5: Replicate scatter plots for the DG75-10/12 cells summarized to metaprobesets before (A,B) and after probe correction (C,D) with the ϵ-criterion using probe noise scores calculated based on the complementarity of RNA fractions.

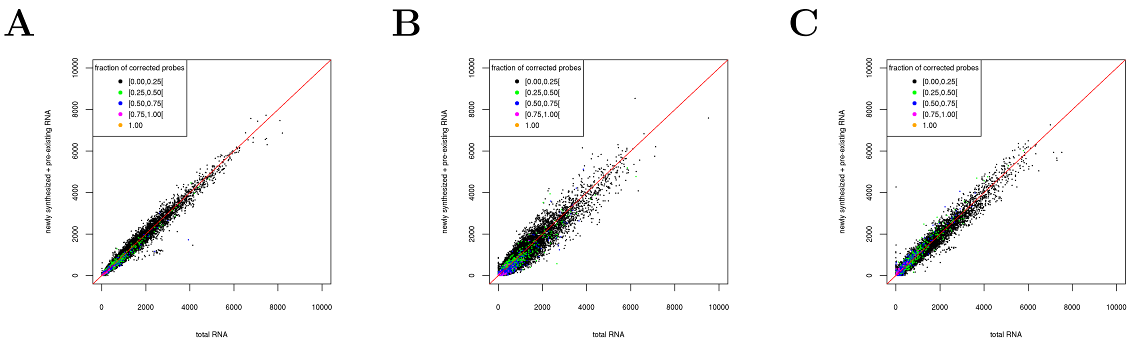


Figure 7.6: Replicate scatter plots for the DG75-10/12 total RNA measurements after probe correction using the ϵ-criterion with probe noise scores based on replicate fold-changes (A: replicate 1, B: replicate 2 and C: replicate 3). To avoid overfitting effects, the total RNA is scattered against the normalized sum of newly-synthesized and pre-existing RNA.

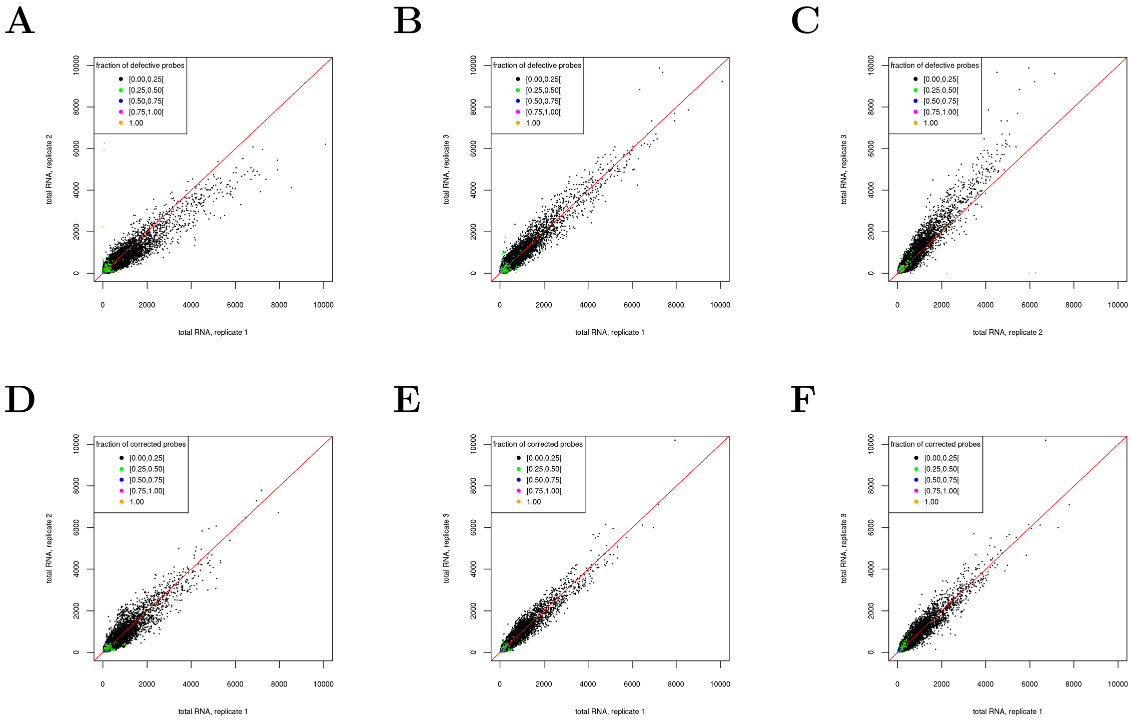


Figure 7.7: Replicate scatter plots for the DG75-eGFP cells before (A-C) and after (D-F) probe correction with the ϵ-criterion with noise scores based on the complementarity of RNA fractions. Here, replicates 1 and 3 were free of artifacts and replicate 2 showed a small stain which results in deviations for this replicate. Here, replicate scatter plots for each pair of replicates are shown. Accordingly, deviations are observed in all replicate scatter plots for which replicate 2 has been used (A,C) but not in the comparison of the artifact-free replicates 1 and 3 (B). After probe correction no such deviation is observed even for replicate 2.

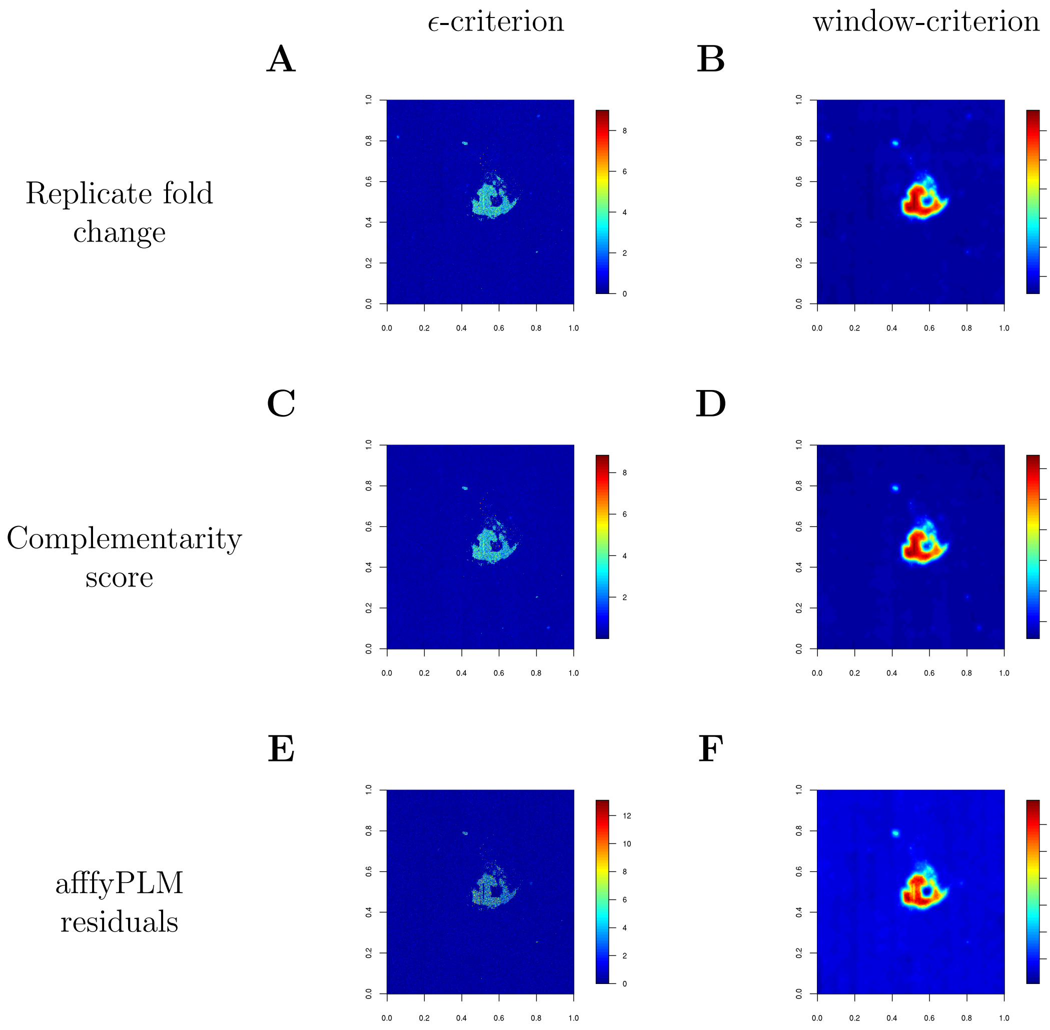


Figure 7.8: Probe noise plots for the spiked Gene ST arrays using the the probe scores based on the comparison of replicates (A-B), the comparison of total RNA against the normalized sum of newly transcribed and pre-existing RNA (Complementarity score) (C-D) and the affyPLM residuals (E-F).

Acknowledgements

This work would not have been possible without many other people, that I owe thanks to.

I thank Robert Küffner for many valuable lessons, discussions, and for becoming a close friend. I'd like to thank Ralf Zimmer, Gergely Csaba, Caroline Friedel, Evi Berchtold, Konrad Schreiber, Lukas Windhager, Florian Erhard, Simone Gigl, Frank Steiner, Franziska Schneider and all colleagues at the LFE Bioinformatik for creating the best working environment one could wish for and for great collaboration.

I would like to thank Ralf Zimmer for an open ear to all the needs a young scientist may have and his ever valuable feedback and constructive criticism. I thank Karsten Borgwardt and Ralf Zimmer for reviewing this work and inspiring my decisions as a scientist.

I thank my mother and parents-in-law for providing the opportunity to finalize this work and caring for our children during long days and nights of work.

Finally, I would like to thank my wife Marisa for constant support, patience and love. Without you I would not be what I am today.

Bibliography

[1] Abbas A, Wolslegel K, Seshasayee D, Modrusan Z and Clark H [2009]: 'Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus', *PLOS ONE*, **4**(7): e6098.

[2] Abdelmoula WM, Balluff B, Englert S, Dijkstra J, Reinders MJT, Walch A, McDonnell LA and Lelieveldt BPF [2016]: 'Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data.', *Proceedings of the National Academy of Sciences*, **113**: 12244–12249, ISSN 1091-6490, `http://dx.doi.org/10.1073/pnas.1510227113`.

[3] Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço AB, dos Santos SC, Cabrito TR, Francisco AP, Madeira SC, Aires RS *et al.* [2011]: 'YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface.', *Nucleic Acids Research*, **39**(Database issue): D136–D140, `http://dx.doi.org/10.1093/nar/gkq964`.

[4] Ackland ML, Zou L, Freestone D, van de Waasenburg S and Michalczyk AA [2007]: 'Diesel exhaust particulate matter induces multinucleate cells and zinc transporter-dependent apoptosis in human airway cells.', *Immunology and Cell Biology*, **85**(8): 617–622, `http://dx.doi.org/10.1038/sj.icb.7100109`.

[5] Ambroise J, Robert A, Macq B and Gala JL [2012]: 'Transcriptional network inference from functional similarity and expression data: a global supervised approach.', *Statistical Applications in Genetics and Molecular Biology*, **11**(1): 1–24, `http://dx.doi.org/10.2202/1544-6115.1695`.

[6] Anscombe FJ [1973]: 'Graphs in Statistical Analysis', *The American Statistician*, **27**(1): 17–21, ISSN 00031305, `http://www.jstor.org/stable/2682899`.

[7] Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR and Korsmeyer SJ [2002]: 'MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia', *Nature Genetics*, **30**: 41–7, `http://www.ncbi.nlm.nih.gov/pubmed/11731795`.

[8] Arteaga-Salas JM, Harrison AP and Upton GJG [2008]: 'Reducing spatial flaws in oligonucleotide arrays by using neighborhood information.', *Statistical Applications in Genetics and Molecular Biology*, **7**(1): Article29, http://dx.doi.org/10.2202/1544-6115.1383.

[9] Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJG and Harrison AP [2008]: 'An overview of image-processing methods for Affymetrix GeneChips.', *Briefings in Bioinformatics*, **9**(1): 25–33, http://dx.doi.org/10.1093/bib/bbm055.

[10] Balluff B, Hanselmann M and Heeren RMA [2017]: 'Mass spectrometry imaging for the investigation of intratumor heterogeneity.', *Advances in cancer research*, **134**: 201–230, ISSN 0065-230X, http://dx.doi.org/10.1016/bs.acr.2016.11.008.

[11] Bansal M, Belcastro V, Ambesi-Impiombato A and di Bernardo D [2007]: 'How to infer gene networks from expression profiles.', *Molecular Systems Biology*, **3**: 78, http://dx.doi.org/10.1038/msb4100120.

[12] Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA *et al.* [2003]: 'Computational discovery of gene modules and regulatory networks.', *Nature Biotechnology*, **21**(11): 1337–1342, http://dx.doi.org/10.1038/nbt890.

[13] Barla A, Mosci S, Rosasco L and Verri A [2008]: 'A method for robust variable selection with significance assessment.', in 'ESANN', pages 83–88.

[14] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM *et al.* [2011]: 'NCBI GEO: archive for functional genomics data sets–10 years on.', *Nucleic Acids Research*, **39**(Database issue): D1005–D1010, http://dx.doi.org/10.1093/nar/gkq1184.

[15] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA *et al.* [2009]: 'NCBI GEO: archive for high-throughput functional genomic data.', *Nucleic Acids Research*, **37**(Database issue): D885–D890, http://dx.doi.org/10.1093/nar/gkn764.

[16] Barzel B and Barabási AL [2013]: 'Network link prediction by global silencing of indirect correlations.', *Nature Biotechnology*, **31**(8): 720–725, http://dx.doi.org/10.1038/nbt.2601.

[17] Benjamini Y and Hochberg Y [1995]: 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1): 289–300.

[18] Benjamini Y and Yekutieli D [2001]: 'The control of the false discovery rate in multiple testing under dependency', *Annals of statistics*, pages 1165–1188.

[19] Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al.* [2004]: 'Global identification of human transcribed sequences with genome tiling arrays.', *Science*, **306**(5705): 2242–2246, http://dx.doi.org/10.1126/science.1103388.

[20] Bhattacharya B and Habtzghi D [2002]: 'Median of the p value under the alternative hypothesis', *The American Statistician*, **56**(3): 202–206.

[21] Bi R, Zhou Y, Lu F and Wang W [2007]: 'Predicting Gene Ontology functions based on support vector machines and statistical significance estimation', *Neurocomputing*, **70**(4-6): 718–725, http://dx.doi.org/10.1016/j.neucom.2006.10.006.

[22] Blainey P, Krzywinski M and Altman N [2014]: 'Points of significance: replication', *Nature Methods*, **11**(9): 879–880.

[23] Bleakley K and Yamanishi Y [2009]: 'Supervised prediction of drug-target interactions using bipartite local models.', *Bioinformatics*, **25**(18): 2397–2403, http://dx.doi.org/10.1093/bioinformatics/btp433.

[24] Bolstad B, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry R and Speed T [2005]: 'Quality assessment of Affymetrix GeneChip data', in Gail M, Krickeberg K, Samet J, Tsiatis A, Wong W, Gentleman R, Carey VJ, Huber W, Irizarry RA and Dudoit S (Editors), 'Bioinformatics and Computational Biology Solutions Using R and Bioconductor', volume 10 of *Statistics for Biology and Health*, pages 33–47, Springer New York, ISBN 978-0-387-29362-2, http://dx.doi.org/10.1186/1471-2105-10-445.

[25] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al.* [2001]: 'Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.', *Nature Genetics*, **29**(4): 365–371, http://dx.doi.org/10.1038/ng1201-365.

[26] Breunig MM, Kriegel HP, Ng RT and Sander J [2000]: 'LOF: identifying density-based local outliers', in 'ACM sigmod record', volume 29, pages 93–104, ACM.

[27] Brons JF, De Jong M, Valens M, Grivell LA, Bolotin-Fukuhara M and Blom J [2002]: 'Dissection of the promoter of the HAP4 gene in S. cerevisiae unveils a complex regulatory framework of transcriptional regulation.', *Yeast*, **19**(11): 923–932, http://dx.doi.org/10.1002/yea.886.

[28] Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M and Haussler D [2000]: 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proceedings of the National Academy of Sciences*, **97**(1): 262–267, http://www.pnas.org/content/97/1/262.full.pdf+html, http://dx.doi.org/10.1073/pnas.97.1.262.

[29] Bryne JC, Valen E, Tang MHE, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B and Sandelin A [2008]: 'JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.', *Nucleic Acids Research*, **36**(Database issue): D102–D106, `http://dx.doi.org/10.1093/nar/gkm955`.

[30] Burger L and van Nimwegen E [2008]: 'Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.', *Molecular Systems Biology*, **4**: 165, `http://dx.doi.org/10.1038/msb4100203`.

[31] Burger L and van Nimwegen E [2010]: 'Disentangling direct from indirect co-evolution of residues in protein alignments.', *PLoS Computational Biology*, **6**(1): e1000633, `http://dx.doi.org/10.1371/journal.pcbi.1000633`.

[32] Burridge K and Wennerberg K [2004]: 'Rho and Rac take center stage', *Cell*, **116**(2): 167–179.

[33] Butte AJ, Tamayo P, Slonim D, Golub TR and Kohane IS [2000]: 'Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.', *Proceedings of the National Academy of Sciences*, **97**(22): 12182–12186, ISSN 0027-8424, `http://dx.doi.org/10.1073/pnas.220392197`.

[34] Castelo R and Roverato A [2009]: 'Reverse engineering molecular regulatory networks from microarray data with qp-graphs.', *Journal of Computational Biology*, **16**(2): 213–227, `http://dx.doi.org/10.1089/cmb.2008.08TT`.

[35] Cerulo L, Elkan C and Ceccarelli M [2010]: 'Learning gene regulatory networks from only positive and unlabeled data.', *BMC Bioinformatics*, **11**: 228, `http://dx.doi.org/10.1186/1471-2105-11-228`.

[36] Chang C and Lin C [2001]: *LIBSVM: a library for support vector machines.*, software available at `http://www.csie.ntu.edu.tw/cjlin/libsvm`.

[37] Chang C and Lin C [2011]: 'LIBSVM: A library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, **2**: 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[38] Chaussabel D, Semnani RT, McDowell MA, Sacks D, Sher A and Nutman TB [2003]: 'Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites.', *Blood*, **102**(2): 672–681, `http://dx.doi.org/10.1182/blood-2002-10-3232`.

[39] Chen Y, Wang W, Zhang P, Zhang W, Liu J and Ma X [2009]: 'Expression of genes psma6 and slc25a4 in patients with acute monocytic leukemia.', *Journal of Experimental Hematology/Chinese Association of Pathophysiology*, **17**(5): 1168.

[40] Chen Z, Li J and Wei L [2007]: 'A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue.', *Artificial Intelligence in Medicine*, **41**(2): 161–175, ISSN 0933-3657, `http://dx.doi.org/10.1016/j.artmed.2007.07.008`.

[41] Cho KH, Choo SM, Jung SH, Kim JR, Choi HS and Kim J [2007]: 'Reverse engineering of gene regulatory networks.', *IET Systems Biology*, **1**(3): 149–163.

[42] Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, Chan ET, Frey BJ, Andrews BJ, Boone C and Hughes TR [2006]: 'Identifying transcription factor functions and targets by phenotypic activation.', *Proceedings of the National Academy of Sciences*, **103**(32): 12045–12050, `http://dx.doi.org/10.1073/pnas.0605140103`.

[43] Chuang HY, Lee E, Liu YT, Lee D and Ideker T [2007]: 'Network-based classification of breast cancer metastasis', *Molecular Systems Biology*, **3**: 140, `http://dx.doi.org/10.1038/msb4100180`.

[44] Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, Agarwal A, Huang W, Parkurst CN, Muratet M *et al.* [2012]: 'A validated regulatory network for Th17 cell specification.', *Cell*, **151**(2): 289–303, `http://dx.doi.org/10.1016/j.cell.2012.09.016`.

[45] Cooke EJ, Savage RS and Wild DL [2009]: 'Computational approaches to the integration of gene expression, ChIP-chip and sequence data in the inference of gene regulatory networks.', *Seminars in Cell and Developmental Biology*, **20**(7): 863–868, `http://dx.doi.org/10.1016/j.semcdb.2009.08.004`.

[46] Cortes C and Vapnik V [1995]: 'Support-vector networks', *Machine Learning*, **20**(3): 273–297.

[47] Danowski K, Gross J, Gellrich K, Petri T, Van Dorland H, Bruckmaier R, Reichenbach H, Zimmer R, Meyer H, Schwarz F *et al.* [2013]: 'Metabolic status and oestrous cycle in dairy cows', *International Journal of Livestock Production*, **4**(9): 135–147, `http://dx.doi.org/DOI:10.5897/IJLP12.006`.

[48] Darwin C [1859]: *On the origin of species*, John Murray.

[49] Davis J and Goadrich M [2006]: 'The relationship between Precision-Recall and ROC curves', in 'Proceedings of the 23rd international conference on Machine learning', pages 233–240, ACM.

[50] de la Fuente A *et al.* [2004]: 'Discovery of meaningful associations in genomic data using partial correlation coefficients.', *Bioinformatics*, **20**(18): 3565–3574, `http://dx.doi.org/10.1093/bioinformatics/bth445`.

[51] De Smet R and Marchal K [2010]: 'Advantages and limitations of current network inference methods.', *Nature Reviews. Microbiology*, **8**(10): 717–729, `http://dx.doi.org/10.1038/nrmicro2419`.

[52] Deming WE [1943]: *Statistical adjustment of data.*, Wiley.

[53] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA [2003]: 'DAVID: Database for Annotation, Visualization, and Integrated Discovery.', *Genome Biology*, **4**(5): P3.

[54] Dimitriadou E, Hornik K, Leisch F, Meyer D and Weingessel A [2011]: *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, R package version 1.6, `http://CRAN.R-project.org/package=e1071`.

[55] Dintner S, Staron A, Berchtold E, Petri T, Mascher T and Gebhard S [2011]: 'Coevolution of ABC transporters and two-component regulatory systems as resistance modules against antimicrobial peptides in Firmicutes Bacteria.', *Journal of Bacteriology*, **193**(15): 3851–3862, `http://dx.doi.org/10.1128/JB.05175-11`.

[56] Dölken L, Malterer G, Erhard F, Kothe S, Friedel CC, Suffert G, Marcinowski L, Motsch N, Barth S, Beitzinger M *et al.* [2010]: 'Systematic analysis of viral and cellular microRNA targets in cells latently infected with human γ-herpesviruses by RISC immunoprecipitation assay', *Cell Host & Microbe*, **7**(4): 324–334.

[57] Dölken L, Ruzsics Z, Rädle B, Friedel CC, Zimmer R, Mages J, Hoffmann R, Dickinson P, Forster T, Ghazal P *et al.* [2008]: 'High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay.', *RNA*, **14**(9): 1959–1972, `http://dx.doi.org/10.1261/rna.1136108`.

[58] Dorogovtsev SN and Mendes JF [2003]: *Evolution of networks: From biological nets to the Internet and WWW*, Oxford University Press.

[59] Dougherty J, Tabus I and Astola J [2008]: 'Inference of gene regulatory networks based on a universal minimum description length.', *EURASIP Journal of Bioinformatics and Systems Biology*, **vol. 2008**: ID 482090, `http://dx.doi.org/10.1155/2008/482090`.

[60] Dunn SD, Wahl LM and Gloor GB [2008]: 'Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.', *Bioinformatics*, **24**(3): 333–340, `http://dx.doi.org/10.1093/bioinformatics/btm604`.

[61] Dhaeseleer P, Liang S and Somogyi R [2000]: 'Genetic network inference: from co-expression clustering to reverse engineering', *Bioinformatics*, **16**(8): 707–726, `http://dx.doi.org/10.1093/bioinformatics/16.8.707`.

[62] Edgar R, Domrachev M and Lash AE [2002]: 'Gene Expression Omnibus: NCBI gene expression and hybridization array data repository', *Nucleid Acids Research*, **30**(1): 207–210, `http://dx.doi.org/10.1093/nar/30.1.207`.

[63] Efron B, Hastie T, Johnstone I and Tibshirani R [2004]: 'Least angle regression', *The Annals of Statistics*, **32**(2): 407–499.

[64] Eisen MB, Spellman PT, Brown PO and Botstein D [1998]: 'Cluster analysis and display of genome-wide expression patterns.', *Proceedings of the National Academy of Sciences*, **95**(25): 14863–14868.

[65] Ekeberg M, Lövkvist C, Lan Y, Weigt M and Aurell E [2013]: 'Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models', *Physical Review E*, **87**: 012707, `http://dx.doi.org/10.1103/PhysRevE.87.012707`.

[66] Elati M and Rouveirol C [2011]: *Unsupervised Learning for Gene Regulation Network Inference from Expression Data: A Review*, pages 955–978, John Wiley & Sons, Inc., ISBN 9780470892107, `http://dx.doi.org/10.1002/9780470892107.ch41`.

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

[68] Ernst J, Beg QK, Kay KA, Balázsi G, Oltvai ZN and Bar-Joseph Z [2008]: 'A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli.', *PLoS Computational Biology*, **4**(3): e1000044, `http://dx.doi.org/10.1371/journal.pcbi.1000044`.

[69] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al.* [2011]: 'Mapping and analysis of chromatin state dynamics in nine human cell types.', *Nature*, **473**(7345): 43–49, `http://dx.doi.org/10.1038/nature09906`.

[70] Ernst M and Haesbroeck G [2017]: 'Comparison of local outlier detection techniques in spatial multivariate data', *Data Mining and Knowledge Discovery*, **31**(2): 371–399.

[71] Fabregat I [2009]: 'Dysregulation of apoptosis in hepatocellular carcinoma cells', *World Journal of Gastroenterology: WJG*, **15**(5): 513.

[72] Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ and Gardner TS [2008]: 'Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.', *Nucleic Acids Research*, **36**(Database issue): D866–D870, `http://dx.doi.org/10.1093/nar/gkm815`.

[73] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ and Gardner TS [2007]: 'Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.', *PLoS Biology*, **5**(1): e8, `http://dx.doi.org/10.1371/journal.pbio.0050008`.

[74] Fang Y, Shi C, Manduchi E, Civelek M and Davies P [2010]: 'MicroRNA-10a regulation of proinflammatory phenotype in athero-susceptible endothelium in vivo and in vitro', *Proceedings of the National Academy of Sciences*, **107**(30): 13450.

[75] Feizi S, Marbach D, Médard M and Kellis M [2013]: 'Network deconvolution as a general method to distinguish direct dependencies in networks.', *Nature Biotechnology*, **31**(8): 726–733, http://dx.doi.org/10.1038/nbt.2635.

[76] François J and Parrou JL [2001]: 'Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae.', *FEMS Microbiology Reviews*, **25**(1): 125–145.

[77] Freue GVC, Hollander Z, Shen E, Zamar RH, Balshaw R, Scherer A, McManus B, Keown P, McMaster WR and Ng RT [2007]: 'MDQC: a new quality assessment method for microarrays based on quality control reports.', *Bioinformatics*, **23**(23): 3162–3169, http://dx.doi.org/10.1093/bioinformatics/btm487.

[78] Friedman JH, Hastie T and Tibshirani R [2010]: 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software*, **33**(1): 1–22, ISSN 1548-7660, http://www.jstatsoft.org/v33/i01.

[79] Fujita A, Sato JaR, Demasi MAA, Sogayar MC, Ferreira CE and Miyano S [2009]: 'Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis.', *Journal of Bioinformatics and Computational Biology*, **7**(4): 663–684.

[80] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M and Haussler D [2000]: 'Support vector machine classification and validation of cancer tissue samples using microarray expression data.', *Bioinformatics*, **16**(10): 906–914, ISSN 1367-4803, http://dx.doi.org/10.1093/bioinformatics/16.10.906.

[81] Furlotte NA, Kang HM, Ye C and Eskin E [2011]: 'Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity', *Bioinformatics*, **27**(13): i288–i294.

[82] Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S *et al.* [2006]: 'Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.', *BMC Genomics*, **7**: 325, http://dx.doi.org/10.1186/1471-2164-7-325.

[83] Garrett-Mayer E, Parmigiani G, Zhong X, Cope L and Gabrielson E [2008]: 'Cross-study validation and combined analysis of gene expression microarray data', *Biostatistics*, **9**: 333–354.

[84] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R *et al.* [2012]: 'Architecture of the human regulatory network derived from ENCODE data.', *Nature*, **489**(7414): 91–100, http://dx.doi.org/10.1038/nature11245.

[85] Geurts P [2011]: 'Learning from positive and unlabeled examples by enforcing statistical significance', in 'JMLR: Workshop and Conference Proceedings', volume 15.

[86] Ghosh A, Wang X, Klein E and Heston WDW [2005]: 'Novel role of prostate-specific membrane antigen in suppressing prostate cancer invasiveness.', *Cancer Research*, **65**(3): 727–731.

[87] Gilad Y and Mizrahi-Man O [2015]: 'A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations].', *F1000Research*, **4**: 121, `http://dx.doi.org/10.12688/f1000research.6536.1`.

[88] Gillis J and Pavlidis P [2011]: 'The impact of multifunctional genes on "guilt by association" analysis.', *PLOS ONE*, **6**(2): e17258, `http://dx.doi.org/10.1371/journal.pone.0017258`.

[89] Gillis J and Pavlidis P [2012]: '"Guilt by association" is the exception rather than the rule in gene networks.', *PLoS Computational Biology*, **8**(3): e1002444, `http://dx.doi.org/10.1371/journal.pcbi.1002444`.

[90] Gloor GB, Martin LC, Wahl LM and Dunn SD [2005]: 'Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.', *Biochemistry*, **44**(19): 7156–7165, `http://dx.doi.org/10.1021/bi050293e`.

[91] Goeman JJ and Bühlmann P [2007]: 'Analyzing gene expression data in terms of gene sets: methodological issues.', *Bioinformatics*, **23**(8): 980–987, `http://dx.doi.org/10.1093/bioinformatics/btm051`.

[92] Goeman JJ and Solari A [2014]: 'Multiple hypothesis testing in genomics', *Statistics in Medicine*, **33**(11): 1946–1978.

[93] Goeman JJ, Van De Geer SA, De Kort F and Van Houwelingen HC [2004]: 'A global test for groups of genes: testing association with a clinical outcome', *Bioinformatics*, **20**(1): 93–99.

[94] Goh CS, Bogan AA, Joachimiak M, Walther D and Cohen FE [2000]: 'Co-evolution of proteins with their interaction partners.', *Journal of Molecular Biology*, **299**(2): 283–293, `http://dx.doi.org/10.1006/jmbi.2000.3732`.

[95] Golub GH and Van Loan CF [1980]: 'An analysis of the total least squares problem', *SIAM Journal on Numerical Analysis*, **17**(6): 883–893.

[96] Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S and Szustakowski JD [2011]: 'Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples.', *PLOS ONE*, **6**: e27156, ISSN 1932-6203, `http://dx.doi.org/10.1371/journal.pone.0027156`.

[97] Gouveia-Oliveira R and Pedersen AG [2007]: 'Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation.', *Algorithms for Molecular Biology*, **2**: 12, `http://dx.doi.org/10.1186/1748-7188-2-12`.

[98] de la Grange P, Gratadou L, Delord M, Dutertre M and Auboeuf D [2010]: 'Splicing factor and exon profiling across human tissues.', *Nucleic Acids Research*, **38**(9): 2825–2838, `http://dx.doi.org/10.1093/nar/gkq008`.

[99] Gravetter FJ and Wallnau LB [2016]: *Statistics for the behavioral sciences*, Cengage Learning.

[100] Greenbaum D, Colangelo C, Williams K and Gerstein M [2003]: 'Comparing protein abundance and mRNA expression levels on a genomic scale', *Genome Biology*, **4**(9): 117, ISSN 1474-760X, `http://dx.doi.org/10.1186/gb-2003-4-9-117`.

[101] Greenfield A, Hafemeister C and Bonneau R [2013]: 'Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks.', *Bioinformatics*, **29**(8): 1060–1067, `http://dx.doi.org/10.1093/bioinformatics/btt099`.

[102] Grunkemeier GL, Wu Y and Furnary AP [2009]: 'What is the Value of a p Value?', *The Annals of Thoracic Surgery*, **87**(5): 1337 – 1343, ISSN 0003-4975, `http://dx.doi.org/http://dx.doi.org/10.1016/j.athoracsur.2009.03.027`.

[103] Gurvitz A and Rottensteiner H [2006]: 'The biochemistry of oleate induction: transcriptional upregulation and peroxisome proliferation.', *Biochimica et Biophysica Acta*, **1763**(12): 1392–1402, `http://dx.doi.org/10.1016/j.bbamcr.2006.07.011`.

[104] Gustafsson M and Hörnquist M [2010]: 'Gene expression prediction by soft integration and the elastic net-best performance of the DREAM3 gene expression challenge.', *PLOS ONE*, **5**(2): e9134, `http://dx.doi.org/10.1371/journal.pone.0009134`.

[105] Gutiérrez-Ríos RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR and Collado-Vides J [2003]: 'Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles.', *Genome Research*, **13**(11): 2435–2443, `http://dx.doi.org/10.1101/gr.1387003`.

[106] Gygi SP, Rochon Y, Franza BR and Aebersold R [1999]: 'Correlation between protein and mRNA abundance in yeast', *Molecular and Cellular biology*, **19**(3): 1720–1730.

[107] Hall N [2007]: 'Advanced sequencing technologies and their wider impact in microbiology.', *The Journal of experimental biology*, **210**: 1518–1525, ISSN 0022-0949, `http://dx.doi.org/10.1242/jeb.001370`.

[108] Han B, Kang HM and Eskin E [2009]: 'Rapid and accurate multiple testing correction and power estimation for millions of correlated markers', *PLoS Genetics*, **5**(4): e1000456.

[109] Hand DJ [2009]: 'Measuring classifier performance: a coherent alternative to the area under the ROC curve', *Machine Learning*, **77**(1): 103–123.

[110] Hanisch D, Zien A, Zimmer R and Lengauer T [2002]: 'Co-clustering of biological networks and gene expression data.', *Bioinformatics*, **18 Suppl 1**: S145–S154, ISSN 1367-4803.

[111] Haq O, Levy RM, Morozov AV and Andrec M [2009]: 'Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease.', *BMC Bioinformatics*, **10 Suppl 8**: S10, `http://dx.doi.org/10.1186/1471-2105-10-S8-S10`.

[112] Hasegawa S, Furukawa Y, Li M, Satoh S, Kato T, Watanabe T, Katagiri T, Tsunoda T, Yamaoka Y and Nakamura Y [2002]: 'Genome-wide analysis of gene expression in intestinal-type gastric cancers using a complementary DNA microarray representing 23,040 genes', *Cancer Research*, **62**(23): 7012.

[113] Hastie T, Tibshirani R and Friedman JH [2001]: *The elements of statistical learning: data mining, inference, and prediction*, New York: Springer-Verlag.

[114] Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P and Botstein D [1999]: 'Imputing Missing Data for Gene Expression Arrays', Technical report, Stanford Statistics Department.

[115] Hautmann SH, Huland E and Huland H [1999]: 'Local intratumor immunotherapy of prostate cancer with interleukin-2 reduces tumor growth.', *Anticancer Research*, **19**(4A): 2661–2663.

[116] Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H and Brent MR [2013]: 'Mapping functional transcription factor networks from gene expression data.', *Genome Research*, **23**(8): 1319–1328, `http://dx.doi.org/10.1101/gr.150904.112`.

[117] Hill AV [1910]: 'Proceedings of the Physiological Society', *The Journal of Physiology*, **40**(Suppl): i–vii, `http://jp.physoc.org/content/40/supplement/i.short`.

[118] Hiltunen JK, Mursula AM, Rottensteiner H, Wierenga RK, Kastaniotis AJ and Gurvitz A [2003]: 'The biochemistry of peroxisomal beta-oxidation in the yeast Saccharomyces cerevisiae.', *FEMS Microbiology Reviews*, **27**(1): 35–64.

[119] Hirsch HA, Iliopoulos D, Joshi A, Zhang Y, Jaeger SA, Bulyk M, Tsichlis PN, Liu XS and Struhl K [2010]: 'A transcriptional signature and common gene networks

link cancer with lipid metabolism and diverse human diseases.', *Cancer Cell*, **17**(4): 348–361, http://dx.doi.org/10.1016/j.ccr.2010.01.022.

[120] Holloway DT, Kon M and DeLisi C [2008]: 'Classifying transcription factor targets and discovering relevant biological features.', *Biology Direct*, **3**: 22, http://dx.doi.org/10.1186/1745-6150-3-22.

[121] Holm S [1979]: 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics*, pages 65–70.

[122] Hong D, Lee J, Hong S, Yoon J and Park S [2008]: *Extraction of Informative Genes from Integrated Microarray Data*, Springer Berlin/Heidelberg.

[123] Hu J, Li H, Waterman M and Zhou X [2006]: 'Integrative missing value estimation for microarray data', *BMC Bioinformatics*, **7**: 449.

[124] Hu Z, Killion PJ and Iyer VR [2007]: 'Genetic reconstruction of a functional transcriptional regulatory network.', *Nature Genetics*, **39**(5): 683–687, http://dx.doi.org/10.1038/ng2012.

[125] Huang DW, Sherman BT and Lempicki RA [2009]: 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.', *Nature Protocols*, **4**(1): 44–57, http://dx.doi.org/10.1038/nprot.2008.211.

[126] Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K *et al.* [2016]: 'A new view of the tree of life', *Nature Microbiology*, **1**: 16048.

[127] Hulsman M, Mentink A, van Someren EP, Dechering KJ, de Boer J and Reinders MJ [2010]: 'Delineation of amplification, hybridization and location effects in microarray data yields better-quality normalization.', *BMC Bioinformatics*, **11**: 156, http://dx.doi.org/10.1186/1471-2105-11-156.

[128] Ibberson D, Benes V, Muckenthaler MU and Castoldi M [2009]: 'RNA degradation compromises the reliability of microRNA expression profiling', *BMC Biotechnology*, **9**(1): 102, ISSN 1472-6750, http://dx.doi.org/10.1186/1472-6750-9-102.

[129] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y and Barkai N [2002]: 'Revealing modular organization in the yeast transcriptional network.', *Nature Genetics*, **31**(4): 370–377, http://dx.doi.org/10.1038/ng941.

[130] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP [2003]: 'Summaries of Affymetrix GeneChip probe level data.', *Nucleic Acids Research*, **31**(4): e15.

[131] Jeffery IB, Higgins DG and Culhane AC [2006]: 'Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.', *BMC Bioinformatics*, **7**: 359, http://dx.doi.org/10.1186/1471-2105-7-359.

[132] Jones DT, Buchan DWA, Cozzetto D and Pontil M [2012]: 'PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.', *Bioinformatics*, **28**(2): 184–190, http://dx.doi.org/10.1093/bioinformatics/btr638.

[133] Joo JWJ, Hormozdiari F, Han B and Eskin E [2016]: 'Multiple testing correction in linear mixed models', *Genome Biology*, **17**(1): 62.

[134] Kauffmann A, Gentleman R and Huber W [2009]: 'arrayQualityMetrics–a bioconductor package for quality assessment of microarray data.', *Bioinformatics*, **25**(3): 415–416, http://dx.doi.org/10.1093/bioinformatics/btn647.

[135] Kenzelmann M, Maertens S, Hergenhahn M, Kueffer S, Hotz-Wagenblatt A, Li L, Wang S, Ittrich C, Lemberger T, Arribas R *et al.* [2007]: 'Microarray analysis of newly synthesized RNA in cells and animals.', *Proceedings of the National Academy of Sciences*, **104**(15): 6164–6169, http://dx.doi.org/10.1073/pnas.0610439104.

[136] Khan I, Murphy P, Casciotti L, Schwartzman J, Collins J, Gao J and Yeaman G [2001]: 'Mice lacking the chemokine receptor CCR1 show increased susceptibility to Toxoplasma gondii infection', *Journal of Immunology*, **166**(3): 1930.

[137] Kim H, Golub GH and Park H [2005]: 'Missing value estimation for DNA microarray gene expression data: local least squares imputation', *Bioinformatics*, **21**(2): 187–198, http://dx.doi.org/10.1093/bioinformatics/bth499.

[138] Kitano H [2002]: 'Computational systems biology.', *Nature*, **420**: 206–210, ISSN 0028-0836, http://dx.doi.org/10.1038/nature01254.

[139] Kostka D and Spang R [2004]: 'Finding disease specific alterations in the co-expression of genes', *Bioinformatics*, **20**(Suppl 1): i194–199.

[140] Kramer R, Wolf S, Petri T, von Soosten D, Dänicke S, Weber EM, Zimmer R, Rehage J and Jahreis G [2013]: 'A commonly used rumen-protected conjugated linoleic acid supplement marginally affects fatty acid distribution of body tissues and gene expression of mammary gland in heifers during early lactation.', *Lipids in Health and Disease*, **12**: 96, http://dx.doi.org/10.1186/1476-511X-12-96.

[141] Kriegel HP, Kröger P, Schubert E and Zimek A [2009]: 'LoOP: local outlier probabilities', in 'Proceedings of the 18th ACM conference on Information and knowledge management', pages 1649–1652, ACM.

[142] Kriegel HP, Kröger P and Zimek A [2009]: 'Clustering High-dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering', *ACM Trans. Knowl. Discov. Data*, **3**(1): 1:1–1:58, ISSN 1556-4681, `http://dx.doi.org/10.1145/1497577.1497578`.

[143] Küffner R, Petri T, Tavakkolkhah P, Windhager L and Zimmer R [2012]: 'Inferring gene regulatory networks by ANOVA.', *Bioinformatics*, **28**(10): 1376–1382, `http://dx.doi.org/10.1093/bioinformatics/bts143`.

[144] Küffner R, Petri T, Windhager L and Zimmer R [2010]: 'Petri Nets with Fuzzy Logic (PNFL): reverse engineering and parametrization.', *PLOS ONE*, **5**(9), `http://dx.doi.org/10.1371/journal.pone.0012807`.

[145] Langdon WB, Upton GJG, da Silva Camargo R and Harrison AP [2010]: 'A survey of spatial defects in Homo Sapiens Affymetrix GeneChips.', *IEEE/ACM Trans Comput Biol Bioinform*, **7**(4): 647–653, `http://dx.doi.org/10.1109/TCBB.2008.108`.

[146] Lapedes AS, Giraud BG, Liu L and Stormo GD [1999]: 'Correlated mutations in models of protein sequences: phylogenetic and structural effects', *Statistics in Molecular Biology*, **33**: 236–256.

[147] Lapuk A, Marr H, Jakkula L, Pedro H, Bhattacharya S, Purdom E, Hu Z, Simpson K, Pachter L, Durinck S *et al.* [2010]: 'Exon-level microarray analyses identify alternative splicing programs in breast cancer.', *Molecular Cancer Research*, **8**(7): 961–974, `http://dx.doi.org/10.1158/1541-7786.MCR-09-0528`.

[148] Lee C, Atanelov L, Modrek B and Xing Y [2003]: 'ASAP: the alternative splicing annotation project', *Nucleic Acids Research*, **31**(1): 101.

[149] Lee HK, Hsu AK, Sajdak J, Qin J and Pavlidis P [2004]: 'Coexpression analysis of human genes across many microarray data sets.', *Genome Research*, **14**(6): 1085–1094, `http://dx.doi.org/10.1101/gr.1910904`.

[150] Li J and Ji L [2005]: 'Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix', *Heredity*, **95**(3): 221.

[151] Liaw A and Wiener M [2002]: 'Classification and Regression by randomForest', *R News*, **2**(3): 18–22, `http://CRAN.R-project.org/doc/Rnews/`.

[152] Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC *et al.* [2014]: 'Comparison of the transcriptional landscapes between human and mouse tissues.', *Proceedings of the National Academy of Sciences*, **111**(48): 17224–17229, `http://dx.doi.org/10.1073/pnas.1413624111`.

[153] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al.* [1996]: 'Expression monitoring by hybridization to high-density oligonucleotide arrays.', *Nature Biotechnology*, **14**(13): 1675–1680, `http://dx.doi.org/10.1038/nbt1296-1675`.

[154] Lockstone HE [2011]: 'Exon array data analysis using Affymetrix power tools and R statistical software.', *Briefings in Bioinformatics*, **12**(6): 634–644, `http://dx.doi.org/10.1093/bib/bbq086`.

[155] Lorenz MC and Heitman J [1998]: 'Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains.', *Genetics*, **150**(4): 1443–1457.

[156] Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK and Zhou J [2007]: 'Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory.', *BMC Bioinformatics*, **8**: 299, `http://dx.doi.org/10.1186/1471-2105-8-299`.

[157] MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD and Fraenkel E [2006]: 'An improved map of conserved regulatory sites for Saccharomyces cerevisiae.', *BMC Bioinformatics*, **7**: 113, `http://dx.doi.org/10.1186/1471-2105-7-113`.

[158] Madar A, Greenfield A, Ostrer H, Vanden-Eijnden E and Bonneau R [2009]: 'The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models.', *Conf Proc IEEE Eng Med Biol Soc*, **2009**: 5448–5451, `http://dx.doi.org/10.1109/IEMBS.2009.5334018`.

[159] Maier T, Gell M and Serrano L [2009]: 'Correlation of mRNA and protein in complex biological samples.', *FEBS letters*, **583**: 3966–3973, ISSN 1873-3468, `http://dx.doi.org/10.1016/j.febslet.2009.10.036`.

[160] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, DREAM5 Consortium, Kellis M, Collins JJ *et al.* [2012]: 'Wisdom of crowds for robust gene network inference.', *Nature Methods*, **9**(8): 796–804, `http://dx.doi.org/10.1038/nmeth.2016`.

[161] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD and Califano A [2006]: 'ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.', *BMC Bioinformatics*, **7 Suppl 1**: S7, `http://dx.doi.org/10.1186/1471-2105-7-S1-S7`.

[162] Mason SJ and Graham NE [2002]: 'Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation', *Quarterly Journal of the Royal Meteorological Society*, **128**(584): 2145–2166.

[163] Meinshausen N and Bhlmann P [2006]: ‘High-Dimensional Graphs and Variable Selection with the Lasso’, *The Annals of Statistics*, **34**(3): 1436–1462, ISSN 00905364, `http://www.jstor.org/stable/25463463`.

[164] Mellmann A, Andersen SP, Bletz S, Friedrich AW, Kohl TA, Lilje B, Niemann S, Prior K, Rossen JW and Harmsen D [2017]: ‘High Interlaboratory Reproducibility and Accuracy of Next-Generation-Sequencing-Based Bacterial Genotyping in a Ring Trial’, *Journal of Clinical Microbiology*, **3**(55): 908–913.

[165] Michoel T, Smet RD, Joshi A, de Peer YV and Marchal K [2009]: ‘Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks.’, *BMC Systems Biology*, **3**: 49, `http://dx.doi.org/10.1186/1752-0509-3-49`.

[166] Minucci S and Pelicci P [2006]: ‘Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer’, *Nature Reviews Cancer*, **6**(1): 38–51.

[167] Moffitt RA, Yin-Goen Q, Stokes TH, Parry RM, Torrance JH, Phan JH, Young AN and Wang MD [2011]: ‘caCORRECT2: Improving the accuracy and reliability of microarray data in the presence of artifacts.’, *BMC Bioinformatics*, **12**: 383, `http://dx.doi.org/10.1186/1471-2105-12-383`.

[168] Monteith G, McAndrew D, Faddy H and Roberts-Thomson S [2007]: ‘Calcium and cancer: targeting Ca2+ transport’, *Nature Reviews Cancer*, **7**(7): 519–530.

[169] Moody DB, Robinson JC, Ewing CM, Lazenby AJ and Isaacs WB [1994]: ‘Interleukin-2 transfected prostate cancer cells generate a local antitumor effect in vivo.’, *Prostate*, **24**(5): 244–251.

[170] Morano KA, Grant CM and Moye-Rowley WS [2012]: ‘The response to heat shock and oxidative stress in Saccharomyces cerevisiae.’, *Genetics*, **190**(4): 1157–1195, `http://dx.doi.org/10.1534/genetics.111.128033`.

[171] Mordelet F and Vert JP [2008]: ‘SIRENE: supervised inference of regulatory networks.’, *Bioinformatics*, **24**(16): i76–i82, `http://dx.doi.org/10.1093/bioinformatics/btn273`.

[172] Mordelet F and Vert JP [2010]: ‘A bagging SVM to learn from positive and unlabeled examples’, Technical report, Cornell University Library.

[173] Murata T [1989]: ‘Petri nets: Properties, analysis and applications’, *Proceedings of the IEEE*, **77**(4): 541–580.

[174] Myers CL, Barrett DR, Hibbs MA, Huttenhower C and Troyanskaya OG [2006]: ‘Finding function: evaluation methods for functional genomic data.’, *BMC Genomics*, **7**: 187, `http://dx.doi.org/10.1186/1471-2164-7-187`.

[175] Naeem H, Zimmer R, Tavakkolkhah P and Küffner R [2012]: ‘Rigorous assessment of gene set enrichment tests.’, *Bioinformatics*, **28**(11): 1480–1486, `http://dx.doi.org/10.1093/bioinformatics/bts164`.

[176] Narendra V, Lytkin NI, Aliferis CF and Statnikov A [2011]: ‘A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks.’, *Genomics*, **97**(1): 7–18, `http://dx.doi.org/10.1016/j.ygeno.2010.10.003`.

[177] NCBI Resource Coordinators [2016]: ‘Database resources of the National Center for Biotechnology Information’, *Nucleic Acids Research*, **44(Database issue)**: D7–D19, `http://dx.doi.org/doi:10.1093/nar/gkv1290.`

[178] Nelder JA [1999]: ‘From Statistics to Statistical Science’, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **48**(2): 257–269, ISSN 1467-9884, `http://dx.doi.org/10.1111/1467-9884.00187`.

[179] Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E and Stamatoyannopoulos JA [2012]: ‘Circuitry and dynamics of human transcription factor regulatory networks.’, *Cell*, **150**(6): 1274–1286, `http://dx.doi.org/10.1016/j.cell.2012.04.040`.

[180] Neu HC [1992]: ‘The crisis in antibiotic resistance.’, *Science*, **257**(5073): 1064–1073.

[181] Nicholls N [2001]: ‘Commentary and analysis: The Insignificance of Significance Testing’, *Bulletin of the American Meteorological Society*, **82**(5): 981–986, `https://doi.org/10.1175/1520-0477(2001)082<0981:CAATIO>2.3.CO;2`, `http://dx.doi.org/10.1175/1520-0477(2001)082<0981:CAATIO>2.3.CO;2`.

[182] Noble W [2006]: ‘What is a support vector machine?’, *Nature Biotechnology*, **24**(12): 1565–7.

[183] Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beißbarth T and Gaedcke J [2010]: ‘Impact of RNA degradation on gene expression profiling’, *BMC Medical Genomics*, **3**(1): 36, ISSN 1755-8794, `http://dx.doi.org/10.1186/1755-8794-3-36`.

[184] Ordonez-Moran P, Larriba M, Palmer H, Valero R, Barbachano A, Dunach M, De Herreros A, Villalobos C, Berciano M, Lafarga M *et al.* [2008]: ‘RhoA-ROCK and p38MAPK-MSK1 mediate vitamin D effects on gene expression, phenotype, and Wnt pathway in colon cancer cells’, *Journal of Cell Biology*, **183**(4): 697.

[185] Otter WD, Jacobs JJL, Battermann JJ, Hordijk GJ, Krastev Z, Moiseeva EV, Stewart RJE, Ziekman PGPM and Koten JW [2008]: ‘Local therapy of cancer with free IL-2.’, *Cancer Immunology*, **57**(7): 931–950, `http://dx.doi.org/10.1007/s00262-008-0455-z`.

[186] Ozcan S and Johnston M [1995]: 'Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose.', *Molecular and Cellular Biology*, **15**(3): 1564–1572.

[187] Palmer AC and Kishony R [2013]: 'Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance.', *Nature Reviews. Genetics*, **14**(4): 243–248, `http://dx.doi.org/10.1038/nrg3351`.

[188] Pavlidis P and Gillis J [2013]: 'Progress and challenges in the computational prediction of gene function using networks: 2012-2013 update [version 1; referees: 2 approved]', *F1000Research*, **2**: 230, `http://dx.doi.org/10.12688/f1000research.2-230.v1`.

[189] Pazos F and Valencia A [2001]: 'Similarity of phylogenetic trees as indicator of protein-protein interaction.', *Protein Engineering*, **14**(9): 609–614.

[190] Pazos F and Valencia A [2008]: 'Protein co-evolution, co-adaptation and interactions.', *EMBO Journal*, **27**(20): 2648–2655, `http://dx.doi.org/10.1038/emboj.2008.189`.

[191] Pearl J [2009]: *Causality*, Cambridge University Press, 2$^{\text{nd}}$ edition.

[192] Petri T, Altmann S, Geistlinger L, Zimmer R and Küffner R [2015]: 'Addressing false discoveries in network inference.', *Bioinformatics*, **31**(17): 2836–2843, `http://dx.doi.org/10.1093/bioinformatics/btv215`.

[193] Petri T, Berchtold E, Zimmer R and Friedel CC [2012]: 'Detection and correction of probe-level artefacts on microarrays.', *BMC Bioinformatics*, **13**: 114, `http://dx.doi.org/10.1186/1471-2105-13-114`.

[194] Petri T, Küffner R and Zimmer R [2011]: 'Experiment specific expression patterns.', *Journal of Computational Biology*, **18**: 1423–1435, ISSN 1557-8666, `http://dx.doi.org/10.1089/cmb.2011.0159`.

[195] Petri T, Küffner R and Zimmer R [2011]: 'Experiment specific expression patterns', in Bafna V and Sahinalp SC (Editors), 'Research in Computational Molecular Biology', volume 15 of *LNBI 6577*, pages 339–354, Springer, `http://dx.doi.org/10.1007/978-3-642-20036-6_32`.

[196] Petricka JJ and Benfey PN [2011]: 'Reconstructing regulatory network transitions.', *Trends in Cell Biology*, **21**(8): 442–451, `http://dx.doi.org/10.1016/j.tcb.2011.05.001`.

[197] Prasad AS, Mukhtar H, Beck FWJ, Adhami VM, Siddiqui IA, Din M, Hafeez BB and Kucuk O [2010]: 'Dietary zinc and prostate cancer in the TRAMP mouse model.', *Journal of Medicinal Food*, **13**(1): 70–76, `http://dx.doi.org/10.1089/jmf.2009.0042`.

[198] Press WH, Teukolsky SA, Vetterling WT and Flannery BP [2007]: *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press.

[199] Prieto C, Rivas MJ, Sánchez JM, López-Fidalgo J and Rivas JDL [2006]: 'Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes.', *Bioinformatics*, **22**(9): 1103–1110, `http://dx.doi.org/10.1093/bioinformatics/btl053`.

[200] Qian J, Lin J, Luscombe NM, Yu H and Gerstein M [2003]: 'Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data.', *Bioinformatics*, **19**(15): 1917–1926.

[201] Quinlan JR [1996]: 'Bagging, Boosting, and C4.5', in 'In Proceedings of the Thirteenth National Conference on Artificial Intelligence', pages 725–730, AAAI Press.

[202] Radovanović M, Nanopoulos A and Ivanović M [2010]: 'On the existence of obstinate results in vector space models', in 'Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval', pages 186–193, ACM.

[203] Rahmann S, Müller T and Vingron M [2003]: 'On the power of profiles for transcription factor binding site detection.', *Statistical Applications in Genetics and Molecular Biology*, **2**: Article7, `http://dx.doi.org/10.2202/1544-6115.1032`.

[204] Ramani AK and Marcotte EM [2003]: 'Exploiting the co-evolution of interacting proteins to discover interaction specificity.', *Journal of Molecular Biology*, **327**(1): 273–284.

[205] Rasmussen C and Williams C [2006]: 'Gaussian processes for machine learning, vol. 1', *The MIT Press, Cambridge, doi*, **10**: S0129065704001899.

[206] Reimers M and Weinstein JN [2005]: 'Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.', *BMC Bioinformatics*, **6**: 166, `http://dx.doi.org/10.1186/1471-2105-6-166`.

[207] Reményi A, Schöler HR and Wilmanns M [2004]: 'Combinatorial control of gene expression.', *Nature Structural and Molecular Biology*, **11**(9): 812–815, `http://dx.doi.org/10.1038/nsmb820`.

[208] Rhodes DR, Barrette TR, Rubin MA, Ghosh D and Chinnaiyan AM [2002]: 'Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer', *Cancer Research*, **62**(15): 4427–4433, `http://cancerres.aacrjournals.org/cgi/content/abstract/62/15/4427`.

[209] Roderick HL and Cook SJ [2008]: 'Ca2+ signalling checkpoints in cancer: remodelling Ca2+ for cancer cell proliferation and survival.', *Nature Reviews Cancer*, **8**(5): 361–375, `http://dx.doi.org/10.1038/nrc2374`.

[210] Romano JP, Shaikh AM and Wolf M [2008]: ‘Control of the false discovery rate under dependence using the bootstrap and subsampling’, *Test*, **17**(3): 417–442.

[211] Rozanov D, Savinov A, Williams R, Liu K, Golubkov V, Krajewski S and Strongin A [2008]: ‘Molecular signature of MT1-MMP: transactivation of the downstream universal gene network in cancer’, *Cancer Research*, **68**(11): 4086.

[212] Salomonis N, Schlieve CR, Pereira L, Wahlquist C, Colas A, Zambon AC, Vranizan K, Spindler MJ, Pico AR, Cline MS *et al.* [2010]: ‘Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation.’, *Proceedings of the National Academy of Sciences*, **107**(23): 10514–10519, `http://dx.doi.org/10.1073/pnas.0912260107`.

[213] Schäfer J and Strimmer K [2005]: ‘A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.’, *Statistical applications in Genetics and Molecular biology*, **4**: Article32, ISSN 1544-6115, `http://dx.doi.org/10.2202/1544-6115.1175`.

[214] Schölkopf B and Smola AJ [2001]: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*, The MIT Press, ISBN 0262194759.

[215] Schölkopf B, Tsuda K and Vert J [2004]: *Kernel Methods in Computational Biology*, The MIT Press, ISBN 0262195097.

[216] Schug A, Weigt M, Onuchic JN, Hwa T and Szurmant H [2009]: ‘High-resolution protein complexes from integrating genomic information with molecular simulation.’, *Proceedings of the National Academy of Sciences*, **106**(52): 22124–22129, `http://dx.doi.org/10.1073/pnas.0912100106`.

[217] Schüller HJ [2003]: ‘Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae.’, *Current Genetics*, **43**(3): 139–160, `http://dx.doi.org/10.1007/s00294-003-0381-8`.

[218] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U and Gaul U [2008]: ‘Predicting expression patterns from regulatory sequence in Drosophila segmentation.’, *Nature*, **451**(7178): 535–540, `http://dx.doi.org/10.1038/nature06496`.

[219] Segal E, Shapira M, Regev A, Pe’er D, Botstein D, Koller D and Friedman N [2003]: ‘Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.’, *Nature Genetics*, **34**(2): 166–176, `http://dx.doi.org/10.1038/ng1165`.

[220] Shalon D, Smith SJ and Brown PO [1996]: ‘A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.’, *Genome Research*, **6**(7): 639–645.

[221] Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM and Butte AJ [2010]: 'Cell type-specific gene expression differences in complex tissues.', *Nature Methods*, **7**(4): 287–289, `http://dx.doi.org/10.1038/nmeth.1439`.

[222] Shimamura T, Imoto S, Yamaguchi R, Fujita A, Nagasaki M and Miyano S [2009]: 'Recursive regularization for inferring gene networks from time-course gene expression profiles.', *BMC Systems Biology*, **3**: 41, `http://dx.doi.org/10.1186/1752-0509-3-41`.

[223] Simon N, Friedman JH, Hastie T and Tibshirani R [2011]: 'Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent', *Journal of Statistical Software*, **39**(5): 1–13, ISSN 1548-7660, `http://www.jstatsoft.org/v39/i05`.

[224] Simpson E [1951]: 'The interpretation of interaction in contingency tables', *Journal of the Royal Statistical Society. Series B (Methodological)*, **13**(2): 238–241.

[225] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP *et al.* [2002]: 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell*, **1**(2): 203–209, `http://dx.doi.org/10.1016/S1535-6108(02)00030-2`.

[226] Smith JJ, Ramsey SA, Marelli M, Marzolf B, Hwang D, Saleem RA, Rachubinski RA and Aitchison JD [2007]: 'Transcriptional responses to fatty acid are coordinated by combinatorial control.', *Molecular Systems Biology*, **3**: 115, `http://dx.doi.org/10.1038/msb4100157`.

[227] Smith P, Hobisch A, Lin D, Culig Z and Keller E [2001]: 'Interleukin-6 and prostate cancer progression', *Cytokine & Growth Factor Reviews*, **12**(1): 33–40.

[228] Smola A and Schölkopf B [1998]: 'A tutorial on support vector regression', Technical Report NC2-TR-1998-030, NeuroCOLT2.

[229] Smola A and Schölkopf B [2004]: 'A tutorial on support vector regression', *Statistics and Computing*, **14**(3): 199–222.

[230] Song JS, Maghsoudi K, Li W, Fox E, Quackenbush J and Liu XS [2007]: 'Microarray blob-defect removal improves array analysis.', *Bioinformatics*, **23**(8): 966–971, `http://dx.doi.org/10.1093/bioinformatics/btm043`.

[231] Soranzo N, Bianconi G and Altafini C [2007]: 'Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data.', *Bioinformatics*, **23**(13): 1640–1647, `http://dx.doi.org/10.1093/bioinformatics/btm163`.

[232] Storey JD and Tibshirani R [2003]: 'Statistical significance for genomewide studies', *Proceedings of the National Academy of Sciences*, **100**(16): 9440–9445.

[233] Suárez-Fariñas M, Pellegrino M, Wittkowski KM and Magnasco MO [2005]: 'Harshlight: a "corrective make-up" program for microarray chips.', *BMC Bioinformatics*, **6**: 294, `http://dx.doi.org/10.1186/1471-2105-6-294`.

[234] The Gene Ontology Consortium [2010]: 'The Gene Ontology in 2010: extensions and refinements.', *Nucleic Acids Research*, **38**(Database issue): D331–D335, `http://dx.doi.org/10.1093/nar/gkp1018`.

[235] Tibshirani R [1994]: 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B*, **58**: 267–288.

[236] Tibshirani R [2011]: 'Regression shrinkage and selection via the lasso: a retrospective', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3): 273–282.

[237] Turcotte B, Liang XB, Robert F and Soontorngun N [2010]: 'Transcriptional regulation of nonfermentable carbon utilization in budding yeast.', *FEMS Yeast Research*, **10**(1): 2–13, `http://dx.doi.org/10.1111/j.1567-1364.2009.00555.x`.

[238] Ucar D, Neuhaus I, Ross-MacDonald P, Tilford C, Parthasarathy S, Siemers N and Ji RR [2007]: 'Construction of a reference gene association network from multiple profiling data: application to data analysis.', *Bioinformatics*, **23**(20): 2716–2724, `http://dx.doi.org/10.1093/bioinformatics/btm423`.

[239] Upton GJG and Lloyd JC [2005]: 'Oligonucleotide arrays: information from replication and spatial structure.', *Bioinformatics*, **21**(22): 4162–4168, `http://dx.doi.org/10.1093/bioinformatics/bti668`.

[240] Vapnik VN [1998]: *Statistical Learning Theory*, Wiley-Interscience, ISBN 0471030031.

[241] Veer VL, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al.* [2002]: 'Gene expression profiling predicts clinical outcome of breast cancer.', *Nature*, **415**(6871): 530–536, `http://dx.doi.org/10.1038/415530a`.

[242] Venet D, Dumont JE and Detours V [2011]: 'Most random gene expression signatures are significantly associated with breast cancer outcome.', *PLoS Computational Biology*, **7**(10): e1002240, `http://dx.doi.org/10.1371/journal.pcbi.1002240`.

[243] Venet D, Pecasse F, Maenhaut C and Bersini H [2001]: 'Separation of samples into their constituents using gene expression data.', *Bioinformatics*, **17 Suppl 1**: S279–S287, ISSN 1367-4803.

[244] Vert JP [2010]: *Reconstruction of Biological Networks by Supervised Machine Learning Approaches*, chapter 7, pages 163–188, John Wiley & Sons, Inc., ISBN 9780470556757, `http://dx.doi.org/10.1002/9780470556757.ch7`.

[245] Vert JP, Qiu J and Noble WS [2007]: 'A new pairwise kernel for biological network inference with support vector machines.', *BMC Bioinformatics*, **8 Suppl 10**: S8, `http://dx.doi.org/10.1186/1471-2105-8-S10-S8`.

[246] Vogelstein B and Kinzler KW [2004]: 'Cancer genes and the pathways they control.', *Nature Medicine*, **10**(8): 789–799, `http://dx.doi.org/10.1038/nm1087`.

[247] Wang SE, Wu FY, Chen H, Shamay M, Zheng Q and Hayward GS [2004]: 'Early activation of the Kaposi's sarcoma-associated herpesvirus RTA, RAP, and MTA promoters by the tetradecanoyl phorbol acetate-induced AP1 pathway.', *Journal of Virology*, **78**(8): 4248–4267.

[248] Wang X, Li A, Jiang Z and Feng H [2006]: 'Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme', *BMC Bioinformatics*, **7**(1): 32.

[249] Wang Z, Gerstein M and Snyder M [2009]: 'RNA-Seq: a revolutionary tool for transcriptomics.', *Nature Reviews. Genetics*, **10**(1): 57–63, `http://dx.doi.org/10.1038/nrg2484`.

[250] Weigt M, White RA, Szurmant H, Hoch JA and Hwa T [2009]: 'Identification of direct residue contacts in protein-protein interaction by message passing.', *Proceedings of the National Academy of Sciences*, **106**(1): 67–72, `http://dx.doi.org/10.1073/pnas.0805923106`.

[251] Weisstein EW [2016], ''Correlation Coefficient.' From MathWorld–A Wolfram Web Resource. ', `http://mathworld.wolfram.com/CorrelationCoefficient.html`.

[252] Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF and Hampton GM [2001]: 'Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer', *Cancer Research*, **61**(16): 5974–5978, ISSN 0008-5472, `http://cancerres.aacrjournals.org/content/61/16/5974.full.pdf`, `http://cancerres.aacrjournals.org/content/61/16/5974`.

[253] Wen C, Wu L, Qin Y, Van Nostrand JD, Ning D, Sun B, Xue K, Liu F, Deng Y, Liang Y *et al.* [2017]: 'Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform.', *PLOS ONE*, **12**: e0176716, ISSN 1932-6203, `http://dx.doi.org/10.1371/journal.pone.0176716`.

[254] Wilcoxon F [1945]: 'Individual comparisons by ranking methods', *Biometrics Bulletin*, **1**(6): 80–83, `http://dx.doi.org/10.2307/3001968`.

[255] Wilson CL and Miller CJ [2005]: 'Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis.', *Bioinformatics*, **21**(18): 3683–3685, `http://dx.doi.org/10.1093/bioinformatics/bti605`.

[256] Winston PH [1992]: *Artificial Intelligence, 3rd edition*, Addison Wesley.

[257] Wu M and Chan C [2012]: ‘Learning transcriptional regulation on a genome scale: a theoretical analysis based on gene expression data.’, *Briefings in Bioinformatics*, **13**(2): 150–161, `http://dx.doi.org/10.1093/bib/bbr029`.

[258] Wu WS, Li WH and Chen BS [2007]: ‘Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data.’, *BMC Bioinformatics*, **8**: 188, `http://dx.doi.org/10.1186/1471-2105-8-188`.

[259] Yamanishi Y, Vert JP and Kanehisa M [2004]: ‘Protein network inference from multiple genomic data: a supervised approach.’, *Bioinformatics*, **20 Suppl 1**: i363–i370, `http://dx.doi.org/10.1093/bioinformatics/bth910`.

[260] Yeoh EJ, Ross ME, Shurtleff SA, Williams KW, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A *et al.* [2002]: ‘Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling’, *Cancer Cell*, **1**(2): 133–143.

[261] Yeung KY, Haynor DR and Ruzzo WL [2001]: ‘Validating clustering for gene expression data’, *Bioinformatics*, **17**(4): 309–318, `http://dx.doi.org/10.1093/bioinformatics/17.4.309`.

[262] Yoon H, Liyanarachchi S, Wright FA, Davuluri R, Lockman JC, de la Chapelle A and Pellegata NS [2002]: ‘Gene expression profiling of isogenic cells with different TP53 gene dosage reveals numerous genes that are affected by TP53 dosage and identifies CSPG2 as a direct target of p53.’, *Proceedings of the National Academy of Sciences*, **99**(24): 15632–15637, `http://dx.doi.org/10.1073/pnas.242597299`.

[263] Zheng W, Zhao H, Mancera E, Steinmetz LM and Snyder M [2010]: ‘Genetic analysis of variation in transcription factor binding in yeast.’, *Nature*, **464**(7292): 1187–1191, `http://dx.doi.org/10.1038/nature08934`.

[264] Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I and Enard W [2017]: ‘Comparative Analysis of Single-Cell RNA Sequencing Methods.’, *Molecular cell*, **65**: 631–643.e4, ISSN 1097-4164, `http://dx.doi.org/10.1016/j.molcel.2017.01.023`.

[265] Zimek A, Schubert E and Kriegel HP [2012]: ‘A survey on unsupervised outlier detection in high-dimensional numerical data’, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **5**(5): 363–387.

[266] Zou H and Hastie T [2005]: ‘Regularization and variable selection via the Elastic Net’, *Journal of the Royal Statistical Society, Series B*, **67**: 301–320.