Doctoral thesis
for a doctoral degree at the
Graduate School of Life Sciences,
Julius-Maximilians-Universität,
Würzburg

Integrated functional analysis of biological networks

Integrierte funktionelle Analyse biologischer Netzwerke

Submitted by
Daniela Beisser

from
Hattingen

December 7, 2011

Submitted on: .....

Members of the Promotionskomitee:

Chairperson:	Prof. Dr. Wolfgang Rössler
Primary Supervisor:	Prof. Dr. Thomas Dandekar
Supervisor (Second):	Prof. Dr. Roy Gross
Supervisor (Third):	Prof. Dr. Jörg Schultz

Internal Supervisors:

Primary Supervisor:	Dr. Tobias Müller
Supervisor (Second):	Dr. Dr. Marcus Dittrich

Date of Public Defence: .....

Date of Receipt of Certificates: .....

Acknowledgements

This dissertation was conducted at the Department of Bioinformatics at the Julius-Maximilians-Universität Würzburg. During the course of my dissertation, I was a member of the Graduate School of Life Sciences and part of the Graduate College Würzburg-Nice "Signal Transduction: Where Cancer and Infection Converge".

Foremost, I want to express my special gratitude to the members of my supervisory committee Prof. Thomas Dandekar, Prof. Roy Gross, and Prof. Jörg Schultz for the advises and support I received from them during the conductance of my research. I would like to thank especially Thomas Dandekar for giving me the opportunity to work in such a stimulating research context.
I am grateful for the funding by the BMBF Funcrypta project and DFG project SPP1316 and the prosperous collaboration with its members.

This thesis would not have been possible without my advisor Tobias Müller, who set up the foundations of the project and guided me in an outstanding way through the entire process of developing own research ideas, performing statistical analyses, presenting and discussing results, writing publications, grants and referee comments. Thanks for the never-ending source of ideas and hours of time and effort to answer all my questions.
I am also particularly grateful to Marcus Dittrich for laying out the theoretical and methodological foundations of my research in his doctoral thesis and the lively discussions in our group meetings. Following this, I would like to show my gratitude to all other current and former member of the Networking group, most notably Desi, Santosh and Benny in Würzburg and Gunnar Klau in Amsterdam.

I want to thank in general all members of the department for creating such a friendly and stimulating context that widened my scientific understanding of the life sciences in many aspects.
In particular, I want to thank Gaby for the helpful and constructive comments on my dissertation, Christian for all the help with the cluster and linux and the great time at work and afterwards, during bike tours, EMBL visits, snowboard weekends and countless other activities. And not to forget the India-crew, with which I enjoyed an incredible time and journey.

Thanks Christian, Gaby, Freda, Desi, Santosh, Astrid and Torben for making

Würzburg feel like home and a great place to live for the last three years.

Lastly, and surely the most, I would like to thank my parents and Achim. They were always supporting and encouraging me in all my pursuits.

We are drowning in information, while starving for wisdom.

(Edward O. Wilson, 1998)

Inhaltsverzeichnis

Zusammenfassung

In den letzten Jahren haben Hochdurchsatz-Experimente gewaltige Mengen an molekularbiologischen Daten geliefert, angefangen mit dem ersten sequenzierten Genom von *Haemophilus influenzae* im Jahr 1995 und dem menschlichen Genom im Jahr 2001. Mittlerweile umfassen die resultierenden Daten neben der Genomik die Bereiche der Transkriptomik, Proteomik und Metabolomik. Die Analyse der Daten mithilfe von bioinformatischen Methoden hat sich entsprechend mit verändert und weiterentwickelt. Durch neuartige, systembiologische Ansätze versucht man zu verstehen, wie Gene und die aus ihnen resultierenden Proteine, biologische Formen und Funktionen entstehen lassen. Dabei interagieren sie miteinander und mit anderen Molekülen in hoch komplexen Strukturen, welche durch neue Ansätze der Netzwerkbiologie untersucht werden. Das tiefgreifende Wissen über einzelne Moleküle, verfügbar durch Hochdurchsatz-Technologien, kann komplementiert werden durch die Architektur und dynamischen Interaktionen molekularer Netzwerke und somit ein umfassenderes Verständnis biologischer Prozesse ermöglichen.
Die vorliegende Dissertation stellt Methoden und statistische Analysen zur Integration molekularer Daten in biologische Netzwerke, Identifikation robuster, funktionaler Subnetzwerke sowie die Anwendung auf verschiedenste biologische Daten vor. Der integrative Netzwerkansatz wurde als ein Softwarepaket, BioNet, in der statistischen Programmiersprache R implementiert. Das Paket beinhaltet statistische Verfahren zur Integration transkriptomischer und funktionaler Daten, die Gewichtung von Knoten und Kanten in biologischen Netzwerken sowie Methoden zur Suche signifikanter Bereiche, Module, und deren Visualisierung. Der exakte Algorithmus wird ausführlich in einer Simulationsstudie getestet und übertrifft heuristische Methoden zur Lösung dieses NP-vollständigen Problems in Genauigkeit und Robustheit. Die Variabilität der resultierenden Lösungen wird bestimmt anhand von gestörten integrierten Daten und gestörten Netzwerken, welche zufällige und verzerrende Einflüsse darstellen, die die Daten verrauschen. Ein optimales, robustes Modul kann durch einen Konsensusansatz bestimmt werden. Basierend auf einer wiederholten Stichprobennahme der integrierten Daten, wird ein Ensemble von Lösungen erstellt, aus welchem sich das robuste und optimale Konsensusmodul berechnen lässt. Zusätzlich erlaubt dieser Ansatz eine Schätzung der Variabilität des Konsensusmoduls und die Berechnung von Konfidenzwerte für Knoten und Kanten.

Der Ansatz wird anschließend auf zwei Genexpressionsdatensätze angewandt. Die erste Anwendung untersucht Genexpressionsdaten für akute lymphoblastische Leukämie (ALL) und analysiert Unterschiede in Subgruppen mit und ohne BRC/ABL Genfusion. Das resultierende funktionale Modul identifiziert Gensignaturen, von denen einzelne Gene bereits mit der Erkrankung in Zusammenhang gebracht wurden oder zur Klassifikation der Subgruppen benutzt wurden, zum Beispiel die Tyrosinkinasegene ABL, FYN und YES1. Das Modul zeigt eine Anreicherung biologischer Prozesse, welche durch die Translokation entstehen, beispielsweise die Peptidyl-Tyrosin Phosphorylierung. Bei einem schwachen Signal, wie es in diesen Genexpressionsprofilen vorhanden ist, hat das Konsensusmodul den Vorteil, dass es robuste Bereiche des Moduls durch hohe Konfidenzwerte identifiziert. Regionen mit niedrigen Konfidenzwerten stammen wahrscheinlicher von verrauschten Daten als von einem Signal in den Genexpressionsdaten. Der Konsensusansatz detektiert zusätzliche Gene, welche auf Grund des schwachen Signals und der verzerrenden Einflüsse nicht in dem ursprünglichen Modul erkannt wurden, aber ebenfalls mit einer Behandlung von ALL oder Klassifikation der analysierten Subtypen assoziiert sind.
Die zweite Anwendung wertet Genexpressions- und Lebenszeitdaten für diffuse großzellige B-Zell Lymphome (DLBCL) aus, beruhend auf molekularen Unterschieden zwischen zwei DLBCL Subtypen mit unterschiedlicher Malignität. Das berechnete Module beinhaltet zwei bekannte Signaturen, die Proliferations-Signatur und die NFκB-Signatur. Robuste Bereiche des Konsensusmoduls erweitern die gefundene NFκB-Signatur um ein weiteres NFκB-Zielgen.
Die resultierenden Module identifizieren und erweitern bekannte Genlisten und Signaturen um weitere signifikante Gene und deren intermolekularen Interaktionen. Die wichtigste Neuerung ist dabei, dass die Gene im Kontext ihrer Interaktionen bestimmt und visualisiert werden statt als Auflistung einzelner und voneinander unabhängiger Transkripte.
In einer dritten Anwendung wird der integrierte Netzwerkansatz benutzt, um Veränderungen im Metabolismus von Tardigraden aufzuspüren und Signalwege zu identifizieren, welche für die extreme Anpassungsfähigkeit an wechselnde Umweltbedingungen und Überdauerung in einem inaktiven Tönnchenstadium verantwortlich sind. Zum ersten Mal wird dafür ein metabolischer Netzwerkansatz vorgeschlagen, der metabolische Veränderungen durch die Integration von metabolischen und transkriptomischen Daten bestimmt. Metabolische Profile werden zur Gewichtung der Knoten des Netzwerks benutzt. Dafür wird ein statistisches Konzept etabliert, welches Metabolite mit einem signifikanten U-förmigen Trend in der Zeit identifiziert. Die Kanten hingegen werden anhand von Enzymen gewichtet, die in exprimierten Sequenzen (ESTs) identifiziert wurden. Durch die Kombination der Daten wird ein Subnetzwerk bestimmt, das die gemeinsamen Änderungen der metabolischen Signalwege erklärt. Das Modul beschreibt den Rückgang eines

messbaren Metabolismus während der Ausbildung eines Tönnchenstadiums, die Produktion von Speichermolekülen und biologischen Schutzstoffen, wie DNA-Stabilisierern, und den Aufbau von Aminosäuren und Zellkomponenten aus Monosacchariden während der Umwandlung zurück in eine aktive Lebensform. Die integrierte Netzwerkanalyse ist in diesem Fall eine geeignete Methode zur Kombination und gemeinsamen Analyse der spärlich verfügbaren Daten.
Abschließend ist zu bemerken, dass die präsentierte integrierte Netzwerkanalyse eine adäquate Technik ist, um experimentelle Daten aus Hochdurchsatz-Methoden, die spezialisiert auf eine Molekülart sind, mit ihren intermolekularen Wechselwirkungen und Abhängigkeiten in Verbindung zu bringen. Sie ist flexibel in der Anwendung auf verschiedenste Daten, von der Analyse von Genexpressionsveränderungen, über Metabolitvorkommen bis zu Proteinmodifikationen, in Kombination mit einem geeigneten molekularen Netzwerk. Der exakte Algorithmus ist akkurat und robust in Vergleich zu heuristischen Methoden und liefert eine optimale, robuste Lösung in Form eines Konsensusmoduls mit zugewiesenen Konfidenzwerten. Durch die Integration verschiedenster Informationsquellen und gleichzeitige Betrachtung eines biologischen Ereignisses von diversen Blickwinkeln aus, können neue und vollständigere Erkenntnisse physiologischer Prozesse gewonnen werden.

Abstract

In recent years high-throughput experiments provided a vast amount of data from all areas of molecular biology, including genomics, transcriptomics, proteomics and metabolomics. Its analysis using bioinformatics methods has developed accordingly, towards a systematic approach to understand how genes and their resulting proteins give rise to biological form and function. They interact with each other and with other molecules in highly complex structures, which are explored in network biology. The in-depth knowledge of genes and proteins obtained from high-throughput experiments can be complemented by the architecture of molecular networks to gain a deeper understanding of biological processes.
This thesis provides methods and statistical analyses for the integration of molecular data into biological networks and the identification of functional modules, as well as its application to distinct biological data. The integrated network approach is implemented as a software package, termed `BioNet` , for the statistical language R. The package includes the statistics for the integration of transcriptomic and functional data with biological networks, the scoring of nodes and edges of these networks as well as methods for subnetwork search and visualisation. The exact algorithm is extensively tested in a simulation study and outperforms existing heuristic methods for the calculation of this NP-hard problem in accuracy and robustness. The variability of the resulting solutions is assessed on perturbed data, mimicking random or biased factors that obscure the biological signal, generated for the integrated data and the network. An optimal, robust module can be calculated using a consensus approach, based on a resampling method. It summarizes optimally an ensemble of solutions in a robust consensus module with the estimated variability indicated by confidence values for the nodes and edges. The approach is subsequently applied to two gene expression data sets. The first application analyses gene expression data for acute lymphoblastic leukaemia (ALL) and differences between the subgroups with and without an oncogenic BCR/ABL gene fusion. Several genes contained in the resulting module, e.g. tyrosine kinase genes (ABL, FYN and YES1,) have been identified previously to be important for the discrimination of the subgroups. The module shows an enrichment for biological processes like peptidyl-tyrosine phosphorylation, which arise due to the translocation. For a weak signal, as in these expression profiles, the consensus module has the advantage of identifying robust parts of the module with high confidence values. Regions

with low confidence values could arise from the noisy data, rather than a signal in the gene expression profiles. The consensus approach discovers additional genes, which are not recognized in the original module due to the high level of noise, but which are also associated with the treatment of ALL and the classification of the analysed subtypes.
In a second application gene expression and survival data from diffuse large B-cell lymphomas are examined. The DLBCL module includes two established signatures for the difference between two DLBCL subtypes with varying malignancies, the proliferation signature and the NFκB signature. Robust parts of the consensus module determined for the DLBCL dataset extend the NFκB signature by an additional NFκB target gene.
Thereby, the identified modules include and extend already existing gene lists and signatures by further significant genes and their interactions. The most important novelty is that these genes are determined and visualised in the context of their interactions as a functional module and not as a list of independent and unrelated transcripts.
In a third application the integrative network approach is used to trace changes in tardigrade metabolism to identify pathways responsible for their extreme resistance to environmental changes and endurance in an inactive tun state. For the first time a metabolic network approach is proposed to detect shifts in metabolic pathways, integrating transcriptome and metabolite data. The metabolite profiles are used to score the nodes of the network by applying a novel statistical framework to identify metabolites with a significant trend over time. The edges are scored according to information on enzymes from expressed sequences (ESTs). Using this combined information, a key subnetwork of concerted changes in metabolic pathways is identified. The module resembles the cessation of a measurable metabolism during the tun formation, the production of storage metabolites and bioprotectants, such as DNA stabilisers, and the generation of amino acids and cellular components from monosaccharides as carbon and energy source during the transition back into an active form. With sparse and diverse data available on tardigrades, the integrated network analysis is a convenient method to bring together all available data and analyse it in a joint manner.
Concluding, the presented integrated network approach is an adequate technique to unite high-throughput experimental data for single molecules and their intermolecular dependencies. It is flexible to apply on diverse data, ranging from gene expression changes over metabolite abundances to protein modifications in a combination with a suitable molecular network. The exact algorithm is accurate and robust in comparison to heuristic approaches and delivers an optimal, robust solution in form of a consensus module with confidence values. By the integration of diverse sources of information and a simultaneous inspection of a molecular event from different points of view, new and exhaustive insights into biological processes can be acquired.

Chapter 1

Introduction

> "The goal of network systems biology is to mine knowledge on the basis of the network data generated from high-throughput techniques by exploiting special features of the biological system, and gain biological insight by further interpreting them in a systematic matter." (Chen et al., 2009)

The above quotation gives a convenient definition of the goals of network systems biology. To this I would like to add for *integrated functional network analysis*, that it does not mine knowledge and gain new insights solely on the basis of the network data, but uses the network to interpret further biological data. Like the discovery of the DNA structure by Watson and Crick did not reveal directly its functioning or the function of individual parts of it, the structure of a cellular network alone will not unveil the operating principles of the cell. A network of physical interactions between proteins, of correlated genes, or of reactions between metabolites gives information about the connectivity and reciprocal influence of its constituents, but can furthermore be used to analyse several sources of molecular data in a combined manner.

High-throughput experimental methods in molecular biology have resulted in the last decade in tremendous amounts of data and are still generating ever more data with an accelerating speed. With the development of microarrays to analyse the transcriptome and recently the even more sophisticated methods of second generation and third generation sequencing to analyse the genome, transcriptome or epigenetic modifications, data are generated faster then it can be processed. Different aspects of a molecular mechanism are continued to be examined with these techniques, without the possibility to combine them straightforwardly. Hence, it is crucial to develop methods

and tools to analyse the obtained data in a systems approach.

This thesis investigates how large-scale biological data can be analysed with the help of network biology. By combining information about the interplay of the molecules that make up the cell such as proteins and metabolites with measured molecular data, the integrated network analysis extends the sole analysis of networks as well as the independent analysis of molecular data. It is well-known that the proteins, genes, transcripts or metabolites studied in biology are not independent of each other and should not be analysed as such. Rather the biologist is interested in the complete picture, how the single parts of the cell form a functioning unit, a complete organism and even a communicating group of individuals. As well as what might go wrong during this process. Network biology is a move towards this understanding, beginning as a first step with the analysis of molecules in dependence on their functional and physical interactions.

The background necessary for the conductance of this study is introduced in Chapter 2. The ideas behind it stem from biology, statistics and computer science and are presented in detail to make them understandable for biologists and bioinformaticians with a diverse range of backgrounds. First, graph theory is introduced as the methodological framework for network analysis with important definitions and concepts. I proceed with the application of graph theory to biology with the presentation of the different types of biological networks. Of course ultimately they should be combined in one large network, describing the complete interactions in a cell, but until then, they are split into networks covering protein interactions, transcription and metabolites separately. A review of the most important results from topological analysis and integrated analysis is given thereupon.
The next part covers the analysis of gene expression data and the required statistical methods in great detail, as this are the data I mostly worked with. It should provide an introduction to non-experts, but also give a comprehensive summary of the methodology for me and the readers of this thesis.

Chapter 3 presents several methods which are used for the descriptive analysis of the molecular data, analysis of obtained results and testing for trends in time series data. In addition the biological material analysed and used in this thesis is described. It covers on the one hand the high-throughput data used for integration: gene expression profiles, expressed sequence tags and mass spectrometry profiles for different applications and on the other hand the interaction networks.

The main results of the thesis are presented and discussed in Chapter 4. First, the integrated network analysis is presented including the statistical approach to convert p-values obtained from statistical tests into scores, the scoring of the network and calculation of a maximum-scoring subnetwork (module) thereupon. Two algorithms are presented to calculate an

exact solution of the maximum-scoring subnetwork problem and a heuristic approach. I implemented all methods necessary for the analysis in the R package `BioNet`, which is described subsequently, including a case study. The major results of this part have been published in Dittrich et al. (2010), describing the integrated network approach and in Beisser et al. (2010), describing the software package and its application to microarray data.
The next part deals with the evaluation of the presented approach and comparisons to established methods in an extensive simulation study. The methods are examined with regard to accuracy and robustness of the obtained solution. To determine the variance of the solutions and different sources of variability, modules are calculated using perturbed molecular data and perturbed networks. This analysis shows that the modules differ in parts greatly due to the noise in the biological data. It is therefore necessary to assess this variability by statistical methods and indicate robust parts of the module. The technique to do this is presented in the next part of the results. It explains how to calculate confidence values for modules using a resampling technique and how to summarise the results from the resampling in a consensus module.
The presented approach to calculate modules and consensus module is thereupon applied to gene expression profiles. An integrated network analysis is performed for acute lymphoblastic leukaemia and diffuse large B-cell lymphoma. The results and advantages of the methodology with respect to the analysed diseases are discussed afterwards. The major results of this part are in submission for publication.
In the last part of the results the metabolic changes of *M.tardigradum* during transition from an active state to an inactive state are examined using integrated network analysis. This includes a novel statistical test for trends in metabolic profiles as well as the calculation of a maximum-scoring subnetwork using node and edge scores. This new type of integrated analysis of metabolic changes, as described here, is also in submission for publication.

Chapter 5 discusses general points and implications of the presented approaches to other fields of biology as well as further development and future trends for integrated network analysis. Some further methodological techniques investigated during my PhD studies are touched upon.

Chapter 2

Background

2.1 Graph Theory

Biological networks are represented formally as graphs and analysed by applying concepts and mechanisms from graph theory. The history of graph theory reaches back to 1736 and Leonard Euler's "Königsberg bridge problem". In Königsberg seven bridges cross the river Pregel to connect four parts of the city (Figure 2.1 A). The question was whether it is possible to visit all parts of the town by crossing all bridges only once. The problem was solved by Euler (showing that it's not possible) by applying methods today known as graph theory.

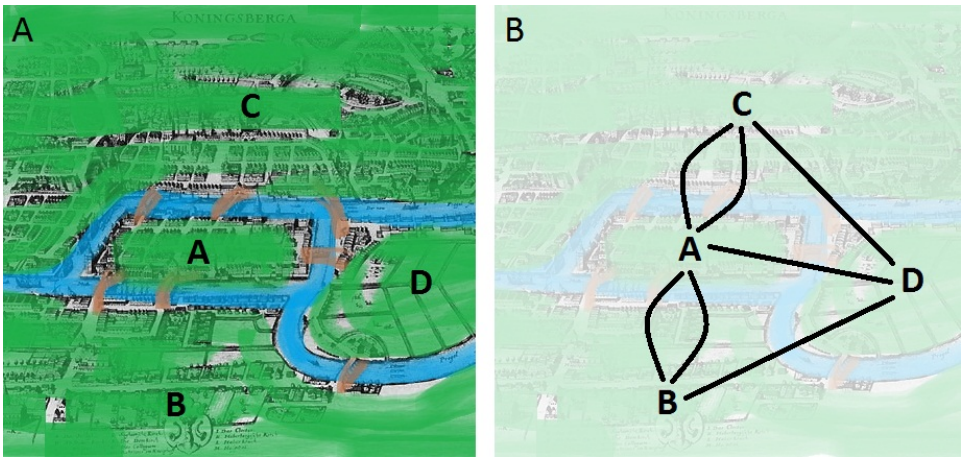


Figure 2.1: The Königsberg bridge problem, modified picture after (Merian-Erben, 2006) and (Junker and Schreiber, 2008). (A) Shows the town of Königsberg with four parts of the city connected via seven bridges. (B) shows a graph representation of (A).

Figure 2.1 B shows Königsberg represented as a graph. A graph $G = (V, E)$ consists of a set of vertices V and a set of edges E, connecting the vertices. In the Königsberg example the nodes represent the parts of the city and the edges the bridges connecting them. Nodes u and v are adjacent or neighbours if they are connected through an edge e, respectively incident with e. An edge beginning and ending in the same vertex is called a self-loop. Since the bridges can be crossed in both directions, the edges are chosen to be undirected. Examples of directed graphs and undirected graphs are given in Figure 2.2. Here the nodes A, B, D form a cycle in the graph, graphs without such cycle are called acyclic graphs and if directed: directed acyclic graphs (DAG).

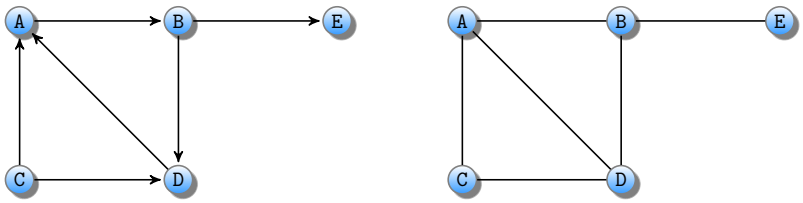


Figure 2.2: Examples for an directed and an undirected graph.

All of the examples above consist of graphs with one connected component. Figure 2.3 depicts a graph with two components, where the nodes A, B, C, D and their edges form the largest connected component of this graph.

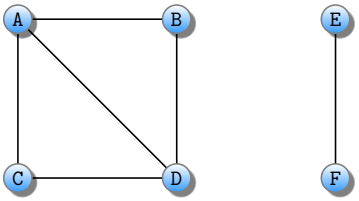


Figure 2.3: Examples for a graph with two components.

A part of a graph G defined by $G' = (V', E')$ is called a subgraph, with an edge set E' connecting the subset of vertices V'. If all edges E' that connect the vertices V' are the same as in the original edge set E, the subgraph is called an induced subgraph. The differences are depicted in Figure 2.4.

Some important definitions and topological measures applied to graphs are given in the following:

- The *shortest path* between two vertices is a sequences of nodes and edges $(v_0, e_1, v_1, ..., e_n, v_n)$ with distinct edges that have the shortest length. For example, the shortest path between the vertex A and E in Figure 2.2 is ({A,B}, {B,E}) with a distance of 2.

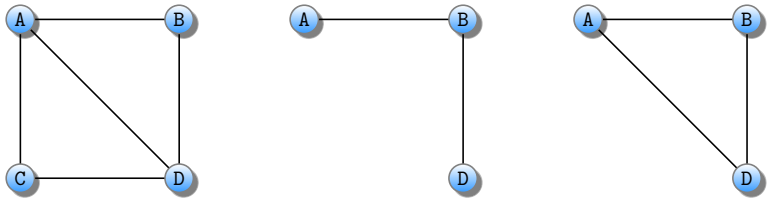


Figure 2.4: A graph G, a subgraph G' and the induced subgraph G'_i

- The *distance* d_{ij} is the length of the shortest path between vertices i and j. In Figure 2.2 $d_{AE} = 2$.
- The *diameter* $d_m = max(d_{ij})$ is defined as the longest shortest path in the graph. In the right graph of Figure 2.2 $d_m = 3$.
- The *average path length* $d = \langle d_{ij} \rangle$ is the average distance between all pairs of nodes.
- The *degree* k_i is given by the number of edges adjacent to vertex. In the right graph of Figure 2.2 $k_A = 3$.
- A *hub* is a highly connected node in the graph. Hubs are of peculiar importance, since the failure of hub nodes causes a breakdown of the network into isolated clusters, while failure of random nodes mostly affects small-degree nodes (Albert and Barabási, 2002).
- The *average degree* of a graph is given by $\langle k \rangle = \frac{2E}{V}$.
- The *degree distribution* $p(k)$ of a graph gives the probability that the degree of a vertex equals k. The degree distribution has become an important characteristic in topological network analysis, by which different network models can be tested. Depending on the shape of the distribution, a network is assigned to a e.g. scale-free, random, or hierarchical network model. Since the presented study does not cover the topological analysis of biological network, I will not go into detail here. More information on topological analysis and network models can be found elsewhere (Barabási and Bonabeau, 2003; Barabási and Oltvai, 2004; Junker and Schreiber, 2008).
- The *local clustering coefficient* $C_i = \frac{2E_i}{k_i(k_i-1)}$ measures the local clustering in the graph by calculating the ratio of connected neighbours of a node i to the maximum possible pairwise interactions between all neighbours of node i.
- The *global clustering coefficient* $C = \langle C_i \rangle$ is the average of the local clustering coefficients.
- The *modularity* quantifies the quality of divisions of the network into modules. By this definition good modules are those with dense internal

connections between the nodes within modules and sparse connections between different modules. The definition of functional modules is different to this one and given later (Section 4.1.1).

Most networks are represented in a form, in which there is only one type of node and one type of edge. An example for that is the protein-protein-interaction network, where the proteins are represented as the nodes and the interactions between them as the edges of the graph. But some networks are represented as bipartite graphs. These graphs possess two distinct types of nodes. Exemplary for these graphs are metabolic networks. Here, the nodes often represent enzymes as well as metabolites, which alternate within the graph. A substrate is followed by an enzyme, which is in turn followed by a product of the reaction.
Another special type of graph is a tree. It is an undirected, connected, acyclic graph. In network analysis one kind of tree is often of importance, which is the (minimum) spanning tree (MST). A spanning tree T is composed of all the vertices of a graph G, but contains a minimal set of edges to connect all vertices (with minimal cost). A spanning tree reduces the complexity of the graph and is often used for approximate solutions for hard algorithmic problems.

2.2 Biological Network Analysis

2.2.1 Biological Networks

High-throughput experimental methods in molecular biology yield an enormous amount of data in diverse areas, termed -omics. The basic components of cell biology consisting of genes, mRNA, proteins and metabolites constitute the genome, transcriptome, proteome and metabolome. The relations and interactions between them account for the functioning of the cell. These relationships can be described in networks, where the elements are the vertices and the relationships are the edges of the network. This approach aims at a "systems-level [or wholistic] understanding of biological systems" (Kitano, 2002).
Different kind of networks can be generated at different levels of biological processes using various high-throughput datasets. Starting at the genome level, transcription regulatory networks describe the interactions between transcription factors (TF) and DNA (Figure 2.5). TFs bind to TF-binding sites on the DNA and act as transcriptional activators or repressors by influencing transcription via interactions with the RNA polymerase or associated proteins, or by altering the chromatin structure. TFs themselves are proteins and the translational product of genes, therefore their action can also be described on the level of gene regulatory networks. The elements of gene regulatory networks are genes and their interactions show which transcriptional product influences the expression of which other genes.
Another way of regulation takes place on the protein level. By the formation of protein complexes, posttranslational modifications such as phosphorylations, structural changes due to the binding of other proteins, proteins can act and influence each other. This level is characterised by protein-protein interaction (PPI) networks. A special group of proteins, enzymes, catalyse metabolic reactions, by converting substrate metabolites into product metabolites. Enzymatic reactions e.g. degrade food molecules and produce energy in catabolism and generate macromolecules such as nucleic acids and fatty acids by consuming energy in anabolic processes. The biological reactions in a cell are depicted in metabolic networks, that link the substrates and the products of a reaction or use bipartite graphs that have two distinct kinds of nodes, one for the metabolites and the other for the reaction or enzyme catalysing the reaction.

The -omics datasets that are used to construct the different biological networks are described in the following subsections. Additional kinds of networks exist, which are not a direct result of experimental data, such as correlation networks, which are inferred from -omics data, but do not necessarily represent direct interactions or causal relationships between the constituents.

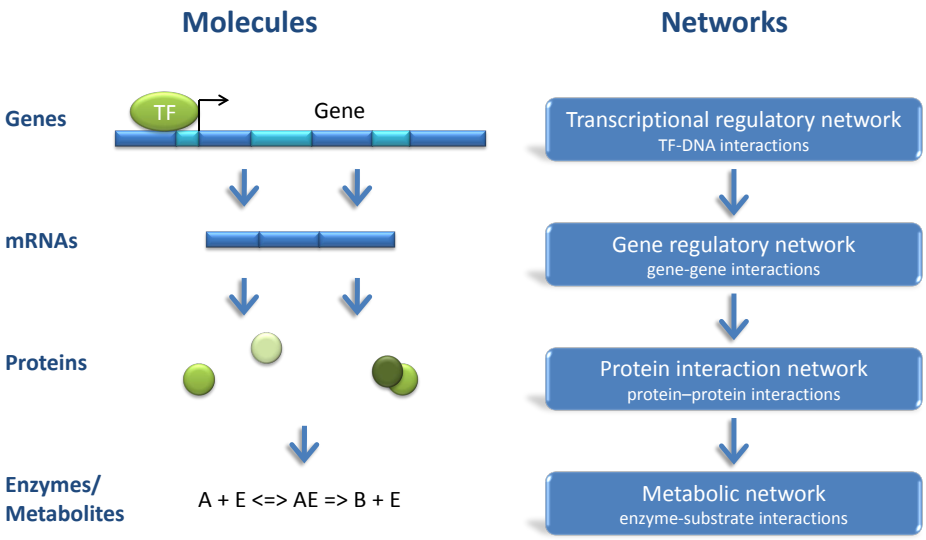


Figure 2.5: The molecules (left panel) that constitute the networks (right panel) are described for different molecular levels: genomic, transcriptomic, proteomic and metabolomic. The figure was created after Chen et al. (2009).

Transcription and Gene Regulatory Networks

Transcription profiling was one of the first large scale approaches developed to analyse the gene expression of thousands of genes in parallel. Common techniques for this purpose are microarrays (Section 2.3), serial analysis of gene expression (SAGE), and nowadays RNA sequencing (RNA-Seq). From these data conclusions can be drawn about the regulatory interaction between genes, especially after knock-outs of known transcription factors and systematic gene perturbation experiments.

Chromatin immunoprecipitation with microarrays technology (ChIP-chip) or subsequent sequencing (ChIP-Seq) is used to determine protein-DNA interactions on a genome-wide basis. Crosslinked protein-target DNA complexes are immunoprecipitated after digestion into fragments of 300-1,000 base pairs. The purified bound DNA from the complexes is identified using microarrays or high-throughput sequencing techniques. The recognised regulatory regions and transcription factor binding sites are used to specify TF-DNA interactions in transcription regulatory networks.

Protein-Protein Interaction Networks

The high-throughput analysis of the proteome has shown to be more difficult than the analysis of the transcriptome, but is now beginning to be systematically explored (Chen et al., 2009) to determine the function of all proteins and their interactions. Protein complexes are identified by two-dimensional gel electrophoresis (2-DE), by which the first electrophoresis separates the complexes by size and the second electrophoresis separates the proteins that constitute the complex by mass.
Another method to identify and quantify proteins is mass spectrometry (MS), often in combination with tandem affinity purification (TAP-MS) to study protein-protein interactions (Rigaut et al., 1999). A TAP-tagged protein is washed through two affinity columns and the binding partners of the protein are examined by MS after the purification.
Other immunoprecipitation and pull-down methods to identify protein complexes such as co-immunoprecipitation (Co-IP) exist, that are antibody-based techniques to determine the interaction partners of the isolated protein.
Several large-scale yeast two-hybrid screens have provided a multitude of interaction data. This technique is based on two fusion-proteins, called bait and prey, to identify interactions between them. The bait protein is fused to a binding domain of a transcription factor, the prey protein to the activating domain of the TF. Only upon interaction of the bait and prey proteins the TF is able to bind to the TF-binding site and activate the transcription of a reporter gene (e.g. LacZ). The screens are conducted in mutant yeast strains, introduced with the bait and prey plasmids.

Metabolic Networks

Mass spectrometry and nuclear magnetic resonance (NMR) are used to quantitatively analyse the metabolites of tissues and cells of many organisms. But it is not yet an advanced enough technique to apply it to new organisms in an -omics approach. The sequencing of complete genomes makes it possible to reconstruct the network of biochemical reactions in many organisms, some of which are available online in databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000), BioCyc (Caspi et al., 2010), BRENDA (Scheer et al., 2011) or Reactome (Croft et al., 2011).

2.2.2 Topological Analysis

The analysis of networks was applied to many other disciplines including social sciences, physics, economics and computer science, before moving into the field of biology. With the advent of high-throughput technologies and the massive amount of publicly available data, the interest in network biology gained significantly. Around the turn of the century, the first cellular networks were analysed and compared. These investigations focused on topological properties of the networks to deduce from these general principles and assertions about the networks.
Initial studies on biological networks focused on topological properties of the network (Hartwell et al., 1999; Barabási and Oltvai, 2004; Ravasz et al., 2002). These included calculations of the degree of the nodes, highly connected nodes (hubs), the degree distribution of the network, local and global clustering coefficients, motifs etcetera.
An important finding was that cellular networks exhibit a scale-free topology. Scale-free networks are characterised by a power-law degree-distribution, where few nodes with many connections (hubs) connect many nodes with only few links. This was previously identified in social and technological networks and first shown for biological networks in metabolic networks for different organisms by Jeong et al. (2000) and Wagner and Fell (2001). The scale-free topology presumably arises from gene duplications and the acquisition of many interactions by the older nodes of the networks, which was shown for protein interaction networks by Wagner (2003) and Eisenberg and Levanon (2003) and for metabolic networks by Jeong et al. (2000) and Wagner and Fell (2001). They also revealed, that metabolic networks are ultra-small, connecting most metabolites via three or four reactions.
Another finding was the high average clustering coefficient $< C >$ in protein interaction networks (Wagner and Fell, 2001) and metabolic networks. A high $< C >$ hints to the presence of modularity or highly interconnected nodes which likely share a similar function, since the components of a biological system do not act independently from each other, but are organised into functional units (Hartwell et al., 1999). Several methodologies were proposed to identify complexes, cliques and functional modules in PPI networks (Xu et al., 2004; Qi et al., 2008; Chua et al., 2008). In addition, topological modules are often organised hierarchically with small modules that assemble to larger ones (Ravasz and Barabási, 2003).
The high clustering of networks leads to a set of commonly seen subgraphs, termed motifs. Motifs consist of locally highly clustered subgraphs with few nodes, forming e.g. pentagons, squares and triangles. In transcription-regulatory networks triangular motifs were identified as feed-forward loops (Milo et al., 2002; Shen-Orr et al., 2002).

Despite the important findings of conserved topological features and general

mechanisms in biological networks, which showed that most biological networks can be assigned to have a small-world, scale-free, modular network architecture, the applied aspects of this work to biological questions is still small. One way to gain more insight and use network biology for real-world questions is to integrate cellular networks with other available molecular data.

2.2.3 Integrated Analysis

Multiple genome-scale datasets allow today to model the cell as an intricate network of molecular interactions. Research in systems biology has shifted accordingly, now focusing on network inference from high-throughput genomic data (Friedman, 2004; Schlitt and Brazma, 2007; Djebbari and Quackenbush, 2008; Soong et al., 2008; van Steensel et al., 2010), data integration (Hanisch et al., 2002; Ofran et al., 2006; Bergholdt et al., 2007; Emig et al., 2008; Huttenhower et al., 2009; Bergholdt et al., 2009; Brorsson et al., 2009), network alignments and comparative analysis of networks (Kelley et al., 2004; Bork et al., 2004; Flannick et al., 2006; Sharan and Ideker, 2006; Kalaev et al., 2008), pathway predictions (Rahnenführer et al., 2004; Scott et al., 2006; Karp et al., 2010; Dale et al., 2010) and visualisation of large networks (Goldovsky et al., 2005; Cline et al., 2007).
Reaching beyond the analysis of mere topological questions, integrated network analysis tries to incorporate additional molecular datasets into the network. These can be either used to improve the network by adding further data to confirm interactions between molecules, or to see the network as a context of functional or physical interactions in which the genes or proteins reside for further analysis. For gene expression data integrated approaches are used to search for pathways, functional modules or gene signatures containing differentially expressed genes in the context of genetic interaction networks or PPI networks.
Ofran et al. (2006) use the reliability of interactions based on gene ontology (GO) annotations and subcellular localisation to analyse pathways and processes in an integrated approach. Likewise Huttenhower et al. (2009) use supporting data for the interactions between genes from evolutionary conservation and nucleosome positioning. In an Bayesian approach they integrate the supporting data with primary data from gene expression and DNA sequences to calculate regulatory modules of co-regulated genes. Others combine genetic interactions or SNP genotyping data with protein interaction networks to search for enriched subnetworks for complex diseases (Bergholdt et al., 2007, 2009; Brorsson et al., 2009). Emig et al. (2008) integrate expression data with domain interaction networks. This approach allows to study alternative transcripts and their corresponding domain architecture. Hanisch et al. (2002) also combine information from expression

data and biological networks. Based on this they compute a joint clustering of genes and vertices of the network.

Dittrich et al. (2008) recently developed a method to integrate multiple genomic datasets into biological networks, by transforming p-values for each gene derived from statistical tests into node scores of a given network and searching for functional modules. The node scores reflect the functional relevance of each node; the higher the score, the larger the node's significance. An integer-linear programming algorithm is used to find the highest-scoring subgraph, a problem that has been proven to be NP-hard (Ideker et al., 2002). Various heuristic approaches have been proposed (Rajagopalan and Agarwal, 2005; Nacu et al., 2007; Sohler et al., 2004; Cabusora et al., 2005; Guo et al., 2007; Scott et al., 2006), most of them inspired by the seminal work from Ideker et al. (2002) that used a simulated annealing heuristic to identify high-scoring subgraphs in integrated networks.

2.3 Gene Expression Analysis

Biological information about gene function, regulation and interactions can be determined by measuring which genes are induced or repressed under certain conditions. Genes whose expression level rise and fall under the same conditions are likely to have a related biological function and perhaps a common regulatory relationship (Mount, 2004). Various techniques have been developed to take a snapshot of gene expression levels of thousands of genes in a single experiment. These include serial analysis of gene expression (Velculescu et al., 1995), oligonucleotide arrays (Lockhart et al., 1996) and cDNA microarrays (Schena et al., 1995, 1996) and new second-generation sequencing techniques, like RNA-Seq (Wang et al., 2009). They provide rapid, parallel surveys of gene-expression patterns for hundreds or thousands of genes in a single assay (Quackenbush). The most common and easiest experimental design looks at the transcription profiles of mRNA under two conditions and searches for genes that are significantly up- or down-regulated between these different groups of samples, e.g. tumour and normal tissue.

2.3.1 Microarrays

For chip-based techniques thousands of genes of one organism are represented by oligonucleotide sequences immobilised on a high-density array on a glass slides or wafer. The extracted mRNA is reverse transcribed into cDNA and amplified using PCR, fluorescent-labelled and hybridised to the slide. The slide is in the next step scanned with a microscope to detect the fluorescent signal for each spot on the array. The amount of label is

assumed to be proportional to the concentration of the mRNA in the sample and therefore used for quantification. Gene expression microarrays have been invented in the Pat Brown laboratory in 1995 (Schena et al., 1995) and by Affymetrix in 1996 (Lockhart et al., 1996). Different microarray chips and techniques exist, the most frequently used are spotted cDNA (two-colour) arrays and oligonucleotide (one-colour) arrays.

One-colour and Two-colour Microarrays

The differences between spotted cDNA arrays and oligonucleotide arrays are illustrated in Figure 2.6. cDNA arrays are arrays spotted with genomic cDNA. To these spots the reverse transcribed and labelled mRNA is hybridised. To measure the intensities of two samples on one array, the samples from different conditions are labelled with different fluorescent dyes, Cy3 and Cy5. Cy3 dyes are green ($\sim$550 nm excitation, $\sim$570 nm emission and therefore appear green), while Cy5 is fluorescent in the red region ($\sim$650/670 nm). Afterwards the labelled mRNA is pooled and hybridisation to the chip. Each spot is scanned and the intensity of the two dyes are measured in different channels.
Oligonucleotide arrays have a more complex chip design. For each gene a series of 11-16 short ($\sim$25bp) oligonucleotides are synthesised directly onto the array. The oligonucleotides, also called probes, are combined into probe pair sets (2 x 11-16 probes) consisting of perfect-match (PM) and mismatch (MM) probes. These probe sets represent one transcript. The reverse transcribed cDNA is labelled with the same dye and hybridised to separate arrays for the measurement of the intensities. The chip layout into probes and probes set with PM and MM probes is used for the normalisation and correction of artefacts that arise from the hybridisation, mRNA decay and dye bleaching. The most frequently used one-colour microarrays are Affymetrix GeneChip Arrays.

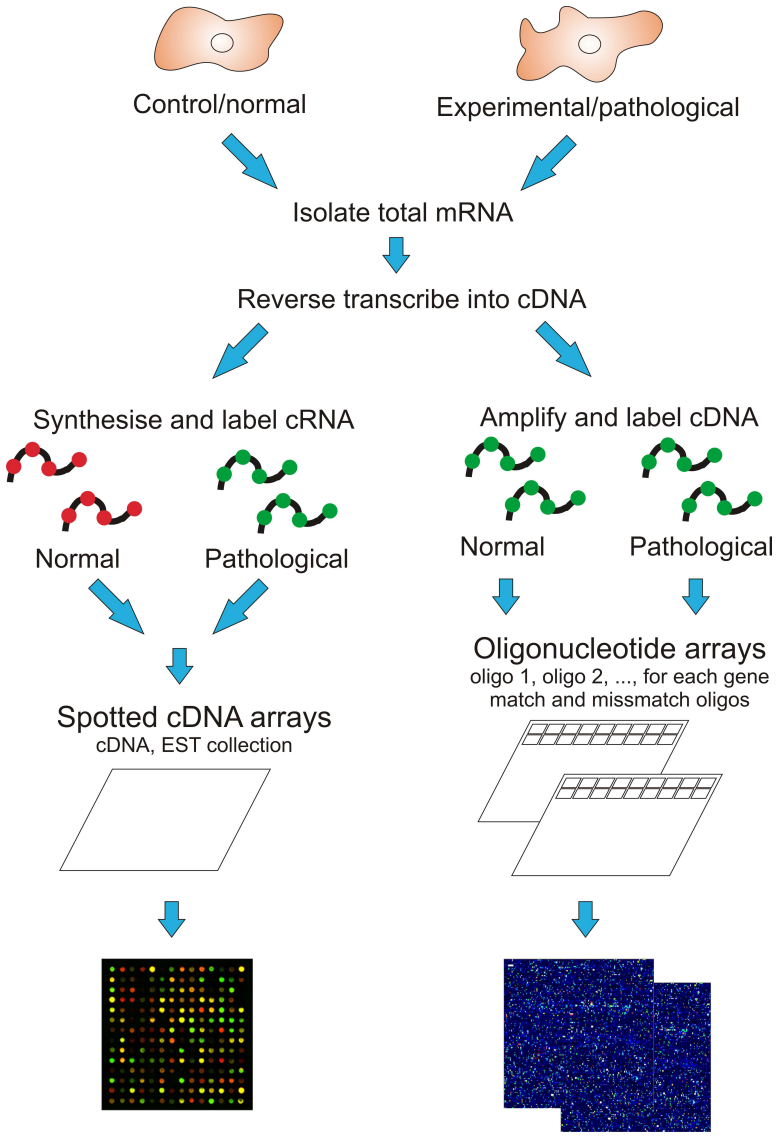


Figure 2.6: cDNA arrays (left panel) are arrays spotted with genomic cDNA. Onto this the reverse transcribed and labelled mRNA is hybridised. The different conditions that are to be analysed are labelled with different fluorescent markers and pooled for the hybridisation. Each spot is scanned and the intensity of the two dyes are measured in different channels. Oligonucleotide arrays (right panel) have a more complex chip design, consisting of perfect-match and mismatch probes that are combined into probesets. The oligos are synthesised directly on the slide. The reverse transcribed cDNA is labelled with the same dye and hybridised to separate arrays for the measurement of the intensities. The figure was created after Mount (2004); Brand and Barton (2002).

2.3.2 Preprocessing

Before testing for differential gene expression or performing any other analysis on the microarrays, the microarrays have to be preprocessed and normalised for a proper comparison. The quality of the gene expression data is influenced by the experimental design, the quality of the samples, the probes and the array platform, printing, controls, RNA extraction, amplification, labelling and hybridisation etcetera. The aim of preprocessing is to obtain a comparable expression matrix. The overall process of microarray analysis is depicted in Figure 2.7. It consists of 3 steps: image analysis, preprocessing of the raw data, including quality assessment, background correction and normalisation and finally analysis of gene expression. The first quantitative values are obtained from the scanner, providing files of pixel intensities. Image analysis is used to obtain a single overall intensity per probe or spot, background estimates and quality measures. Normally, especially for cDNA arrays, this step is included in the scanner software and the raw files already comprise probe- or spot-wise intensity values.

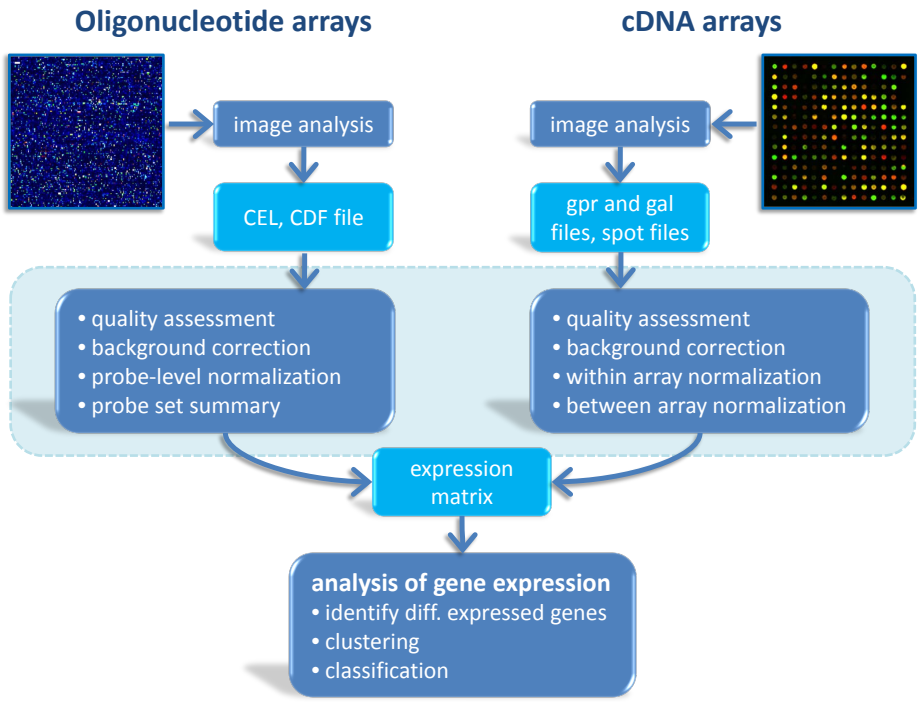


Figure 2.7: Microarray preprocessing for oligonucleotide arrays (left panel) and cDNA arrays (right panel).

Quality Control

Before and after the correction and normalisation steps, the raw data and normalised data are inspected, respectively. Several ways of visualising these datasets have been developed to reveal artefacts and systematic and technical biases. The grossest artefacts, like smear and scratches on the array or strong dye effects, can be identified by looking at the raw images of the gene expression data. These strong effect often can not be corrected for to make the arrays comparable and the affected arrays should be removed before further normalisation and analysis. Additional diagnostic plots include among others density plots and boxplots of the overall intensities of each array and for two-colour arrays of the two channels per array. The MA plot visualises differential expression (M: Minus) on the y-axis against the absolute intensity on the x-axis (A: Average). M and A are defined by

$$M = logR - logG \tag{2.1}$$

$$A = \frac{1}{2}(logR + logG) \tag{2.2}$$

for two-colour arrays with red and green intensities (Dudoit et al., 2002). Similarly, for one-colour arrays the pairwise means and differences of log-expression values from at least two microarrays are used for the MA plot (Bolstad et al., 2003). This rotated and rescaled scatterplot shows subtle imbalances in the distribution of the observed gene expression ratios from a value of zero. Often a locally weighted regression line fitted to the data points displays this effect more clearly. More sophisticated plots exist for oligonucleotide arrays, e.g. the RNA digestion plot gives a qualitative measure of the amount of RNA degradation that occurred during the RNA preparation.

Background Correction

Images of spotted arrays contain information of foreground and background intensities. The background intensity is considered as noise and can be removed from the overall intensity of the spot to obtain the signal. Several local, global and moving background correction methods exist, the usual background subtraction method simply subtracts the local background intensities from the foreground intensities. Because of the high-density packaging of probes on oligonucleotide arrays, here the background information is difficult to obtain and not commonly used (Parmigiani, 2003), an exception is the RMA convolution for Affymetrix chips in RMA normalisation developed by Irizarry et al. (2003). In this procedure, the PM values are

corrected for each array using a global model for the distribution of probe intensities. The PM probes are modelled by a mixture of an exponential signal component $S \sim Exp(\alpha)$ and a normal noise component $B \sim N(\nu, \sigma^2)$, truncated at zero, which allows the adjustment of the observed intensity.

Normalisation

In contrast to cDNA arrays preprocessing of Affymetrix arrays is usually performed on the probe-level. Therefore the background correction and normalisation is followed by a summarisation step that yields an estimate of expression on the gene level. The most common normalisation methods are listed in the following.

Robust Multichip Average Robust Multichip Average (RMA) (Irizarry et al., 2003) is a normalisation method consisting of three steps: a convolution background correction (described above), quantile normalisation and finally a median-polish summarisation.

GCRMA GCRMA is similar to RMA, but uses a different background adjustment model taking sequence information into account to account for non-specific binding. Because of this the background intensity is often underestimated in RMA.

Quantile Quantile normalisation makes arrays comparable by imposing the same empirical distribution of intensities to each array (Bolstad et al., 2003). When plotting the quantiles of two identical distributions against each other, the plot shows a diagonal line from (0,0) to (1,1). Derived from this, the quantile normalisation transforms the data by projecting each point from two data vectors plotted in a quantile-quantile plot onto the 45-degree diagonal line. This gives the same distribution to both data vectors.

Variance Stabilising Variance Stabilising Normalisation (vsn) combines background correction and normalisation into one single step and shares information across arrays to estimate the background correction parameters. Vsn applies a normalisation transformation using a generalised logarithm to adjust the data from different arrays and make the variances across replicates approximately independent of the mean. This behaviour can often be observed in the MA plot before normalisation, where the variance of the M values changes in dependence on the A values.

Loess The global loess method (Yang et al., 2001, 2002) uses MA-plots to normalise Cy5 and Cy3 channel intensities on cDNA microarrays. Loess, or lowess (locally weighted scatterplot smoothing) is a locally weighted polynomial regression, where at each point in the dataset a low-degree polynomial is fitted to a subset of the data. Subsequently the M-values are adjusted by subtracting the loess fit. For single-channel arrays in contrast, pairs of arrays are normalised to each other. The cyclic loess method normalises intensities for a set of arrays in a pairwise manner. The procedure cycles through all pairwise combinations of arrays, repeating the normalisation process until convergence.

Summarisation

The summarisation methods for oligonucleotide arrays include averaging on the log2 scale, log of the average on the natural scale, median on the log scale, log of the median on the natural scale, tukey biweight: a robust average on the log2 scale, lm: a linear model fit on the log2 scale and two robust linear models: rlm and median polish.

2.3.3 Testing for Differential Gene Expression

The main aim of analysing gene expression data is the search for genes whose patterns of expression differ according to phenotype or experimental condition. In the easiest application it reduces to the most basic statistical question: a two-sample comparison. Wrongly, this implies an independence between the analysed genes. But measurements on different genes are in general not independent, due to their interactions and reciprocal regulations. However, the high dimension of gene expression space and the high number of unknown relations between genes prohibits a more accurate exploration and will therefore reduce to a gene-by-gene approach. The simplest strategy is to select genes based on average changes in expression across groups, but this ignores the variability in the measurement of the gene and does not allow the assessment of significance of expression changes. Hence, a statistical test is needed. Most analyses use a one-gene approach for testing the null hypothesis of no differential expression, such as the t-test or its nonparametric counterparts and adjust for multiple testing by correcting the p-values.

Two statistical approaches for the analysis of gene expression data are presented here; for a simple experimental design of a two group comparison. More sophisticated experimental designs and ways of their analysis are described elsewhere (Parmigiani, 2003; Smyth et al., 2003; Gentleman et al., 2005).

T-test The t-test is a statistical hypothesis test in which the test statistic follows a Student's t distribution under the null hypothesis. For microarray analysis one is usually interested in the differences between two conditions or treatments, to which the two sample t-test can be applied. It tests whether under the null hypothesis the means of two normally distributed populations are equal.

The t statistic to test for equal means of the populations is as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{2.3}$$

where σ^2 is the variance of the groups and n the group size.

A p-value can be calculated from the corresponding t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

The assumptions for the t-test are, that the two populations being compared should follow a normal distribution, should have the same variance and the data should be sampled independently. A non-parametric alternative to the paired Student's t-test is the Wilcoxon signed-rank test. It is applied when the population cannot be assumed to be normally distributed or the data are on the ordinal scale.

The t-test was implemented in R for the study at hand or an implementations of the t-test and Wilcoxon test from the R-package stats or genefilter were used (Gentleman et al., 2008a; R Development Core Team, 2008).

One problem of the t-test is, that it needs enough data to estimate the variances accurately. Many researchers have therefore proposed alternative statistics that use information from all genes to improve the estimate of gene-specific variances. These statistics are referred to as modified or moderated t-statistics. One example of such a modified t-statistic, which is based on an empirical Bayes approach is implemented in the limma package by Smyth (2004).

Limma The idea behind limma is to fit a general linear model to the expression data for each gene. An empirical Bayes approach is used to borrow information across genes and shrink the estimated sample variances towards a pooled estimate, making the analyses stable even for experiments with a small number of arrays (Smyth, 2004). Therefore, the moderate t-statistic uses the posterior residual standard deviations instead of ordinary standard deviations.

It is assumed that a set of n microarrays yields a response vector $\mathbf{y}_g^T = (y_{g1}, ..., y_{gn})$ for the gth gene. For oligonucleotide arrays, the probes are assumed to have been normalised and summarised, so that y_{gi} is the expression value for each gene g on each array i. $E(y_g) = X\alpha_g$ can be expressed as a linear model with X the design matrix of the microarrays experiment and a coefficient vector α_g. Certain contrasts β_g of the coefficients are of biological interest and these are defined by $\beta_g = C^T\alpha_g$. The statistical test examines whether individual contrast values β_g are equal to zero.

As an example, in a hypothetical experiment with 6 oligonucleotide arrays one is interested in differential gene expression between normal (3 samples) and tumour (3 samples) tissues. A possible linear model for this question is:

$$Expression\ gene\ i\ = \beta_1 * \ tumour + \beta_2 * \ normal.$$

In this case, the contrast of interest is $\beta_2 - \beta_1$. To test whether gene i is differentially expressed the null hypothesis $H_0 : \beta_2 - \beta_1 = 0$.

For this simple example the design matrix and the contrast matrix would be chosen as follows:

Table 2.1: Design matrix

samples	Tumour	Normal
Array 1	1	0
Array 2	1	0
Array 3	1	0
Array 4	0	1
Array 5	0	1
Array 6	0	1

Table 2.2: Contrast matrix

Tumour	Normal
-1	1

The moderate t-statistic in limma is given by:

$$\tilde{t_{gj}} = \frac{\bar{\beta}_{gj}}{\tilde{s}_g\sqrt{v_{gj}}}, \tag{2.4}$$

with the gene-specific posterior variance $\tilde{s}_g$, the estimators of the contrast j $\bar{\beta}_{gj}$ and the covariance v_{gj}.

The moderated t follows a t-distribution under the null hypothesis H_0: $\beta_{gj} = 0$ with the original degrees of freedom d_0 and added degrees of freedom d_g, estimated from the data.

The linear model and test statistic from the R-package limma was used (Smyth, 2004) for this project.

2.3.4 Survival Analysis

Survival analysis deals with the modelling of time to event data. In the medical application of survival analysis, the event is most often the death of the patient (in other applications also disease recurrence, recovery etcetera) and analysed is the survival time. To describe the realisation of a survival analysis, several terms have to be introduced: (i) the survival function, (ii) the hazard function and (iii) the Cox regression model.

(i) The survival function $S(t)$ gives the probability that a subject survives longer than time t.

$$S(t) = \Pr(T > t) \tag{2.5}$$

The survival function can be expressed as a Kaplan-Maier curve and tested for equivalence between two groups with a log-rank test if the predictor variable is categorical. For continuous variables, such as gene expression data, Cox proportional hazards regression analysis (Cox PH, Cox model) is used (Cox, 1972).

(ii) The hazard function $h(t)$ is a function of the probability of an event in the time interval $[t, t+i]$, given that the individual has survived up to time t. Or in other words: the risk per time interval to die at time t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} \tag{2.6}$$

(iii) The Cox proportional hazard model:

$$h(t; X) = h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_m X_m) = h_0(t) \exp(\beta' X), \tag{2.7}$$

gives the hazard at time t for an individual with explanatory variables $X_1, ..., X_m$ and the coefficients to estimate $\beta_1, ..., \beta_m$ (Cox, 1972).

Analogously to other regression models, Cox regression is used to analyse the effect of several influencing variables on a dependent variable, where the dependent variable is the censored survival time. Censored observations are observations of patients without an event during the observed time period and with unknown progress. A requirement of Cox PH is that the effect of the influencing variable is constant over time (therefore: proportional

hazard). The coefficients of the Cox PH model are determined by maximum likelihood.

For gene expression data the effect of each gene on the survival of the patients, determined by the coefficient β, is estimated. A likelihood-ratio statistics can be used to test the null hypothesis that all of the βs are zero and to calculate corresponding p-values.

In this study the implementation of the survival analysis of Andersen and Gill (1982) in the R package *survival* (Therneau et al., 1990) was used.

Chapter 3

Material and Methods

3.1 GO Term Enrichment

Gene Ontology (GO) (Ashburner et al., 2000) term enrichment was used to test for a significant enrichment of GO terms in a set of tested terms against a background distribution. In general, ontologies provide controlled, representational vocabularies to describe concepts and relations (Gruber, 1993). In biological ontologies the tested terms are specific for a set of genes that are annotated by them. The GO terms correspond to an unique entity of one of the three branches from the GO ontology: cellular component (CC), biological process (BP) or molecular function (MF). Each branch of the ontology is represented by an directed acyclic graph, with the terms CC, BP, or MF as their roots. The further from the root, the more specific the GO terms get, as exemplified in Figure 3.1.

A statistical test checks for an over- or underrepresentation of GO terms from the tested set against an universe set. Several similar tests and implementations exist, most of them using a hypergeometric test to answer the question:

When sampling n genes (test set) out of N genes (universe set or reference set), what is the probability that k or more of these genes belong to a functional category shared by m of the N genes in the reference set?

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{3.1}$$

All analyses of this thesis were conducted using the R package GOstats (Falcon and Gentleman, 2007). It implements the hypergeometric test and

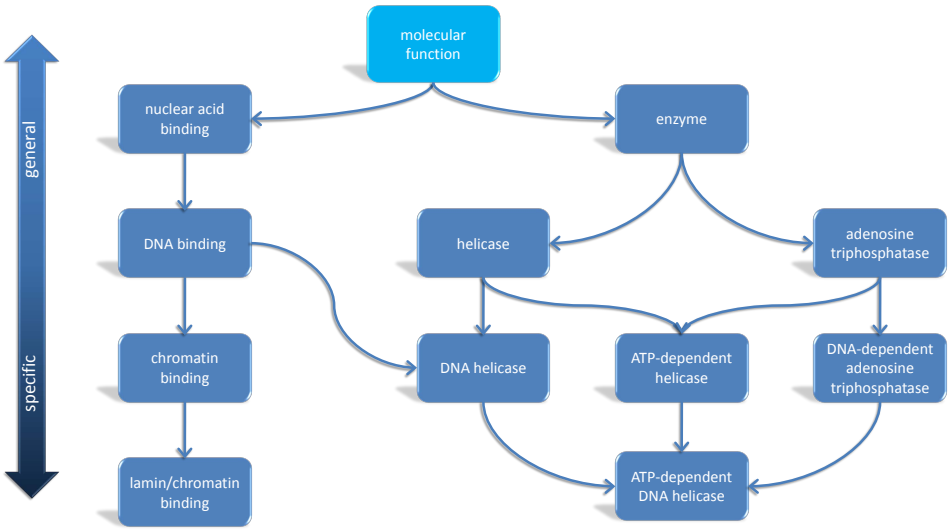


Figure 3.1: Examples of Gene Ontology modified after Ashburner et al. (2000). The example illustrates the structure and style used by GO to represent the gene ontologies and to associate genes with nodes within an ontology. The ontologies are built from a structured, controlled vocabulary. The molecular function ontology reflects conceptual categories of gene-product function.

a conditional hypergeometric test. The conditional hypergeometric test uses the relationships among the GO terms to address concerns that arise due to the hierarchical structure of GO and the dependence of categories.

For functional characterisation of genes or enzymes contained in the functional modules a GO term enrichment against the complete network was performed. This identified the GO categories that were significantly overrepresented in this set of genes.

3.2 Correspondence Analysis

Correspondence analysis (CA) is a dimensionality reduction technique developed by Jean-Paul Benzécri (Benzécri, 1973), which is used to obtain a low-dimensional projection of high-dimensional data.

For microarrays which contain thousands of genes or metabolite data with hundreds of metabolites, it is difficult to visualise the underlying structure of the data. It is therefore desirable to reduce the dimensions of the data and look at the dimensions which contain most information. This is ac-

complished by applying correspondence analysis on the expression data and transforming the data by basis transformation using singular value decomposition (SVD; core of all projection methods), so that the first two components contain most information. The microarray data can then be visualised in a biplot according to the first two components. Each axis reveals a specific characteristic of the dataset and observations/variables having high similarity with respect to this characteristic have similar coordinates in the plot.

A nice overview of correspondence analysis for microarray data is given by Busygin and Pardalos (2007), which is summarised in the following.

A dataset is normally given as a rectangular matrix $A = (a_{ij})_{m\times n}$ of n observations (columns) and m variables (rows). In the case of microarray data, rows represent genes and the value a_{ij} shows the expression level of gene i in sample j. From the data matrix A a correspondence matrix $P = (p_{ij})_{m\times n}$ is constructed:

$$p_{ij} = \frac{a_{ij}}{a_{++}} \tag{3.2}$$

where $a_{++} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}$ is the grand total of A.

The mass of the ith row is $r_i = a_{i+}/a_{++}$ and the mass of the jth column is defined as $c_j = a_{+j}/a_{++}$, where $a_{i+} = \sum_{j=1}^{n} a_{ij}$ and $a_{j+} = \sum_{i=1}^{m} a_{ij}$.

Then the matrix $S = (s_{ij})_{m\times n}$, to which SVD is applied, is formed:

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \tag{3.3}$$

The SVD of $S = U\Lambda V^T$ decomposes S into the product of three specific matrices. Columns of the matrix $U = (u_{ij})_{m\times n}$ are orthonormal vectors spanning the columns of S, columns of the matrix $V = (v_{ij})_{n\times n}$ are orthonormal vectors but spanning the rows of S and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ is a diagonal matrix of non-negative singular values of S having an increasing order.

Due to specific properties of the matrix S, the columns and rows of the original data matrix A can be represented in one low-dimensional space of dimensionality $K < n$. The coordinates of gene i in the new space are given by

$$f_{ik} = \frac{\lambda_k u_{ik}}{\sqrt{r_i}}, \quad k = 1, 2, ..., K \tag{3.4}$$

and of sample j by

$$g_{jk} = \frac{\lambda_k v_{jk}}{\sqrt{c_j}}, \quad k = 1, 2, ..., K \tag{3.5}$$

Only the first two or three coordinates are plotted, e.g. for a biplot $K = 2$ is selected.

Correspondence analysis was performed using the R package vegan (Dixon and Palmer, 2003).

3.3 Resampling Procedures

The statistical method of jackknifing was first introduced by Quenouille (1956) and Tukey (1958) by deleting one observation to estimate the bias and variance of a statistic of interest. For the more general delete-j observations jackknife the bootstrap can be seen as an approximation of it, which was described by Efron (1979). The difference between these resampling approaches is, that the bootstrap is a random resampling procedure with replacement, jackknife draws random subsets of the data without replacement by deleting j observations.

Often one is interested in the standard error or the confidence interval of statistical estimator $\hat{t}$ for a parameter of interest which is given as function T of the data points $x_1, x_2, \ldots, x_n$

$$\hat{t} = T(x_1, x_2, \ldots, x_n) \ . \tag{3.6}$$

Drawing J times randomly a subset of $n - j$ values from the observed data $x_1, x_2, \ldots, x_n$, J jackknife pseudo-replicates of $n-j$ data points are obtained. For each sample the estimates

$$\hat{t}^{(i)} = T(x_1^{(i)}, x_2^{(i)}, \ldots, x_{n-j}^{(i)}), \quad i = 1, \ldots, J \ , \tag{3.7}$$

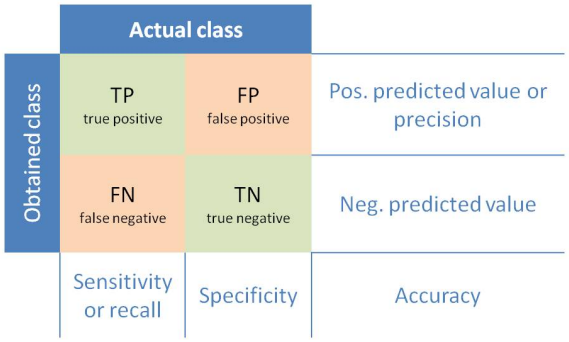
are calculated. Based on this jackknifed distribution of the estimator the standard error and confidence intervals can be estimated. To obtain a similar variance in the jackknife resamples as for bootstrapping, a 50% jackknife can be used and half of the observations dropped as recommended by Felsenstein (1985, 2004).

3.4 Accuracy and Variance Measures

For the evaluation of resulting functional modules, the terms 'true positives' (TP), 'true negatives' (TN), 'false positives' (FP) and 'false negatives' (FN)

were used to compare the obtained class of a node with the actual class, where the classes were either "part of the module" or "not part of the module" (Table 3.1). By this the modules were evaluated against simulated reference modules (Section 4.2.1) in a Receiver Operating Characteristic (ROC) analysis in terms of recall, precision and the Jaccard coefficient.

Table 3.1: Accuracy measures



Recall, precision and the Jaccard coefficient are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

$$J = \frac{TP}{TP + FP + FN} = \frac{A_v \cap B_v}{A_v \cup B_v}, \tag{3.10}$$

where A_v, B_v are the node sets of the modules under consideration.

The Area Under the Curve (AUC) values of the recall-precision curves served as an overall performance measure of the considered algorithms. Similarity between modules were assigned by the Jaccard coefficient between the set of nodes in the modules (Rivera et al., 2010). In contrast to other measures, the Jaccard coefficient does not vary with varying true negative counts, which is always large for small simulated modules and bias the measure. Furthermore, the variance of the solutions was assessed. Similar to the mean squared error which assesses the quality of an estimator in terms of its variation and unbiasedness, not only the accuracy of an resulting module, but also the variation in the obtained solutions should be measured. The variation was determined by running an algorithm on a network with perturbed data by a resampling procedure (Section 3.3) and comparing the different obtained solutions. The number of times a gene appeared in the varying modules was specified and the variance in AUC and Jaccard coefficient of the solutions of different algorithms was compared.

3.5 Trend Tests

A variety of distribution-free test procedures are available to test the hypothesis that k samples originate from a common distribution. Depending on the alternative hypothesis, different statistical tests can be applied. The simplest alternative is that at least two of the distributions have a different mean, which can be examined using the Kruskall-Wallis test. For ordered alternatives, thus trends in the data (Figure 3.2), the Jonckheere-Terpsta and the Umbrella test are presented in the following.

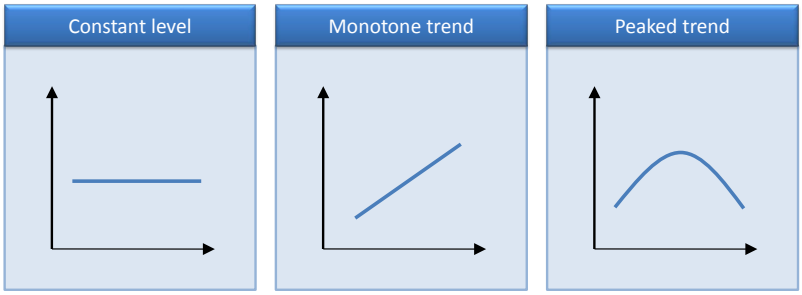


Figure 3.2: Possible trends (constant, monotonic trend or peaked trend) in the data that can be tested for with the Jonckheere-Terpstra test and the Umbrella test.

Kruskal-Wallis Test The Kruskal-Wallis test or H-test (Kruskal and Wallis, 1952) is a parameter-free, rank-based statistical test, used to test for differences in means of at least two of k groups. The null-hypothesis H_0 states that all samples are drawn from the same population G_i with identical distribution: $H_0 : G_1 = G_2 = ... = G_k$. The corresponding test statistic is given by:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}, \tag{3.11}$$

where n_i is the number of observations in group i, r_{ij} is the rank (among all observations) of observation j from group i, N is the total number of observations across all groups, $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$, $\bar{r} = \frac{1}{2}(N+1)$ is the average of all the r_{ij}.

Under H_0 the test statistic follows a chi-square distribution with $df = k-1$ degrees of freedom. The null hypothesis of equal population medians would then be rejected if $H \geq \chi^2_{\alpha:k-1}$.

Jonckheere-Terpstra Test The Jonckheere-Terpstra (JT) test is a parameter-free, rank-based test to analyse monotonic trends between groups by Jonckheere (Jonckheere, 1954) and Terpstra (Terpstra, 1952). The null-hypothesis H_0 states that all samples are drawn from the same population G_i with identical distribution: $H_0 : G_1 = G_2 = ... = G_k$. The alternative hypothesis states: $H_1 : G_1 <= G_2 <= ... <= G_k$ with at least one strict inequality. The corresponding test statistic is given by:

$$J = \sum_{i<j}^{k} U_{ij} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij} \tag{3.12}$$

with $U_{ij} = \sum_{s=1}^{n_i} \sum_{t=1}^{n_j} \Psi(X_{jt} - X_{is})$ and

$$\Psi(u) = \begin{cases} 1 & \text{wenn } u > 0 \\ 1/2 & \text{wenn } u = 0 \\ 0 & \text{wenn } u < 0 \end{cases}$$

Under H_0 the test statistic J follows a normal distribution with the expected value μ_J and variance σ_J given by:

$$\mu_J = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4} \tag{3.13}$$

and

$$\sigma_J = \sqrt{\frac{N^2(2N+3) - \sum_{i=1}^{k} n_i^2(2n_i+3)}{72}} \tag{3.14}$$

The resulting variable Z is standard normal distributed:

$$Z = \frac{J - \mu_J}{\sigma_J} \tag{3.15}$$

From the standard normal distribution it can be calculated when to reject H_0 for a given alpha and the corresponding p-value can be obtained.

An implementation of the JT test from the R-package SAGx was used in the study (Broberg, 2009).

Umbrella Test The umbrella test is a generalisation of the JT test by Mack and Wolfe (Mack and Wolfe, 1981), which is used to test for trends with a peak, also known as umbrella shaped data. The null-hypothesis H_0

states that all samples are drawn from the same population G_i with identical distribution: $H_0 : G_1 = G_2 = ... = G_k$. The alternative hypothesis states: $H_1 : G_1 <= G_2 <= ... <= G_l >= G_{l+1} >= ... >= G_k$ where l is the peak position, with at least one strict inequality. The corresponding test statistic is given by:

$$MW = \sum_{r=1}^{l-1} \sum_{s=r+1}^{l} U_{rs} + \sum_{r=l}^{k-1} \sum_{s=r+1}^{k} U_{sr} \tag{3.16}$$

Where U_{rs} and U_{sr} for the r^{th} and s^{th} group with $1 \leq r < s \leq k$ are defined as:

$$U_{rs} = \sum_{h=1}^{n_r} \sum_{i=1}^{n_s} \Psi(X_{rh} - X_{si}) \tag{3.17}$$

and

$$U_{sr} = n_r n_s - U_{rs} \tag{3.18}$$

with

$$\Psi(u) = \begin{cases} 1 & \text{wenn } u > 0 \\ 1/2 & \text{wenn } u = 0 \\ 0 & \text{wenn } u < 0 \end{cases}$$

Under H_0 the test statistic MW follows a normal distribution with the expected value μ_{MW} and variance σ_{MW} given by:

$$\mu_{MW} = \frac{m_1^2 + m_2^2 - \sum_{i=1}^{k} n_i^2 - n_l^2}{4} \tag{3.19}$$

and

$$\sigma_{MW} = \sqrt{\frac{2(m_1^3 + m_2^3) + 3(m_1^2 + m_2^2) - \sum_{i=1}^{k} n_i^2(2n_i + 3) - n_l^2(2n_l + 3) + 12n_l m_1 m_2 - 12n_l^2 N}{72}} \tag{3.20}$$

with $N = \sum_{i=1}^{k} n_i, \quad m_1 = \sum_{i=1}^{l} n_i \quad \text{and } m_2 = \sum_{i=l}^{k} n_i.$

The resulting variable Z is standard normal distributed:

$$Z = \frac{MW - \mu_{MW}}{\sigma_{MW}} \tag{3.21}$$

From this it can be calculated when to reject H_0 for a given alpha and the corresponding p-value can be obtained.

The Umbrella test was implemented in R for the analysis and will be made available in the R package `BioNet`.

3.6 Gene Expression Profiles

Large B-cell Lymphoma Expression data were taken from a study on diffuse large B-cell lymphoma (DLBCL) from Rosenwald et al. (2002). The DLBCL data comprised 194 samples on custom microarrays from 112 tumours with the germinal center B-like phenotype (GCB DLBCL) and from 82 tumours with the activated B-like phenotype (ABC DLBCL). These tumour subtypes differ in their malignancy as well as in the treatment options for the patients. Expression profiling was performed on the *Lymphochip* including 12,196 cDNA probe sets corresponding to 3,583 genes. In addition, survival information from 190 patients were available.

Acute Lymphocytic Leukaemia The acute lymphocytic (or lymphocytic) leukaemia (ALL) microarray dataset from Chiaretti et al. (2004) was used in this study. The ALL dataset contained 128 samples on Affymetrix hgu95av2 gene chips with 12,625 probesets corresponding to 8,799 different genes.

3.7 Interaction Networks

Protein-protein interactions were taken from the Human Protein Reference Database (HPRD) (Mishra et al., 2006), release of 2006. The interactome was included as a graph object in the R package DLBCL, which accompanied as an experiment data package the software package `BioNet` (Section 4.1.3). This dataset constituted a network of 9,386 proteins and 36,504 interactions. Mapping the DLBCL gene expression data to the PPI network resulted in a largest connected component of 2,034 genes and 7,756 interactions, which was used as an interaction network for the analysis of this dataset.
Mapping of the ALL microarray data resulted in a largest connected component of 5,301 genes and 21,369 interactions.

3.8 Tardigrade Cultures

The study was carried out with the eutardigrade *Milnesium tardigradum* Doyère 1840, (Eutardigrada, Apochela, Milnesidae) which was originally collected in Tübingen, Germany and is kept in laboratory culture by the collaborator group of R. Schill (University Stuttgart). The animals were cultured on petri-dishes (ø9,4cm) with a layer of agarose (3%) (peqGOLD Universal Agarose, peqLAB, Erlangen Germany) covered with a thin layer of Volvic-water (Danone Waters, Wiesbaden, Germany) at 20° C. The animals were fed bdelloid rotifers, *Philodina citrina*, which had been raised on the green algae of the species *Chlorogonium elongatum* (Ehrenberg, 1832). For this study *Milnesium tardigradum* was starved over two days before harvested to avoid contamination with food-organisms. After repeated washing with clean water, animals were transferred into microliter tubes (400 individuals per tube). By using a micropipette, the animals surrounding water was reduced by careful aspiration to approximately 1.5 +/- 0.5 μL. The open microliter tubes were then exposed for dehydration to 85% relative humidity (RH) in a small chamber containing a saturated solution of KCl (Roth, Karlsuhe, Germany). The first time point of the dehydration period was sampled and frozen in liquid nitrogen after 1 h. Further samples were taken after 720, 900, 1020, 1080, 1110, 1140, 1170, 1185 min and finally after 1200 min after which the cryptobiotic tun formation was completed in all individuals. To produce different states of rehydration 10 μL Volvic-water was added to each microliter tube. The first rehydration state was immediately sampled and frozen in liquid nitrogen. Further samples during rehydration were taken 5 min, 10 min, 15 min, 20 min, 60 min, 90 min, 150 min, 210 min and 270 min.

3.9 EST Sanger Sequences

Two EST libraries were used, (i) Expressed Sequence Tags (ESTs) annotated with EC numbers from a current study (sequences are available in the tardigrade workbench at `http://waterbear.bioapps.biozentrum.uni-wuerzburg.de` (Förster et al., 2009, 2011b)) and (ii) ESTs from the active and inactive state of *Milnesium tardigradum* (dEST: differential ESTs) available at NCBI (for details see (Mali et al., 2010), raw traces were submitted to the trace archive). The following preparatory work was performed by M. Grohme (TH Wildau). Base calling on Sanger sequencing traces was carried out using phred (version 0.071220.b) (Ewing et al., 1998; Ewing and Green, 1998). Adapters and vector sequences were masked using cross_match (version 1.090518) with parameters -minmatch 8 -minscore 20 (Green) and trimmed using SnoWhite (version 1.1.3) (Dlugosch). The

dEST sequences were clustered into 4,422 clusters using CD-HIT-EST with a similarity threshold of 90% including forward and reverse matches (Li and Godzik, 2006).

3.10 Mass Spectrometry Profiles

Metabolome analysis was performed using GC-MS based metabolite profiling (Fiehn et al., 2000) of a whole organism extract with modifications for automated GC-time of flight (TOF)-MS application (Erban et al., 2007) by J. Kopka (MPI Golm). Pools of 400 *M. tardigradum* individuals were homogenised by sonification on ice with 150 μL 100% methanol which was pre-cooled to -20°C. Homogenised samples were extracted with internal standard (Erban et al., 2007) for 15 min at 70°C, cooled to room temperature, re-extracted with 100 μL CHCl3 for 5 min at 3770°C, mixed with 200 μL water and centrifuged 5 min at 14,000rpm to separate liquid phases and to remove debris. The upper methanol/water phase was collected and dried in a speed vac. Metabolite extracts were stored and shipped as was described previously (Erban et al., 2007). The dried metabolite extract was chemically derivatised and subjected to GC-TOF-MS analysis by sequential methoxyamination and trimethylsilylation and GC-TOF-MS profiling (Erban et al., 2007). The GC-TOF-MS chromatograms were processed by TagFinder-Software (Luedemann et al., 2008). Four technical repeats were analysed. Compounds were identified under manual supervision by matching to the reference library of mass spectra and retention indices of the Golm Metabolome Database, `http://gmd.mpimp-golm.mpg.de` (Hummel et al., 2010). Retention index thresholds for compound matching were according to Strehmel and co-authors (Strehmel et al., 2008). Averaged normalised mass detector responses of each metabolite observed in the four technical repeats were calculated using specific and selective mass features (Luedemann et al., 2008). Relative changes of metabolite pool sizes were assessed. normalisation was performed by the mass detector response of the internal standard and the number of individuals per sample. Metabolites that could not be separated on the basis of chromatographic elution patterns were eliminated from subsequent analysis. By this criteria, the initial list of 132 metabolites measured on the platform was narrowed to abundances for 92 metabolites.

Chapter 4

Results and Discussion

4.1 Integrated Network Analysis using Heinz and BioNet

4.1.1 Functional Modules in Molecular Networks

The following section is based on a book chapter that originated as part of my doctoral studies (Dittrich et al., 2010). It was published in "Systems Biology for Signalling Networks" with equal contributions from all authors. Parts of the text and figures are taken or adapted slightly without explicit additional markings from this chapter, with permission from Springer. It describes a novel approach to integrate molecular data with biological networks and calculate modules comprising important genes for the disease under study.

The increasing amount of molecular data from various high-throughput technologies calls for new methods to analyse it in an integrative manner. Gene expression profiling technologies provide a plenitude of information on gene expression in various tissues and under diverse experimental conditions. Combining this information with the knowledge of interactions of the gene products in a protein-protein interaction network generates a meaningful biological context in terms of functional and regulatory association of differentially expressed genes.

To assess the influence of specific genes on a disease phenotype, microarray technologies are commonly used, especially in tumour biology. Here, the identification of differentially expressed genes in diverse tissue samples or cancer stages is a well-established method to classify tumours and tumour subtypes. Ordinary gene expression analysis yields a number of genes that are up- or down-regulated given a predefined threshold, but it can neither re-

veal causal effects, nor functional associations between these genes. Another application of microarrays is the assessment of disease-relevant genes in survival analysis. Survival analysis allows to determine the survival-relevance of certain genes to use these as predictors for the progression of the disease and for classification. The combined analysis of expression profiles, protein-protein interaction data and further information of the influence of specific genes on a disease-specific pathophysiology thus allows the detection of previously unknown dysregulated modules not recognisable by separate analysis of each of the datasets.

Several approaches exist to identify modules in an integrative network analysis, using gene-specific patient data in combination with a biological network. These are based on p-values derived from the analysis of gene expression data. Ideker et al. (2002) firstly devised a function for the scoring of networks by transferring p-values into scores for the nodes of a network and searching for high-scoring subnetworks. The primary combinatorial problem for the identification of maximum-scoring subnetworks has been proven to be *NP*-hard. Therefore, the authors introduced a heuristic approach based on simulated annealing. Heuristic approaches in general can not guarantee to identify the global maximum-scoring subgraph and are often computationally demanding. Furthermore, they tend to result in large high-scoring networks, which may be difficult to interpret afterwards in a meaningful biological sense.

This section describes an exact approach by Dittrich et al. (2008) that delivers provably optimal and suboptimal solutions to the maximal-scoring subnetwork problem by integer linear programming in acceptable running time. It includes a novel scoring function for the nodes of the network that is based on p-values and allows the integration of multivariate p-values using order statistics. The scoring function is based on a signal-noise decomposition of the p-value distribution using a beta-uniform mixture model (BUM). An adjustment parameter that can be statistically interpreted as false discovery rate (FDR) to control the resulting size of the subnetwork and therefore to extract smaller modules of interpretable size.

Scoring of the network by decomposing the p-value distribution and aggregating multiple p-values into one p-value of p-values, will be outlined in detail in the following sections. A step-wise analysis is performed based on gene expression and survival data. This explains in detail how an adequate node score can be derived from the p-values of different analyses and will give an overview of the search strategy conducted to find maximal-scoring subnetworks.

Data Integration

The following briefly introduces the exemplary datasets that will be analysed in the remainder of the section. This will exemplify how an integrated network analysis can be performed using PPI, microarray and clinical survival data as examples for molecular datasets. To illustrate the approach a network is analysed obtained by combining gene expression data from two different cancer subtypes with survival data and a comprehensive interactome network derived from the Human Protein Reference Database.

The Integrated Data The methodological approach presented here is based on the integration of two distinct molecular datasets. Microarray data from 2 groups of a disease (e.g. cancer subtypes) are used to test for differential expression between these groups. In addition survival data are abandon for these patients, allowing the determination of genes associated with the risk of relapse.
After normalisation of the gene expression data, the significance of differential expression between the two subtypes is assessed by using an appropriate statistical test (e.g. t-test, limma). This results in an uncorrected p-value for differential expression for each gene. To assess the influence of each gene on the survival time, survival analysis is subsequently performed by fitting a gene-wise univariate Cox proportional hazard model to the expression data. For each gene p-values are obtained from the likelihood ratio test of the regression coefficient, denoting the association with survival. The p-values from both analyses correspond to differential gene expression and to risk association respectively. In the next step these two p-values are concatenated into one p-value for each gene. From this aggregated p-value a score is derived.

The Interaction Network For the network data the literature-curated human protein-protein interactions from HPRD are used. The entire interactome network comprises 36,504 interactions between 9,392 proteins. A chip-specific interactome network as the vertex-induced subgraph is extracted by the subset of genes for which expression data exist on the microarray (Figure 4.1).

The problem of identifying functional modules can be decomposed into two separate subproblems. The first part of the problem is to define a scoring function, which captures the information of the experimental data and maps them onto the nodes of the network. After scoring each node of the network, the second problem is to use an adequate algorithm to search for the maximal scoring connected subgraph.

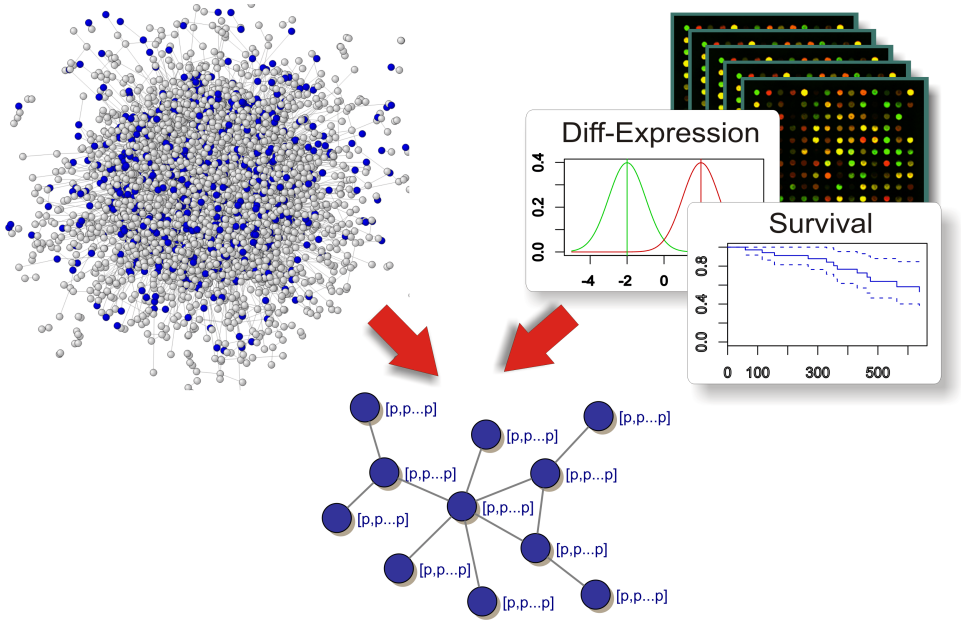


Figure 4.1: Integration of PPI, microarray and clinical data into a common framework. First the vertex-induced subgraph of all genes on the array is extracted (dark nodes). Differential expression can be calculated using a standard t-test or limma test and relapse risk association for each gene is estimated by Cox regression. All nodes in the network are annotated with the vector of p-values derived from these analyses. There is practically no limitation to the number of p-values (here two) that can be integrated by the presented approach.

Having annotated each node of the interaction network with experimentally derived p-values (from differentially expressed genes and the p-values of the Cox-regression model indicating risk association for each gene), the next step is to aggregate these p-values into an adequate score for each node and search the maximum-scoring subnetwork. A four step solution is used to solve this problem:

1. The vector of p-values is aggregated into one p-value using an order statistic
2. The resulting p-value distribution is decomposed into a signal and noise component and a beta-uniform mixture model is fitted to the data
3. A score is defined as a log ratio of signal to noise
4. The maximum-scoring subnetwork is extracted

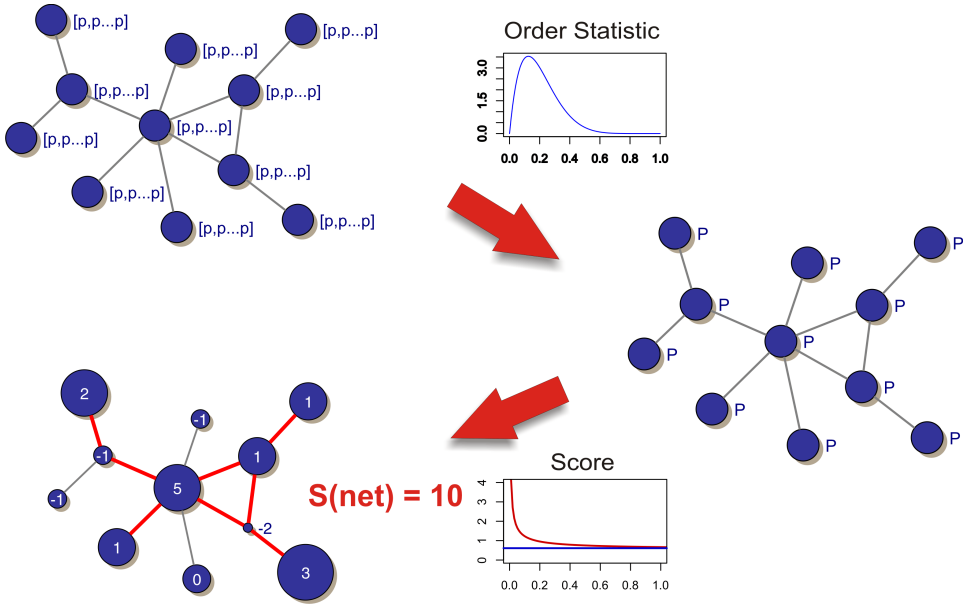


Figure 4.2: Definition of a node scoring function. First all p-values are aggregated using an order statistic. Subsequently, a signal-noise decomposition is performed based on a beta-uniform mixture model. The score is derived by a log ratio of signal and noise component. The final subnetwork score is calculated by the summing up all of the considered node scores. Finding highest-scoring networks is described in Section 4.1.1.

P-value aggregation

The ith order statistic of the associated p-values is used to obtain one p-value of p-values. By definition p-values are uniformly distributed under the null hypothesis (Wasserman, 2005). This means, by considering the p-value as a random variable p, by definition the following equation holds

$$P(p \leq x) = x \tag{4.1}$$

for all $x \in [0, 1]$. Hence, the distribution function of the random variable p is the identity function, and thus equal to the distribution function of the uniform distribution. In general, the probability density function of the ith smallest observation $x_{(i)}$ is given by

$$f(x_{(i)}) = \frac{n!}{(n-i)!(i-1)!} f(x) F(x)^{i-1} (1 - F(x))^{n-i} \tag{4.2}$$

for distribution function $F(x)$ and density function $f(x)$ of a random variable X and for $i \in 1, \ldots, n$ (Lindgren, 1993). Thus equation (4.2) can be applied

with $f(x) = 1$ and $F(x) = x$ to get

$$f(x_{(i)}) = \frac{n!}{(n-i)!(i-1)!} \cdot 1 \cdot x^{i-1}(1-x)^{n-i} \quad 0 \leq x \leq 1 \qquad (4.3)$$

or, in other words, the ith order statistic $x_{(i)}$ is distributed according to the beta distribution $B(i, n-i+1)$ with the associated cumulative distribution function

$$F(x_{(i)}) = \frac{n!}{(n-i)!(i-1)!} \int_0^{x_{(i)}} z^{i-1}(1-z)^{n-i}\, dz. \qquad (4.4)$$

For example, considering a vector x of $n = 4$ ordered p-values $(0.001, 0.05, 0.1, 0.5)$. The p-value of the second order statistic is derived by applying equation (4.4):

$$\begin{aligned} F(x_{(2)}) &= \frac{4!}{2!1!} \int_0^{x_{(2)}} z(1-z)^2\, dz \\ &= 12 \left(\frac{z^2}{2} - \frac{2z^2}{3} + \frac{z^4}{4} \right) \Bigg|_0^{x_{(2)}} \\ &= 6{x_{(2)}}^2 - 8{x_{(2)}}^3 + 3{x_{(2)}}^4 \end{aligned} \qquad (4.5)$$

For $x_{(2)} = 0.05$ a significant p-value of 0.014 is obtained. All order statistics, from the $1st$ to the 4^{th}, yield p-values of $(0.004, 0.014, 0.0037, 0.0625)$.

Signal-Noise Decomposition

Following Pounds and Morris (2003) the distribution of the p-values is considered as a mixture of a noise and a signal component. Visual inspection of the empirical p-value distribution as displayed in Figure 4.5 suggests that the signal component can be appropriately modelled by a $B(a, 1)$-distribution whereas the noise component is naturally modelled by a $B(1, 1) =$ uniform distribution. Thus the family of beta distributions comprises both the signal distribution as well as the noise distribution (Figure 4.3).

The $B(a, b)$ distribution is given by

$$B(a, b)(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}, \qquad (4.6)$$

where $\Gamma(\cdot)$ denotes the Gamma function with $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Thus the distribution f_{mix} of the mixture model with mixture parameter λ and shape parameter a reduces to

$$\begin{aligned} f_{\text{mix}}(a, \lambda)(x) &= \lambda B(1,1)(x) + (1-\lambda)B(a,1)(x) \\ &= \lambda + (1-\lambda)ax^{a-1} \qquad \text{for } x, a \in [0,1], \end{aligned} \qquad (4.7)$$

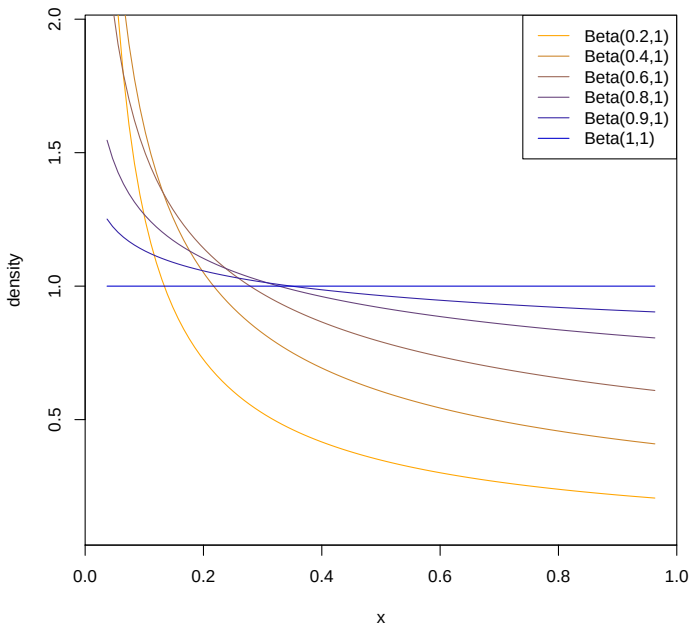


Figure 4.3: Shapes of beta distribution with varying parameter a and fixed parameter $b = 1$. For $a = 1$ the beta distribution is equal to the uniform distribution.

because $\frac{\Gamma(a+1)}{\Gamma(a)\Gamma(1)} = a$. For given data $x = x_1 \dots x_n$ the log likelihood is

$$\log \mathcal{L}(\lambda, a; x) \quad = \quad \sum_{i=1}^{n} \log(\lambda + (1-\lambda) a x_i^{a-1}) \ , \tag{4.8}$$

and consequently the maximum-likelihood estimations of the unknown parameters are given by $[\hat{\lambda}, \hat{a}] = \text{argmax}_{\lambda,a} \mathcal{L}(\lambda, a; x)$.

Both parameters are obtained by a numerical optimisation method (L-BFGS-B method (Byrd et al., 1995) as implemented in R). Applying this optimisation to the presented dataset delivers value of 0.536 for the mixture parameter λ and 0.276 for the shape parameter a of the beta distribution as depicted in Figure 4.4.

As stated above, p-values are by definition uniformly distributed under the null hypothesis while the true signal distribution is not known a priori. Therefore a uniform distribution will adequately model the noise component. Modelling the signal component by a beta distribution is an assumption that has to be justified. The fitted density function of the mixture model fits the data very well as demonstrated in the left plot of Figure 4.5 and

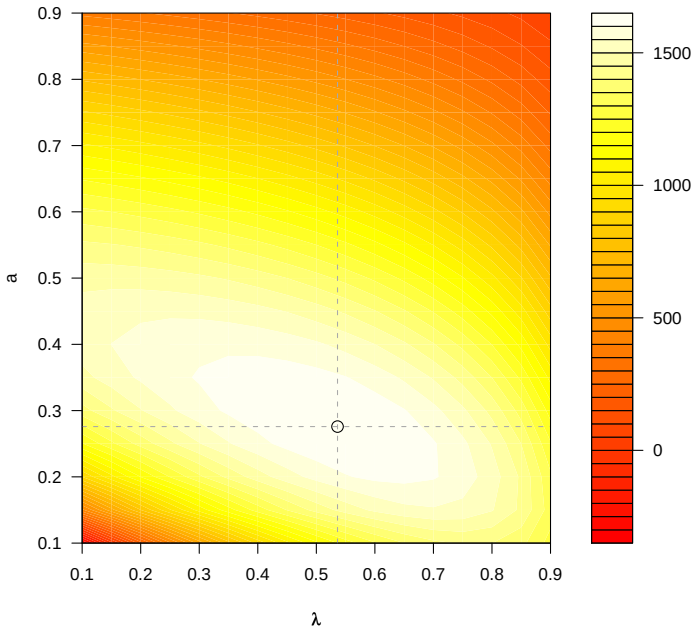


Figure 4.4: Log likelihood surface of the mixture parameter λ (x-axis) and the shape parameter a (y-axis) of the beta-uniform mixture model for the B-cell lymphoma dataset. The numerically determined optimal parameter pair is indicated by the cross lines. The maximum-likelihood estimates are $\widehat{\lambda} = 0.536$ and $\widehat{a} = 0.276$.

a quantile-quantile plot of the fitted density function versus the empirical distribution function (right panel). This shows that the signal component is well captured by the beta distribution.

Scoring the Nodes of the Network

A scoring function can be defined by the log ratio of the signal to the noise component. In the BUM-Model the signal component is equal to the $B(a, 1)$ while the noise component is given by $B(1, 1)$, which is equivalent to the uniform distribution. This simplifies the denominator in the score function to the constant 1:

$$\begin{aligned} S(x) &= \log\left(\frac{Signal}{Noise}\right) = \log\left(\frac{B(a,1)(x)}{B(1,1)(x)}\right) = \log\left(\frac{ax^{a-1}}{1}\right) \\ &= \log(a) + (a-1)\log(x) \ . \end{aligned} \tag{4.9}$$

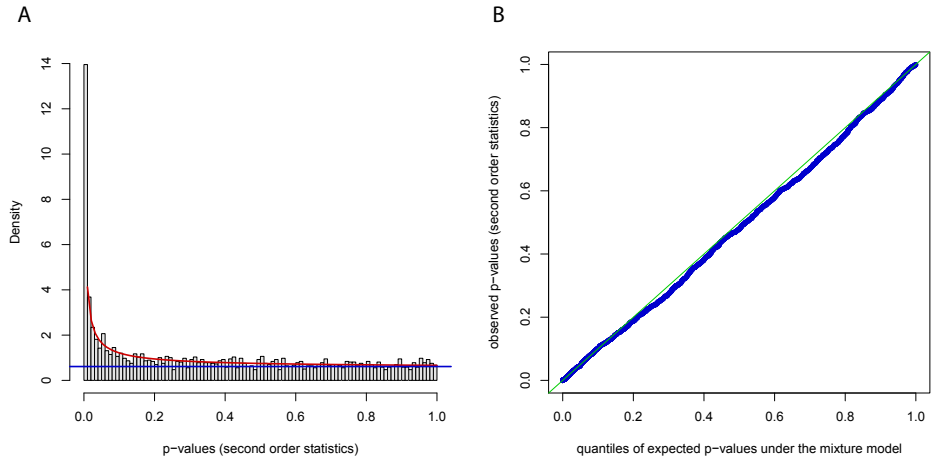


Figure 4.5: (A) The histogram depicts the fitted mixture model (curved line) to the empirical distribution of the p-values. The optimal parameters for the model are $a = 0.276$ and $\lambda = 0.563$. The horizontal line indicates the upper bound π for the fraction of noise. (B) Quantile-quantile plot of the fitted distribution and the empirical distribution. The straight line indicates that the shape of the fitted model coincides with the shape of the empirical distribution.

This function transforms a given set of p-values to a real-valued score where positive scores indicate significant p-values and negative scores denote non-significant p-values. As detailed in Pounds and Morris (2003) the BUM model allows the estimation of the false discovery rate. This can be used to fine-tune the zero value threshold: Incorporating a threshold p-value $\tau(\text{FDR})$ into the scoring function an adjusted log likelihood ratio score is given by

$$S^{\text{FDR}}(x) = \log\left(\frac{ax^{a-1}}{a\tau^{a-1}}\right) = (a-1)\left(\log(x) - \log(\tau(\text{FDR}))\right) \ . \qquad (4.10)$$

The adjusted and unadjusted scores differ only by an additive offset dependent on the parameter τ. With the adjusted score p-values above the threshold $\tau(\text{FDR})$ are considered to be noise and will be assigned a negative score, whereas p-values below the threshold are considered significant and will thus obtain a positive score.

Due to the logarithmic scale of the scoring function a network score can consequently be defined by the sum over all node scores in the subnetwork T:

$$S^{\mathrm{FDR}}(T) \;:=\; \sum_{x_i \in T} S^{\mathrm{FDR}}(x_i) \;. \tag{4.11}$$

In the context of the presented cancer study with p-values from the t-test and Cox regression model this score combines the information on differential expression with that on risk association. Genes that are differentially expressed between the subgroups and are simultaneously associated with overall survival, will obtain a positive score. Thus modules with a high score in the integrated PPI network are searched that contain genes that are differentially expressed and risk associated. These can be identified by calculating the maximal-scoring subnetwork.

Calculation of the Maximum-scoring Subnetwork

The task is to find an subnetwork

$$T^* = \operatorname*{argmax}_{T \in \mathcal{T}} S^{\mathrm{FDR}}(T) \;, \tag{4.12}$$

where $\mathcal{T}$ is the set of all connected subgraphs of the protein-protein interaction network, that maximises the score.

Since the network score is additive with respect to the network nodes, this combinatorial search problem is exactly the Maximum-Weight Connected Subgraph Problem (MWCS). The MWCS problem is transformed into instances of the prize-collecting Steiner tree problem (PCST). This problem tries to connect profits (nodes) by choosing connections with the lowest costs (edges). A mathematical programming-based algorithm for PCST by Ljubić et al. (2006) is used to find solutions of (4.12). It is shown in the computational results in Chapter 4.4.2 that this approach finds provably optimal and suboptimal subnetworks in short computation time for biologically relevant instance sizes. Astonishingly, it also outperforms the heuristic methods in terms of speed. The details of the method and the transformation can be found in Dittrich et al. (2008).

In contrast to Ideker et al. (2002) who approach this problem of finding a MWCS heuristically, techniques from mathematical programming are used to solve it to provable optimality. Starting from an integer linear programming (ILP) formulation modelling the problem under consideration, as a linear program with integer variables, techniques like cutting plane methods or Lagrangian relaxation can be combined with branch-and-bound to generate provably optimal solutions. Of course, these methods do not guarantee polynomial running time in the general case. For many practically relevant instances, however, they work astonishingly well.

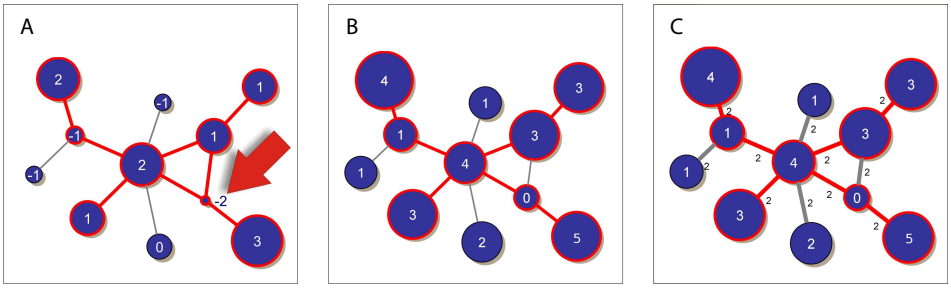


Figure 4.6: Transformation of MWCS to PCST. (A) Example of an MWCS instance. The minimum weight is $S' = -2$. (B) Vertex profits in PCST result from subtracting S' from every node weight. (C) Finally, all edge weights are set to $-S' = 2$. Optimal solutions are marked in black. The maximum-weight connected subgraph T has weight $S(T) = 7$, the optimal prize-collecting Steiner tree has profit $P(T) = 23 - 14 = 9$. Observe that $P(T) = S(T) - S'$.

In addition to computing the best solution to (4.12) the approach is also able to compute a list of promising solutions. In the ILP approach, binary variables x_v determine the presence of nodes in a subgraph $T = (V_T, E_T)$, that is, $x_v = 1$ if $v \in V_T$ and $x_v = 0$ otherwise. Now let $T^* = (V_{T^*}, E_{T^*})$ be an optimal subnetwork as identified by the branch-and-cut algorithm. Adding the Hamming distance-like inequality

$$\sum_{v \in V_{T^*}} (1 - x_v) \geq \alpha |V_{T^*}|$$

with $\alpha \in [0, 1]$ and re-optimising leads to a best solution differing in at least $\alpha |V_{T^*}|$ nodes from T^*. This procedure can be iterated k times. By this the user can determine the number k of suboptimal solutions that should be reported and adjust the variety of solutions via the percentage of nodes that should be different to the optimal solution (parameter: α).

4.1.2 Heuristic to Maximum-Scoring Subnetwork Problem

Formulation

Since the MWCS problem was shown to be NP-hard, the calculation of the solution is formulated as an integer linear programming algorithm, which is then run using the CPLEX optimisation library. To avoid the use of an commercial software and make the algorithm available in an R package, a heuristic algorithm is implemented to find an approximate solution to the

MWCS, based on the same scoring of the network. This heuristic implementation allows the calculation of very large subnetworks in faster running time. The heuristic is, like the optimal algorithm, based on the transformation of node to edge scores and uses the efficient calculation of MSTs to obtain an approximate solution.

The following depicts an outline of the algorithm (Figure 4.7):

1. In the first step all positive connected nodes are aggregated into meta-nodes (B).
2. By defining an edge weight based on the negative node scores, the node scores are transferred to the edges. $w(a, b) = |Score_a/k_a + Score_b/k_b|$ (C).
3. On these edge scores a minimum spanning tree is calculated (D).
4. All paths between positive meta-nodes are calculated based on the MST to obtain the negative nodes between the positives (E).
5. The MST of the largest component of the negative nodes (F) is used to search for the highest scoring path between all positive meta-nodes.
6. The path with the highest score, regarding node scores of negative nodes and the positive meta-nodes they connect, gives the resulting approximated module (G) as an induced subnetwork from the original network (H).

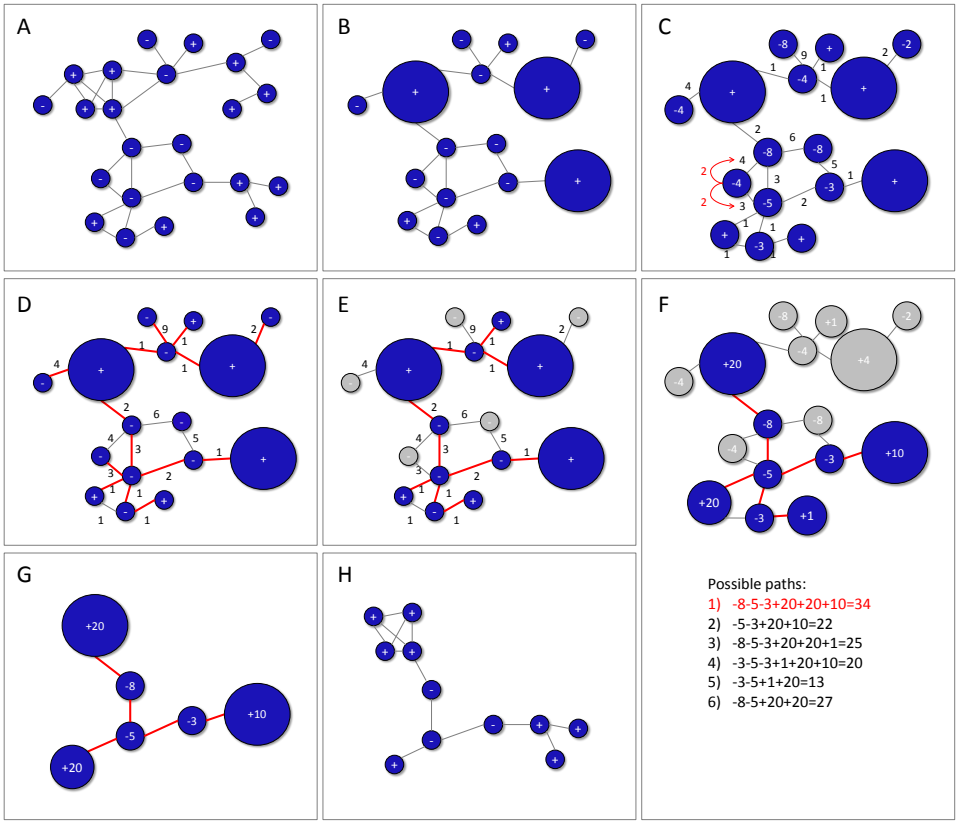


Figure 4.7: Heuristic to maximum-scoring subnetwork problem. The heuristic consists of 6 steps depicted in the panels of the figure. The single steps are described in the outline of the algorithm. (G) and (H) show the resulting module.

Validation

To validate the performance of the heuristic an artificial signal modules is simulated. Therefore, an induced subnetwork of the HPRD-network comprising the genes present on the hgu133a Affymetrix chip is used. Within this network artificial signal modules of biological relevant sizes of 30 and 150 nodes are set, respectively; the remaining genes are considered as background noise. For all considered genes microarray data are simulated and analysed subsequently analogously to real gene expression analysis. A large range of FDRs between 0 and 0.8 is scanned and evaluated regarding the obtained solutions in terms of recall and precision. The heuristic solutions are compared to the optimal solution and a heuristic implemented in the Cytoscape plugin *jActiveModules* (Ideker et al., 2002). The results of the heuristic implemented in the `BioNet` (Section 4.1.3) package closely resemble

the optimal solutions and perform superior to the results of the other heuristical approach. Especially for a strong signal with 150 genes the heuristic yields a good approximation of the maximum-scoring subnetwork.

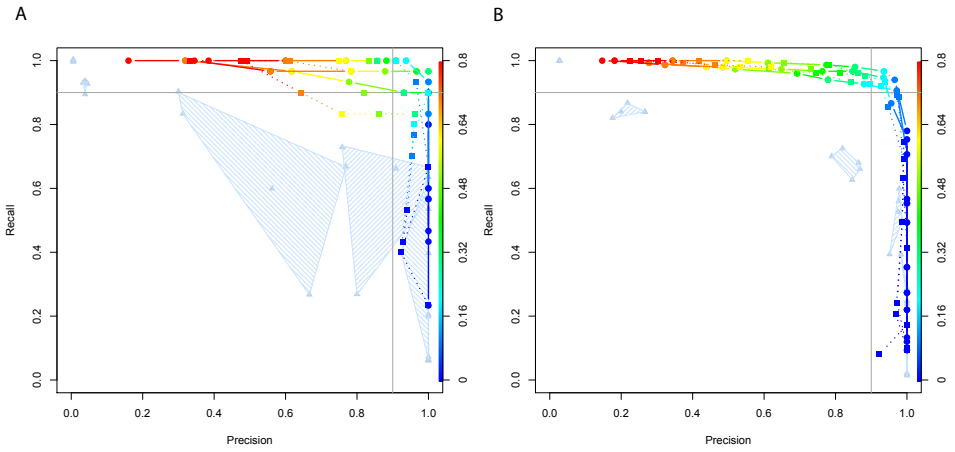


Figure 4.8: Performance validation. Plot of the recall vs. precision of a batch of solutions calculated for a wide range of FDRs (colouring scheme) with three replications each, for the exact solution and two heuristics. For the algorithm by Ideker et al. (2002) the 6 convex hulls (triangles) of solutions (solutions 5 and 6 partially overlap) are displayed obtained by applying it recursively to five independent simulations. Two different signal component sizes (30, left plot and 150, right plot) are evaluated with the same procedure. Clearly, the presented exact approach (solid line) captures the signal with high precision and recall over a relatively large range of FDRs. The results of the `BioNet` heuristic algorithm (dotted line) are much closer to the optimal solution over the entire range of FDRs compared to the *jActiveModules* heuristic; in particular, in the important region of high recall and precision.

4.1.3 BioNet, an R-package for the Functional Analysis of Biological Networks

To make the methodology presented in Section 4.1.1 available to the community, it is implemented as a software package in the open-source statistical language R. The following section is based on the publication Beisser et al. (2010), that originated from my doctoral studies, describing the software `BioNet` and its application to microarray data. Parts of the text and figures are taken with permission from Oxford Journals. The `BioNet` package provides an extensive framework for integrated network analysis in R, including an exact and a heuristic approach to identify functional modules. It comprises the statistics for the integration of transcriptomic and func-

tional data with biological networks, the scoring of nodes as well as methods for network search and visualisation. The `BioNet` package is available from `http://bionet.bioapps.biozentrum.uni-wuerzburg.de` and the Bioconductor website `http://www.bioconductor.org/`.

Integrated analysis of microarray data in the context of biological networks such as protein-protein interaction networks has become a major technique in systems biology. The primary objective is the identification of functional modules (significantly differentially expressed subnetworks) within large networks. This can be achieved by computing a score for each node in the network reflecting its functional relevance. Subsequently, a network search algorithm is required to find the highest-scoring subgraph. In fact, this problem has been proven to be NP-hard (Ideker et al., 2002). Recently, an algorithm (*heinz*, heaviest induced subgraph) that computes provably optimal and suboptimal solutions to the maximal-scoring subgraph (MSS) problem in reasonable running time using integer linear programming was devised by Dittrich et al. (2008). In extension to this, the R package BioNet implements methods for (i) integrating multiple p-values obtained from different experiments, (ii) scoring the nodes of the network by a modular scoring function, (iii) calculating provably optimal and suboptimal solutions to the MSS problem, (iv) calculating high-scoring solutions with a novel heuristic and (v) 2D and 3D visualisation of network solutions, see for example Figure 4.9.

Description

The `BioNet` package provides a comprehensive set of methods for the integrated analysis of gene expression data and biological networks. P-values are distributed uniformly under null hypotheses. Therefore, as a first step, multiple p-values derived from the analysis of different experiments (e.g. t-test or regression models) can be aggregated using a uniform order statistics (`aggrPvals`). The resulting distribution of combined p-values can be considered as a mixture of signal and noise, where the signal component is modelled to be Beta(a,1) distributed (Pounds and Morris, 2003). The model fit can be verified by the provided diagnostic plots (`plot.bum, hist.bum`). By fitting a beta-uniform mixture model (`fitBumModel`), the maximum-likelihood estimates for the mixture model can be obtained. These parameters are subsequently used to score the nodes of the network (`scoreNodes, scoreFunction`). The adjusted node score is given by $(a-1)\left(\log(x) - \log(\tau(\mathrm{FDR}))\right)$, where τ denotes the threshold for a given false discovery rate. The optimal and heuristic solutions of the MSS can be calculated by `runHeinz, runFastHeinz`. Bioconductor data structures and classes of the graph packages graph, RBGL as well as igraph are supported (Gentleman et al., 2004; Carey et al., 2005; Csardi and Nepusz, 2006). Networks can be imported and ex-

ported in different formats, allowing a smooth data exchange with standard network analysis tools like Cytoscape (Shannon et al., 2003) and further analysis.

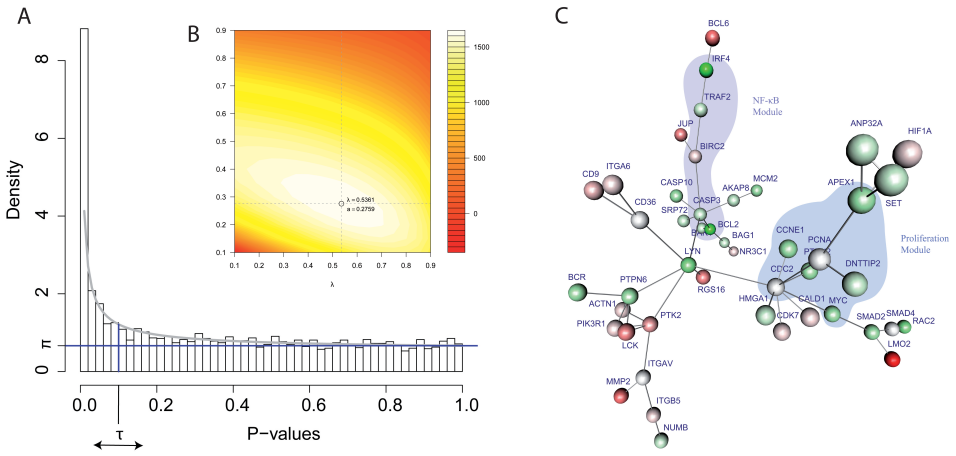


Figure 4.9: Node scoring and network solution for the DLBCL dataset. (A) Fitted mixture model and empirical p-value distribution: π indicates the upper bound for the fraction of noise and τ the significance threshold according to a given FDR. (B) Log-likelihood surface for the mixture parameter λ (x-axis) and the shape parameter a (y-axis). The derived scores from the p-value distribution are used subsequently to calculate the MSS. (C) A 3D visualisation of the identified optimal scoring module. Differential expression is depicted by node colouring for the disease under study (diffuse large B-cell lymphoma). Disease-relevant modules (shaded) (Rosenwald et al., 2002) are captured and extended by the network analysis (Section 4.4.2).

4.1.4 Case Study for BioNet

This case study exemplifies how an integrated network analysis can be conducted using the `BioNet` package. In the first part gene expression data from different lymphoma subtypes and clinical survival data are integrated with a comprehensive protein-protein interaction network based on HPRD. This is shown first in a quick start and later in a more detailed analysis. The second part focuses on the integration of gene expression data from Affymetrix single-channel microarrays with the human PPI network.

Quick Start

The quick start section gives a short overview of the essential `BioNet` methods and their application. A detailed analysis of the same dataset of diffuse large B-cell lymphomas is presented and discussed in Section 4.4.2.
The major aim of the presented integrated network analysis is to identify modules, which are differentially expressed between two different lymphoma subtypes (ABC and GCB) and simultaneously are risk associated (measured by the survival analysis).

First of all, the `BioNet` package and the required datasets, containing a human protein-protein interaction network and p-values derived from differential expression and survival analysis are loaded into the workspace.

```
> library(BioNet)
> library(DLBCL)
> data(dataLym)
> data(interactome)
```

Then these two p-values need to be aggregated into one p-value.

```
> pvals <- cbind(t = dataLym$t.pval, s = dataLym$s.pval)
> rownames(pvals) <- dataLym$label
> pval <- aggrPvals(pvals, order = 2, plot = FALSE)
```

Next a subnetwork of the complete network is derived, containing all the proteins which are represented by probesets on the microarray. Self-loops are removed from the network.

```
> subnet <- subNetwork(dataLym$label, interactome)
> subnet <- rmSelfLoops(subnet)
> subnet
```

```
A graphNEL graph with undirected edges
Number of Nodes = 2559
Number of Edges = 7788
```

To score each node of the network a Beta-uniform mixture model is fitted to the p-value distribution and subsequently the parameters of the model are used for the scoring function (Dittrich et al., 2008). A false-discovery rate of 0.001 is chosen.

```
> fb <- fitBumModel(pval, plot = FALSE)
> scores <- scoreNodes(subnet, fb, fdr = 0.001)
```

Here the fast heuristic approach to calculate an approximation to the optimal scoring subnetwork is used. An optimal solution can be calculated using

the `heinz` algorithm (Dittrich et al., 2008) requiring a commercial CPLEX license, see Section 4.1 and 4.1 for installation.

```
> module <- runFastHeinz(subnet, scores)
> logFC <- dataLym$diff
> names(logFC) <- dataLym$label
```

Both 2D and 3D module visualisation procedures are available in `BioNet`. For a 3D visualisation, see Section 4.1. Alternatively, the network can be easily exported in Cytoscape format, see Section 4.1.

```
> plotModule(module, scores = scores, diff.expr = logFC)
```

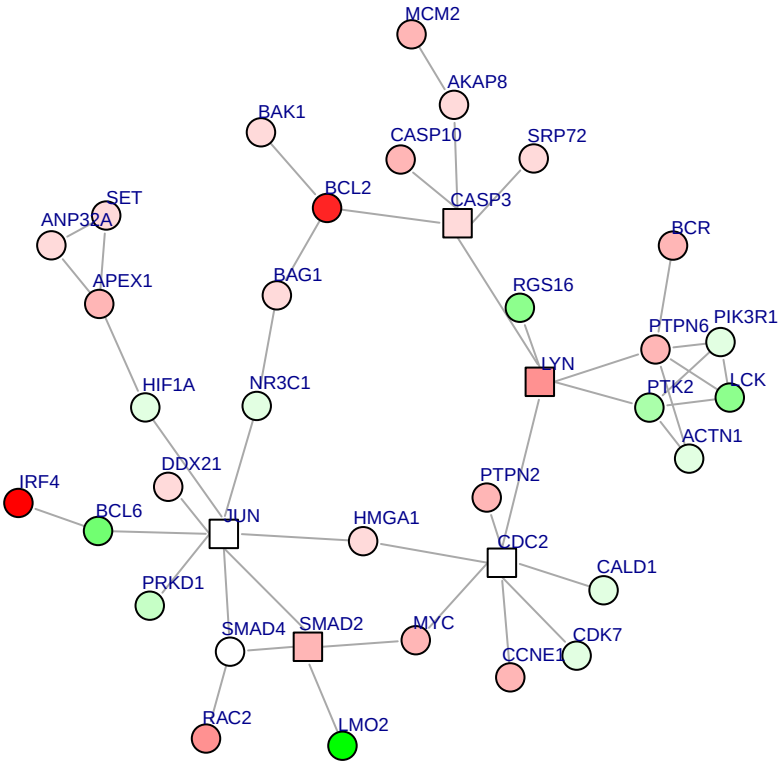


Figure 4.10: Resultant functional module. Differential expression between ABC and GCB B-cell lymphoma is coloured in red and green, where green shows an up-regulation in ACB and red an up-regulation in GBC. The shape of the nodes depicts the score: rectangles indicate a negative score, circles a positive score.

Diffuse Large B-cell Lymphoma Study

Integrated network analysis not only focuses on the structure (topology) of the underlying graph but integrates external information in terms of node and edge attributes. This part gives a detailed example how an integrative network analysis can be performed using a protein-protein interaction network, microarray and clinical (survival) data, for details see Section 4.1.1.

The Data First, the microarray data and interactome data which are available as expression set and a graph object from the `BioNet` package are loaded into the workspace. The graph objects can be either in the *graphNEL* format, which is used in the package `graph` and `RBGL` (Gentleman et al., 2008b; Carey et al., 2005) or in the *igraph* format from the package `igraph` (Csardi and Nepusz, 2006).

```
> library(BioNet)
> library(DLBCL)
> data(exprLym)
> data(interactome)
```

The published gene expression dataset from diffuse large B-cell lymphomas is used (Rosenwald et al., 2002). In particular, gene expression data from 112 tumours with the germinal center B-like phenotype and from 82 tumours with the activated B-like phenotype are included in this study. The expression data have been pre-compiled in an `ExpressionSet` structure.

```
> exprLym
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 3583 features, 194 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: Lym432 Lym431 ... Lym274 (194 total)
  varLabels: Subgroup IPI ... Status (5 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation:
```

For the network data literature-curated human protein-protein interactions obtained from HPRD are used (Prasad et al., 2009). Altogether the entire network used here comprises 9,386 nodes and 36,504 edges.

```
> interactome
```

```
A graphNEL graph with undirected edges
Number of Nodes = 9386
Number of Edges = 36504
```

From this a *Lymphochip*-specific interactome network as the vertex-induced subgraph is extracted by the subset of genes for which expression data are available on the *Lymphochip*. This can easily be done, using the `subNetwork` command.

```
> network <- subNetwork(featureNames(exprLym), interactome)
> network
```

```
A graphNEL graph with undirected edges
Number of Nodes = 2559
Number of Edges = 8538
```

Since the aim is to identify modules as connected subgraphs the largest connected component is obtained.

```
> network <- largestComp(network)
> network
```

```
A graphNEL graph with undirected edges
Number of Nodes = 2034
Number of Edges = 8399
```

So finally a *Lymphochip* network is derived which comprises 2,034 nodes and 8,399 edges.

Calculating the p-values

Differential Expression The next step uses `rowttest` from the package `genefilter` to analyse differential expression between the ABC and GCB subtype after imputation of missing values:

```
> library(genefilter)
> library(impute)
> expressions <- impute.knn(exprs(exprLym))$data
```

```
Cluster size 3583 broken into 2332 1251
Cluster size 2332 broken into 1628 704
Cluster size 1628 broken into 1 1627
Done cluster 1
Cluster size 1627 broken into 727 900
Done cluster 727
Done cluster 900
Done cluster 1627
```

```
Done cluster 1628
Done cluster 704
Done cluster 2332
Done cluster 1251
```

```
> t.test <- rowttests(expressions, fac = exprLym$Subgroup)
```

```
> t.test[1:10, ]
```

The result looks as follows:

Table 4.1: Result of rowttest.

	statistic	dm	p.value
MYC(4609)	-3.38	-0.41	0.00
KIT(3815)	1.37	0.12	0.17
ETS2(2114)	0.51	0.04	0.61
TGFBR3(7049)	-2.89	-0.43	0.00
CSK(1445)	-3.61	-0.26	0.00
ISGF3G(10379)	-2.94	-0.25	0.00
RELA(5970)	-0.95	-0.07	0.34
TIAL1(7073)	-2.03	-0.09	0.04
CCL2(6347)	-0.94	-0.16	0.35
SELL(6402)	-2.18	-0.26	0.03

Survival Analysis The survival analysis implemented in the package *survival* can be used to assess the risk association of each gene and calculate the associated p-values. As this is a time-consuming calculation p-values are pre-calculated in the dataLym object.

```
> data(dataLym)
> ttest.pval <- t.test[, "p.value"]
> surv.pval <- dataLym$s.pval
> names(surv.pval) <- dataLym$label
> pvals <- cbind(ttest.pval, surv.pval)
```

The complete survival analysis is shown in the following. First of all the data have to be centered for the analysis.

```
> mapped.lym = exprs(exprLym)
> gmeans = rowMeans(mapped.lym, na.rm=T)
> exprs.cent = sweep(mapped.lym,1,gmeans)
> expdat = as.data.frame(t(exprs.cent))
```

Feature data are extracted and combined to create a survival object.

```
> features = as.matrix(t(rbind(pData(exprLym)[colnames(mapped.lym),
+     "Status"], pData(exprLym)[colnames(mapped.lym), "FollowUpYears"])))
> time1 = as.numeric(features[,2])
```

```
> status = features[,1]
> status[which(status == "Alive")] = 0
> status[which(status == "Dead")] = 1
> status = as.numeric(status)
> pat = !is.na(status)
> sv = Surv(time1[pat], status[pat])
```

The univariate Cox PH regression is run over all genes to obtain p-values.

```
> calc.srv <- function(expdat, pat)
> {
>   srv.pval = mat.or.vec(dim(expdat[pat,])[2], 1)
>   srv.coeff = mat.or.vec(dim(expdat[pat,])[2], 1)
>   for(i in 1:dim(expdat[pat,])[2])
>   {
>     cox = summary(coxph(as.formula(paste("sv ~ `",
+          colnames(expdat[pat,])[i] ,"`", sep="")), data=expdat[pat,][i]))
>     srv.pval[i] = cox$logtest[3]
>     srv.coeff[i] = cox$coef[1]
>     print(i)
>   }
>   return(list(srv.pval, srv.coeff))
> }
>
> names(srv.pval) = colnames(expdat[pat,])
> names(srv.coeff) = colnames(expdat[pat,])
```

Calculation of the Score Two p-values for each gene are obtained, for differential expression and survival relevance. Next, they are aggregated for each gene into one p-value of p-values using order statistics. The function `aggrPvals` calculates the second order statistic of the p-values for each gene.

```
> pval <- aggrPvals(pvals, order = 2, plot = FALSE)
```

Now, a Beta-uniform mixture model is fitted to the distribution of the aggregated p-values. The following plot shows the fitted Beta-uniform mixture model in a histogram of the p-values.

```
> fb <- fitBumModel(pval, plot = FALSE)
> fb

Beta-Uniform-Mixture (BUM) model

3583 pvalues fitted

Mixture parameter (lambda):        0.537
```

```
shape parameter (a):                    0.276
log-likelihood:                         1651.0
```

```
> par(mfrow = c(1, 2))
> hist(fb)
> plot(fb)
```

```
windows
      2
```

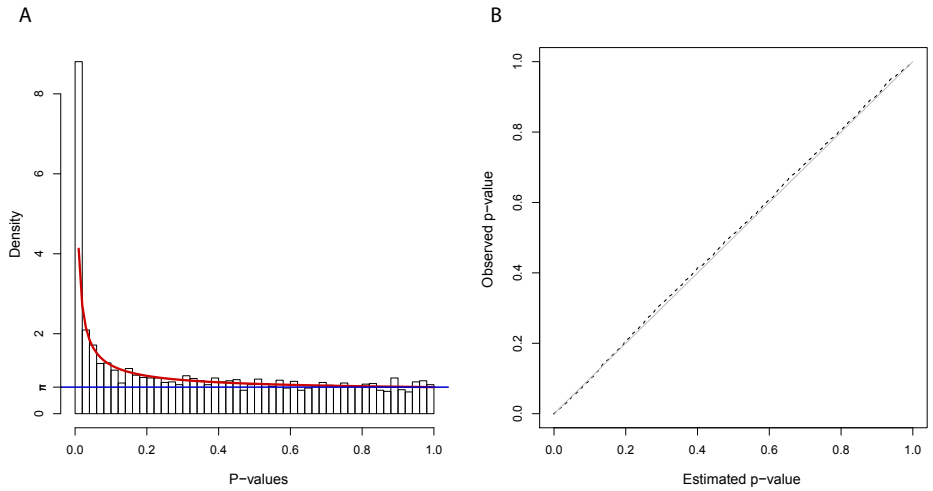


Figure 4.11: Histogram of p-values, overlayed by the fitted BUM model coloured in red and the π-upper bound displayed as a blue line. The right plot shows a quantile-quantile plot, indicating a nice fit of the BUM model to the p-value distribution.

The quantile-quantile plot indicates that the BUM model fits nicely to the p-value distribution. A plot of the log-likelihood surface can be obtained with `plotLLSurface`. It shows the mixture parameter λ (x-axis) and the shape parameter a (y-axis) of the Beta-uniform mixture model. The circle in the plot depicts the maximum-likelihood estimates for λ and a.

```
> plotLLSurface(pval, fb)
```

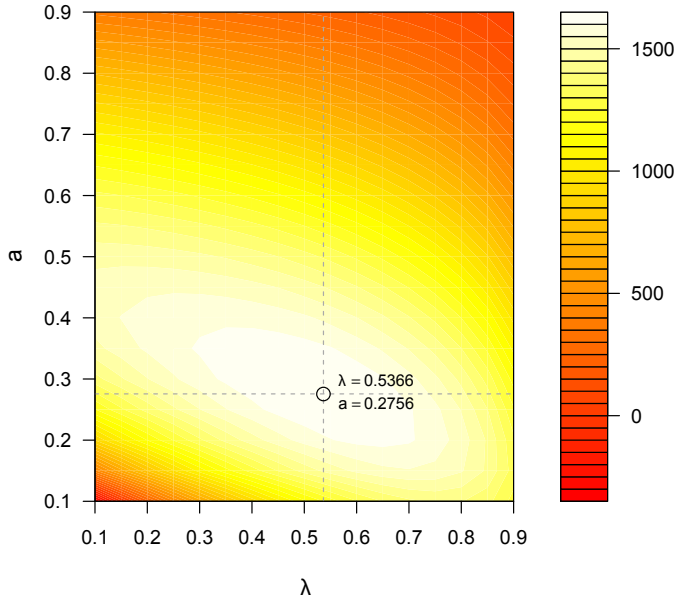


Figure 4.12: Log-likelihood surface plot. The range of the colours shows an increased log-likelihood from red to white. Additionally, the optimal parameters λ and a for the BUM model are highlighted.

The nodes of the network are now scored using the fitted BUM model and a FDR of 0.001.

```
> scores <- scoreNodes(network = network, fb = fb,
+     fdr = 0.001)
```

In the next step the network with the scores and edges is written to a file and the `heinz` algorithm is used to calculate the maximum-scoring subnetwork. In order to run `heinz` self-loops have to be removed from the network.

```
> network <- rmSelfLoops(network)
> writeHeinzEdges(network = network, file = "lymphoma_edges_001",
+     use.score = FALSE)
```

```
[1] TRUE
```

```
> writeHeinzNodes(network = network, file = "lymphoma_nodes_001",
+     node.scores = scores)
```

```
[1] TRUE
```

Calculation of the Maximum-Scoring Subnetwork In the following the `heinz` algorithm is started using the `heinz.py` python script. This starts

the integer linear programming optimisation and calculates the maximum-scoring subnetwork using `CPLEX`.

The command is: "heinz.py -e lymphoma_edges_001.txt -n lymphoma_nodes_001.txt -N True -E False" or `runHeinz` on a linux machine with CPLEX installed.

The output is pre-calculated in `lymphoma_nodes_001.txt.0.hnz` and `lymphoma_edges_001.txt.0.hnz` in the subdirectory "extdata" of the R `BioNet` library directory.

```
> datadir <- file.path(.path.package("BioNet"),
+     "extdata")
> dir(datadir)
```

```
[1] "ALL_edges_001.txt.0.hnz"
[2] "ALL_nodes_001.txt.0.hnz"
[3] "cytoscape.sif"
[4] "lymphoma_edges_001.txt.0.hnz"
[5] "lymphoma_nodes_001.txt.0.hnz"
[6] "n.weight.NA"
[7] "weight.EA"
```

The output is loaded as a graph and plotted with the following commands:

```
> module <- readHeinzGraph(node.file = file.path(datadir,
+     "lymphoma_nodes_001.txt.0.hnz"), network = network)
> diff <- t.test[, "dm"]
> names(diff) <- rownames(t.test)
```

```
> plotModule(module, diff.expr = diff, scores = scores)
```

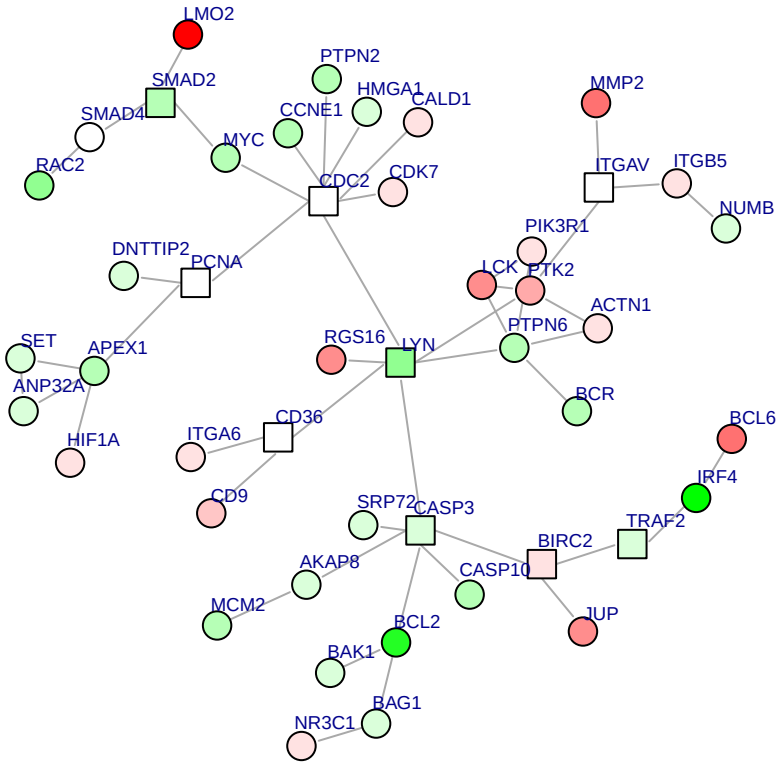


Figure 4.13: Resultant functional module for the lymphoma dataset. Differential expression between ABC and GCB B-cell lymphoma is coloured in red and green, where green shows an up-regulation in GCB and red an up-regulation in ABC. The shape of the nodes depicts the score: rectangles indicate a negative score, circles a positive score.

The log fold-changes are visualised by the colouring of the nodes, the shape of the nodes depicts the score (positive=circle, negative=square). It is also possible to visualise the module in 3D with the function `plot3dModule`, but therefore the rgl package, a 3D real-time rendering system, has to be installed. The plot can be saved to pdf-file with the function `save3dModule`. And the resulting module would look as follows.

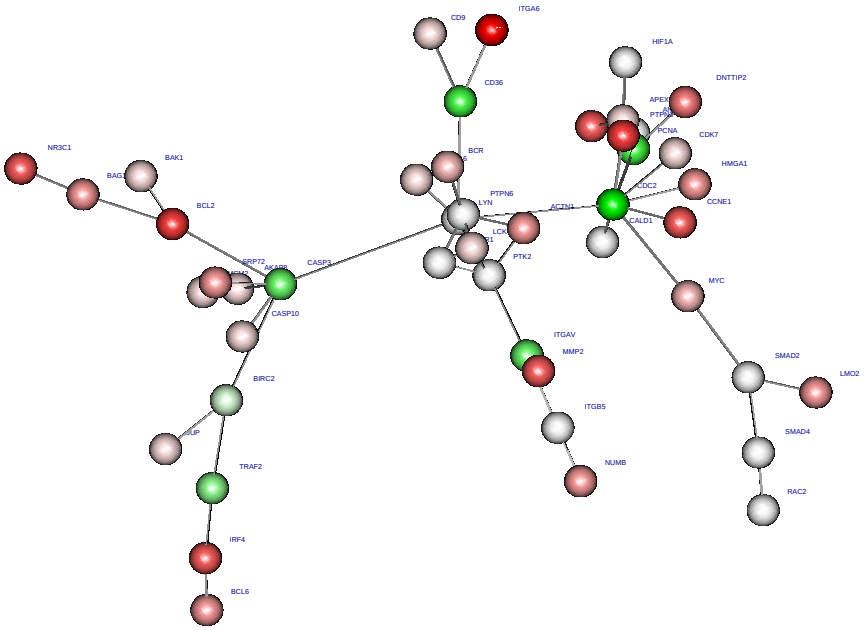


Figure 4.14: 3D visualisation of the same functional module shown in Figure 4.13. Here, only the scores are depicted by the colouring of the nodes (positives: red, negatives: green.

The resulting subnetwork consists of 46 nodes and 50 edges. It has a cumulative sum of the scores of 71.07 from 37 positive (coloured in red) and 9 negative nodes (coloured in green).

```
> sum(scores[nodes(module)])

[1] 71.07341

> sum(scores[nodes(module)] > 0)

[1] 37

> sum(scores[nodes(module)] < 0)

[1] 9
```

The captured interactome module that has been described to play a major biological role in the GCB and ABC DLBCL subtypes. It includes for example, the proliferation module which is more highly expressed in the ABC DLBCL subtype (Rosenwald et al., 2002) comprising the genes: MYC, CCNE1, CDC2, APEX1, DNTTIP2, and PCNA. Likewise, genes IRF4, TRAF2, and BCL2, which are associated with the potent and oncogenic NFκB pathway.

ALL Study

This section describes the integrated network approach applied to the analysis of Affymetrix microarray data. In addition to the previous section, the data are analysed for differential expression using the package `limma` (Smyth, 2005). The resulting module can be exported in various formats and for example displayed with Cytoscape (Shannon et al., 2003).

The Data First, the microarray data are load and the human interactome data, which are available as a graph object from the `BioNet` package. The popular Acute Lymphoblastic Leukaemia dataset (Chiaretti et al., 2004) with 128 arrays is used as an example for an Affymetrix single-channel microarray. These data are available in the package `ALL` (Huber and Gentleman, 2006) as a normalised `ExpressionSet`.

```
> library(BioNet)
> library(DLBCL)
> library(ALL)
> data(ALL)
> data(interactome)
```

The microarray data include expression values from 128 samples of patients with T-cell ALL or B-cell ALL using Affymetrix hgu95av2 arrays. Aim of the integrated analysis is to capture significant genes in a functional module, all of which are potentially involved in acute lymphoblastic leukaemia and show a significant difference in expression between the B- and T-cell samples.

```
> ALL
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 128 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: 01005 01010 ... LAL4 (128 total)
  varLabels: cod diagnosis ... date last seen (21
    total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
  pubMedIds: 14684422 16243790
Annotation: hgu95av2
```

The same network data from HPRD are used (Prasad et al., 2009), comprising 9,386 nodes and 36,504 edges.

```
> interactome
```

```
A graphNEL graph with undirected edges
Number of Nodes = 9386
Number of Edges = 36504
```

In the next step the Affymetrix identifiers are mapped to the protein identifiers of the PPI network. Since several probesets represent one gene, one of these has to be selected or they can be concatenated into one gene. One possibility is to use the probeset with the highest variance for each gene. This is accomplished for the `ExpressionSet` with the `mapByVar` function. It also maps the Affymetrix IDs to the identifiers of the network using the chip annotations and network geneIDs, which are unique, and returns the network names in the expression matrix. The expression matrix is reduced to the genes which are present in the network.

```
> mapped.eset = mapByVar(ALL, network = interactome,
+     attr = "geneID")
> mapped.eset[1:5, 1:5]
```

```
                 01005     01010    03002    04006    04007
MAPK3(5595)   7.597323  7.479445 7.567593 7.384684 7.905312
TIE1(7075)    5.046194  4.932537 4.799294 4.922627 4.844565
CYP2C19(1557) 3.900466  4.208155 3.886169 4.206798 3.416923
BLR1(643)     5.903856  6.169024 5.860459 6.116890 5.687997
DUSP1(1843)   8.570990 10.428299 9.616713 9.937155 9.983809
```

The dataset is reduced to 6,274 genes. To find out how many genes are contained in the human interactome the intersect is calculated.

```
> length(intersect(rownames(mapped.eset), nodes(interactome)))
```

```
[1] 6274
```

Since the human interactome contains 6,274 genes from the chip one can either extract a subnetwork with the method `subNetwork` or proceed with the whole network. Automatically the negative expectation value is used later when deriving the scores for the nodes without intensity values. Here, the subnetwork is extracted. Again the largest connected component of the network is used and existing self-loops are removed.

```
> network = subNetwork(rownames(mapped.eset), interactome)
> network
```

```
A graphNEL graph with undirected edges
Number of Nodes = 6274
Number of Edges = 25211
```

```
> network <- largestComp(network)
> network <- rmSelfLoops(network)
> network
```

```
A graphNEL graph with undirected edges
Number of Nodes = 5762
Number of Edges = 23546
```

So finally, a *chip-specific* network which comprises 5,762 nodes and 23,546 edges is derived.

Calculating the p-values

Differential Expression In the next step `limma` (Smyth, 2005) is used to analyse differential expression between the B-cell and T-cell groups.

```
> library(limma)
> design = model.matrix(~-1 + factor(c(substr(unlist(ALL$BT),
+     0, 1))))
> colnames(design) = c("B", "T")
> contrast.matrix <- makeContrasts(B - T, levels = design)
> contrast.matrix
```

```
      Contrasts
Levels B - T
     B     1
     T    -1
```

```
> fit <- lmFit(mapped.eset, design)
> fit2 <- contrasts.fit(fit, contrast.matrix)
> fit2 <- eBayes(fit2)
```

The p-values are calculated and the corresponding scores thereupon.

```
> pval = fit2$p.value[, 1]
```

Calculation of the Score The p-values are used to fit the Beta-uniform mixture model to their distribution. Figure 4.15 shows the fitted Beta-uniform mixture model in a histogram of the p-values. The quantile-quantile plot indicates that the BUM model fits to the p-value distribution. Although the data show a slight deviation from the expected values, the analysis is continued with the fitted parameters.

```
> fb <- fitBumModel(pval, plot = FALSE)
> fb
```

```
Beta-Uniform-Mixture (BUM) model

6274 pvalues fitted

Mixture parameter (lambda):         0.453
shape parameter (a):                  0.145
log-likelihood:                         12108.1
```

```
> par(mfrow = c(1, 2))
> hist(fb)
> plot(fb)
```

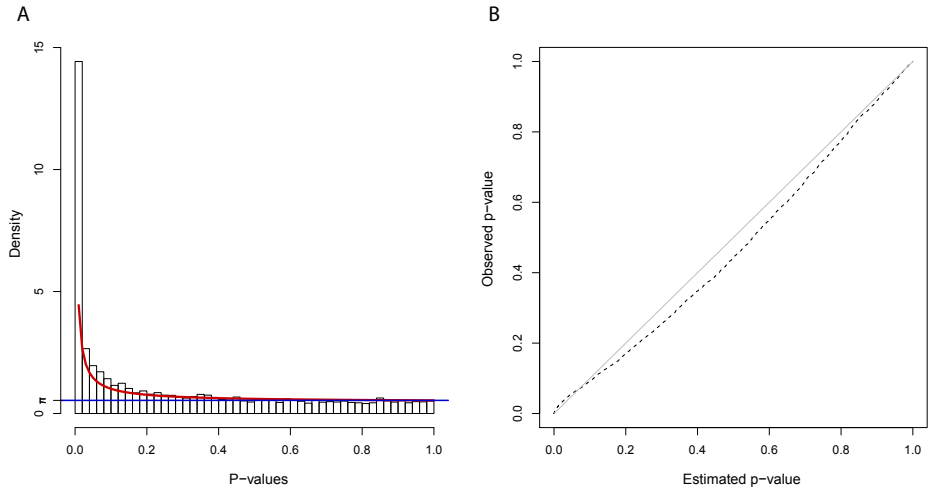


Figure 4.15: Histogram of p-values, overlayed by the fitted BUM model in red and the π-upper bound displayed as a blue line. The right plot shows a quantile-quantile plot, in which the estimated p-values from the model fit deviate slightly from the observed p-values.

The nodes of the network are scored using the fitted BUM model and a FDR of 1e-14. Such a low FDR is chosen to obtain a small module, which can be visualised appropriately.

```
> scores <- scoreNodes(network = network, fb = fb,
+     fdr = 1e-14)
```

In the next step the network with the scores and edges is written to file and the `heinz` algorithm is used to calculate the maximum-scoring subnetwork.

```
> writeHeinzEdges(network = network, file = "ALL_edges_001",
+     use.score = FALSE)
```

```
[1] TRUE
```

```
> writeHeinzNodes(network = network, file = "ALL_nodes_001",
+     node.scores = scores)
```

```
[1] TRUE
```

Calculation of the Maximum-Scoring Subnetwork In the following the `heinz` algorithm is started using the `heinz.py` python script. This starts the integer linear programming optimisation and calculates the maximum-scoring subnetwork using `CPLEX`.

The command is: "heinz.py -e ALL_edges_001.txt -n ALL_nodes_001.txt -N True -E False" or `runHeinz` on a linux machine with CPLEX installed.

The output is pre-calculated in `ALL_nodes_001.txt.0.hnz` and `ALL_edges_001.txt.0.hnz` in the R `BioNet` directory, subdirectory extdata.

The output is loaded as a graph with the following commands:

```
> datadir <- file.path(.path.package("BioNet"),
+     "extdata")
> module <- readHeinzGraph(node.file = file.path(datadir,
+     "ALL_nodes_001.txt.0.hnz"), network = network)
```

Attributes are added to the module, to depict the difference in expression and the score later.

```
> nodeDataDefaults(module, attr = "diff") <- ""
> nodeData(module, n = nodes(module), attr = "diff") <-
+     fit2$coefficients[nodes(module), 1]
> nodeDataDefaults(module, attr = "score") <- ""
> nodeData(module, n = nodes(module), attr = "score") <-
+     scores[nodes(module)]
> nodeData(module)[1]
```

```
$`BTK(695)`
$`BTK(695)`$geneID
[1] "695"

$`BTK(695)`$geneSymbol
[1] "BTK"

$`BTK(695)`$diff
[1] 0.919985

$`BTK(695)`$score
[1] -8.900234
```

The module is saved as XGMML file and look at with the software Cytoscape (Shannon et al., 2003), colouring the node by their "diff" attribute and changing the node shape according to the "score".

```
> saveNetwork(module, file = "ALL_module", type = "XGMML")

[1] "...adding nodes"
[1] "...adding edges"
[1] "...writing to file"
```

The resulting network with 31 nodes and 32 edges and coloured nodes, looks as follows:

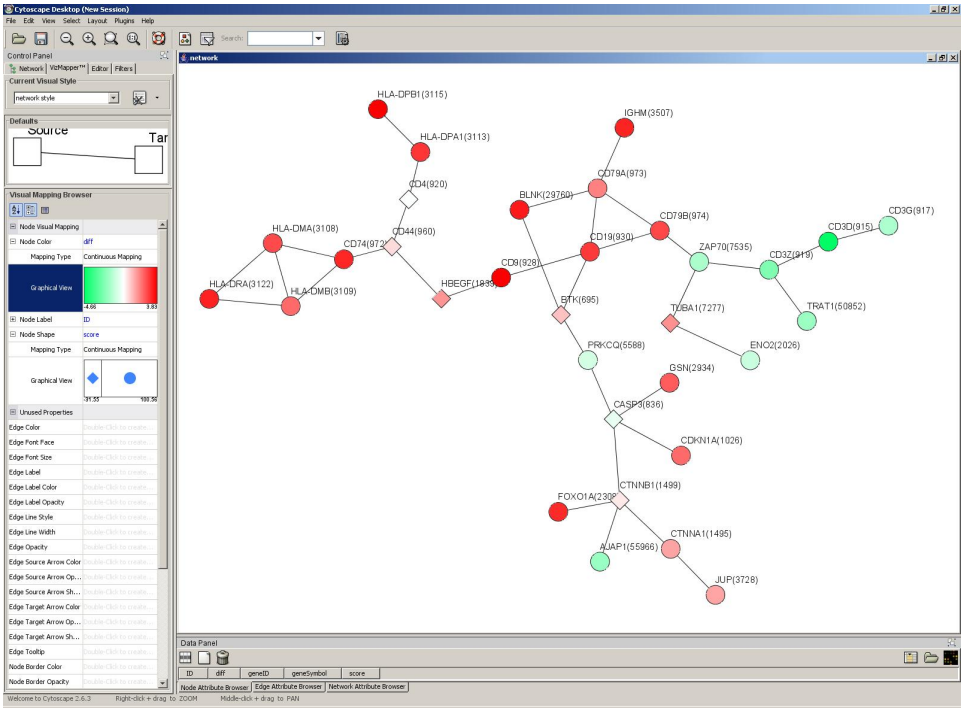


Figure 4.16: Resultant module visualised in Cytoscape. Significantly up-regulated genes are coloured in red, genes that show significant down-regulation are coloured in green, for the contrast B vs. T cells. The score of the nodes is shown by the shape of the nodes, circles indicate a positive score, diamonds a negative score.

The module comprises several parts, one part showing a high up-regulation in the B-T contrast (CD79A, BLNK, CD19, CD9, CD79B), participates in B cell activation and differentiation and response to stimuli according to their GO annotation. While the other large up-regulated part is involved in antigen processing and presentation and immune response (HLA-DMA, HLA-DPA1, CD4, HLA-DMB, HLA-DRB5, HLA-DPB1). T cell/leukocyte activating genes (CD3D, CD3G, CD3Z, ENO2, TRAT1, ZAP70) are coloured

in green. The lower middle part is involved in negative regulation of apoptosis, developmental processes and programmed cell death. Most of them are involved in overall immune system processes and as expected, mostly B and T cell specific genes comprise the resulting module, as this contrast is used for the test of differential expression.

Installation

The `BioNet` Package The `BioNet` package is freely available from Bioconductor at http://www.bioconductor.org.

External Code to Call CPLEX The algorithm to identify the optimal scoring subnetwork is based on the software dhea (district heating) from Ljubić et al. (2006). The C++ code is extended in order to generate suboptimal solutions and is controlled over a Python script. The dhea code uses the commercial CPLEX callable library version 9.030 by ILOG, Inc. (Sunnyvale,CA). In order to calculate the optimal solution a CPLEX library is needed. The other routines, the dhea code and heinz.py Python script (current version 1.63) are publicly available for academic and research purposes within the heinz package of the open source library LiSA (http://www.planet-lisa.net). The dhea code has to be included in the same folder as heinz.py, in order to call the routine by the Python code. To calculate the maximum-scoring subnetwork without an available CPLEX license a heuristic is included in the `BioNet` package, see `runFastHeinz`.

4.1.5 Discussion

Systematic analyses of cellular systems are gaining more and more importance in biological and medical research. Recent advances in experimental high-throughput technologies make large amounts of data on genomic, transcriptomic, proteomic and phenomic level available. Microarray-based technologies allow highly parallelized measurements of mRNA or DNA levels. In the field of proteomics, mass spectrometry techniques are becoming capable of analysing the entire cellular proteome on a quantitative level. In addition, large-scale data on protein-protein interactions are accumulating and permit the generation of large cellular interaction networks.

In contrast to reductionist approaches, which usually focus on specific isolated parts, systems biology analyses the entire system as a whole. Classical approaches for network analysis investigate the topological structure of the network. This implicitly assumes that all nodes and edges in the network are

alike and thus can be treated as equivalent. Structural analysis of interaction networks may successfully recover complexes as highly interconnected regions in the network or can describe global network properties as, for example, scale-free network architecture. In biological networks, however, analysis of network structure alone can only deliver limited insight into the functioning of a cellular system. In general, different proteins in a cell, corresponding to different network nodes, take over different functions in different cellular processes, are expressed in different amounts or may be localised in different subcompartments of a cell. Disregarding this kind of information considerably reduces the biological insight to be gained from network analysis. Integrated network analysis, in contrast, allows the superpositioning of biological information onto the network structure; the subsequent analysis can then be performed in this biological context.

A major strength of the presented methodology is its flexibility and its generalisability. Indeed any data and analysis method resulting in p-values as a measurement of significance are, by nature, a suitable input for the scoring function and can thus be used to score network nodes for subsequent module searching. Although p-values derived from order statistics of p-values from t-tests or Cox regressions of gene expression data can generally be modelled appropriately with a BUM model, it is always advisable to check the applicability of the model for observed p-value distribution. A quantile-quantile plot of the observed vs. the theoretical distribution of the BUM model visualises potential deviation from the theoretical model with high sensitivity. Depending on signal content of the data, the flexible score allows to fine-tune sensitivity and specificity and thereby the size of the resulting modules by choice of an appropriate FDR. The module can eventually be calculated either by using the exact approach or by using the included heuristic, according to one's computational means.
All areas of biological sciences can benefit from an integrated analysis of molecular networks, considering its straightforward applicability to a wide range of experimental data and the intuitive concept of a functional module due to the diagnostically conclusive visualisation of the resulting subnetwork. Since more and more functional and molecular data will become available in the future, the analysis, as exemplified in this section, will certainly become an important standard tool in the analysis of biological large-scale data.

4.2 Assessing Accuracy and Robustness of Functional Modules

Extending the topological analysis of biological networks, integrated network analysis incorporates additional genomic data into the network. High-throughput genomic data provide a wealth of information on genomic, transcriptomic and proteomic level, that is widely used in integrated network analysis. For gene expression data integrated approaches are used to search for pathways, functional modules or gene signatures containing differentially expressed genes in the context of genetic interaction networks or PPI networks (Scott et al., 2006; Dittrich et al., 2008; Ideker et al., 2002; Ulitsky and Shamir, 2007). The introduced exact approach (Section 4.1.1) resolves the subnetwork-finding problem to optimality using integer linear programming. All of the above mentioned methods generate a resulting module derived from expression data and a biological network. Besides the accuracy of the solution, which is intricate to assess in biological data, another question concerns the variability of the solution taking into consideration that molecular data are noisy. Therefore, the objective is on the one hand to produce an accurate solution and on the other hand to obtain a robust module. The first step to approach this problem is to assess the accuracy and variability of the solutions in a simulation study. This is done by comparing the exact approach to established methods for module detection. The different sources of variability are investigated by assessing the accuracy and robustness using (i) unperturbed data, (ii) perturbed integrated data and (iii) perturbed networks. The different perturbations mimic the noise present in the underlying gene expression data and the high uncertainty of edge information based on different measurements of protein-protein interactions. Due to these technical difficulties, the underlying PPI may contain a relatively high amount of false positive as well as false negative edges (von Mering et al., 2002).

4.2.1 Simulation of Reference Modules

To evaluate the performance of the proposed algorithm and the improvement over other methods, a simulation framework is created on the basis of the input microarray data. An induced network from HPRD is used as source of the protein-protein interactions, contained in the `BioNet` package, with 2,034 genes which are existent on the DLBCL microarray. To compare the resulting modules to the true solution, reference modules G_{sub} are created as follows:

1. Start with a given graph $G = (V, E)$ and an empty graph $G_{sub} = (W = \emptyset, F = \emptyset)$ with $W \subseteq V$ and $F \subseteq E$

2. Select random seed node $v \in V$ and include node in W
3. Expand G_{sub} by adding a node u from the $neighborhood_G(v \in V \backslash W)$ for which the average path length $l_{G_{sub}} \approx l_G$
4. Repeat step 3 until given size $|W(G_{sub})|$ is reached

The average path length is chosen as a characteristic network measure in the algorithm, which remains approximately constant between the network and all extracted modules in real datasets; in contrast to other considered measures like the degree, degree distribution, diameter, average path-length or clustering coefficient. The subnetwork is termed signal module in the following, since for the genes contained in the module expression values are generated which show differential expression between two groups. Expression values for genes not contained in the signal module are drawn from a normal distribution $N(\mu, \sigma^2)$, with a variance of 1 and with the same mean μ for both groups. These expression values correspond to random noise. In contrast, genes of the signal component are set according to a normal distribution with the same variance of 1, but with different means (μ_1, μ_2) for both groups. In this study varying signal strengths (differences in means) are applied to increase the difficulty to detect the signal module. Subsequently, the simulated gene expression data are analysed as detailed in Section 4.1.1.

4.2.2 Perturbation of the Integrated Data

A resampling procedure (Section 3.3) is used to perturb the integrated microarray data. A 50% jackknife is used and half of the observations dropped as recommended by Felsenstein (1985, 2004). For microarray data the jackknifing procedure works as follows (also see Figure 4.17). The samples are drawn without replacement from the gene expression matrix . With a 50% jackknife applied, randomly 50% of the samples in the gene expression matrix are drawn from each group. This produces a different gene expression set each time, called the pseudo-replicate.

4.2.3 Perturbation of the Network

Functional modules are calculated on integrated protein-protein interaction network with respect to three types of perturbations. The perturbations considered are

- Random deletion of edges
- Random addition of edges
- Random rewiring of edges

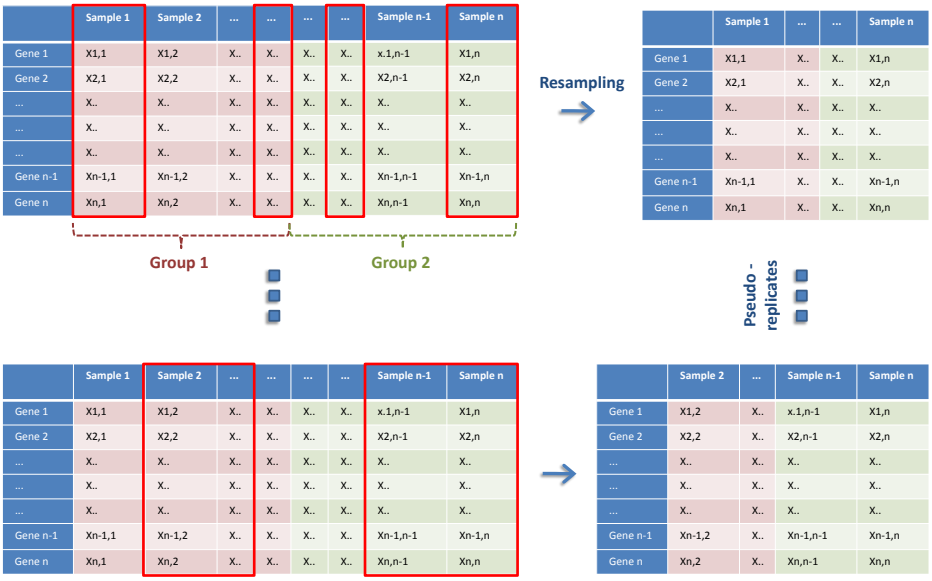


Figure 4.17: The microarray gene expression data are perturbed by jackknife resampling. Thereby samples are drawn without replacement from the gene expression matrix. A 50% jackknife is applied, drawing randomly 50% of the samples in the gene expression matrix from each group. This produces a different gene expression set each time.

for 10%, 25% and 50% of all edges. For the rewiring two edges are chosen randomly in each step. The nodes the edges connect are permuted in order to form new connections between them. The step is iterated until the selected number of edges is rewired. This procedure is also applied in other applications to generate a null distribution of the network for statistical testing. One advantage of the rewiring over other randomisation methods is that it keeps the degree distribution unchanged. For the creation of a null distribution the rewiring is applied thousands of times to achieve a complete randomisation of the network.

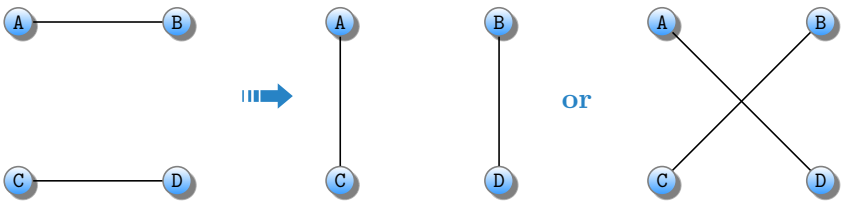


Figure 4.18: Rewiring of the network by permuting two edges.

4.2.4 Implementation Details and Parameter Settings of other Methods

The DEGAS algorithm (Ulitsky et al., 2008, 2010) of the program Matisse identifies minimal connected subnetworks in a PPI network in which the number of dysregulated genes from expression profiles exceeds a certain threshold. jActiveModules is another heuristic approach to identify maximum scoring subnetworks based on expression p-values by transforming p-values into scores and assigning each protein of a PPI network a score. The DEGAS algorithm is applied throughout the study with the following parameters: UP regulated, significant threshold=1, l parameter (number of outlier cases)=1, heuristic=ExpandingGreedy, k-steps=1, k parameter (number of significant genes per case) is varied in steps of 10 from 1 to 70 to obtain modules of varying sizes. For jActiveModules the number of modules is set to 1 and it is run iteratively on the previous solution until the smallest possible module size is reached.

4.2.5 Accuracy

Comparison to Heuristic Method

To validate the performance of exact algorithm artificial signal modules and microarray data are simulated and analysed with the presented algorithm. Therefore, an induced subnetwork of the HPRD-network comprising the genes present on the hgu133a affymetrix chip is obtained. Within this network artificial signal modules are set of biological relevant size of 30 and 150 nodes, respectively; the remaining genes are considered as background noise. For all considered genes microarray data are simulated as described in Section 4.2.1. A set of arrays (20) is divided into two groups of ten replicates each. In this validation study a signal strength of 2 is applied. Subsequently, the simulated gene expression data are analysed analogously to normal gene expression analysis.

A large range of FDRs (0-0.8) is scanned in order to evaluate solutions of different sizes, reflecting the fine-tuning of the signal-noise decomposition. The modules are evaluated in terms of recall and precision, see Figure 4.19. Of special interest is the upper right region of the plots, that covers an area with a recall and precision higher than 90%. A large number of solutions, in the FDR-range between 0.1 and 0.4, pass through it. The performance of the optimal approach is contrasted to that of Ideker et al. (2002) implemented in the *jActiveModule* Cytoscape plugin. Since it provides no adjustable scoring function the proposal of Ideker et al. (2002) is followed and their algorithm is recursively applied to the obtained solution several times for five

independent simulations. By this, six discrete solution spaces are obtained for different module sizes visualised as shaded polygons representing their convex hulls in Figure 4.19. The obtained module sizes decrease from large subnetworks with a poor precision to smaller ones. Although, after several recursive iterations the number of false positives reduces substantially and the resultant subnetworks are considerably smaller, they never fall within the region of high precision and recall in the upper right corner. However, these solutions display a large variance especially for the smaller simulated signal modules. An overall similar behaviour is observed for the larger module, but with a smaller variance since the signal in the network is stronger. In contrast to *jActiveModules* the optimal approach is very robust with respect to recall and precision for different module sizes.

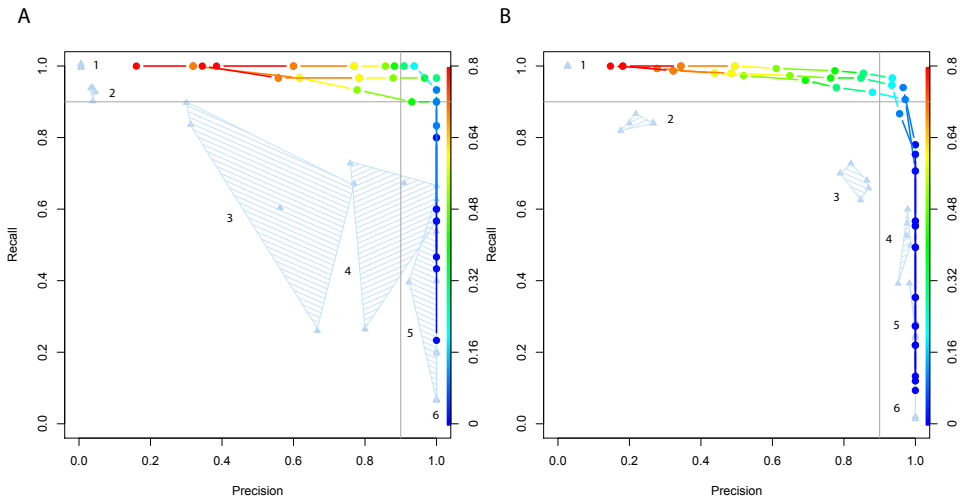


Figure 4.19: Evaluation of simulated datasets. For a wide range of FDRs (colour scheme) and three replications modules are calculated based on simulated microarray data. These are evaluated regarding recall vs. precision and contrasted to the algorithm by Ideker et al. (2002), for which 6 convex hulls (triangles) of solutions (solutions 5 and 6 partially overlap) are displayed for the recursive applications of the algorithm on 5 independent simulated datasets. Two different signal component sizes (30, left panel and 150, right panel) are evaluated with the same procedure. The presented exact approach captures the signal with high precision and recall (over 90% for both) for a large range of FDRs. In contrast, none of the solutions delivered by the heuristic approach falls within the upper right region of high precision and high recall. Data points are jittered for a better visualisation.

Comparisons using Perturbed Molecular Data

To later assess the robustness of module identification approaches to minor changes in the integrated data, different algorithms are analysed on simulated perturbed microarray datasets. The exact algorithms as well as two heuristic approaches are used, a program called Matisse (Module Analysis via Topology of Interactions and Similarity SEts) (Ulitsky and Shamir, 2007) and the module finding plug-in for Cytoscape, jActiveModules (Ideker et al., 2002), to calculate the subnetworks. The comparisons are performed on a 50 node signal module with a signal strengths of 1 between the two conditions in the microarray data. A 50% jackknife is used to generate 20 datasets of perturbed microarray data, as inputs for all three methods. Furthermore, the same PPI network derived from HPRD, consisting of the genes from the microarray, are used for all algorithms. Different module sizes are obtained from the programs by either changing parameter or iteratively applying the method on the resulting subnetwork, see Section 4.2.4. This allows to assess the performance and variability in a receiver operating characteristic analysis in terms of recall and precision and area under the curve calculation.

The resulting recall-precision curves of the 20 resampled datasets for varying modules sizes are depicted in Figure 4.20 for jActiveModules (A), Matisse (B) and `heinz` (C), respectively. Depicted are the alpha-convex hulls of the 20 recall-precision curves. The thickened points in A, B and C show the points used to fit a lowess regression, which is extended through the points (0,1) and (1,0). D depicts the lowess curves for the three above mentioned methods and shows the differences in the methods more clearly. In E the areas under the curves of the three methods are shown for the 20 resamples. It is apparent that the optimal solution achieves the highest AUC values as well as the smallest variance in AUCs, as a measure of variance of the obtained solutions. The difference in variance between jActiveModules and `heinz` is significant, when testing for differences in variances. A p-value of 0.02296 in the Brown-Forsythe version of the Levene-type test for equal variances.

Comparisons using Perturbed Network

The perturbation of the PPI network is investigated for two heuristic methods for module-identification: jActiveModules and Matisse. And the exact method `heinz` as well as the heuristic fastHeinz. For all methods the same simulated unperturbed microarray datasets are used as source for the integration. Simulated networks are perturbed by randomly adding, deleting and rewiring 0%, 10%, 25% and 50% of all edges. The perturbations are applied 5 times to obtain 5 independent solutions for each perturbation level.

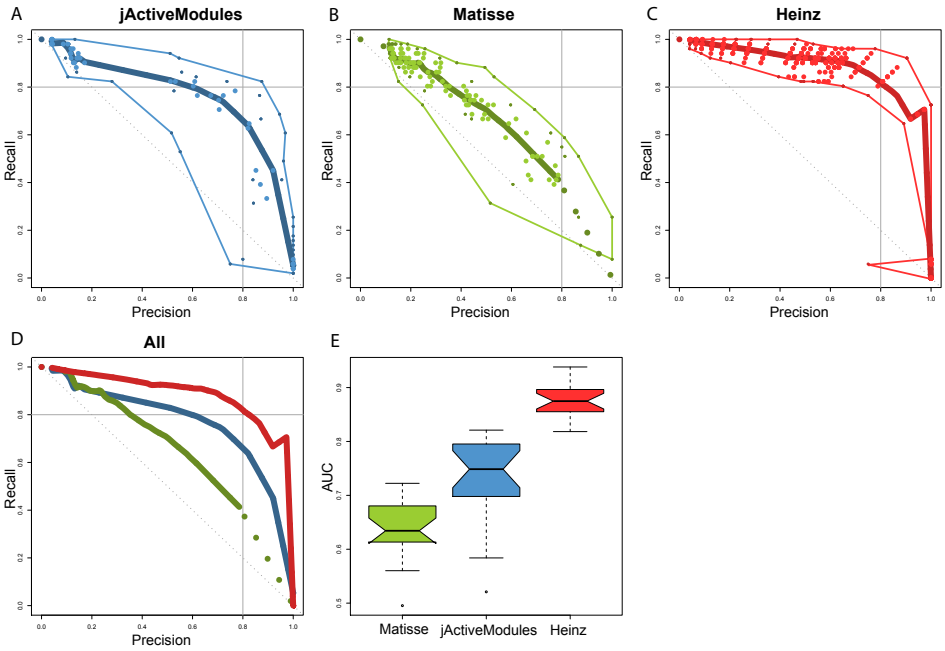


Figure 4.20: Comparative analysis of module detection methods. On 20 simulated resamples of microarray data, with an simulated module size of 50 and a differences in mean of 1, three methods are run to detect functional modules, e.g. jActiveModules (A), Matisse (B) and `heinz` (C). The accuracy is analysed in precision-recall plots by varying the size of the resulting module over parameter settings or iterative runs. Depicted are the alpha-convex hulls of the 20 recall-precision curves. The thickened points in (A), (B) and (C) show the points used to fit a lowess regression, which is extended through the points (0,1) and (1,0). (D) depicts the lowess curves for all above mentioned methods. In (E) the areas under the curves of the three methods are shown for the 20 resamples. The optimal solution achieves the highest AUC values as well as the smallest variance in AUCs.

The effect of deleting, adding and rewiring random edges from the network is shown in Figure 4.20, by plotting precision vs. recall of the obtained functional modules. The solutions of 0% perturbation resemble the modules for the unperturbed networks. The higher the recall and precision values, the better the identified modules. With varying degree of perturbation, from 10% to 50%, the performance of the methods decreases. Here, `heinz` (as well as its fast heuristic implementation) is most stable in performance when changing the underlying network, even with 50% of the edges removed. The other methods are more sensible to destruction of the network and show a decreasing performance upon deletion and rewiring of edges.

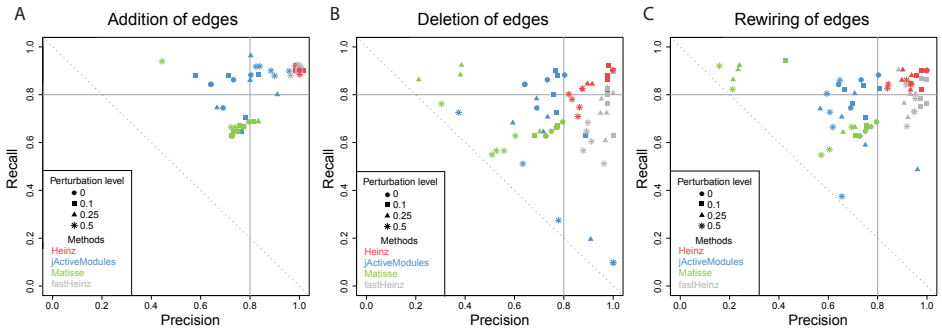


Figure 4.21: Simulated networks are perturbed by randomly adding, deleting and rewiring 0%, 10%, 25% and 50% of all edges. (A), (B), (C) show recall-precision plots of the calculated modules after applying the different perturbations (symbols) on the network for the methods: `heinz` (red), jActiveModules (blue), Matisse (green) and fastHeinz (gray). The perturbations are applied 5 times to obtain 5 independent solutions for each perturbation level.

4.2.6 Robustness

Comparisons using Perturbed Molecular Data

The objective of this study is not only to find a module which obtains a good accuracy but also yields results that are robust to minor changes in the underlying data. The robustness and variance of obtained solutions from Section 4.2.5 of the simulation study are assessed in the following. The 20 best solutions of each algorithm in terms of recall and precision are chosen to analyse how they perform regarding robustness and variance (Figure 4.22 A, B, C). Figure 4.22 D depicts a histogram of how often a node is found in a module. *Unstable* indicates, a node is found only in one or two modules. *Stable* denotes, the node is found in 19 or 20 modules. These are the most interesting cases. Methods with many *stable* nodes and few *unstable* nodes have robust solutions with a low variance, whereas methods with opposite characteristics are non-robust and give very different solutions for each re-sampled dataset. `heinz` is the most robust method with a high frequency if nodes appearing in all solutions. Since the simulated module contains 50 nodes, around 50 nodes should be included in all solutions (stable). `heinz` comes close to this number, but the other methods obviously contain only parts of the simulated module and aggregate lots of nodes without signal.
Another measure of variance is the Jaccard coefficient. The pairwise Jaccard coefficients of the 20 best solutions result in 190 comparisons for each method (Figure 4.22 E). These comparisons show the similarity between the

resulting modules, whereas overall large Jaccard coefficients illustrate a high similarity between all 20 modules and a low variance in the different nodes of the modules. The differences of mean between the Jaccard coefficients of `heinz` and Matisse and `heinz` and jActiveModules are significant with a p-value of $5.016254e-64$ and $2.836045e-64$ respectively (two sample Wilcoxon test).

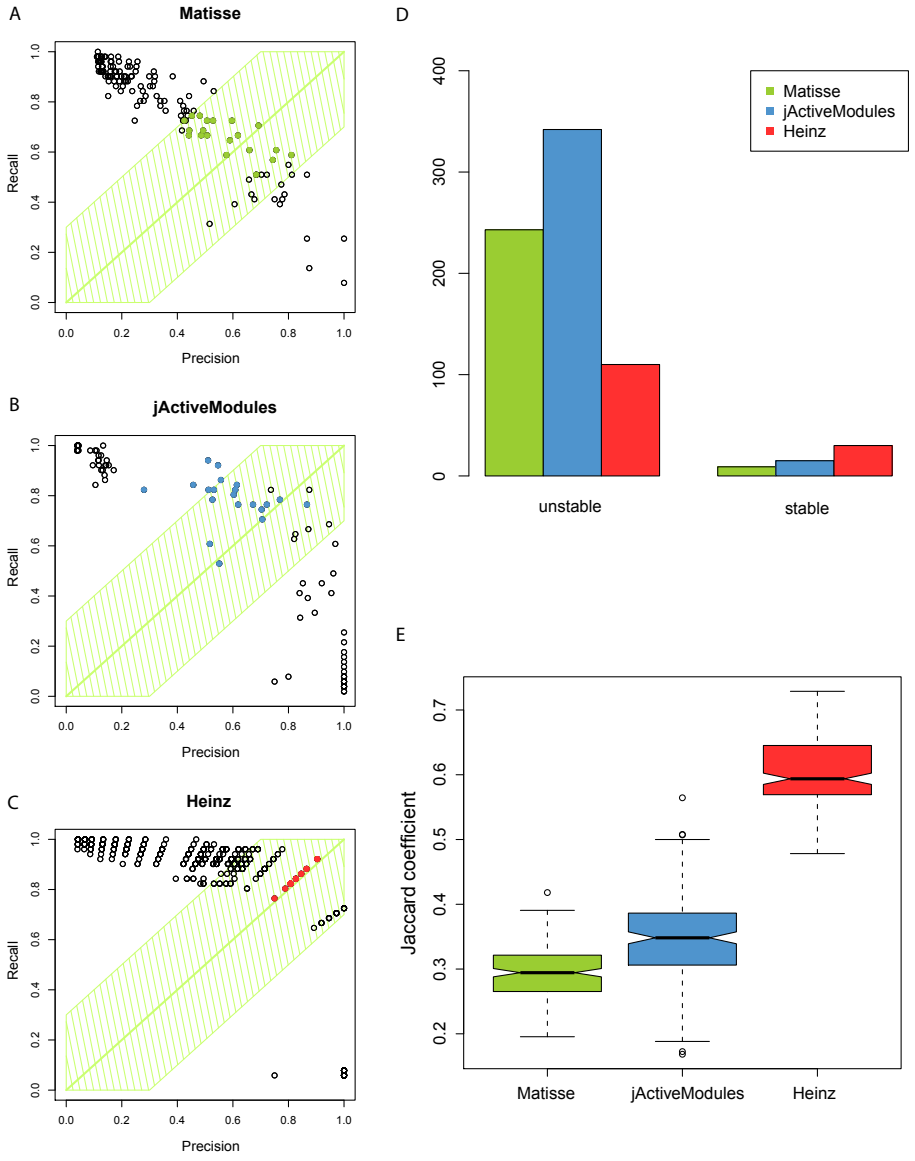


Figure 4.22: Comparative analysis of best solutions. (A), (B) and (C) depict the best solutions regarding recall and precision for each of the 20 simulated resamples as shown in Figure 4.20. These solutions are compared in (D) and (E). (D) depicts a histogram of how often a node is found in a module. *Unstable* indicates, a node is found only in one or two modules. *Stable* denotes, the node is found in 19 or 20 modules. Methods with many *stable* nodes and few *unstable* nodes have robust solutions with a low variance, whereas methods with fewer *stable* nodes and many *unstable* nodes are non-robust and give very different solutions for each resample. In (E) the pairwise Jaccard coefficients of the 20 best solutions are shown, resulting in 190 comparisons for each method.

Comparisons using Perturbed Network

The results of the perturbation of the PPI network are investigated regarding the consistency of the solutions. By calculating the pairwise Jaccard coefficient of 5 independent solutions for each level of perturbation and each method (jActiveModules, Matisse, `heinz` and fastHeinz) the similarity of the identified functional modules is analysed (Figure 4.23). `heinz` (and its heuristic implementation) is very robust in identifying the same solutions and only drops to a Jaccard coefficient around 0.6 when 50% of the edges are deleted. For this scenario the other methods perform considerably worse, reaching Jaccard coefficients of around 0.1 (jActiveModules) and around 0.3 (Matisse).

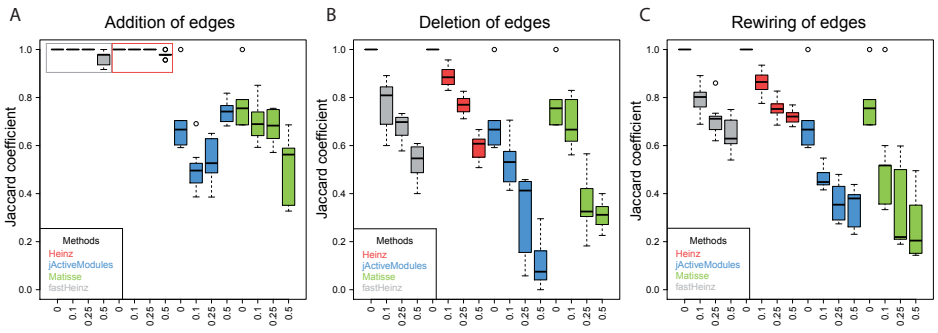


Figure 4.23: Simulated networks are perturbed by randomly adding (A), deleting (B) and rewiring (C) 0%, 10%, 25% and 50% of all edges. The perturbations are applied 5 times to obtain 5 independent solutions for each perturbation level. The similarity and variance of the obtained modules for each method are compared by calculating the all vs. all Jaccard coefficients. The different methods are: `heinz` (red), jActiceModules (blue), Matisse (green) and fastHeinz (gray).

4.2.7 Discussion

The previous section introduced a new statistical method to evaluate modules identified in biological networks. There are many module identification techniques proposed in network analysis, however only rarely these methods are assessed in terms of accuracy and robustness. Obviously, a good module identification algorithm for integrated network analysis should have a high accuracy and show little variation. An estimator which is only highly accurate but not robust, or, vice versa, an estimator which is highly robust, but not accurate may only be of limited practical use.
Regarding the calculation of modules as estimators of the true optimal modules, one is interested in the most efficient unbiased estimator, which has less

variability and is such more likely to make an estimate close to the true optimal solution.

There are three source of variability existent: (i) variability of the (heuristic) module identification method, (ii) variability of integrated data (e.g. gene expression data) and (iii) variability of the underlying network (e.g. PPI network). The performance and robustness of the exact algorithm was tested against other methods in an extensive simulation study, which highlighted the advantage of using an exact approach over heuristic approaches in accuracy and robustness. This holds true for the original integrated data with an PPI network as well as for simulations based on perturbations of the molecular data and perturbations on the network itself.

(i) Several heuristics were described in this study (Matisse, jActiveModules), which showed inherent variation. This variation is artificial, arising from parameter settings and random number generations. The exact `heinz` algorithm allows to evaluate the remaining biological variation. These methodological variabilities were assessed by the ROC analysis. Comparison of the exact approach to the two established heuristics evidenced not only a higher accuracy but also a higher stability of the resulting modules when calculating an exact solution.

(ii) A resampling-based approach was established for the evaluation of robustness of modules obtained from integrated network analysis. By resampling the integrated data, the variability of the modules arising from the molecular data was assessed. While on general the module provides new information about the interplay of significantly differentially expressed genes, it so far only gives the optimal solution on the assumption that the gene expression profiles measure the exact biological changes. Therefore it is highly desirable to have a reliability measure for the presence of an identified node and edge in the solution with respect to small perturbations of the underlying gene expression data.

(iii) Another important source of noise is the high uncertainty of edge information based on the different measurements of protein-protein interactions. The underlying PPI network may certainly contain false positive edges, but even more false negative edges. To mimic this effect for network analysis, not only the underlying gene expression data were perturbation but also the underlying network structure itself. Comparisons of the accuracy and robustness of the resulting modules revealed that some effects (e.g. removal of edges) had a dramatical effect, while including false positive edges did not really hamper the accuracy of the optimal module. This means, that for integrated network analysis an

over conservative edge confidence (e.g. from PPI detection methods or due to rigorous STRING interaction confidence scores (von Mering et al., 2005)) may not advantage, but in contrast, may drastically reduce the overall accuracy.

While jActiveModules showed a good performance in finding the functional module, it lacked in consistency of the identified solutions and showed a relatively large variance in the identified genes when repeatedly run on slightly perturbed datasets. The same holds true for the Matisse algorithm, its solutions varied notably in their composition of nodes as measured by the Jaccard coefficient in a pairwise comparison of the modules. In addition, its accuracy was the lowest of all tested algorithms.

4.3 Assigning Confidence Values to Functional Modules

The original functional module gives new information about the interplay of significantly differentially expressed genes, but it does not provide information about the robustness of the solution or of the individual nodes and edges. As it was shown in the previous section, the resulting modules vary in parts heavily depending on the noise in the molecular data. The objective is therefore, on the one hand to produce an accurate solution and on the other hand to obtain a robust module in which robust parts are indicated by support values. For this purpose a novel concept of a consensus module is proposed here, based on jackknife resampling.
In phylogeny, Felsenstein (1985) introduced the concept of bootstrapping to define a confidence measure for phylogenetic trees using bootstrap resampling. Analogously, resampling procedures can be used to assess the robustness of functional modules in integrated network analysis. In this study jackknife resampling of the integrated data is used in order to construct a consensus module and to assign confidence values to individual nodes and edges of the module.

4.3.1 Consensus Modules from Consensus Scores

Necessary Improvements of the Heinz Algorithm

The original implementation of the `heinz` algorithm is an integer linear programming-based approach to solve a Steiner tree problem. For the incorporation of edge weights and computation of consensus modules of a fixed size the following changes are made. The formulation of the maximisation problem is changed to incorporate edge scores to: For an undirected, vertex- and edge-weighted graph $G = (V, E, w)$ with weights $w : V \cup E \to \mathbb{R}$ find a subtree $T = (V_t, E_T)$ of $G, V_T \in T, E_T \in E$ that maximises $score(T) = \sum_{v \in V_T} w(v) + \sum_{e \in E_T} w(v)$. Furthermore, the constraint $\sum_{v \in V} x_v$ is added to the ILP formulation to constrain the search space to contain only modules of size k.

The Consensus Algorithm

Here a jackknife procedure is proposed to assess the robustness and variability of the network modules as depicted in Figure 4.24.

Two approaches exist to assign confidence values to modules. Either support values are calculated from the resampling procedure and appointed to the

optimal module, or the support values are used to derive a new score for the network and calculate a novel module, the consensus module. Briefly, the algorithm for the calculation of the consensus modules consists of the following steps:

1. Resampling of microarray data using jackknife for J pseudo-replicates (Figure 4.24 first line)
2. Scoring the nodes of the network and calculating the maximum-scoring subnetwork of the size of the original module for each replicate (Figure 4.24 middle part)
3. Calculating the frequency of nodes and edges in the resulting jackknifed modules
4. Re-scoring the network with a consensus score derived from the frequency of edges and nodes (Figure 4.24 lower part)
5. Calculating the maximum-scoring subnetwork of the size equal to the size of the original module, this constitutes the consensus module

In more detail, the algorithm starts with the generation of J jackknife samples of the expression data. The jackknife pseudo-replicates are treated in the same manner as the original data and a standard t-test can be performed to obtain gene-wise p-values. Under the null hypothesis the p-values are uniformly distributed, when containing a signal, the p-values are a mixture of a beta-uniform distribution as expected from the original data samples. For each of these jackknife replicates a node score for all genes in the network is calculated as described in Section 4.1.3 and the optimal module is calculated, resulting in a set of slightly differing modules due to the resampling of the expression data. The frequency of each gene in the resulting J modules is used to define a consensus score for each node

$$S_J(v, \rho) = \sum_{i=1}^{J} I(v \in V_i) - \rho \ ,$$

and each edge

$$S_J(e, \rho) = \sum_{i=1}^{J} I(e \in E_i) - \rho \ ,$$

with V_i being the vertex set and E_i being the edge set of the resultant module for each of the J jackknife sets and a given cut-off $\rho \in [0, J]$. The original network is subsequently re-scored with the consensus score for the nodes and edges and the maximum-scoring subnetwork is calculated with the size set to

the size of the original module. This resultant optimal scoring subnetwork is then defined as the consensus module.

The frequencies of the nodes and edges are used to depict confidence values in the optimal and the consensus module. Confidence scores are given as percentage of the frequency of an edge and node in the jackknifed modules. These scores are depicted in the plot of the modules by the node sizes and edge widths. The more frequently an edge (or node) occurs in any of the perturbed modules, the more likely it is, that this edge (or node) is a robust part of the functional module and should be taken into consideration for further research.

Additionally, according to the central limit theorem the empirical distribution of the score, derived from the jackknife modules, is approximately normal and can be used for the construction of confidence intervals of the score. Suboptimal modules within the confidence interval of the score should be regarded and are analysed in Section 4.3.2.

Implementation Details of the Consensus Algorithm

The consensus algorithms is implemented in the following with $J = 100$ jackknife pseudo-replicates and $\rho = 50$. A delete-half jackknife is used, which means half of the observations are dropped in each resample as recommended by Felsenstein (1985, 2004). These parameter settings are used throughout the study.

4.3.2 Score Distribution of Maximum-Scoring Subnetworks

Minor changes in the gene expression data have an apparent influence on the resulting subnetwork (Section 4.2), therefore the maximum-scoring subnetwork might be not the best solution from a biological point of view. The resulting question is, up to which score should suboptimal modules also be regarded. To identify a confidence interval of similarly good modules, jackknifing is used to approximate the distribution of the scores of the modules. The distribution of the score of the jackknife resamples is approximately normal distributed for the simulated data, which can be seen in Figure 4.25 (inset).

Since the scores of the resampled modules are biased towards lower values, the distribution is corrected by shifting the medium of the distribution to the score of the optimal module. The score of the optimal and consensus module is indicated as a red line and a dashed red line in Figure 4.25, respectively.

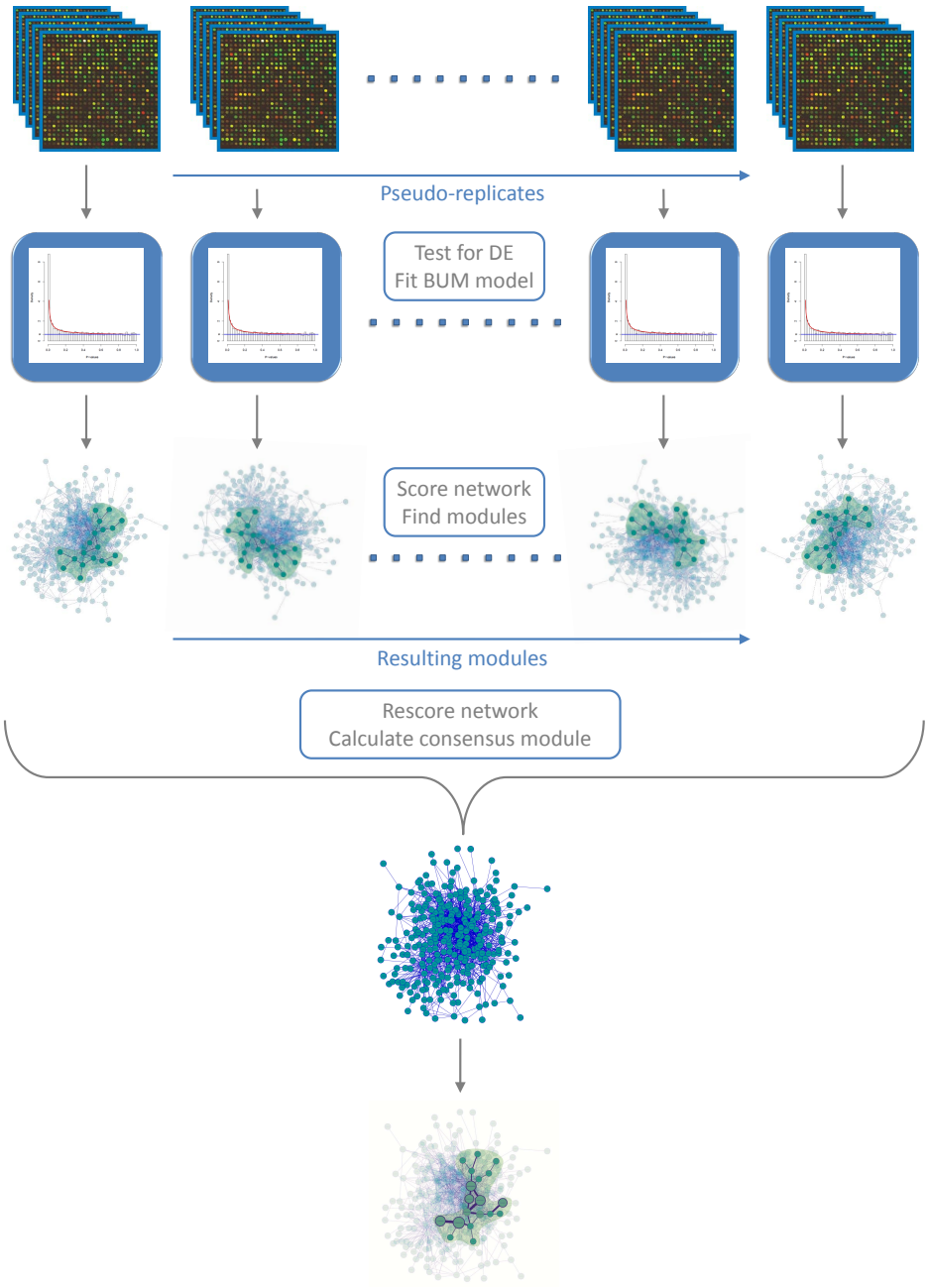


Figure 4.24: Workflow of the resampling procedure for integrated network analysis. First, the microarray data are resampled, i.e. for each gene the expression values are resampled using jackknife, yielding J jackknife samples per gene. Differential expression is tested on all jackknife replicates. A BUM model is fitted to the resulting p-value distribution. Based on the estimated parameters of the distribution, scores are calculated for all nodes of the network. For each scored network the maximum-scoring subnetwork is calculated. Each of the resulting modules differ slightly due to the resampling. The resulting set of modules are subsequently used as a basis for deriving consensus node and edge scores. The original network is re-scored and a consensus module is calculated.

100 suboptimal modules are sampled from the solution space of suboptimal modules with a Hamming distance from 1-100. These are depicted as vertical blue-green lines in Figure 4.25. The first and second standard deviation of the distribution are given by the green shaded background, with a thicker shading of the first standard deviation. Within the region of two standard deviations reside 14 unique suboptimal modules for the simulated data. These suboptimal solutions represent a subset of possible equally biologically meaningful suboptimal solutions within the range of a predefined confidence interval of two standard deviations.

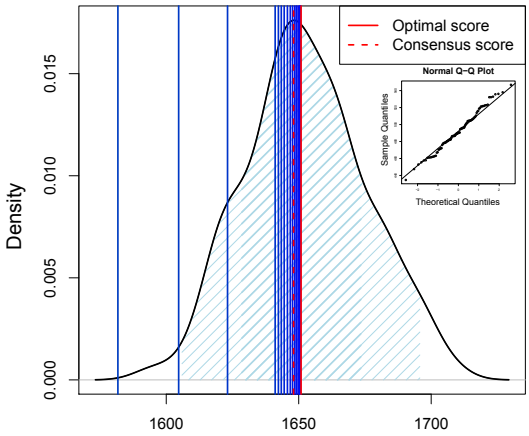


Figure 4.25: Score distribution derived from jackknife resampling. The distribution of the subnetwork score is derived from the jackknife resampling. The resampled scores are approximately normal distributed and the median is shifted to the optimal score. It shows the distribution for a simulated network. Coloured in red is the original score and dashed the score of the consensus module. 100 suboptimal solutions are calculated, which are depicted as vertical blue-green lines. Shaded in green is the fist and second standard deviation. Since minor changes in the gene expression data have an influence on the resulting network, all solutions obtaining a score within two standard deviations are considered as similarly good solutions. 14 unique suboptimal modules are within this region for the simulated data.

4.3.3 Discussion

In the last section a novel computational technique was developed based on jackknife resampling to assess the robustness of a functional module and assign confidence values to the contained nodes and edges in the module. The method is aimed at providing a robust subnetwork which yields a high

accuracy and robustness in terms of a consensus module. In particular, on the biological side highly robust submodules within the solutions are likely to have a common biological function, which may provide deeper inside into the substructure of the module.

The consensus module shows two advantages to the original module:

(i) The robustness of the module is increased by simultaneously keeping the accuracy of the original module. The robust parts are indicated by support values from the jackknife resampling and allow an identification of the most reliable parts of the module.

(ii) The consensus approach is based on a novel scoring of the network and therefore enables the incorporation of new nodes into the module, not identified before in the original module. The consensus module itself can be considered as an overall jackknife estimate of the functional module.

The jackknife procedure was additionally used to construct a score distribution and confidence intervals of the score. The median of the distribution was defined to be the optimal score. Around the optimal score the confidence interval specifies a range of scores within which a good solution is estimated to lie. Modules that reach scores inside this range should also be regarded as valid solutions.
Using the score distribution one could enumerate all solutions from the optimal score to a lower bound of the confidence interval and analyse the differences in the solutions. The problem with this approach is the number of possible modules within the confidence interval. It is not feasible to calculate these within an appropriate execution time and with valid memory usage. An alternative is to sample the best suboptimal modules within the confidence interval with a predefined Hamming distance from the optimal module. This solution was chosen here.

This presented study describes a novel and important method to identify and evaluate functional modules and their robust components. The approach to consider both, the accuracy of an module-detection algorithm as well as a low variability of the obtained solutions, is a substantial advantage over existing methods. Additionally, the confidence interval of the score provides a lower limit for the consideration of alternative suboptimal modules.

4.4 Application to Gene Expression Profiles

4.4.1 Acute Lymphoblastic Leukaemia

Introduction

Acute lymphocytic leukaemia is an acute leukaemia arising from degenerated malignant lymphoblasts. ALL occurrence has a bimodal distribution, with a childhood form between the age of 2-5 and an adult form around 50. ALL cells are known to be derived from either B- or T-cell precursors. In B-lineage ALL, malignant cells often have additional specific genetic abnormalities, which have a significant impact on the clinical course of the disease (Chiaretti et al., 2004). Several cytogenetically defined prognostic subgroups were identified some of which have a worse prognosis than others, e.g. a translocation between chromosomes 9 and 22, known as the Philadelphia chromosome. The defect in the Philadelphia chromosome is a translocation between the long arms of chromosomes 9 and 22, which results in a shortened chromosome 22 (Faderl et al., 1998). The result of the translocation is the oncogenic BCR/ABL gene fusion, located on chromosome 22. Depending on the location of the fusion the molecular weight of the translated protein can range from 185 to 230 kD. In comparison to the normal c-ABL protein (p145) the fusion protein has a significantly increased tyrosine phosphokinase activity. The differences in gene expression between patients with BCR/ABL gene fusion and without translocation are investigated in the following.

Integrated Data

A functional network is calculated by combining the gene expression data from two subgroups of acute lymphoblastic leukaemia (Chiaretti et al., 2004) with a comprehensive interactome network derived from the Human Protein Reference Database. In particular the differential expression between the B-cells with and those without a translocation between chromosomes 9 and 22, resulting in the BCR/ABL gene fusion, is investigated.

The expression data are available as an rma normalised expressionSet object in the R package ALL (Smyth, 2004). The significance of differential expression between the two subtypes is assessed using robust statistics based on linear models and a moderated t-test (limma). This yields an uncorrected p-value for differential expression for each gene. The p-values constitute a quantitative measurement describing the significance of differential expression for each gene.

For the network data a dataset of literature-curated human protein-protein interactions, which was obtained from HPRD, is used. The entire interactome network assembled from these data consists of 36,504 interactions between 9,386 different proteins. From this a *ALL*-specific interactome network is derived as a vertex-induced subgraph extracted by the subset of genes for which expression data exist on the microarray. Since the module should be a connected subgraphs, the largest connected component of the network is used. The resulting network comprises 2,034 different gene products and 7,756 interactions.

Optimal Module

Applying the exact method to the ALL network the optimal-scoring subnetwork, shown in Figure 4.26 A, is obtained, for scores using a restrictive FDR of 0.02. The resultant module comprises 33 nodes of which 25 are positive and 8 possess a negative score. The overall subnetwork score sums up to 38.43, 60.03 from the positive nodes and -21.60 from the negative nodes. The module is enriched in genes responsible for peptidyl-tyrosine phosphorylation such as ABL1, CAV1, FYN, ITGA5, EGFR, PRKCA and TNK2, as a result from the chromosomal translocation. Several other genes are involved in actin reorganisation and cell adhesion (MARCKS, ACTN1, TES, EGFR, LPP, TNN, ABL1...). Three tyrosine kinase genes (ABL, FYN and YES1) were also identified as differentially expressed between the subgroups by Chiaretti et al. (2004).
Juric et al. (2007) calculated a top-scoring network of interactions among the differentially expressed genes in BCR/ABL-positive versus negative ALL. Their subnetwork contains 4 genes (FYN, LCP2, ITGA5 and CRADD), which are also part of the functional module, calculated with the `heinz` algorithm. Here, FYN (a non-receptor tyrosine kinase) is an important hub in the module and ITGA5 is part of cell adhesion and invasion processes and angiogenesis.

Consensus Module

Using an FDR of 0.02 a consensus module from 100 jackknife resamples of the ALL microarray data is calculated (Figure 4.26 C). Support values are determined for the nodes and edges contained in the module, i.e, edges with high jackknife support values represent interactions between genes/proteins that appear often together in resampled subnetworks. The relatively weak signal of this dataset is reflected by lower jackknife support values and the number of genes differing between the optimal and the consensus solution. This was already mentioned by Chiaretti et al. (2004), who found out that

the negative ALL samples had a more similar gene expression signature to BCR/ABL than to all other rearrangements and the classification between them turned out to be difficult due to the weak signal. Despite this few differences between the original and the consensus module can be seen. 15% of the genes from the original module are obviously an unrobust signal and appear too infrequently in the jackknifed modules to be contained in the consensus. These are replaced in the consensus by genes that are picked up more often in the jackknifed modules (TUBA1, CNN3, FHL1, FHL2 and ITGB5). Three of them (TUBA1, CNN3, FHL1) have previously been linked to ALL and even have been used for classification. Ross et al. (2003) classified paediatric acute lymphocytic leukaemia using a a gene expression subset to discriminate genetic subtypes. The gene set to distinguish BCR/ABL includes among others CASP10, TUBA1, CNN3, ABL1 and FYN. Of these TUBA1 and CNN3 are solely identified in the consensus and are part of the robust component discussed in the next section. In addition, it is known that TUBA1 is a target of vincristine and vinblastin in ALL treatment. They modify alpha-tubulin and inhibit growing microtubules (Verrills et al., 2003). Furthermore, it was pointed out previously that FHL1 is also aberrantly expressed in a significant proportion of ALL cell lines (Rice et al., 2008; Yeoh et al., 2002). These examples underline that the consensus module contains additional important genes, which are involved in the analysed process.

GO Enrichment of the Robust Component

GO term enrichment analysis is performed on the nodes of the original module, the consensus module and robustly connected components with jackknife support of the edges and nodes greater than 25% for both modules. The GO enrichment of the robust components results in more significant p-values than the GO enrichment of the overall modules, also yielding slightly different categories. Identified biological processes include e.g. peptidytyrosine phosphorylation (CAV1, ITGA5, EGFR, PRKCA and TNK2), cellular component movement and positive regulation of cyclin-dependent protein kinase activity during G1/S. Theses are known processes resulting from the presence of the BCR/ABL fusion transcript which results in a higher tyrosine kinase activity and increased transformation potential to a cancer cell with enhanced cell adhesion, invasion and angiogenesis. See Table 4.2 for the GO enrichment of the robust component of the consensus module.

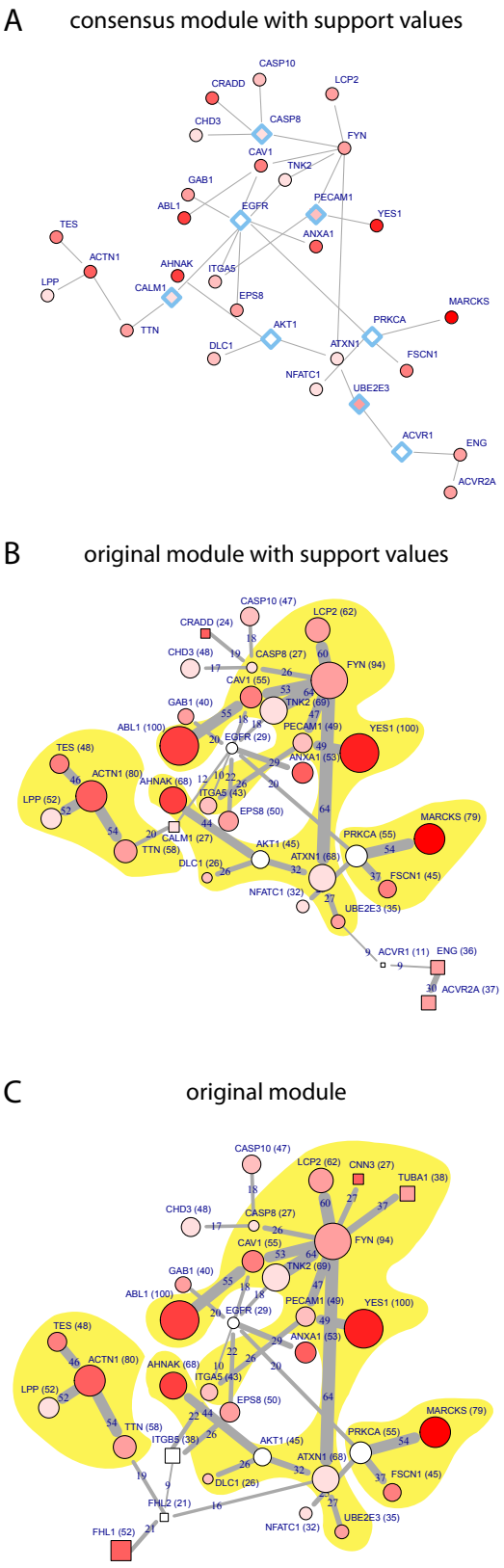


Figure 4.26: Resulting modules for ALL. The thickened blue diamonds displayed in (A) emphasise nodes with a negative score in the original network, while circular nodes depict positive scores. Nodes only present in either the original module or the corresponding consensus module in (B) and (C) are depicted by squared node symbols. Colouring of the nodes represents differential expression of the genes (red: up-regulated in samples with the BCR/ABL chromosomal translocation, green: down-regulated). For the ALL dataset the optimal solution (A, B) is calculated with an FDR of 0.02 and for the consensus module (C) a threshold $\rho = 50$ is used. Node and edge jackknife support values are indicated by the sizes of the nodes and width of the edges and edge labels. In (B) the support values are highlighted in the original module, in (C) support values are highlighted in the consensus module which is calculated using a consensus score based on the support for nodes and edges. The yellow area highlights the network areas with at least 25% support values.

Score Distribution

To identify additional biologically meaningful modules suboptimal modules are calculated with the `heinz` algorithm. All suboptimal solution within a confidence interval from the optimal score are taken into account. Jackknifing is used to approximate the distribution of the scores of the modules, to estimate confidence intervals. Since the scores of the resampled modules is biased towards lower values, the distribution is corrected by shifting the medium of the distribution to the score of the optimal module. The score of the optimal and consensus module is indicated as a red line and a dashed red line in Figure 4.27. The vertical cyan lines in Figure 4.27 depict the scores of 10 suboptimal modules. These are sampled from the solution space of suboptimal modules with a Hamming distance from 1-100 to the maximum-scoring subnetwork. The first and second standard deviation of the distribution are given by the blue shaded background. Within the region of two standard deviations reside 24 unique suboptimal modules for the ALL data. They represent a subset of possible suboptimal solutions within a confidence interval from the optimal score.
A histogram of the nodes contained in these 24 suboptimal modules shows that nodes which appear frequently (at least 20 times), correspond to the most robust nodes in the consensus module (Figure 4.28). All of the nodes have a support value between 55 and 100 in the consensus module.

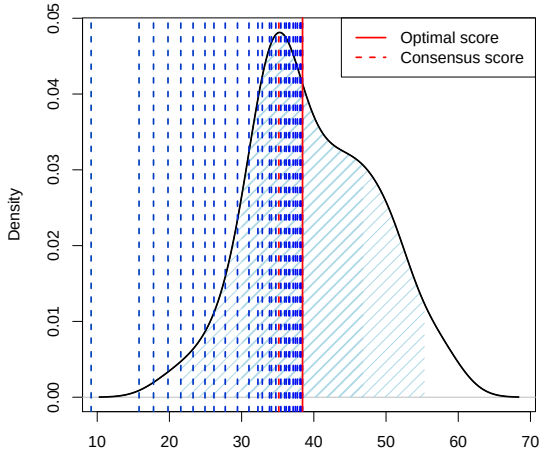


Figure 4.27: Score distribution derived from jackknife resampling for the ALL module. The resampled scores are approximately normal distributed and the median is shifted to the optimal score. Coloured in red is the original score and dashed the score of the consensus module. 100 suboptimal solutions are calculated, which are depicted as vertical blue-green lines. Shaded in green is the fist and second standard deviation. 24 unique suboptimal solutions reside within two standard deviations.

Table 4.2: GO enrichment for the robust part of the consensus module of the ALL dataset. It comprises nodes and edges of the consensus module with support values of at least 25% (Figure 4.26 C). All GO term with a p-value < 0.001 are shown.

GO ID	P-value	GO Term	Genes
GO:0018212	1.32e-07	peptidyl-tyrosine modification	ABL1, CAV1, FYN, ITGA5, EGFR, PRKCA, TNK2
GO:0050730	6.37e-06	regulation of peptidyl-tyrosine phosphorylation	CAV1, ITGA5, EGFR, PRKCA, TNK2
GO:0045937	7.99e-06	positive regulation of phosphate metabolic process	CAV1, ITGA5, DLC1, EGFR, PRKCA, TNK2
GO:0051174	3.41e-05	regulation of phosphorus metabolic process	AKT1, ATXN1, CAV1, ITGA5, TTN, DLC1, EGFR, PRKCA, TNK2
GO:0007169	6.23e-05	transmembrane receptor protein tyrosine kinase signalling pathway	AKT1, ATXN1, EPS8, ITGA5, EGFR, PRKCA, LCP2
GO:0043687	7.28e-05	post-translational protein modification	ABL1, AKT1, ANXA1, CAV1, FYN, ITGA5, TTN, DLC1, EGFR, UBE2E3, PRKCA, TNK2, YES1
GO:0031659	1.22e-04	positive regulation of cyclin-dependent protein kinase activity involved in G1/S	AKT1, EGFR
GO:0070141	1.22e-04	response to UV-A	AKT1, EGFR
GO:0030036	1.31e-04	actin cytoskeleton organisation	ABL1, CNN3, EPS8, FSCN1, TTN, DLC1
GO:0009628	1.38e-04	response to abiotic stimulus	AKT1, ATXN1, CASP8, CAV1, FYN, EGFR, PRKCA
GO:0006928	1.39e-04	cellular component movement	ACTN1, AKT1, ANXA1, FYN, ITGA5, PECAM1, DLC1, EGFR, PRKCA

Table 4.2: GO enrichment (cont.)

GO ID	P-value	GO Term	Genes
GO:0045471	1.46e-04	response to ethanol	CASP8, EPS8, FYN, PRKCA
GO:0042325	1.78e-04	regulation of phosphorylation	AKT1, ATXN1, CAV1, ITGA5, TTN, EGFR, PRKCA, TNK2
GO:0045428	1.88e-04	regulation of nitric oxide biosynthetic process	AKT1, CAV1, EGFR
GO:0023033	2.09e-04	signalling pathway	ABL1, AKT1, ANXA1, ATXN1, CASP8, CAV1, EPS8, FYN, ITGA5, TTN, DLC1, EGFR, ITGB5, PRKCA, LCP2, TNK2
GO:0043412	2.10e-04	macromolecule modification	ABL1, AKT1, ANXA1, CAV1, FYN, ITGA5, TTN, DLC1, EGFR, UBE2E3, PRKCA, TNK2, YES1
GO:0046777	2.65e-04	protein amino acid autophosphorylation	AKT1, FYN, EGFR, YES1
GO:0032270	3.92e-04	positive regulation of cellular protein metabolic process	AKT1, CAV1, ITGA5, DLC1, PRKCA, TNK2
GO:0006936	4.90e-04	muscle contraction	CAV1, CNN3, TTN, ITGB5, PRKCA
GO:0016043	4.94e-04	cellular component organisation	ABL1, ACTN1, AKT1, ANXA1, CASP8, CAV1, CNN3, EPS8, FSCN1, ITGA5, PECAM1, TTN, DLC1, EGFR, PRKCA, TUBA1
GO:0007611	5.11e-04	learning or memory	ATXN1, FYN, ITGA5, PRKCA
GO:0015749	5.18e-04	monosaccharide transport	AKT1, PRKCA, YES1

Table 4.2: GO enrichment (cont.)

GO ID	P-value	GO Term	Genes
GO:0015758	5.18e-04	glucose transport	AKT1, PRKCA, YES1
GO:0046209	5.18e-04	nitric oxide metabolic process	AKT1, CAV1, EGFR
GO:0051341	5.18e-04	regulation of oxidoreductase activity	ABL1, AKT1, EGFR
GO:0051246	5.39e-04	regulation of protein metabolic process	AKT1, CAV1, ITGA5, DLC1, EGFR, UBE2E3, PRKCA, TNK2
GO:0007044	5.64e-04	cell-substrate junction assembly	ACTN1, ITGA5, DLC1
GO:0016477	6.66e-04	cell migration	AKT1, FYN, ITGA5, PECAM1, DLC1, EGFR, PRKCA
GO:0018105	7.15e-04	peptidyl-serine phosphorylation	AKT1, CAV1, PRKCA
GO:0010035	7.21e-04	response to inorganic substance	CASP8, CAV1, TTN, EGFR, PRKCA
GO:0003008	7.37e-04	system process	ATXN1, CAV1, CNN3, FYN, ITGA5, TTN, EGFR, ITGB5, PRKCA, YES1
GO:0051674	7.65e-04	localization of cell	AKT1, FYN, ITGA5, PECAM1, DLC1, EGFR, PRKCA
GO:0007155	8.18e-04	cell adhesion	ABL1, ACTN1, ITGA5, PECAM1, LPP, DLC1, EGFR, ITGB5
GO:0001934	8.19e-04	positive regulation of protein amino acid phosphorylation	CAV1, ITGA5, PRKCA, TNK2

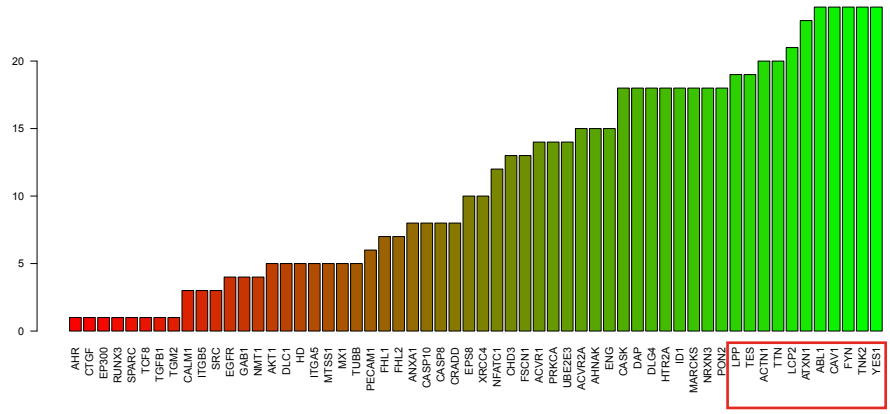


Figure 4.28: Histogram of nodes contained in the 24 unique suboptimal solutions within two standard deviations from the optimal score. Nodes which appear at least 20 times (marked in red), i.e. in at least 20 of the considered suboptimal modules, correspond to the most robust nodes in the consensus module with a support value of at least 52.

4.4.2 Diffuse Large B-cell Lymphoma

Introduction

Diffuse large B-cell lymphoma is the most common histologic subtype of non-Hodgkin lymphoma (NHL) accounting for approximately 40 percent of adult NHL cases (Alizadeh et al., 2000). DLBCL is an aggressive malignancy of mature B lymphocytes, which can be subdivided into three molecular subtypes arising from distinguished B-cells: germinal center B-cell (GCB) DLBCL, activated B-cell (ABC) DLBCL (Alizadeh et al., 2000) and primary mediastinal B-cell lymphoma (PMBL) (Savage et al., 2003). The distinct subtypes cause significantly different survival rates following chemotherapy in the patients, where patients with GCB DLBCL have the best prognosis. In the following the datasets are introduced that will be analysed in the remainder of the chapter. This will exemplify how an integrated network analysis can be performed on multiple sources of integrated data, microarray and clinical survival data. A module is obtained by combining the gene expression data from two different lymphoma subtypes ABC and GCB DLBCL with survival data and an interactome network derived from HPRD.

Integrated Data

The study presented here is based on microarray data from diffuse large B-cell lymphomas (Rosenwald et al., 2002). This comprises gene expression data from 112 tumours with the germinal center B-like phenotype and from 82 tumours with the activated B-like phenotype. These tumour subtypes differ in their malignancy as well as in the treatment options for the patients. Expression profiling has been performed on the *Lymphochip* including 12,196 cDNA probe sets corresponding to 3,583 genes (Rosenwald et al., 2002). In addition, survival information from 190 patients is available (Rosenwald et al., 2002).

As a first step two question arise:

1. Which genes are differentially expressed between the two tumour subtypes?
2. Which genes are associated with the risk of relapse?

After normalisation, the significance of differential expression between the two subtypes ABC and GCB can be assessed by using robust statistics based on linear models and a moderated t-test (limma). This yields an uncorrected p-value for differential expression for each gene. These p-values constitute a quantitative measurement describing the significance of differential expression for each gene.

To assess the risk association of each gene a survival analysis is subsequently performed by fitting a univariate Cox model to the expression data of each gene. From the likelihood ratio test of the regression coefficient, p-values are obtained for each gene denoting the association with survival, independent of the assigned tumour subtype. Thus two p-values are obtained for each gene from both analyses, corresponding to differential expression on the one hand and to risk association on the other hand. In the next step, these two p-values for each gene are combined into one p-value from which the score is derived.

For the network data a set of literature-curated human protein-protein interactions is used that was obtained from the Human Protein Reference Database. The entire interactome network assembled from these data consists of 36,504 interactions between 9,392 different proteins. From this a *Lymphochip*-specific interactome network is derived as the vertex-induced subgraph extracted by the subset of genes for which expression data are available on the *Lymphochip*. The resulting network comprises 2,561 different gene products and 8,538 interactions, with a large connected component of 2,034 proteins (79.4%) and 8,399 interactions (98.4%). The remaining proteins are either non-interacting single nodes in the network (472) or form tiny clusters of a handful of nodes (23). The study focuses on the largest connected component of this network.

Optimal Module

Applying the above described searching procedure to the lymphoma network the optimal-scoring subnetwork is obtained as shown in Figure 4.29 for the combined score using a restrictive FDR of 0.001. The resultant module comprises 46 nodes of which 37 are positive and 9 possess a negative score. The overall subnetwork score sums up to 70.2, 102.9 from the positive and -32.8 from the negative nodes.

Further, the optimal solution contains and extends interactome modules that have been identified previously and described to play major biological roles in the ABC and GCB DLBCL subtypes. The resulting optimal module connects and expands the proliferation module, which is more highly expressed in the ABC subtype (Rosenwald et al., 2002). It includes the genes MYC, CCNE1, CDC2, APEX1, DNTTIP2, and PCNA (highlighted in blue) and parts of the oncogenic NFκB pathway (highlighted in purple) containing the genes IRF4, TRAF2, and BCL2.

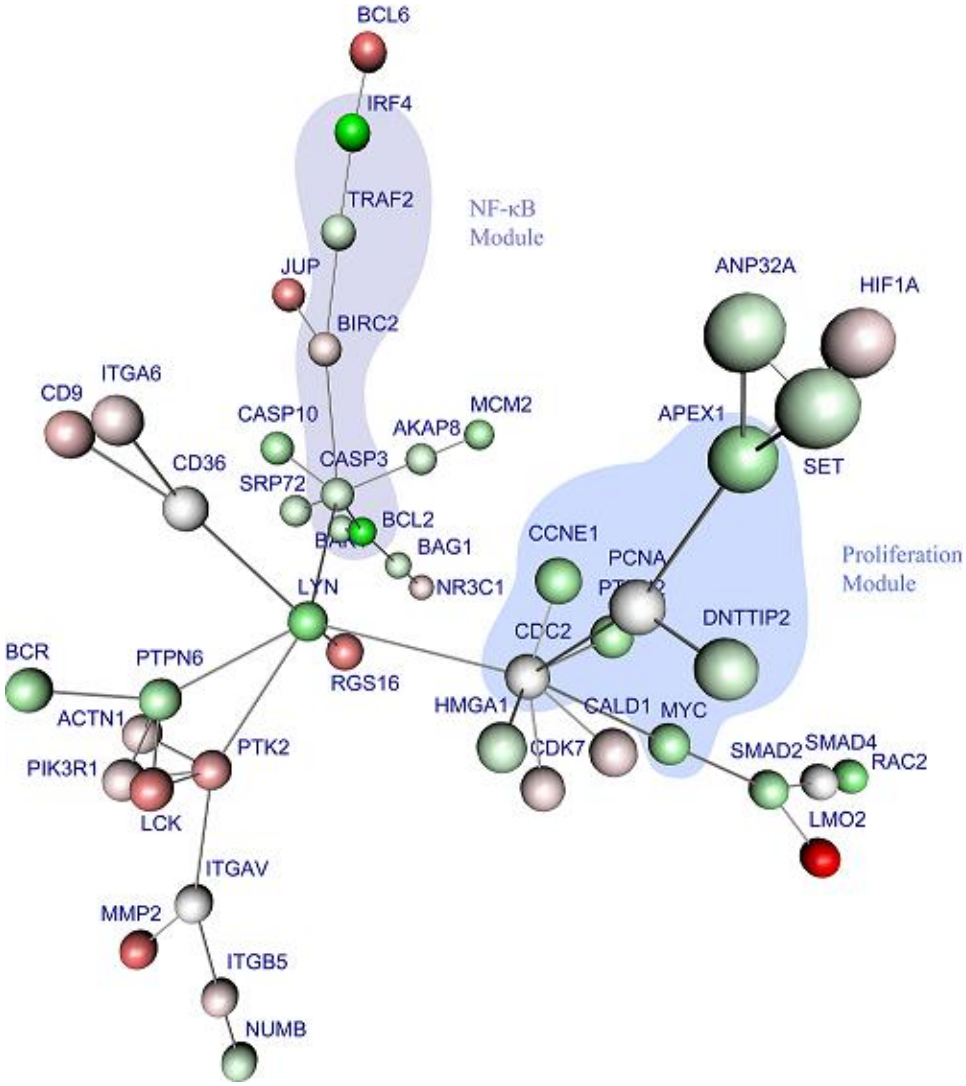


Figure 4.29: 3D visualisation of the optimal subnetwork of DLBCL. It was identified by using a score based on the p-values of a gene-wise two sided t-test, an univariate Cox-regression hazard model and an FDR of 0.001. An over-expression of the proliferation module (MYC, CCNE1, CDC2, APEX1, DNTTIP2, and PCNA) and the NFκB signature (IRF4, TRAF2, and BCL2) can be observed. Proteins are denoted by their Entrez Gene names. The colours correspond to the up- or down-regulation of the genes (green=down-regulated in GCB, red=up-regulated in GCB).

Consensus Module

To identify a robust module the presented jackknife procedure is applied to the DLBCL dataset. In particular, the different solutions obtained from the optimal module and the jackknife consensus module are compared. The optimal module is calculated with an FDR of $1e^{-07}$ (Figure 4.30 A) and the consensus module from 100 jackknife resamples of the DLBCL microarray data. Support values are determined for the nodes and edges contained in the module (node size and edge width in Figure 4.30 B and C), i.e, edges with high jackknife support values represent interactions between genes/proteins that appear often together in resampled subnetworks.

The nodes and interactions of the DLBCL dataset are much more robust and give a more distinct signal than the ALL data, which is indicated by the high support values. Therefore, when using the edges and nodes with at least 50% support values, almost the whole consensus module is selected as robustly connected components (excluded: PTPN2, STAT3 and PRKCQ). 19% of the genes which showed up in the original module, are obviously an unrobust signal and appeared too few times in the jackknifed modules to be contained in the consensus, e.g. TGFBR2, CD44, LCK, PTPN1 and IRS1. Other genes do not turn up in the original optimal module, but only in the consensus, because they are included frequently in the jackknifed modules. This geneset contains e.g. HDAC7A, TRAF2, DCTD, SMAD4 and RAC2. For all of these genes associations to DLBCL or to proliferation and survival pathways are known. These pathways are expected to be differentially regulated between the tumour subgroups, because of the higher malignancy and lower survival rates of ABC than GCB. HDAC7 induces the expression of Nur77 and Nor1, both promoting growth and survival in cancer cell and a high expression of Nor1 shows favourable responses to chemotherapy in DLBCL (Chen et al., 2009). The expression of TRAF2 correlates with poor progression free survival time in ABC-like DLBCL (van Galen et al., 2008). The dCMP deaminase (DCTD) activity is a measure of the clinical aggression of human lymphoid malignancies including leukaemia and lymphoma (Ellims and Medley, 1984) as well as the SMAD4 expression might be associated with the development of some DLBCLs (germinal center) (Go, 2004). Finally, RAC2, a hematopoietic-specific GTPase, is involved in growth factor-induced proliferation and survival in B cells (Gu et al., 2005). The NFκB signature (Rosenwald et al., 2002) containing the genes IRF4, TRAF2, CFLAR, BCL2 and CCND2 is found in the consensus module and partially in the original module. It has been reported that NFκB targets are highly expressed in ABC DLBCL (Davis et al., 2001). The NFκB target gene TRAF2 is absent in the original module and only appears in the consensus module.

A GO term analysis shows, that the p-values are very similar in magnitude between the identified GO categories. Despite this, only one biological process in common is identified between the original module and the 50% consensus module, which is "negative regulation of cell cycle" including the genes BCL2, BCL6, CASP3, CDKN1A and HDAC1. Other biological processes identified for the genes of the consensus module include negative regulation of cell differentiation, size and growth, regulation of cell development, base-excision repair, negative regulation of developmental process and B cell proliferation. In contrast to that, the original modules hints to processes like response to stimuli (e.g. to peptide hormone, organic cyclic substance, insulin, drugs), phosphorylation, homoeostatic process and cellular and organ developmental processes. Probably due to more genes involved in signalling and signal transduction, that are absent in the consensus module, e.g. IRS1, PTPN1, PTPN2, ICAM3, TGFBR2.

Score Distribution

The score distribution for the DLBCL module is derived by jackknife resampling. It is approximately normal distributed and the medium of the distribution is shifted to the optimal score. The distribution is depicted in Figure 4.31, where the vertical red lines show the optimal score as well as the consensus score. The vertical blue lines depict the scores of 100 suboptimal modules. The 100 suboptimal modules are sampled form the solution space of suboptimal modules with a Hamming distance between 1 and 100. For a given confidence interval suboptimal solutions are regarded as equivalent solutions. The confidence interval is chosen to be two standard deviations from the optimal score (Figure 4.31 green shading). Within this region reside 27 unique suboptimal modules.
All suboptimal solutions differ only slightly from the original module by the deletion or addition of few nodes. Most of the time they are subsets of the original module. In a histogram of the nodes contained in the 27 suboptimal modules within two standard deviations, it can be seen that nodes which appear frequently (at least 20 times), correspond to the most robust nodes in the consensus module (Figure 4.32). All of the nodes have very high support values around 100 (lowest: 86) in the consensus module. Nodes that appear only very infrequently, e.g. ICAM3, ADM, BTK, IRS1, are not contained in the consensus module.

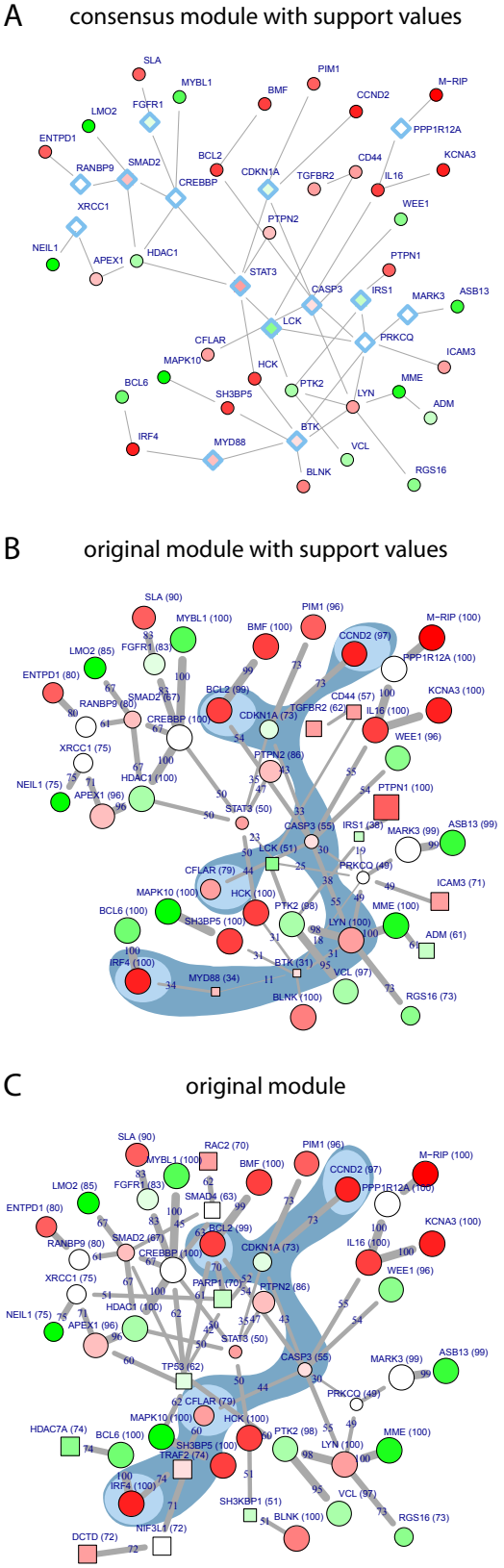


Figure 4.30: Resulting modules for DLBCL. The thickened blue diamonds displayed in (A) emphasise nodes with a negative score in the original network, while circular nodes depict positive scores. Nodes only present in either the original module or the corresponding consensus module in (B) and (C) are depicted by squared node symbols. Colouring of the nodes represents differential expression of the genes (for DLBCL red: up-regulated in ABC, green: down-regulated in ABC). For the DLBCL dataset the optimal solution (A, B) is calculated with an FDR of 10^{-7} whereas for the consensus module (C) a threshold $\rho = 50$ is used. In (B) and (C) the NFκB signature is highlighted in darkblue (containing: IRF4, CFLAR, BCL2, CCND2) and in the consensus additionally TRAF2, highlighted in lightblue. In (B) and (C) the support values are indicated in the original module and in the consensus module for DLBCL, respectively.

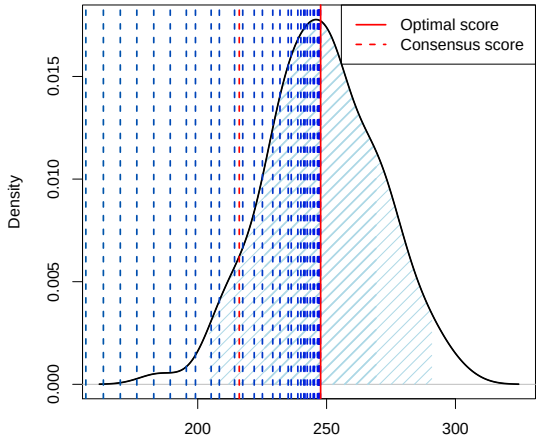


Figure 4.31: The distribution of the subnetwork score for DLBCL, derived by jackknife resampling. It is approximately normal distributed with the median of the distribution shifted to the optimal score. Indicated in red is the original score and as a dashed line the score of the consensus module. 100 suboptimal solutions are calculated, depicted as vertical blue-green lines in the distribution. Shaded in green are the fist and second standard deviation. Similarly good solutions, with a score within two standard deviations, include 27 unique suboptimal modules.

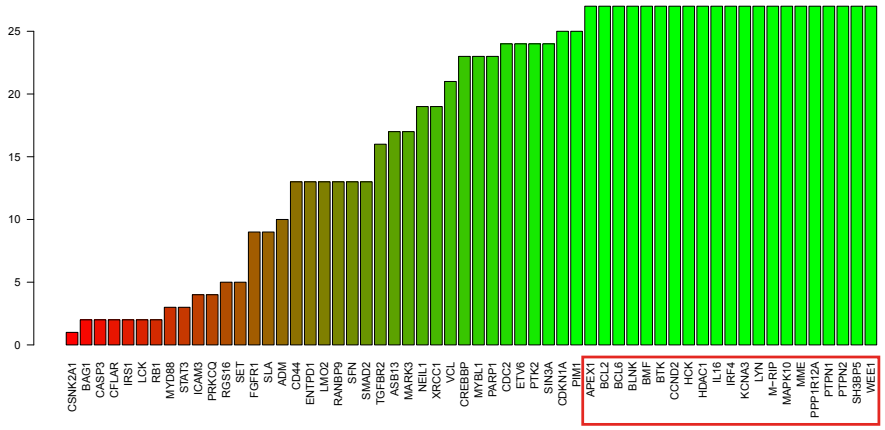


Figure 4.32: Histogram of nodes contained in the 27 unique suboptimal solutions within two standard deviations from the optimal score. Nodes which appear 27 times (marked in red), i.e. in all of the considered suboptimal modules, correspond to the most robust nodes in the consensus module with a support value of at least 86.

4.4.3 Discussion

The integrated network analysis performed on microarray data of acute lymphocytic leukaemia and microarray and survival data of diffuse large B-cell lymphomas identified relevant significant genes in the context of their interactions in the PPI network. These interactions could reveal utilised signalling pathways between the gene products or activating and inhibiting regulatory interactions. In both cases known essential proteins were part of the functional module. For ALL several genes contained in the module, e.g. tyrosine kinase genes (ABL, FYN and YES1,) have been identified previously to be important for the discrimination of the subgroups with and without translocation (Chiaretti et al., 2004). The module showed an enrichment for biologically meaningful processes like peptidyl-tyrosine phosphorylation, which arise from the BCR/ABL translocation. The DLBCL module included 2 established signatures, the proliferation signature and the NFκB signature. Thereby the identified modules include and extend already existing signatures by further significant genes and their interactions for the disease under study.

The presented consensus module approach additionally incorporates the idea of identifying a robust solution, despite several sources of noise in the biological data. In Section 4.3 a novel computational technique was introduced based on jackknife resampling to assess the robustness of a functional module and assign confidence values to the contained nodes and edges. In the last section the consensus algorithm was applied to the diffuse large B-cell lymphomas and acute lymphoblastic leukaemia microarray data. For a weak signal, as in the ALL gene expression profiles, the consensus module has the advantage of identifying robust parts of the module, whereas parts of the module with lower confidence values could arise from the noisy data, rather than a signal in the gene expression profiles. For ALL the consensus approach identified additional genes in comparison to the original module that were previously associated with the treatment of ALL and the classification of the analysed subtypes.
Robust parts of the consensus module determined for the DLBCL dataset included the well-known NFκB signature of up-regulated genes in the more malignant ABC subtype with an additional NFκB target gene.
The importance of the robust regions in the two biological datasets was supported by a significant enrichment of relevant biological processes among the genes contained in these regions.

In general, the analysis of transcriptional control and other regulatory events benefits from the use of biological networks as a framework to integrate the intermolecular dependencies into the study. Different kinds of networks, apart from protein-protein interaction networks, have been used recently

or can be used for a network-based analysis. These include, to name a few, networks based on connections between kinases and target proteins, of transcription factors and target genes, of mRNAs and non-coding RNA regulations or (response) regulators and target genes. With superimposed information on differential expression it is possible to identify and analyse highly significant regions of concerted changes in gene expression for different developmental stages, disease subtypes, infected versus non-infected cells/organisms or changing environmental conditions including temperature, aerobic versus anaerobic conditions and dietary alternations. Similar to the stress-induced changes in metabolism analysed in tardigrades (Section 4.5.1), stress response networks in bacteria from a transcriptional or metabolic perspective would be a field, suitable to analyse with integrated network approaches. Especially intracellular bacteria, e.g. *Salmonella typhimurium*, encounters many environmental stresses such as high temperature, acidic pH etcetera. A complex network of genes are involved in the survival of the bacteria; understanding these cellular networks may answer some of the basic questions about their intracellular survival and yield new starting points for therapies.

4.5 Application to Metabolic Profiles

4.5.1 Metabolic Time Series and EST Data from *M. tardigradum*

Introduction

Tardigrades are multicellular organisms, resistant to extreme environmental changes including desiccation, freezing and radiation. They outlast these conditions in an inactive form, called tun state or cryptobiosis (Spallanzani, 1776; Baumann, 1922; Rahm, 1921; Schill, 2010). All metabolic activity decreases during tun formation up to a complete cessation of measurable metabolism until environmental conditions improve and the tardigrade returns to its active state. Other invertebrate taxa that undergo cryptobiosis to escape damage to cellular structures and cell death are nematodes and rotifers (Clegg, 2001). All of these organisms are apparently able to prevent or repair damage under cryptobiosis. The tardigrade is a striking case as the whole animal phylum (most species) can undergo at least four types of cryptobiosis: anhydrobiosis (lack of water), anoxybiosis (lack of oxygen), cryobiosis (freezing) and osmobiosis (high solute concentration). This study focuses on the metabolic mechanisms of anhydrobiosis. The tardigrade species *Milnesium tardigradum* was analysed during tun formation, which was induced by dehydration. In tardigrades few metabolites have been analysed including carbohydrates that stabilise the membrane in the dry state (Crowe, 1975; Hengherr et al., 2008; Westh and Ramløv, 1988; Westh and Ramlov, 1991; Jönsson and Persson, 2010) or give protection and stress resistance (Jönsson and Schill, 2007; McGee et al., 2004; Ramlov and Westh, 2001; Reuner et al., 2010; Schill et al., 2004; Schokraie et al., 2011; Förster* et al., 2011; Altiero et al., 2012). The emphasis of this study lies on the integrated analysis of metabolism during dehydration and the subsequent rehydration. To examine this, metabolites were measured in a time series by gas chromatography coupled with mass spectrometry (GC-MS), additionally expressed sequence tags were integrated from EST libraries taken from an unassigned state, the active and tun state (Mali et al., 2010; Förster et al., 2011b). Here an integrative network approach is used to trace changes in tardigrade metabolism and identify pathways responsible for their extreme resistance against physical stress using these sources of data.

Similarly to other -omics data analysis, metabolic data analysis can likewise benefit from the integration of network information, e.g. *Cecil et al.* (Cecil et al., 2011) (flux-model oriented) or *Deo et al.* (Deo et al., 2010) (metabolic profile oriented).
Technically extending the latter metabolic analysis, tardigrade metabolism

is characterised using metabolite changes and EST data in the context of metabolic pathways. The analysis is based on a novel statistical approach to identify significantly changing metabolites with a trend in mass spectrometry profiles. This information is used to score the nodes of a metabolic network, constructed for this organism. Moreover, transcriptome information is used additionally to score the edges of the network. For the realisation, new methods are implemented extending the `heinz` framework (Dittrich et al., 2008; Beisser et al., 2010).

This study analyses the changes in metabolism, e.g. energy turnover, biosynthesis of cellular components, necessary for the change from an active state to an inactive state during dehydration and vice versa back into an active state during rehydration. First the metabolite data and EST data for *M. tardigradum* are presented and descriptively analysed. In the next step a metabolic reaction network is constructed, derived from the KEGG reference pathways (Kanehisa and Goto, 2000), as a source for pathway information. Subsequently, the Umbrella trend test is introduced and applied to the metabolite profiles. Based on the different data sources, scores are calculated which serve as a basis for the integrated analysis and for scoring the nodes and edges of the reaction network. Thereupon, a functional module is identified and visualised. This module constitutes a connected subnetwork of significantly altered metabolites and enzymes with expression changes between the active and inactive stage. The specific subnetwork is analysed with respect to the metabolic changes which the tardigrades undergo.

Construction of Metabolic Reaction Network

A metabolic reaction network is created on the basis of KEGG reference pathways. Therefore, the complete KEGG reference pathways were downloaded as XML files (Kanehisa and Goto, 2000) from `http://www.genome.jp/kegg/xml/KGML_v0.7.1` in December 2010. The KEGG pathways are loaded with the R package KEGGgraph (Zhang and Wiemann, 2009) and are processed using `BioNet` (Beisser et al., 2010) and igraph (Csardi and Nepusz, 2006). Using these packages, the pathways are converted into graphs with compounds as nodes and reactions as edges. The graphs resulting from each metabolic pathway are combined into one supergraph. Subsequently, the supergraph is converted into an undirected graph and pool metabolites are removed which form illegitimate shortcuts in the network. The excluded pool metabolites are: H^+, H_20, P, ATP, NAD^+, $NADH$, ADP, CO_2, CoA, $NADP^+$, $NADPH$, NH_3, PP. The resulting network consists of 3,671 nodes (metabolites) and 4,403 edges (reactions).

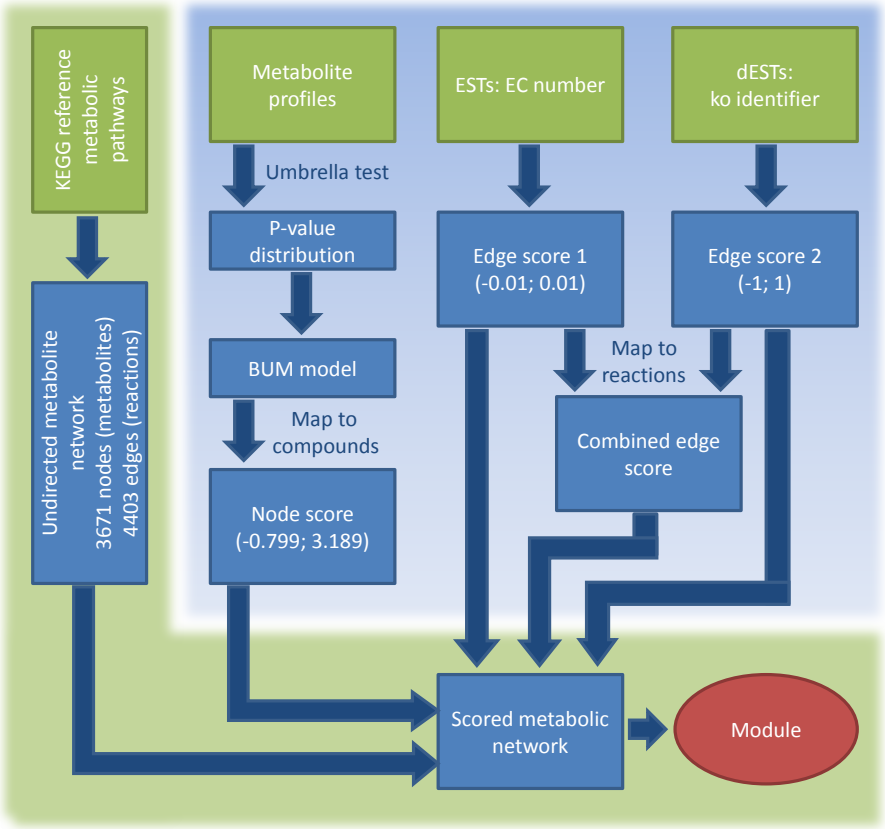


Figure 4.33: Different data sources are integrated to find a functional module explaining the metabolic changes in *M. tardigradum*. On the left hand side the metabolite network is created from KEGG reference pathways. On the top of the right side different sources of molecular data are integrated: the metabolite profiles and the two sets of ESTs. From these node and edge scores are derived which are subsequently used to score the metabolic network. In the last step a module or specific subnetwork is calculated from this scored network using the `heinz` algorithm.

Integrated Data

The aim of this study is to integrate diverse data sources by a network approach and look for region of interest in the network, representing significant metabolic changes. The data used for integration consists of metabolite profiles from (i) mass spectrometry (Section 3.10), (ii) EST data from previous studies (Förster et al., 2009) with mappings to EC numbers and (iii)

differential EST (dESTs) libraries for the active and inactive state of the tardigrades (Section 3.9). The distributions of mapped reactions to ESTs are depicted in Figure 4.34. The two sources of normal and differential ESTs cover by mapping of ko identifier and EC number a total of 1,063 reactions, 301 on common and 128 solely by dESTs and 634 by ESTs respectively.

(i) The metabolic profiles contain time course data for 84 metabolites after removal of unidentified metabolites, which were measured by GC-MS, but could not be assigned explicitly to a specific metabolite. Two distinct phases of tardigrade adaptation were measured, the dehydration phase (10 time points) and rehydration phase (10 time points). Fellenberg et al. (2001) introduced correspondence analysis (CA) to identify principal factors in microarray data. In a similar manner, the CA is applied to metabolite profiles with the strongest variance (top 10%, top 50% and all) to obtain the principal factors which contain most information (Figure 4.35). The first two axis explain 66% of the total variance (45% for top 50 % and 43% of CA with all metabolites), whereas the first axis clearly separates the de- and rehydration process (Figure 4.35, green to blue: dehydration; orange to red: rehydration), while the second axis separates early and late time points, with an exception for time points 3 and 14. This might be due to a similar profile of the metabolites at these time points in early dehydration and early rehydration. The CA not only allows to visualise the time points, but also the metabolites that are most specific for the axis (Figure 4.35, gray metabolites). This hints, e.g., to the importance of trehalose during early rehydration and the production of the amino acids valine and isoleucine during late rehydration. In contrast to the binary effect which is visible in the CA (de- and rehydration), the metabolic time courses are used subsequently to test for trends in the profiles. The most informative signal in the metabolite data is their change over time, which is examined using an Umbrella test, described in the next subsection. Based on the p-values from the trend test the metabolites can be scored and used as node scores for the metabolic network.

(ii) The presence of ESTs mapped to an EC number are used to give a minimal weight of 0.01 to the corresponding edges (reactions), A weight of -0.01 is given to edges without identified enzyme mapping. This favours the use of edges for which it is known, that an enzyme exists over reactions that might not exist in tardigrades, during the module search.

(iii) The dESTs are clustered into 4,422 clusters using CD-HIT-EST. Out of each cluster a representative is used to map KEGG ko identifiers with the KEGG Automatic Annotation Server (KAAS) (Moriya et al., 2007). Representative dESTs are annotated to other well characterised

invertebrate species, the fruit fly *Drosophila melanogaster* and the nematodes *Caenorhabditis elegans*, *Caenorhabditis briggsae* and *Brugia malayi*. Through KAAS 932 KEGG identifiers could be annotated to the clusters. The log2 ratio of active n_a to inactive n_i counts of dESTs is used for edge scoring:

$$S_e(t) = |\log_2(\frac{n_a}{n_i})| - t, \tag{4.13}$$

with threshold $t \geq 0$ to adjust the sensitivity/specificity of S_e similarly to the FDR used for the node scoring (Equation 4.10).

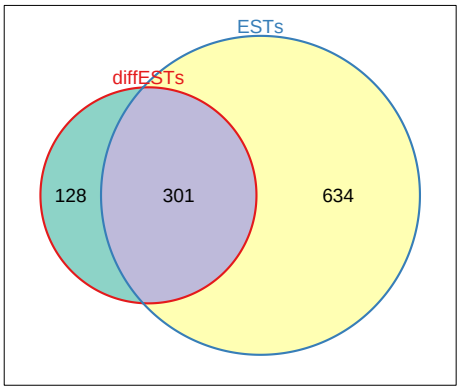


Figure 4.34: The Venn diagram depicts the distribution of mapped reactions to ESTs. The two sources of normal and differential ESTs cover by mapping of ko identifier and EC number a total of 1,063 reactions, 301 on common and 128 solely by dESTs and 634 by ESTs respectively.

Testing for Trends in Metabolic Profiles

The metabolite data are first examined by testing for differences in the means of the measurements for dehydration and rehydration using a Wilcoxon test (Figure 4.36). Despite significant differences for some metabolites, which correspond to the results from the CA, a closer look at the metabolite data shows increasing and decreasing trends in the time courses. This can be explained by the slow transition from the active into the inactive stage with a cessation of metabolism at time point 10 (20 h), where 100% of the tardigrades are in the tun stage. Therefore, the experimental design requires a trend test to analyse the metabolite time course data for metabolites that change the most. Two methods are used subsequently, the Jonckheere-Terpstra test (JT test) and the Umbrella test. Both tests are rank-based,

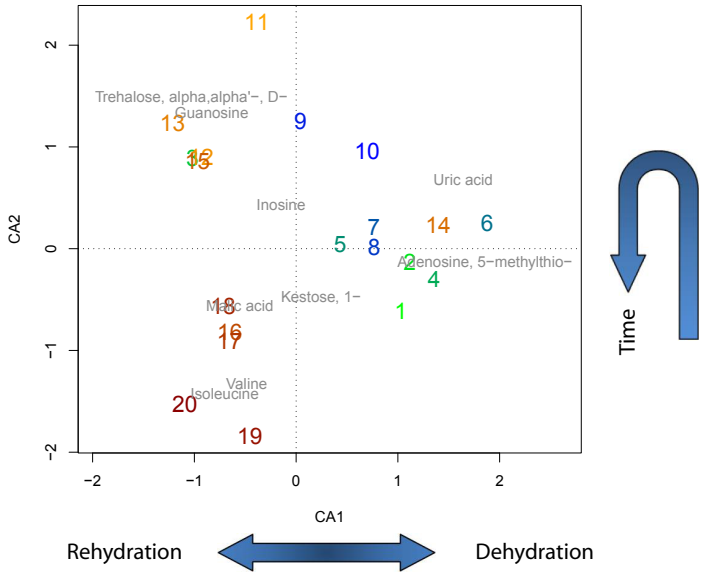


Figure 4.35: The mass spectrometry time points from 1 to 20 are depicted in a correspondence analysis. The CA is performed on the metabolites with the largest variance (gray), responsible for the separation. The colour scheme corresponds to the states of the tardigrades. From green to dark blue for dehydration and orange to red for rehydration. The x-axis clearly separates the two processes, while the y-axis separates early and late time points, with the exception 3 and 14 for both axis.

which is a requirement for the metabolic profiles, since absolute values are not comparable between metabolites and time points are not equally spaced. Both tests consider relative values between time points and an increase or decrease in their ranks.

The differences between the tests (Wilcoxon, JT test and Umbrella test) are shown for simulated time series in Figure 4.37 with the corresponding p-values. The JT test identifies a monotonic trend in the data, either increasing or decreasing. Therefore, the resulting p-values for the JT test are only significant for a monotonic upward trend (Figure 4.37 B and F). All other cases do not yield significant results. In contrast to this, the Umbrella test is used to test for trends with a peak or low-point (Figure 4.37 C, D, G and H).

From a biological point of view it is more reasonable to consider trends with a peak, rather then monotonic trends. Since it is more expected, that the

metabolism changes during the dehydration (time points 1-10) and rehydration (time points 11-20) phase and should be minimal between in the inactive state (time points 10-11). An umbrella form would be expected e.g. for storage metabolites or metabolites necessary for the protection of cellular structures, while for metabolites involved in energy production and cell growth an inverse umbrella shape is likely. Therefore the Umbrella test is applied to the metabolite profiles to calculate significant peaked trends in the metabolites. The turning point of the trend is set to time point 10, where all tardigrades have completed the dehydration process.

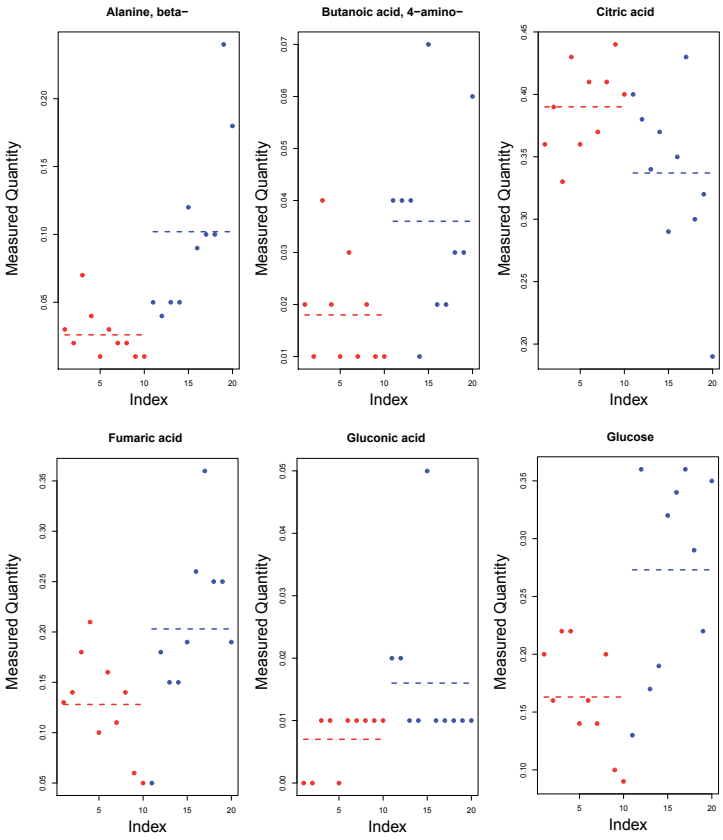


Figure 4.36: The observed key metabolite changes. Shown are key identified metabolites as measured by GC-MS. Given are raw intensities according to detection (y- axis: Measured quantities). Time steps are as follows: 60-1200min (dehydration); 0-270min (subsequent rehydration). For visualisation the 20 time steps are shown at equal spacing. Red depicts dehydration, blue rehydration, dotted lines show the group means.

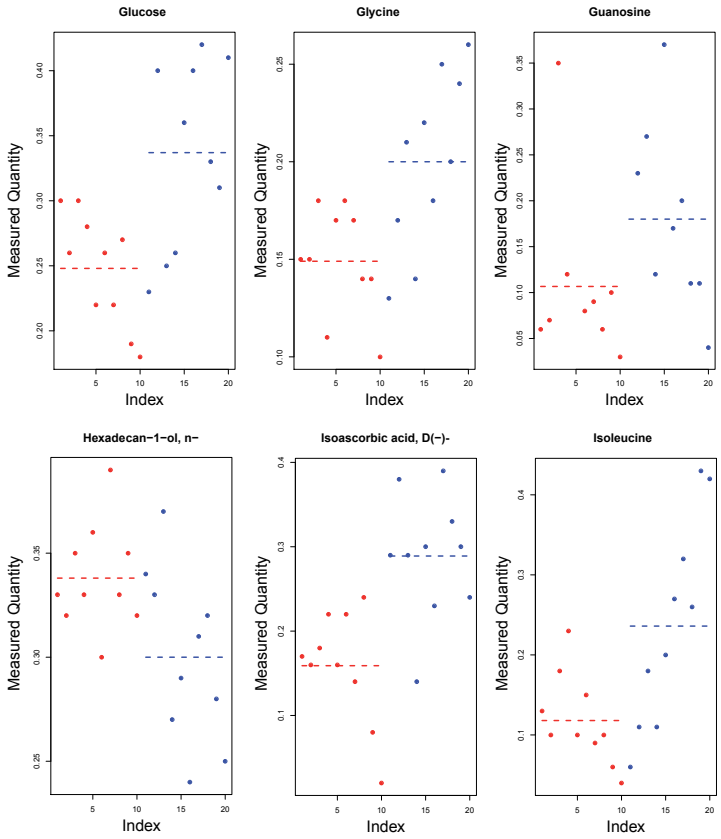


Figure 4.36: The observed key metabolite changes (cont.)

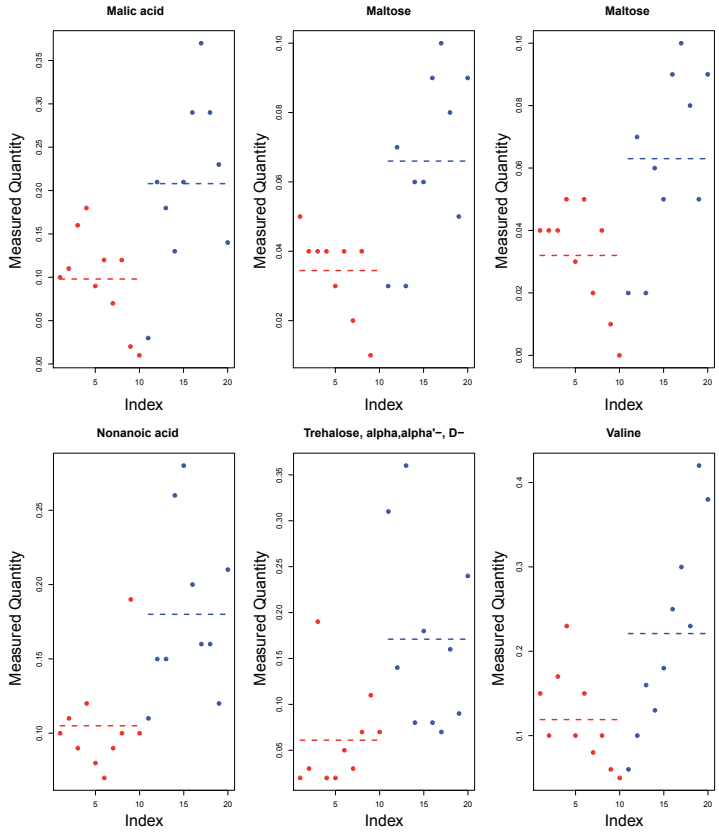


Figure 4.36: The observed key metabolite changes (cont.)

The p-values from the Umbrella test are subsequently used to score the nodes of the metabolite network. To convert the p-values to scores for the network and calculate functional modules the approach by Dittrich et al. (Dittrich et al., 2008) is used. A beta-uniform mixture model is fitted to the p-value distribution and node scores are calculated based on a log ratio of signal to noise. The fitted beta-uniform mixture model is depicted in Figure 4.38 with the corresponding quantile-quantile plot. The computed node scores range from -0.799 to 3.189, whereby significant p-values lead to a positive score, while non-significant p-values lead to a negative score for the corresponding nodes.

Calculation of the Metabolic Module

A functional module is calculated with an algorithm `heinz` (Dittrich et al., 2008; Beisser et al., 2010) using the node and edges scores to find a maximum-scoring subnetwork. The module is calculated in a two-step approach. At

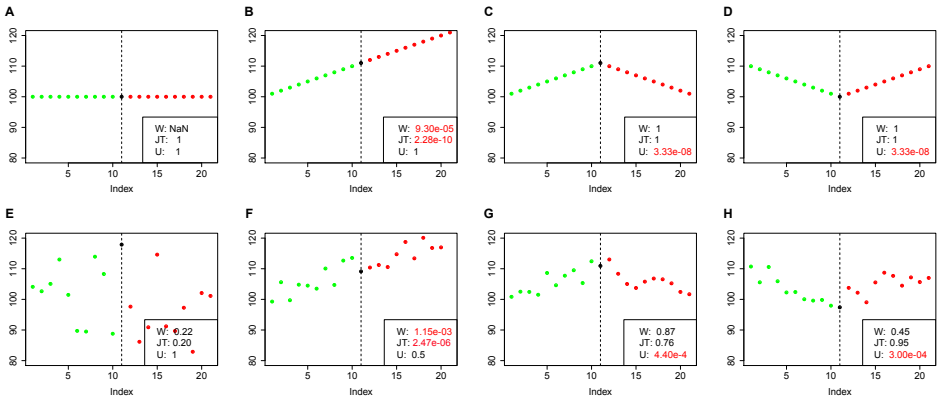


Figure 4.37: Time series are generated containing no trend, an upward or a peak/low point at time point 11. (A-D) show the time series without noise and (E-F) for jittered data. On these time series the JT and Umbrella tests are performed and p-values calculated.

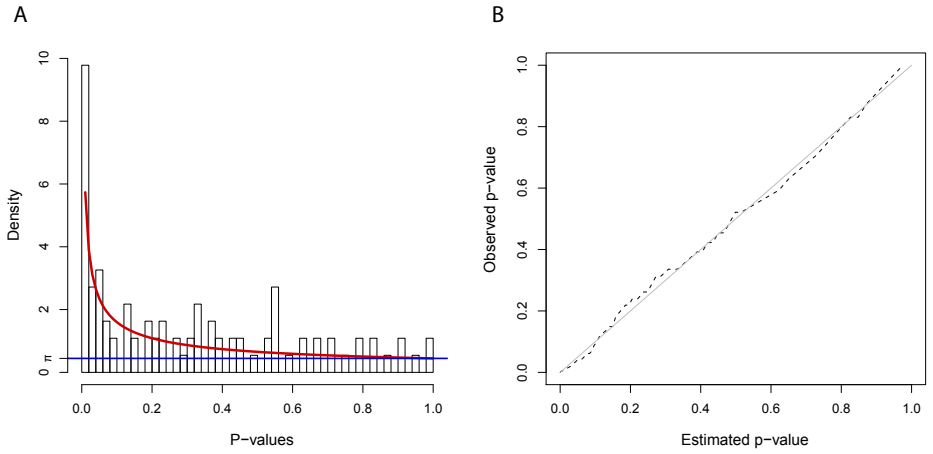


Figure 4.38: (A) depicts the fitted beta-uniform mixture model to the p-value distribution from the Umbrella test on the metabolite data. The goodness of fit is shown in (B) in the quantile-quantile plot of estimated p-values against observed p-values.

first a subnetwork is extracted based solely on the edge scores from the EST data. Edges representing reactions with identified enzymes are scored with 0.01, all other with -0.01. The `heinz` algorithm is applied to identify a maximum-scoring subnetwork using these edge scores. This excludes parts of the metabolite network where no enzymes can be identified in tardigrades. Since the metabolic network is created from the KEGG reference pathways, it contains all possible metabolic reactions and metabolites of which some

might not occur in tardigrades. In the maximum-scoring subnetwork these metabolites and reactions are removed, due to their negative edge scores. The second step uses only this subnetwork to score the metabolites and differentially abundant ESTs. A functional metabolic module is calculated based on the node scoring from the metabolic profiles with and FDR of 0.2 (Equation 4.10) and the log ratio score for the edges from the dESTs with $t = 1$ (Equation 4.13). By integrating the different sources of information, a module is obtained representing significant changes in trend in metabolites between the dehydration and rehydration process as well as changes in EST abundance, connected by reactions for which enzymes are identified in the tardigrades (Section 4.5.1).

Physical Stress-induced Metabolic Module for *M. tardigradum*

During dehydration the metabolism of tardigrades slowly reduces, until a complete cessation of measurable metabolism in the tun stage. The recovery time during rehydration is probably a function of metabolic activities linked to repair of damages caused by desiccation and to restoration of metabolic pathways (Rebecchi et al., 2007). The module reveals these processes by accumulating metabolic pathways involved in glycolysis/gluconeogenesis and carbohydrate metabolism, pentose phosphate pathway and the metabolism/catabolism of certain amino acids starting from pyruvate, including e.g methionine, lysine, phenylalanine, valine, arginine, tyrosine, threonine. Significant changes in these pathways mainly show an umbrella-shaped trend in the metabolic profiles, resembling a catabolic reaction or degradation of the metabolites followed by a restoration and production of amino acids and cellular components from one-carbon sugars as carbon and energy source. These processes are also consistently identified by the GO enrichment analysis, performed on the genes represented as enzymes in the functional module (Table 4.3). The inverse metabolic trend with a peaked shape is less common and would be expected for storage metabolites or bioprotectants. The module includes 4 metabolites with significant changes in this direction: putrescine, spermidine, D-glycerate and sn-glycerol 3-phophate. Sn-glycerol could potentially increase during dehydration to produce triglyceride as a highly efficient energy storage. Putrescine is a diamine created by the decarboxylation of ornithine. The addition of two propylamine residues yields spermidine, an essential growth factor. Other probable biological functions of spermidine are the stabilisation of DNA by association of the amino-groups with the phosphate residues of the DNA, increase of RNA synthesis and enhancement of stability of tRNAs and ribosomes (Michal, 1998).

Especially the information on differentially abundant enzymes is valuable,

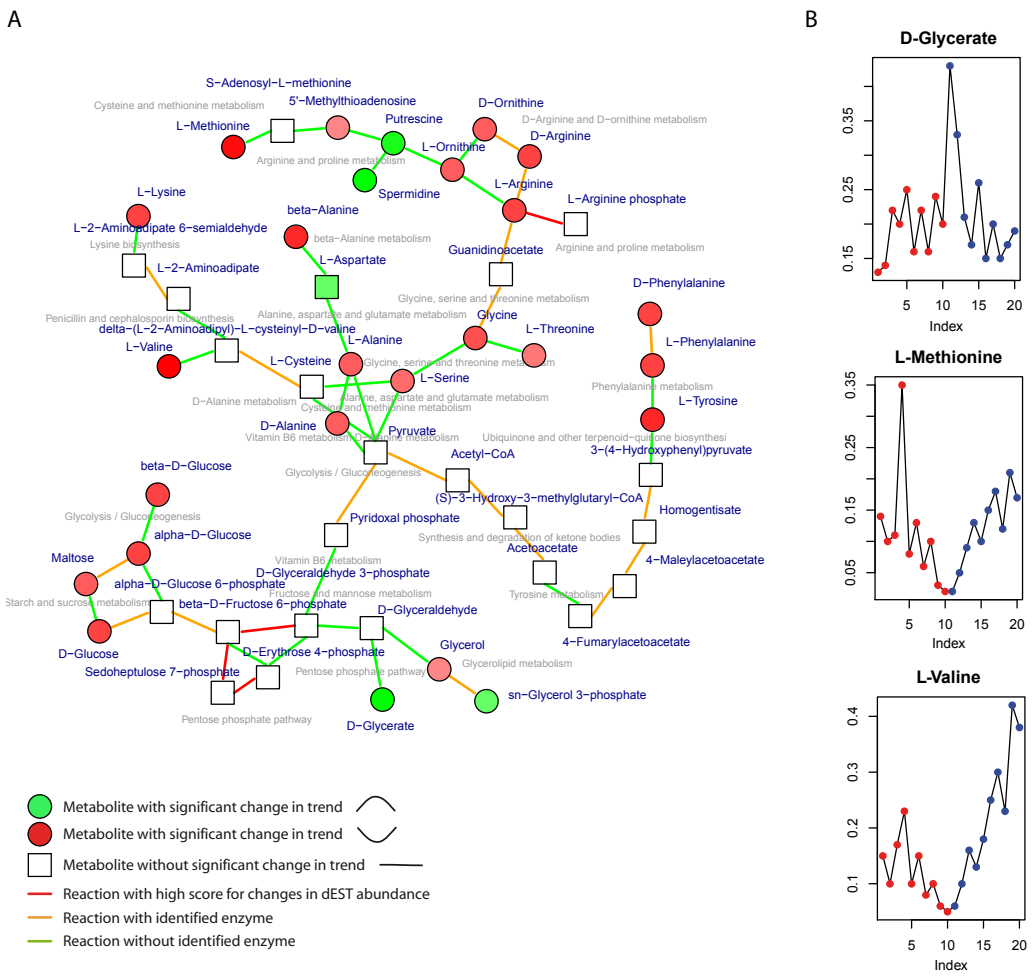


Figure 4.39: Shown is in (A) a functional module based on node and edge scores from the dESTs, calculated on a subnetwork based on edges scores from the ESTs. Circles depict nodes with positives scores, squares with negative score. The colouring of the nodes show the Z score (high: red, low: green) of the Umbrella test. For the edges the colours indicate a differential abundance of the enzyme (red), presence (orange) or absence (green) of the enzyme responsible for the metabolic reaction. (B) shows representative metabolites with a significant trend. The time points are depicted equally spaced for better visualisation.

since by this, reactions between metabolites are identified, that would normally not be taken into account, e.g. the link between the pentose phosphate pathway to glycolysis by the transaldolase reaction: sedoheptulose 7-phosphate + glyceraldehyde 3-phosphate $\rightleftharpoons$ erythrose 4-phosphate + fruc-

tose 6-phosphate. This reaction seems to be highly important for tardigrade metabolism since it shows strong differences in EST counts between the active and inactive stage. Unfortunately only few data on the abundance of the enzymes from dESTs are available and can be used to identify further commonly used reactions. This will improve as further stage-specific data become available.

Table 4.3: A GO enrichment is performed on the genes contained in the module derived from node and combined edge scores against the complete network. The following biological processes are enriched in the module depicted in Figure 4.39

	GOBPID	P-value	Term
1	GO:0009072	0.00001	aromatic amino acid family metabolic process
2	GO:0019439	0.00068	aromatic compound catabolic process
3	GO:0006559	0.00178	L-phenylalanine catabolic process
4	GO:0006570	0.00178	tyrosine metabolic process
5	GO:0009063	0.00504	cellular amino acid catabolic process
6	GO:0006730	0.00831	one-carbon metabolic process

4.5.2 Discussion

In this study an integrated approach to analyse sparse, heterogeneous data from a largely unknown organism was presented. For tardigrades (here: *M. tardigradum*) neither the genome is known nor the complete transcriptome or proteome, and up to now only quite incomplete analyses exist on a molecular level. This study represents a first attempt to use the existing data in an integrative approach to complement each other. By integrating metabolic profiles and transcript data into a metabolic network created from KEGG pathways, a relevant metabolic module of concerted changes in metabolism during the process of de- and rehydration was identified.

A two step approach was performed in this integrated network analysis. First of all the integrated data were statistically analysed and in a second step the resulting p-values were used to score the nodes and edges of a constructed metabolite network to search for a maximum-scoring subnetwork.

(i) From a statistical point of view a novel approach was used to analyse metabolite profiles. The umbrella test was implemented to identify significant non-monotonic trends in the metabolite profiles. Furthermore, expressed sequence tags were assigned to enzymes using the KEGG annotation server and analysed quantitatively regarding their differential abundance in the active and inactive form of the tardigrade.

(ii) The test results and information on enzymes were used subsequently to score a constructed metabolic network. It is the first time a metabolic module was calculated using several sources of information (for the edges and nodes). The integrative approach comprised a subnetwork calculation using edge scores derived from the knowledge of identified enzymes in tardigrades, this excluded metabolic pathways without known enzymes. Subsequently, the final metabolic module was calculated using node scores from the metabolic profiles and edges scores for differential abundance of enzymes.

The resulting module represented metabolic pathway for which enzymes were identified in the tardigrades, with coincidental significant changes in trend in metabolites as well as changes in EST abundance during the transitions between the active and inactive stage. Significant changes occurred mainly in pathways involved in glycolysis and sugar metabolism, pentose phosphate pathway and the metabolism of certain amino acids and additional reactions to store energy in triglyceride and protect the DNA. In contrast to already existing studies (e.g. Deo et al. (2010)), the presented approach not only uses the metabolic profiles, but allows the integration of further molecular data for the nodes as well as edges of a metabolic network. This captures the analysed process with more detail and completeness.

A clear benefit of the proposed integrative network approach is the feasible identification of key processes and pathways changing during the transition phases, despite the few data available for tardigrades so far. Functional modules identify relationships among changed metabolites and highlight the importance of specific enzymes and reactions in the metabolic changes. With sparse and diverse data at hand for tardigrades, the network approach is a convenient technique to integrate all available data and analyse them in a combined manner. As more molecular data become available, especially the whole genome and measurements of less common metabolites, this analysis will provide a deeper and more detailed insight into tardigrade metabolic processes and adaptations.

In general, metabolite networks can be used with a variety of additional superimposed information. This includes information on transcription, enzyme activity, metabolite abundance or isotopologically labelled glucose degradation. Including all available molecular data allows a more specific and detailed analysis of the metabolic process under study. In future studies, as more such data become available, this kind of integrated network approach will certainly become increasingly popular and an indispensable method in systems biology.

Chapter 5

General Discussion and Outlook

The presented integrated network approach is a prerequisite step toward the systematic analysis of cellular mechanisms, building upon the individual analysis of diverse molecular data. These data are on the one hand the high-throughput data generated by newly established platforms and techniques to measure the amount of mRNA, metabolites, proteins or modifications of DNA or proteins, and on the other hand the increasing number of known interactions between molecules, be it interactions between proteins, reactions creating a product from a substrate molecule or phosphorylations by which a signal is passed from the membrane into the nucleus. By integrating distinct experimental datasets into a biological network, the biological entities are put into context with their interaction partners and can thus be analysed jointly and with a focus on a specific aspect of interest.
The first part of the thesis described the integration of diverse data sources into a protein-protein interaction network. A statistical framework was implemented in a software package to score the nodes of a network by the significance of the biological data. Therefore, p-values from statistical tests were aggregated and a combined score was calculated based on their distribution. Large positive regions in the network constitute biologically relevant areas. They are searched for using an exact approach based on integer linear programming or a heuristic implemented in the package. The application on microarrays and survival data on diffuse large B-cell lymphoma and microarrays from acute lymphoblastic leukaemia patients demonstrated the ability of the integrated network approach to identify significantly deregulated genes and how these are interrelated in the form of a functional module. The approach is flexible and allows the integration of diverse molecular data, as long as a statistical test can be performed to generate p-values, and the us-

age of various biological networks. This flexibility is underlined by another biological case study which analysed tardigrade metabolism on the basis of metabolic profiles, ESTs and a metabolic network generated from KEGG pathways. In the same manner all of these data could be combined to search for modules of metabolic changes during transitions between the active and inactive stage of tardigrades due to extreme environmental changes. In addition to the original application on genetic data, here a new statistical test had to be developed to analyse the metabolic time courses and generate the necessary p-value distribution. Another extension was the usage of edge scores. Not only the nodes of the network were scored, but also the edges, to integrate further information. In this case information on existing enzymes was added to the analysis. A difficulty that arose with it, was the adjustment of node and edge scores. At the moment they are scored independently and balanced by hand or modules are calculated consecutively on edge and node scores to obtain an overall module that reflects the significance of nodes and edges in equal measure. Preferable is a multivariate scoring of nodes and edges at the same time. A fundamental problem with this is the estimation of a valid covariance matrix that represents the network structure. Several approaches exist that use the correlation or partial correlation estimated from gene expression values as a network (Schaefer et al., 2009), but the large amount of genes in contrast to few samples, makes the estimate rather unreliable. Even with a valid covariance structure and a multivariate test (e.g. T^2-test) the question remains of how to weight the covariance according to the integrated edge data. The basis of these ideas is the balancing of p-values for the nodes and edges, alternatively they could be balanced during the scoring step by fitting a model to a multivariate p-value distribution. This is a future enhancement, which still has to be investigated and evaluated.

The sensitivity and specificity of the method are modifiable, and thereby the size of the resulting modules, by choice of an appropriate FDR or predefined module size. The modulation of the size of the resulting module allows to zoom into the most significant areas and visualise the module; a visualisation is impossible for larger subnetworks, because of the dense network structure. A FDR that represents the signal part of the data can be calculated by first of all calculating the amount of signal and noise that is in the molecular data. The upper bound π for the fraction of noise can be obtained from the beta-uniform mixture model or similar methods to estimate the amount of noise from the uniform distribution of p-values (e.g. Bartholome and Gehring (2010)). The changes in sensitivity and specificity due to changes in FDR and size parameter settings were assessed and compared to established methods in an in-depth simulation study. It was shown that the exact algorithm as well as the heuristic approach obtained very good recall and precision values in contrast to other heuristic approaches. In addition to the accuracy also the robustness was investigated for these methods. As

stated by Lee (2010), a challenge arises "due to the fact that high-throughput biotechnical data and large biological databases are inevitably noisy because biological information and signals of interest are often observed with many other random or biased factors that may obscure main signals and information of interest (Cho and Lee, 2004). Therefore, investigations on large biological data cannot be successfully performed unless rigorous statistical algorithms are developed and effectively utilised to reduce and decompose various sources of error." As he mentions, it is an important task to consider the different sources of variability in the data, e.g. arising in this case from the integrated data or the network. Therefore, the variability of modules, that result from the module detection using perturbed data, was examined. Despite the smaller variance for the exact solution, all modules differed due to slight differences in the biological data. The aim was therefore to estimate this variance, assign confidence values to the nodes and edges of the module and calculate a robust resulting module. This was archived by the computation of a consensus module with support values, based on a jackknife resampling procedure. Similarly to the computation of consensus trees in phylogeny to obtain a robust phylogenetic tree, a resampling and consensus procedure was used here to obtain a summary of possible subnetwork structures. In contrast to phylogeny the resampling was not based on bootstrapping but on jackknifing. The difference is, that bootstrapping is a resampling with replacement and jackknifing is a resampling procedure without replacement. According to Felsenstein (Felsenstein, 1985, 2004) a delete-half jackknife is equivalent to the bootstrap. The bootstrap approach was discarded, because it did not yield values that were t-distributed after applying a t-test and could therefore not be used to calculate p-values for differential expression. Despite testing several correction procedures for the t-test, e.g. variance correction, bootstrapping could not be utilised. Therefore, the valid alternative of the delete-half jackknife was used here.
Finally, the application of the consensus approach to the microarray datasets identified robust, optimal modules with assigned support values to their nodes and edges. This statistical algorithm allowed to reduce and decompose various sources of noise in the biological data, as requested by Lee.

Other fields of biology, that could benefit and gain new insights by an integrated network-based approach, include

Immunology The immune system not only recognises foreign substances, from viruses to parasitic worms, but also identifies and interacts with its own cells and tissues and needs to distinguish these from disease causing agents and abnormal cells. Regarding the interactions as a immune system specific network could yield new ways to analyse this system and broaden the understanding of autoimmunity, organ rejection and HIV. See Hyduke et al.

(2007) and Campbell et al. (2011) for network-based approaches to study immune system-microbial interactions from the microbial (*Escherichia coli*) side and for activated pathways in the immune system during bacterial and allergic responses.

Ecology In ecology food chains have long been taken over by food webs, network representations of feeding connections. Other species interactions include plants-animal pollinators, coevolution of species or symbiont and host, the invasion of new species into networks of interacting species. These network have been analysed regarding their topological properties and what these imply for the studied ecosystem, the stability of the network as well as dynamic processes (Bascompte, 2010).

Public health Network-based approaches are recently used in epidemiology to understand the spreading of diseases and develop effective vaccination strategies (Eubank et al., 2010). Other applications of networks are drug-disease networks (Yildirim et al., 2007; Hu and Agarwal, 2009), which can be used to identify new drug targets, common properties of yet unrelated drug or causes of drug-drug interactions and side-effects. A plethora of information can be integrated to create patient-specific or disease-specific networks to enhance the medical treatment of a patient or against a specific disease (Ideker and Sharan, 2008).

Microbiology In microbiology interactions exist between pathogen and host, symbiont and host or in microbial ecosystems. A combined and interdependent analysis of gene expression between pathogens or symbionts and hosts (Hyduke et al., 2007) under diverse conditions and knock out experiments could reveal regulatory mechanisms in and between the species. The integrated functional analysis could furthermore identify subnetworks of dependent gene expression changes using gene expression profiles and measures of correlation in gene expression between the species. An example for a host-pathogen network analysis is the construction of a *Yersinia pestis*-human protein interaction network to improve the understanding of the bacillus' pathogenesis by Yang et al. (2011).

Development and Future Trends for Integrated Network Analysis

The future application of network biology surely involves two fields, one being the inference of networks from molecular data, the other one being the dynamical analysis of networks, showing the functioning of cellular processes over time.

In the field of network inference a lot has developed over the last 5 to 10 years. Mainly these approaches try to reverse-engineering regulatory interactions among genes or try to relate the expression of a gene to the expression of the other genes in the cell in gene networks using gene expression profiles. Gene networks do not necessarily imply a physical interaction, but can also refer to indirect interactions. Nevertheless, the similar expression patterns between two genes hint to a coregulation. Concerning steady state microarray data, networks are often inferred through calculating coexpression (D'haeseleer et al., 2000), using Gaussian graphical models (Hartemink et al., 2001; Schaefer et al., 2009), Bayesian networks (Friedman et al., 2000; Segal et al., 2003), multiple regression (Gardner et al., 2003), ordinary differential equations (Bansal et al., 2007) or mutual information (Margolin et al., 2006).
Gene networks are divided into correlation networks, based on a correlation (e.g. Pearson correlation) in the expression patterns of the genes, and networks based on conditional independence, these include Gaussian graphical model (GGM), sparse GGMs and Bayesian networks depending on the order of the conditional independence. The problem of correlation networks is, that they can not distinguish direct from indirect dependencies. This problem is relaxed in conditional models (GGM and Bayesian networks), where the correlation between two genes is calculated given some dependencies. This provides information of whether the correlation between two genes can be explained by other genes in the model. Problems with these approaches arise due to the calculation of correlation or covariance matrices that have to be full rank to be inverted. Since the number of samples is usually much smaller than the number of genes on a microarray, the necessary assumptions are not fulfilled. Available software uses approximations by the calculation of a pseudoinverse, shrinkage approaches or simplified models, e.g. see Schaefer et al. (2009).
Gene regulatory networks are mainly predicted from observational data following an intervention. For example gene expression profiles of deletion mutants. Causal effects are predicted from these data to understand cause-effect relationships between the genes. This technique has proven to be useful for small regulatory networks with 10 to 100 genes, as for example supplied from simulations by the DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge, a competition in reverse engineering of gene regulation networks. Recently, Maathuis et al. (2010) demonstrated that it is possible to infer a regulatory network for S. cerevisiae where the number of variables (genes: 5,361) is much larger than the sample size (63) and the variables are substantially disturbed by noise. Despite this progression, network inference techniques still have to be optimised and evidence their applicability and usefulness to real biological data with only few replicates. By now, the analysis I conducted using these kind of networks were not satisfactory and did not provide a benefit over protein-protein interac-

tion networks. The calculation either did not work for the number of genes (e.g. Affymetrix Human Genome Chip hgu95av2 with 12,625 features) or resulted in implausible genetic interactions.
Once these methods are sufficiently advanced, they will certainly be also beneficial for integrated network biology and could be used in applications as a replacement of physically measured networks.

The dynamical analysis of networks is another interesting and important field, integrated network analysis will head for. The shift from a static to a dynamic network analysis is essential to gain new insights into molecular systems. So far, mainly static microarray data are used for the integration. Recently Tang et al. (2011) investigated and compared time course protein interaction networks (TC-PINs) by incorporating time series gene expression data into PPI networks and calculating functional modules using a clustering algorithms. They obtained 36 specific PPI subnetworks corresponding to 36 time points. Similarly, de Lichtenberg et al. (2005) and Wu et al. (2009) analysed the dynamics of protein complexes and hub proteins during the yeast cell cycle by integrating data on protein interactions and gene expression. de Lichtenberg et al. (2005) revealed 29 time-dependent heavily intraconnected modules of periodically and constitutively expressed proteins in a temporal cell cycle context. Wu et al. (2009) contrasted the properties of the hub proteins between an aggregated network of all cell cycle phases and four phase-specific sub-networks (G1, S, G2 and M phase). In addition to these analysis it would be worthwhile to search for specific functional modules that change over time. By now, the proteins cluster, because of the network structure and the presence or absence of a protein in a phase. But one could also investigate significant changes from one phase to the next and visualise a dynamic movement of a resulting functional module through the overall network. Technically the analysis builds upon the analysis described in this thesis and can be performed similarly on time-series microarray data, for example the yeast cell cycle. More intricate is an appropriate interpretable visualisation and analysis of the time-dependent modules, especially the movement throughout a large network. This is one of the problems in integrated network analysis that should be tackled in the near future to overcome the static view on molecular networks.

Bibliography

Albert, R. and Barabási, A. L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, Jan. 2002. doi: 10.1103/RevModPhys.74.47. URL http://dx.doi.org/10.1103/RevModPhys.74.47.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000. doi: 10.1038/35000501. URL http://dx.doi.org/10.1038/35000501.

Altiero, T., Guidetti, R., Boschini, D., and Rebecchi, L. Heat shock proteins in encysted and anhydrobiotic eutardigrades. *J. LImnol.*, 71(1):accepted, 2012.

Andersen, P. and Gill, R. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000. doi: 10.1038/75556. URL http://dx.doi.org/10.1038/75556.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3:78, 2007. doi: 10.1038/msb4100120. URL http://dx.doi.org/10.1038/msb4100120.

Barabási, A.-L. and Bonabeau, E. Scale-free networks. *Sci Am*, 288(5):60–69, May 2003.

Barabási, A.-L. and Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004. doi: 10.1038/nrg1272. URL http://dx.doi.org/10.1038/nrg1272.

Bartholome, K. and Gehring, J. *Gene Set Regulation Index (GSRI)*, 2010. R package version 1.0.0.

Bascompte, J. Ecology. structure and dynamics of ecological networks. *Science*, 329(5993):765–766, Aug 2010. doi: 10.1126/science.1194255. URL http://dx.doi.org/10.1126/science.1194255.

Baumann, H. Die anabiose der tardigraden. *Zool. Jahrb.*, 45:501–556, 1922.

Beisser, D., Klau, G. W., Dandekar, T., Mueller, T., and Dittrich, M. Bionet: an R-package for the functional analysis of biological networks. *Bioinformatics*, 26:1129–1130, Feb 2010a. doi: 10.1093/bioinformatics/btq089. URL http://dx.doi.org/10.1093/bioinformatics/btq089.

Benzécri, J.-P. *L'Analyse des Données. L'Analyse des Correspondences.*, volume II. Dunod, Paris, France, 1973.

Bergholdt, R., Størling, Z. M., Lage, K., Karlberg, E. O., Olason, P. I., Aalund, M., Nerup, J., Brunak, S., Workman, C. T., and Pociot, F. Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. *Genome Biol*, 8(11):R253, 2007. doi: 10.1186/gb-2007-8-11-r253. URL http://dx.doi.org/10.1186/gb-2007-8-11-r253.

Bergholdt, R., Brorsson, C., Lage, K., Nielsen, J. H. H., Brunak, S., and Pociot, F. Expression profiling of human genetic and protein interaction networks in type 1 diabetes. *PloS one*, 4(7):e6250+, July 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006250. URL http://dx.doi.org/10.1371/journal.pone.0006250.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14(3):292–299, Jun 2004. doi: 10.1016/j.sbi.2004.05.003. URL http://dx.doi.org/10.1016/j.sbi.2004.05.003.

Brand, N. J. and Barton, P. J. R. Myocardial molecular biology: an introduction. *Heart*, 87(3):284–293, Mar 2002.

Broberg, P. *SAGx: Statistical Analysis of the GeneChip*, 1.22.0 edition, 2009. URL http://home.swipnet.se/pibroberg/expression_hemsida1.html. R package version 1.22.0.

Brorsson, C., Hansen, N. T., Lage, K., Bergholdt, R., Brunak, S., Pociot, F., and Consortium, D. G. Identification of t1d susceptibility genes within the mhc region by combining protein interaction networks and snp genotyping data. *Diabetes Obes Metab*, 11 Suppl 1:60–66, Feb 2009.

Busygin, S. and Pardalos, P. M. Exploring microarray data with correspondence analysis. In Pardalos, P. M., Boginski, V. L., and Vazacopoulos, A., editors, *Data Mining in Biomedicine*, volume 7 of *Springer Optimization and Its Applications*, pages 25–37. Springer US, 2007. ISBN 978-0-387-69319-4.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. Y. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(6):1190–1208, 1995. URL citeseer.ist.psu.edu/byrd94limited.html.

Cabusora, L., Sutton, E., Fulmer, A., and Forst, C. V. Differential network expression during drug and stress response. *Bioinformatics*, 21(12):2898–2905, Jun 2005. doi: 10.1093/bioinformatics/bti440. URL http://dx.doi.org/10.1093/bioinformatics/bti440.

Campbell, C., Thakar, J., and Albert, R. Network analysis reveals cross-links of the immune pathways activated by bacteria and allergen. *Phys Rev E Stat Nonlin Soft Matter Phys*, 84(3-1):031929, Sep 2011.

Carey, V. J., Gentry, J., Whalen, E., and Gentleman, R. Network structures and algorithms in Bioconductor. *Bioinformatics*, 21(1):135–136, Jan 2005. doi: 10.1093/bioinformatics/bth458. URL http://dx.doi.org/10.1093/bioinformatics/bth458.

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Paley, S., Popescu, L., Pujar, A., Shearer, A. G., Zhang, P., and Karp, P. D. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 38(Database issue):D473–D479, Jan 2010. doi: 10.1093/nar/gkp875. URL http://dx.doi.org/10.1093/nar/gkp875.

Cecil, A., Rikanović, C., Ohlsen, K., Liang, C., Bernhardt, J., Oelschlaeger, T. A., Gulder, T., Bringmann, G., Holzgrabe, U., Unger, M., and Dandekar, T. Modeling antibiotic and cytotoxic effects of the dimeric isoquinoline iq-143 on metabolism and its regulation in staphylococcus aureus, staphylococcus epidermidis and human cells. *Genome Biol*, 12(3):R24, 2011. doi: 10.1186/gb-2011-12-3-r24. URL http://dx.doi.org/10.1186/gb-2011-12-3-r24.

Chen, J., Fiskus, W., Eaton, K., Fernandez, P., Wang, Y., Rao, R., Lee, P., Joshi, R., Yang, Y., Kolhe, R., Balusu, R., Chappa, P., Natarajan, K., Jillella, A., Atadja, P., and Bhalla, K. N. Cotreatment with bcl-2 antagonist sensitizes cutaneous t-cell lymphoma to lethal action of hdac7-nur77-based mechanism. *Blood*, 113(17):4038–4048, Apr 2009. doi: 10.1182/blood-2008-08-176024. URL http://dx.doi.org/10.1182/blood-2008-08-176024.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778, Apr 2004. doi: 10.1182/blood-2003-09-3243. URL http://dx.doi.org/10.1182/blood-2003-09-3243.

Cho, H. and Lee, J. K. Bayesian hierarchical error model for analysis of gene expression data. *Bioinformatics*, 20(13):2016–2025, Sep 2004. doi: 10.1093/bioinformatics/bth192. URL http://dx.doi.org/10.1093/bioinformatics/bth192.

Chua, H. N., Ning, K., Sung, W.-K., Leong, H. W., and Wong, L. Using indirect protein-protein interactions for protein complex prediction. *J Bioinform Comput Biol*, 6(3):435–466, Jun 2008.

Clegg, J. S. Cryptobiosis–a peculiar state of biological organization. *Comp Biochem Physiol B Biochem Mol Biol*, 128(4):613–624, Apr 2001.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2(10):2366–2382, 2007. doi: 10.1038/nprot.2007.324.

Cox, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, B(34):187–220, 1972.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, 39(Database issue):D691–D697, Jan 2011. doi: 10.1093/nar/gkq1018. URL http://dx.doi.org/10.1093/nar/gkq1018.

Crowe, J. H. The physiology of cryptobiosis in tardigrades. *Memorie dell'Istituto Italiano di Idrobiologica*, 32 (Suppl):37–59, 1975.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.sf.net.

Dale, J. M., Popescu, L., and Karp, P. D. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, 11:15, 2010. doi: 10.1186/1471-2105-11-15. URL http://dx.doi.org/10.1186/1471-2105-11-15.

Davis, R. E., Brown, K. D., Siebenlist, U., and Staudt, L. M. Constitutive nuclear factor kappaB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. *J Exp Med*, 194 (12):1861–1874, Dec 2001.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, Feb 2005. doi: 10.1126/science.1105103. URL http://dx.doi.org/10.1126/science.1105103.

Deo, R. C., Hunter, L., Lewis, G. D., Pare, G., Vasan, R. S., Chasman, D., Wang, T. J., Gerszten, R. E., and Roth, F. P. Interpreting metabolomic profiles using unbiased pathway models. *PLoS Comput Biol*, 6(2):e1000692, Feb 2010. doi: 10.1371/journal.pcbi.1000692. URL http://dx.doi.org/10.1371/journal.pcbi.1000692.

D'haeseleer, P., Liang, S., and Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, Aug 2000.

Dittrich, M., Beisser, D., Klau, G. W., and Müller, T. Functional modules in protein–protein interaction networks. In Choi, S. and Choi, S., editors, *Systems Biology for Signaling Networks*, volume 1 of *Systems Biology*, pages 353–369. Springer New York, 2010.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24 (13):i223–i231, Jul 2008. URL http://dx.doi.org/10.1093/bioinformatics/btn161.

Dixon, P. and Palmer, M. W. Vegan, a package of r functions for community ecology. *Journal of Vegetation Science*, 14(6):927–930, Dec. 2003. ISSN 1100-9233. URL http://dx.doi.org/10.1658/1100-9233(2003)014[0927:VAPORF]2.0.CO;2.

Djebbari, A. and Quackenbush, J. Seeded bayesian networks: constructing genetic networks from microarray data. *BMC Syst Biol*, 2:57, 2008. doi: 10.1186/1752-0509-2-57. URL http://dx.doi.org/10.1186/1752-0509-2-57.

Dlugosch, K. M. Snowhite: A cleaning pipeline for roche 454 cdna sequences. URL http://www.kdlugosch.net/software.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002. URL http://www3.stat.sinica.edu.tw/statistica/j12n1/j12n16/j12n16.htm.

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364. doi: 10.2307/2958830. URL http://dx.doi.org/10.2307/2958830.

Eisenberg, E. and Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys Rev Lett*, 91(13):138701, Sep 2003.

Ellims, P. H. and Medley, G. Deoxycytidylate deaminase activity in lymphoproliferative disorders. *Leuk Res*, 8(1):123–128, 1984.

Emig, D., Cline, M. S., Lengauer, T., and Albrecht, M. Integrating expression data with domain interaction networks. *Bioinformatics*, 24(21):2546–2548, Nov 2008. doi: 10.1093/bioinformatics/btn437. URL http://dx.doi.org/10.1093/bioinformatics/btn437.

Erban, A., Schauer, N., Fernie, A. R., and Kopka, J. Nonsupervised construction and application of mass spectral and retention time index libraries from time-of-flight gas chromatography-mass spectrometry metabolite profiles. *Methods Mol Biol*, 358:19–38, 2007.

Eubank, S., Barrett, C., Beckman, R., Bisset, K., Durbeck, L., Kuhlman, C., Lewis, B., Marathe, A., Marathe, M., and Stretz, P. Detail in network models of epidemiology: are we there yet? *J Biol Dyn*, 4(5):446–455, Sep 2010. doi: 10.1080/17513751003778687. URL http://dx.doi.org/10.1080/17513751003778687.

Ewing, B. and Green, P. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res*, 8(3):186–194, Mar 1998.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Res*, 8(3):175–185, Mar 1998.

Faderl, S., Kantarjian, H. M., Talpaz, M., and Estrov, Z. Clinical significance of cytogenetic abnormalities in adult acute lymphoblastic leukemia. *Blood*, 91(11):3995–4019, 1998. URL http://bloodjournal.hematologylibrary.org/content/91/11/3995.short.

Falcon, S. and Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., and Vingron, M. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A*, 98(19):10781–10786, Sep 2001. doi: 10.1073/pnas.181597298. URL http://dx.doi.org/10.1073/pnas.181597298.

Felsenstein, J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791, 1985. ISSN 00143820. doi: 10.2307/2408678. URL http://dx.doi.org/10.2307/2408678.

Felsenstein, J. *Inferring phylogenies*. Sinauer Associates, 2004. ISBN 9780878931774. URL http://books.google.com/books?id=GI6PQgAACAAJ.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. Metabolite profiling for plant functional genomics. *Nat Biotechnol*, 18(11):1157–1161, Nov 2000. doi: 10.1038/81137. URL http://dx.doi.org/10.1038/81137.

Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H., and Batzoglou, S. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res*, 16(9):1169–1181, Sep 2006. doi: 10.1101/gr.5235706. URL http://dx.doi.org/10.1101/gr.5235706.

Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, Feb 2004. doi: 10.1126/science.1094068. URL http://dx.doi.org/10.1126/science.1094068.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620, 2000. doi: 10.1089/106652700750050961. URL http://dx.doi.org/10.1089/106652700750050961.

Förster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., Schnölzer, M., Frohme, M., Müller, T., Schill, R. O., and Dandekar, T. Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as rna elements in tardigrades. *BMC Genomics*, 10:469, 2009. doi: 10.1186/1471-2164-10-469. URL http://dx.doi.org/10.1186/1471-2164-10-469.

Förster, F., Beisser, D., Grohme, M., Liang, C., Mali, B., Reuner, A., Siegl, A. M., Engelmann, J. C., Shkumatov, A., Elham Schokraie and, T. M., Blaxter, M., Schnölzer, M., Schill, R. O., Frohme, M., and Dandekar, T. The compared transcriptomes of hypsibius dujardini and milnesium tardigradum with rna motifs, encoded proteins, resulting clusters and pathways, tardigrade-specific and general stress adaptations. *In submission*, 2011a.

Förster, F., Beisser*, D., Grohme, M., Liang, C., Mali, B., Reuner, A., Siegl, A. M., Engelmann, J. C., Shkumatov, A., Schokraie, E., Müller, T., Blaxter, M., Schnölzer, M., Schill, R. O., Frohme, M., and Dandekar, T. The compared transcriptomes of hypsibius dujardini and milnesium tardigradum with rna motifs, encoded proteins, resulting clusters and pathways, tardigrade-specific and general stress adaptations. page in submission, 2011b.

Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, Jul 2003. doi: 10.1126/science.1081900. URL http://dx.doi.org/10.1126/science.1081900.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387251464.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. *genefilter: genefilter: methods for filtering genes from microarray experiments*, 2008a. R package version 1.30.0.

Gentleman, R., Whalen, E., Huber, W., and Falcon, S. *graph: graph: A package to handle graph data structures*, 2008b. R package version 1.12.1.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004. doi: 10.1186/gb-2004-5-10-r80. URL http://dx.doi.org/10.1186/gb-2004-5-10-r80.

Go, J. Expression pattern of smad proteins in diffuse large b-cell lymphomas. *Korean J Pathol*, 38: 301–305, 2004.

Goldovsky, L., Cases, I., Enright, A. J., and Ouzounis, C. A. Biolayout(java): versatile network visualisation of structural and functional relationships. *Appl Bioinformatics*, 4(1):71–74, 2005.

Green, P. cross_match. URL http://www.phrap.org.

Gruber, T. R. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5:199–220, June 1993. ISSN 1042-8143. doi: 10.1006/knac.1993.1008. URL http://portal.acm.org/citation.cfm?id=173743.173747.

Gu, Y., Jasti, A. C., Jansen, M., and Siefring, J. E. Rhoh, a hematopoietic-specific rho gtpase, regulates proliferation, survival, migration, and engraftment of hematopoietic progenitor cells. *Blood*, 105(4):1467–1475, Feb 2005. doi: 10.1182/blood-2004-04-1604. URL http://dx.doi.org/10.1182/blood-2004-04-1604.

Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D., and Wang, J. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, 23(16):2121–2128, Aug 2007. doi: 10.1093/bioinformatics/btm294. URL http://dx.doi.org/10.1093/bioinformatics/btm294.

Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:S145–S154, 2002.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput*, pages 422–433, 2001.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52, Dec 1999. doi: 10.1038/35011540. URL http://dx.doi.org/10.1038/35011540.

Hengherr, S., Heyer, A. G., Köhler, H.-R., and Schill, R. O. Trehalose and anhydrobiosis in tardigrades - evidence for divergence in responses to dehydration. *FEBS Journal*, 275:281–288, 2008.

Hu, G. and Agarwal, P. Human disease-drug network based on genomic expression profiles. *PLoS One*, 4 (8):e6536, 2009. doi: 10.1371/journal.pone.0006536. URL http://dx.doi.org/10.1371/journal.pone.0006536.

Huber, W. and Gentleman, R. *estrogen: 2x2 factorial design exercise for the Bioconductor short course*, 2006. R package version 1.8.2.

Hummel, J., Strehmel, N., Selbig, J., Walther, D., and Kopka, J. Decision tree supported substructure prediction of metabolites from gc-ms profiles. *Metabolomics*, 6(2):322–333, Jun 2010. doi: 10.1007/s11306-010-0198-7. URL http://dx.doi.org/10.1007/s11306-010-0198-7.

Huttenhower, C., Mutungu, K. T., Indik, N., Yang, W., Schroeder, M., Forman, J. J., Troyanskaya, O. G., and Coller, H. A. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274, Dec 2009. doi: 10.1093/bioinformatics/btp588. URL http://dx.doi.org/10.1093/bioinformatics/btp588.

Hyduke, D. R., Jarboe, L. R., Tran, L. M., Chou, K. J. Y., and Liao, J. C. Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in escherichia coli. *Proc Natl Acad Sci U S A*, 104(20):8484–8489, May 2007. doi: 10.1073/pnas.0610888104. URL http://dx.doi.org/10.1073/pnas.0610888104.

Ideker, T. and Sharan, R. Protein networks in disease. *Genome Res*, 18(4):644–652, Apr 2008. doi: 10.1101/gr.071852.107. URL http://dx.doi.org/10.1101/gr.071852.107.

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

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003. doi: 10.1093/biostatistics/4.2.249. URL http://dx.doi.org/10.1093/biostatistics/4.2.249.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000. doi: 10.1038/35036627. URL http://dx.doi.org/10.1038/35036627.

Jonckheere, A. R. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41:133–145, 1954.

Jönsson, K. I. and Persson, O. Trehalose in three species of desiccation tolerant tardigrades. *The Open Zoology Journal*, 3:1–5, 2010.

Jönsson, K. I. and Schill, R. O. Induction of hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade richtersius coronifer. *Comp Biochem Physiol B Biochem Mol Biol*, 146(4):456–460, Apr 2007. doi: 10.1016/j.cbpb.2006.10.111. URL http://dx.doi.org/10.1016/j.cbpb.2006.10.111.

Junker, B. H. and Schreiber, F. *Analysis of Biological Networks (Wiley Series in Bioinformatics)*. Wiley-Interscience, Mar. 2008. ISBN 0470041447. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASI%N/0470041447.

Juric, D., Lacayo, N. J., Ramsey, M. C., Racevskis, J., Wiernik, P. H., Rowe, J. M., Goldstone, A. H., O'Dwyer, P. J., Paietta, E., and Sikic, B. I. Differential gene expression patterns and interaction networks in bcr-abl-positive and -negative adult acute lymphoblastic leukemias. *J Clin Oncol*, 25 (11):1341–1349, Apr 2007. doi: 10.1200/JCO.2006.09.3534. URL http://dx.doi.org/10.1200/JCO.2006.09.3534.

Kalaev, M., Smoot, M., Ideker, T., and Sharan, R. Networkblast: comparative analysis of protein networks. *Bioinformatics*, 24(4):594–596, Feb 2008. doi: 10.1093/bioinformatics/btm630. URL http://dx.doi.org/10.1093/bioinformatics/btm630.

Kanehisa, M. and Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1): 27–30, Jan 2000.

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., and Caspi, R. Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform*, 11(1):40–79, Jan 2010. doi: 10.1093/bib/bbp043. URL http://dx.doi.org/10.1093/bib/bbp043.

Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. R., and Ideker, T. Pathblast: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32(Web Server issue):W83–W88, Jul 2004. doi: 10.1093/nar/gkh411. URL http://dx.doi.org/10.1093/nar/gkh411.

Kitano, H. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, Mar. 2002. ISSN 1095-9203. doi: 10.1126/science.1069492. URL http://dx.doi.org/10.1126/science.1069492.

Kruskal, W. H. and Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. ISSN 01621459. doi: 10.2307/2280779. URL http://dx.doi.org/10.2307/2280779.

Lee, J. K. *Statistical Bioinformatics: For Biomedical and Life Science Researchers*. Wiley-Blackwell, 2010.

Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, Jul 2006. doi: 10.1093/bioinformatics/btl158. URL http://dx.doi.org/10.1093/bioinformatics/btl158.

Lindgren, W. *Statistical Theory*. Chapman & Hall, New York, 1993.

Ljubić, I., Weiskircher, R., Pferschy, U., Klau, G. W., Mutzel, P., and Fischetti, M. An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem. *Math. Program., Ser. B*, 105(2-3):427–449, 2006.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996. doi: 10.1038/nbt1296-1675. URL http://dx.doi.org/10.1038/nbt1296-1675.

Luedemann, A., Strassburg, K., Erban, A., and Kopka, J. Tagfinder for the quantitative analysis of gas chromatography–mass spectrometry (gc-ms)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737, Mar 2008. doi: 10.1093/bioinformatics/btn023. URL http://dx.doi.org/10.1093/bioinformatics/btn023.

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. Predicting causal effects in large-scale systems from observational data. *Nat Methods*, 7(4):247–248, Apr 2010. doi: 10.1038/nmeth0410-247. URL http://dx.doi.org/10.1038/nmeth0410-247.

Mack, H. and Wolfe, D. K-sample rank tests for umbrella alternatives. *J. Amer. Statist. Ass.*, 76: 175–181, 1981.

Mali, B., Grohme, M. A., Förster, F., Dandekar, T., Schnölzer, M., Reuter, D., Wełnicz, W., Schill, R. O., and Frohme, M. Transcriptome survey of the anhydrobiotic tardigrade milnesium tardigradum in comparison with hypsibius dujardini and richtersius coronifer. *BMC Genomics*, 11:168, 2010. doi: 10.1186/1471-2164-11-168. URL http://dx.doi.org/10.1186/1471-2164-11-168.

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

McGee, B., Schill, R. O., and Tunnacliffe, A. Hydrophilic proteins in invertebrate anhydrobiosis. *Integrative and Comparative Biology*, 44:679–679, 2004.

Merian-Erben. Königsberg 1651, Dec 2006. URL http://www.preussen-chronik.de/_/bild_jsp/key=bild_kathe2.html.

Michal, G. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Wiley-Spektrum, 1 edition, Dec. 1998. ISBN 0471331309. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASI%N/0471331309.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002. doi: 10.1126/science.298.5594.824. URL http://dx.doi.org/10.1126/science.298.5594.824.

Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G. M., Nagini, M., Kumar, G. S. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K. B., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S. K., and Pandey, A. Human protein reference database–2006 update. *Nucleic Acids Res*, 34(Database issue): D411–D414, Jan 2006.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35(Web Server issue):W182–W185, Jul 2007. doi: 10.1093/nar/gkm321. URL http://dx.doi.org/10.1093/nar/gkm321.

Mount, D. W. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2nd edition, Mar. 2004. ISBN 0879696087. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASI%N/0879696087.

Nacu, S., Critchley-Thorne, R., Lee, P., and Holmes, S. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7):850–858, Apr 2007. doi: 10.1093/bioinformatics/btm019. URL http://dx.doi.org/10.1093/bioinformatics/btm019.

Ofran, Y., Yachdav, G., Mozes, E., tsen Soong, T., Nair, R., and Rost, B. Create and assess protein networks through molecular characteristics of individual proteins. *Bioinformatics*, 22(14):e402–e407, Jul 2006. doi: 10.1093/bioinformatics/btl258. URL http://dx.doi.org/10.1093/bioinformatics/btl258.

Parmigiani, G. *The Analysis of Gene Expression Data*. Springer, Berlin, Apr. 2003. ISBN 0387955771. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASI%N/0387955771.

Pounds, S. and Morris, S. W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, Jul 2003.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. Human protein reference database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–D772, Jan 2009. doi: 10.1093/nar/gkn892. URL http://dx.doi.org/10.1093/nar/gkn892.

Qi, Y., Balem, F., Faloutsos, C., Klein-Seetharaman, J., and Bar-Joseph, Z. Protein complex identification by supervised graph local clustering. *Bioinformatics*, 24(13):i250–i258, Jul 2008. doi: 10.1093/bioinformatics/btn164. URL http://dx.doi.org/10.1093/bioinformatics/btn164.

Quackenbush, J. *Nature Reviews Genetics*, June . ISSN 1471-0056. doi: 10.1038/35076576.

Quenouille, M. H. Notes on Bias in Estimation. *Biometrika*, 43(3/4), 1956. URL http://www.jstor.org/stable/2332914.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

Rahm, P. Effect of very low temperatures on the fauna of moss. *Proc K Ned AkadWet Ser C Biol Med Sci*, 23:235–248, 1921.

Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat Appl Genet Mol Biol*, 3:Article16, 2004. doi: 10.2202/1544-6115.1055. URL http://dx.doi.org/10.2202/1544-6115.1055.

Rajagopalan, D. and Agarwal, P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6):788–793, Mar 2005. doi: 10.1093/bioinformatics/bti069. URL http://dx.doi.org/10.1093/bioinformatics/bti069.

Ramlov, H. and Westh, P. Cryptobiosis in the *Eutardigrade Adorybiotus* (Richtersius) coronifer: tolerance to alcohols temperature and de novo protein synthesis. *Zoologischer Anzeiger*, 240:517–523, 2001.

Ravasz, E. and Barabási, A.-L. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(2 Pt 2):026112, Feb 2003.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, Aug 2002. doi: 10.1126/science.1073374. URL http://dx.doi.org/10.1126/science.1073374.

Rebecchi, L., Altiero, T., and Guidetti, R. Anhydrobiosis: the extreme limit of desiccation tolerance. *ISJ*, 4:65–81, 2007.

Reuner, A., Hengherr, S., Mali, B., Förster, F., Arndt, D., Reinhardt, R., Dandekar, T., Frohme, M., Brümmer, F., and Schill, R. O. Stress response in tardigrades: differential gene expression of molecular chaperones. *Cell Stress Chaperones*, 15(4):423–430, Jul 2010. doi: 10.1007/s12192-009-0158-1. URL http://dx.doi.org/10.1007/s12192-009-0158-1.

Rice, K. L., Kees, U. R., and Greene, W. K. Transcriptional regulation of fhl1 by tlx1/hox11 is dosage, cell-type and promoter context-dependent. *Biochem Biophys Res Commun*, 367(3):707–713, Mar 2008. doi: 10.1016/j.bbrc.2007.12.005. URL http://dx.doi.org/10.1016/j.bbrc.2007.12.005.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032, Oct 1999. doi: 10.1038/13732. URL http://dx.doi.org/10.1038/13732.

Rivera, C. G., Vakil, R., and Bader, J. S. Nemo: Network module identification in cytoscape. *BMC Bioinformatics*, 11 Suppl 1:S61, 2010. doi: 10.1186/1471-2105-11-S1-S61. URL http://dx.doi.org/10.1186/1471-2105-11-S1-S61.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Müller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., Staudt, L. M., and Project, L. M. P. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, 346(25):1937–1947, Jun 2002.

Ross, M. E., Zhou, X., Song, G., Shurtleff, S. A., Girtman, K., Williams, W. K., Liu, H.-C., Mahfouz, R., Raimondi, S. C., Lenny, N., Patel, A., and Downing, J. R. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959, Oct 2003. doi: 10.1182/blood-2003-01-0338. URL http://dx.doi.org/10.1182/blood-2003-01-0338.

Savage, K. J., Monti, S., Kutok, J. L., Cattoretti, G., Neuberg, D., de Leval, L., Kurtin, P., Dal Cin, P., Ladd, C., Feuerhake, F., Aguiar, R. C. T., Li, S., Salles, G., Berger, F., Jing, W., and Pinkus, G. The molecular signature of mediastinal large b-cell lymphoma differs from that of other diffuse large b-cell lymphomas and shares features with classical hodgkin lymphoma. *Blood*, 102:3871 – 9, 2003/12/01/ 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12933571.

Schaefer, J., Opgen-Rhein, R., , and Strimmer., K. *GeneNet: Modeling and Inferring Gene Networks*, 2009. URL http://CRAN.R-project.org/package=GeneNet. R package version 1.2.4.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., and Schomburg, D. Brenda, the enzyme information system in 2011. *Nucleic Acids Res*, 39(Database issue):D670–D676, Jan 2011. doi: 10.1093/nar/gkq1089. URL http://dx.doi.org/10.1093/nar/gkq1089.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A*, 93(20): 10614–10619, Oct 1996.

Schill, R. Anhydrobiotic abilities of tardigrades. In Lubzens, E., Cerda, J., and Clark, M., editors, *Dormancy and Resistance in Harsh Environments*, volume 21 of *Topics in Current Genetics*, pages 133–146. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-12421-1. URL http://dx.doi.org/10.1007/978-3-642-12422-8_8.

Schill, R. O., Steinbrück, G. H. B., and Köhler, H.-R. Stress gene (hsp70) sequences and quantitative expression in milnesium tardigradum (tardigrada) during active and cryptobiotic stages. *J Exp Biol*, 207(Pt 10):1607–1613, Apr 2004.

Schlitt, T. and Brazma, A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8 Suppl 6:S9, 2007. doi: 10.1186/1471-2105-8-S6-S9. URL http://dx.doi.org/10.1186/1471-2105-8-S6-S9.

Schokraie, E., Hotz-Wagenblatt, A., Warnken, U., Frohme, M., Dandekar, T., Schill, R. O., and Schnölzer, M. Investigating heat shock proteins of tardigrades in active versus anhydrobiotic state using shotgun proteomics. *Journal of Zoological Systematics and Evolutionary Research*, 49:111–119, 2011. ISSN 1439-0469. doi: 10.1111/j.1439-0469.2010.00608.x. URL http://dx.doi.org/10.1111/j.1439-0469.2010.00608.x.

Scott, J., Ideker, T., Karp, R. M., and Sharan, R. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13(2):133–144, Mar 2006. doi: 10.1089/cmb.2006.13.133. URL http://dx.doi.org/10.1089/cmb.2006.13.133.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, Jun 2003. doi: 10.1038/ng1165. URL http://dx.doi.org/10.1038/ng1165.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003. doi: 10.1101/gr.1239303.

Sharan, R. and Ideker, T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433, Apr 2006. doi: 10.1038/nbt1196. URL http://dx.doi.org/10.1038/nbt1196.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–68, May 2002. doi: 10.1038/ng881. URL http://dx.doi.org/10.1038/ng881.

Smyth, G. limma: Linear Models for Microarray Data. In Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., and Dudoit, S., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, chapter 23, pages 397–420. Springer-Verlag, New York, 2005. ISBN 0-387-25146-4. doi: 10.1007/0-387-29362-0_23. URL http://dx.doi.org/10.1007/0-387-29362-0_23.

Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 (1):Article 3, 2004.

Smyth, G. K., Yang, Y. H., and Speed, T. Statistical issues in cdna microarray data analysis. *Methods Mol Biol*, 224:111–136, 2003. doi: 10.1385/1-59259-364-X:111. URL http://dx.doi.org/10.1385/1-59259-364-X:111.

Sohler, F., Hanisch, D., and Zimmer, R. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521, Jul 2004. doi: 10.1093/bioinformatics/bth112. URL http://dx.doi.org/10.1093/bioinformatics/bth112.

Soong, T.-T., Wrzeszczynski, K. O., and Rost, B. Physical protein-protein interactions predicted from microarrays. *Bioinformatics*, 24(22):2608–2614, Nov 2008. doi: 10.1093/bioinformatics/btn498. URL http://dx.doi.org/10.1093/bioinformatics/btn498.

Spallanzani, L. Opuscoli di fisica animale e vegetabile. In *Modena: Società Tipografica*, page 203–285. 1776.

Strehmel, N., Hummel, J., Erban, A., Strassburg, K., and Kopka, J. Retention index thresholds for compound matching in gc-ms metabolite profiling. *J Chromatogr B Analyt Technol Biomed Life Sci*, 871(2):182–190, Aug 2008. doi: 10.1016/j.jchromb.2008.04.042. URL http://dx.doi.org/10.1016/j.jchromb.2008.04.042.

Tang, X., Wang, J., Liu, B., Li, M., Chen, G., and Pan, Y. A comparison of the functional modules identified from time course and static ppi network data. *BMC Bioinformatics*, 12:339, 2011. doi: 10.1186/1471-2105-12-339. URL http://dx.doi.org/10.1186/1471-2105-12-339.

Terpstra, T. J. The asymptotic normality and consistency of kendall's test against trend, when ties are present in one ranking. *Proc. Kon. Ned. Akad. v. Wetensch.*, 55:327–333, 1952.

Therneau, T., Grambsch, P., and Fleming, T. Martingale-based residuals for survival models. *Biometrika*, 77(1):147, 1990.

Tukey, J. Bias and confidence in not quite large sample. *Ann. Math. Statist.*, 29:614, 1958.

Ulitsky, I. and Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol*, 1:8, 2007. doi: 10.1186/1752-0509-1-8. URL http://dx.doi.org/10.1186/1752-0509-1-8.

Ulitsky, I., Karp, R. M., and Shamir, R. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In *RECOMB'08: Proceedings of the 12th annual international conference on Research in computational molecular biology*, pages 347–359, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 3-540-78838-7, 978-3-540-78838-6.

Ulitsky, I., Krishnamurthy, A., Karp, R. M., and Shamir, R. Degas: de novo discovery of dysregulated pathways in human diseases. *PLoS One*, 5(10):e13367, 2010. doi: 10.1371/journal.pone.0013367. URL http://dx.doi.org/10.1371/journal.pone.0013367.

van Galen, J. C., Muris, J. J. F., Giroth, C. P. E., Vos, W., Ossenkoppele, G. J., Meijer, C. J. L. M., and Oudejans, J. J. Expression of tnf-receptor associated factor 2 correlates with poor progression-free survival time in abc-like primary nodal diffuse large b-cell lymphomas. *Histopathology*, 52 (5):578–584, Apr 2008. doi: 10.1111/j.1365-2559.2008.02970.x. URL http://dx.doi.org/10.1111/j.1365-2559.2008.02970.x.

van Steensel, B., Braunschweig, U., Filion, G. J., Chen, M., van Bemmel, J. G., and Ideker, T. Bayesian network analysis of targeting interactions in chromatin. *Genome Res*, 20(2):190–200, Feb 2010. doi: 10.1101/gr.098822.109. URL http://dx.doi.org/10.1101/gr.098822.109.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.

Verrills, N. M., Walsh, B. J., Cobon, G. S., Hains, P. G., and Kavallaris, M. Proteome analysis of vinca alkaloid response and resistance in acute lymphoblastic leukemia reveals novel cytoskeletal alterations. *J Biol Chem*, 278(46):45082–45093, Nov 2003. doi: 10.1074/jbc.M303378200. URL http://dx.doi.org/10.1074/jbc.M303378200.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002. doi: 10.1038/nature750. URL http://dx.doi.org/10.1038/nature750.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue):D433–D437, Jan 2005. doi: 10.1093/nar/gki005. URL http://dx.doi.org/10.1093/nar/gki005.

Wagner, A. How the global structure of protein interaction networks evolves. *Proc Biol Sci*, 270(1514): 457–466, Mar 2003. doi: 10.1098/rspb.2002.2269. URL http://dx.doi.org/10.1098/rspb.2002.2269.

Wagner, A. and Fell, D. A. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478): 1803–1810, Sep 2001. doi: 10.1098/rspb.2001.1711. URL http://dx.doi.org/10.1098/rspb.2001.1711.

Wang, Z., Gerstein, M., and Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009. doi: 10.1038/nrg2484. URL http://dx.doi.org/10.1038/nrg2484.

Wasserman, L. A. *All of Statistics: A concise course in statistical inference*. Springer, second edition, 2005.

Westh, P. and Ramlov, H. Trehalose accumulation in the tardigrade adorybiotus coronifer during anhydrobiosis. *J Exp Zool*, 258:303–311@ARTICLERamlov2001, author = Ramlov, H. and Westh, P., title = Cryptobiosis in the *Eutardigrade Adorybiotus* (Richtersius) coronifer: tolerance to alcohols temperature and de novo protein synthesis, journal = Zoologischer Anzeiger, year = 2001, volume = 240, pages = 517–523, citeulike–article–id = 422683, keywords = anhydrobiosis, bibtex–import, butanol, ethanol, hexanol, hsp, protein, synthesis, tardigrada, posted–at = 2005–12–05 18:05:09, priority = 2 , 1991.

Westh, P. and Ramløv, H. Cryptobiosis in arctic tardigrades with special attention to the appearance of trehalose. In *Greenland Excursion*. Institute of polar Ecology. Kiel University., 1988.

Wu, X., Guo, J., Zhang, D.-Y., and Lin, K. The properties of hub proteins in a yeast-aggregated cell cycle network and its phase sub-networks. *Proteomics*, 9(20):4812–4824, Oct 2009. doi: 10.1002/pmic.200900053. URL `http://dx.doi.org/10.1002/pmic.200900053`.

Xu, X., Wang, L., and Ding, D. Learning module networks from genome-wide location and expression data. *FEBS Lett*, 578(3):297–304, Dec 2004. doi: 10.1016/j.febslet.2004.11.019. URL `http://dx.doi.org/10.1016/j.febslet.2004.11.019`.

Yang, H., Ke, Y., Wang, J., Tan, Y., Myeni, S. K., Li, D., Shi, Q., Yan, Y., Chen, H., Guo, Z., Yuan, Y., Yang, X., Yang, R., and Du, Z. Insight into bacterial virulence mechanisms against host immune response via the yersinia pestis-human protein-protein interaction network. *Infect Immun*, 79(11):4413–4424, Nov 2011. doi: 10.1128/IAI.05622-11. URL `http://dx.doi.org/10.1128/IAI.05622-11`.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. Normalization for cDNA microarray data. *Microarrays: Optical Technologies and Informatics*, 4266 of Proceedings of SPIE, 2001.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, Feb 2002.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W. E., Naeve, C., Wong, L., and Downing, J. R. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, Mar 2002.

Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., and Vidal, M. Drug-target network. *Nat Biotechnol*, 25(10):1119–1126, Oct 2007. doi: 10.1038/nbt1338. URL `http://dx.doi.org/10.1038/nbt1338`.

Zhang, J. D. and Wiemann, S. Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, 25(11):1470–1471, Jun 2009. doi: 10.1093/bioinformatics/btp167. URL `http://dx.doi.org/10.1093/bioinformatics/btp167`.

Nomenclature

2-DE	two-dimensional gel electrophoresis
ABC DLBCL	activated B-like DLBCL
ALL	acute lymphoblastic (or lymphocytic) leukaemia
AUC	area under the curve
BP	biological process
BUM	beta-uniform mixture model
CA	correspondence analysis
CC	cellular component
ChIP-chip	chromatin immunoprecipitation with chip technology
ChIP-seq	chromatin immunoprecipitation with sequencing
Co-IP	co-immunoprecipitation
Cox PH	Cox proportional hazard
DAG	directed acyclic graph
dhea	district heating algorithm
DLBCL	diffuse large B-cell lymphoma
DREAM	dialogue for reverse engineering assessments and methods
FDR	false discovery rate

FN	false negatives
FP	false positives
GCB DLBCL	germinal center B-like DLBCL
GGM	Gaussian graphical model
GO	gene ontology
heinz	heaviest induced subgraph
HPRD	human protein reference database
ILP	integer linear programming
JT	Jonckheere-Terpstra
KAAS	KEGG automatic annotation server
KEGG	Kyoto encyclopedia of genes and genomes
lowess	locally weighted scatterplot smoothing
Matisse	module analysis via topology of interactions and similarity sets
MF	molecular function
MM	missmatch probes
MS	mass spectrometry
MSS	maximal-scoring subgraph
MST	minimum spanning tree
MWCS	maximum-weight connected subgraph
NHL	non-Hodgkin lymphoma
NMR	nuclear magnetic resonance
PCST	prize-collecting Steiner tree
PM	perfect match probes
PMBL	primary mediastinal B-cell lymphoma

PPI	protein-protein interaction network
RH	relative humidity
RMA	robust multichip average
RNA-Seq	RNA sequencing
ROC	receiver operating characteristic
SAGE	serial analysis of gene expression
SVD	singular value decomposition
TAP-MS	tandem affinity purification followed by mass spectrometry
TC-PIN	time course protein interaction network
TF	transcription factor
TN	true negatives
TP	true positives
vsn	variance stabilising normalisation

Affidavid / Eidesstattliche Erklärung

English:

I hereby confirm that my thesis entitled "Integrated functional analysis of biological networks" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date (Daniela Beisser)

Deutsch:

Hiermit erkläre ich an Eides statt, die Dissertation „Integrierte funktionelle Analyse biologischer Netzwerke" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ort, Datum (Daniela Beisser)

Publications

Beisser, D., Klau, G. W., Dandekar, T., Mueller, T., and Dittrich, M. Bionet: an r-package for the functional analysis of biological networks. *Bioinformatics*, Feb 2010. doi: 10.1093/bioinformatics/btq089. URL http://dx.doi.org/10.1093/bioinformatics/btq089.

Beisser, D., Grohme, M. A., Kopka, J., Frohme, M., Schill, R. O., Hengherr, S., Dandekar, T., Klau, G. W., Dittrich, M., and Müller, T. Integrated pathway modules using time-course metabolic profiles and est data from milnesium tardigradum. In submission.

Beisser, D., Brunkhorst, S., Klau, G. W., Dittrich, M. T., and Müller, T. Robustness and accuracy of functional modules in integrated network analysis. Under review.

Dittrich, M., Beisser, D., Klau, G. W., and Müller, T. Functional modules in protein–protein interaction networks. In Choi, S. and Choi, S., editors, *Systems Biology for Signaling Networks*, volume 1 of *Systems Biology*, pages 353–369. Springer New York, 2010. doi: 10.1007/978-1-4419-5797-9_14. URL http://dx.doi.org/10.1007/978-1-4419-5797-9_14.

Förster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., Schnölzer, M., Frohme, M., Müller, T., Schill, R. O., and Dandekar, T. Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as rna elements in tardigrades. *BMC Genomics*, 10:469, 2009. doi: 10.1186/1471-2164-10-469. URL http://dx.doi.org/10.1186/1471-2164-10-469.

Förster*, F., Beisser*, D., Frohme, M., Schill, R. O., and Dandekar, T. Bioinformatics identifies tardigrade molecular adaptations including the dna-j family and first steps towards dynamical modelling. *Journal of Zoological Systematics and Evolutionary Research*, 49:120–126, 2011. ISSN 1439-0469. doi: 10.1111/j.1439-0469.2010.00609.x. URL http://dx.doi.org/10.1111/j.1439-0469.2010.00609.x.

Förster*, F., Beisser*, D., Grohme, M., Liang, C., Mali, B., Reuner, A., Siegl, A. M., Engelmann, J. C., Shkumatov, A., Schokraie, E., Müller, T., Blaxter, M., Schnölzer, M., Schill, R. O., Frohme, M., and Dandekar, T. The compared transcriptomes of hypsibius dujardini and milnesium tardigradum with rna motifs, encoded proteins, resulting clusters and pathways, tardigrade-specific and general stress adaptations. In preparation.

Liang*, C., Beisser*, D., Förster, F., Kopka, J., Müller, T., Schill, R. O., and Dandekar, T. Milnesium tardigradum metabolic adaptations during active stage to tun transition. In preparation.

* equally contributing first authors