FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Omics profile interpretation on molecular interaction graphs

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der technischen Wissenschaften

eingereicht von

Raul Fechete
Matrikelnummer 0225871

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Univ.-Prof. Dr. Rudolf Freund
Univ.-Doz. Dr. Bernd Mayer

Diese Dissertation haben begutachtet:

(Univ.-Prof. Dr. Rudolf Freund) (Univ.-Doz. Dr. Bernd Mayer)

Wien, 01.10.2012

(Raul Fechete)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Omics profile interpretation on molecular interaction graphs

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der technischen Wissenschaften

by

Raul Fechete
Registration Number 0225871

to the Faculty of Informatics
at the Vienna University of Technology

Advisors:
Univ.-Prof. Dr. Rudolf Freund
Univ.-Doz. Dr. Bernd Mayer

The dissertation has been reviewed by:

(Univ.-Prof. Dr. Rudolf Freund) (Univ.-Doz. Dr. Bernd Mayer)

Wien, 01.10.2012

(Raul Fechete)

Erklärung zur Verfassung der Arbeit

Raul Fechete
Pilgerimgasse 25/13, 1150 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Acknowledgements

I would like to thank my mentors Bernd Mayer and Rudolf Freund for their guidance and expert advice, as well as all my colleagues from emergentec for the great working environment of the past five years.

Thank you.

Abstract

Molecular interaction networks are a core concept in Life Sciences - a field of study with the specific focus on integrative information analysis - and an ideal tool for modeling cellular processes. On the molecular level, cellular processes are the direct cause of phenotype, whether healthy or diseased, and all observable properties of a cell can be traced back to one or more processes.

One of today's main challenges in research is the variability of disease on the process level, meaning similar phenotypes often have different causes. With the advent of the Omics revolution, an enormous amount of data relevant in this context has become available, much of it, however, still pending meaningful interpretation.

Here we demonstrate two approaches to tackle heterogeneity on the process level, based on Omics data integration on molecular interaction graphs. The first one uses a synthetic lethality network to address heterogeneity in cancer, while the second one uses an extended protein-protein interaction network for overcoming variance towards patient stratification.

Our first method demonstrates both in neuroblastoma cell-lines and in human tissue how to find synthetic lethal hubs the knock-down of which would lead to the death of malignant cells. We generalize this method for three additional tumor types and identify relevant hubs including drugs for targeting them.

In our second method we propose a novel interaction network holding validated protein-protein interactions and edges additionally inferred from high quality pathway, ontology and domain data. We use this network to investigate diabetic nephropathy from a clinical perspective based on literature, drug, clinical trial and patent information. Subsequently, we introduce the concept of units towards identifying multi-biomarker panels for patient stratification.

Our results demonstrate that it is possible through information integration to address biological variability issues while at the same improving causative interpretability. We assert that the methods presented in this thesis expand the set of available treatment approaches and will prove in the midterm to be a valuable stepping stone towards Systems Medicine.

Kurzfassung

Molekulare Interaktionsetzwerke stellen ein Kernkonzept in Life Sciences - ein Forschungsgebiet dessen Hauptaugenmerk auf der integrativen Informationsanalyse liegt - dar und sind gleichzeitig ein optimales Werkzeug für die Modellierung von zellulären Prozessen. Auf molekularer Ebene stellen solche Abläufe die Ursache für den Phänotyp dar und alle Merkmale eines Organismus lassen sich zu einem oder mehreren solchen Prozessen zurückverfolgen.

Eine der größten Herausforderungen der heutigen Forschung ist die Variabilität der Erkrankung auf Prozessebene, dh ähnliche Phänotypen haben oft verschiedene Ursachen. Mit dem Advent der Omics Revolution, ist eine Lawine an relevanten Daten verfügbar geworden, denen jedoch zum Großteil noch eine sinnvolle Interpretation fehlt.

In dieser Arbeit werden zwei Ansätze zur Bekämpfung von Heterogenität auf Prozessebene präsentiert, welche auf Omics Datenintegration auf molekularen Interaktionsgraphen basieren. Der erste Ansatz benutzt ein Netzwerk von synthetischen letalen Interaktionen um die Heterogenität in der Krebstherapie in den Griff zu bekommen, während der zweite ein erweitertes Protein-Protein Interaktionsnetzwerk verwendet um biologischer Varianz hinsichtlich Patientenstratifizierung entgegen zu wirken.

Die Ergebnisse der ersten Methode zeigen, dass es möglich ist sowohl in Neuroblastom Zelllinien als auch im menschlichen Gewebe synthetisch letale Hubs zu identifizieren, der Knockdown welcher den Tod maligner Zellen herbeiführen würde. Diese Methode wird anschließend für drei weitere Tumorarten verallgemeinert und relevante Hubs und Medikamente werden identifiziert.

In der zweiten Methode wird ein neues Interaktionsnetzwerk präsentiert, das einerseits validierte Protein-Protein Interaktionen und andererseits aus hochqualitativen Pathway-, Ontologie- und Domänendaten abgeleitete Kanten enthält. Das Netzwerk wird benutzt um die diabetische Nephropathie aus klinischer Sicht zu durchleuchten. Ein Units-Konzept zur Identifikation von Biomarker Kombinationen zwecks Patientenstratifizierung wird abschließend exemplifiziert.

Die Ergebnisse zeigen, dass es durch Informationsintegration möglich ist die biologische Variabilität bei gleichzeitiger Verbesserung der Interpretierbarkeit in den Griff zu bekommen. Wir behaupten, dass die Methoden, die in dieser Arbeit präsentiert wurden, die derzeit verfügbaren Behandlungsansätze erweitern und sich mittelfristig als wertvoller Schritt in Richtung Systemmedizin erweisen werden.

Contents

CHAPTER 1

Introduction

1.1 The Omics revolution: from association to causality

The discovery of the deoxyribonucleic acid (DNA) 1953 by Watson and Crick [1] proved to be a major milestone that would revolutionize biology for the years to come. In time, our understanding of genetics deepened as more became known of the cell. Ribonucleic acids (RNA) and proteins were discovered and the transcription / translation principle of the protein formation was brought to paper.

The initial hypothesis of the DNA being a phenotype (the composite of an organism's observable characteristics) blueprint was validated and with increasing knowledge it became evident that the phenotype was a result of both intrinsic factors such as complex interactions between DNA, RNA, proteins and other compounds in the cell as well as extrinsic, i.e. environmental factors.

The discovery of the link between cellular processes and phenotype opened new avenues in biological research. A disease was no longer a non-descript malfunction but the result of specific process perturbations which could now be acted upon. The best example for this is cancer. The observable phenomenon of cancer is the result of cells with altered processes that no longer act as intended. Hanahan et al. provided in 2000 [2] and in a revised version of manuscript from 2011 [3] an insightful description of the dysfunctional mechanisms necessary for a cell to achieve malignancy. These include among others evading programmed cell death (apoptosis), forced growth of blood vessels (induced angiogenesis), replicative immortality and tissue invasiveness.

While gaining insight into cellular processes allows direct intervention on the causes of the undesired phenotype, the first step, however, still remains locating the malfunction itself. To achieve this, various Omics methods have been developed. The term Omics summarizes a broad spectrum of techniques which measure cell-wide activation of genes, respective proteins, metabolites, etc. Omics allows the quantitative assessment of many thousand cellular components in parallel, and hence provides a system-wide landscape of cellular components being specific e.g. for a tumor cell.

Omics methods can be used e.g. to determine which genes are currently expressed in a cell or which proteins bind to each other. A major drawback of such methods is, however, that the results are mainly discrete snapshots of the cellular status quo with little to no insight into the involved dynamics. Using such data, scientists can attempt to reconstruct the order of events and then validate these hypotheses.

Often when investigating a certain phenotype it is e.g. lists of differentially regulated genes that are the immediate result. Assuming the experiments were performed with sufficient regard to statistical significance, the obtained lists should be characteristic for this particular phenotype, not taking into account variance within the phenotype itself.

While correlation can be determined using such approaches, it is often causality that is mistakingly extrapolated from the data. Put in a nutshell, if two properties A and B of a system are detected simultaneously (correlation) this does not imply e.g. that B is result of A (causality). In a specific example, the process of inflammation is a typical response of the organism to pathogens and features associated with inflammation will often be detected in the context of other afflictions without having a direct connection to the actual cause.

Nevertheless, correlation is a powerful tool for the analysis of cellular processes and a valuable hypothesis source as is shown by Prieto et al. [4] and Yang at el. [5]. The former uses genome-wide expression data to compute a human gene coexpression landscape showing genes having similar regulation. The latter uses coexpression to specifically investigate processes affected by disease.

For an in-depth understanding of the cellular mechanisms, an understanding of the causal relationships linking their components is essential. To achieve this, theoretical models that integrate as much of the presently available knowledge as possible are necessary. This led to the advent of *Systems Biology*.

This new branch of biology has an integrative approach at the core, which, by contrast to the previous reductionist concepts, aims to consolidate the known on the assumption that a system is more than just the sum of its components. The additional properties that arise only through the aggregation of components are called *emergent properties*.

A prototypic representation of multi-level information integration is the interaction network. An emergent property in this context is network robustness, i.e. network resilience to node or edge removal, which for the same number of nodes and edges may vary depending on the underlying topology.

Networks and the integration of information on biological networks mark the scope of this thesis.

1.2 Biological networks: a step towards causality

Networks as the core concept of Systems Biology are a powerful theoretical model for understanding processes and supporting hypothesis generation. Networks can model a wide array of information whether of biological nature such as gene expression correlation, transcriptional regulation, direct protein binding or general scope such as relation direction, strength, order, topology, etc.

The paradigm experienced a strong hype beginning in the early 2000, following Barabasi's

concept of scale-free biological networks [6] and has been a subject of intense scientific debate ever since.

General network concepts

Several characteristics of biological networks are particularly relevant in our context and will be outlined in the following section. The *small world* property is such a characteristic and it states that a random network becomes connected already for a small number of edges (phase transition in the Erdös-Renyi model). Furthermore, each node can be reached from any other by traveling only a small number of edges. This also holds true for biological networks with nodes representing genes and edges representing relations between them.

A further aspect, which was mentioned previously, is network *robustness*. The concept describes the resilience of networks, i.e. the ability to withstand changes with little to no impact on the overall structure. This concept is a direct result of *scale-freedom*. The latter is characterized through an asymmetric node degree distribution, more precisely respecting a power-law distribution, i.e. few nodes with many neighbors, also called hubs, and many node with only few neighbors. An additional restriction imposed by scale-freedom is the inverse relationship between node degree and local clustering, i.e. the neighborhood of a hub is less tightly connected than the neighborhood of a low-degree node.

When putting robustness and scale-freedom in relation to each other, it is easy to see why scale-free graphs are more robust than random ones. For example, when choosing a random node to remove, it is more probable to select one with a low degree, meaning its removal will have little impact on graph cohesion. The same property that makes this type of networks resilient against random attacks, also makes them particularly vulnerable against targeted attacks: deliberately choosing a hub in a scale-free network, will have a much stronger effect than in a random one.

Motifs are specific subnetwork configurations that occur significantly more often than others. Motifs play an important role in biological networks, as they develop dedicated functions through their topology (emergent properties). A typical motif is the feed-forward loop (FFL) in which a gene A regulates the transcription of two other genes B and C. Furthermore, gene B regulates the transcription of C. Assuming gene C needs both A and B to be expressed (AND connection) the motif plays the role of a signal filter. As stoichiometric processes in the cell do not happen instantly, if gene A is up-regulated, gene B will become up-regulated with a certain delay. Gene C will only become active once both A and B are present. If the expression of A was a short-time impulse only, it will disappear before gene B becomes fully up-regulated, so gene C will not be expressed this time. The same principle can be applied the other way around, if A and B share an OR connection, meaning, if A disappears shortly, gene C will not be affected unless enough time passes for B to also become down-regulated.

Other relevant motifs include single input modules (SIM) where one gene regulates several others, dense overlapping regions (DOR) where a set of genes complexly regulates a second one, gene self-regulation, feed-back loops in which a gene inhibits its own expression once a certain level has been reached, etc.

Graph measures often used for characterizing biological networks include the clustering coefficient for investigating edge density, the index of aggregation for determining the size of the largest connected component relative to the size of the whole graph, the node degree distribution

often used in the context of scale-freedom and the characteristic path length for determining the average shortest path length.

Mönks et al. [7] present an overview of the topological characteristics for three different types of biological networks, including metabolic, paralog and physical interaction networks.

Genetic interaction networks

Genetic networks are a class of biological networks in which interactions between genes occur indirectly. The most prominent members in this category are transcriptional regulation and synthetic lethality (synlet) networks. The former describe relations between genes in which the presence of one leads to the expression of the other. The latter describe gene pairs for which the absence of both is lethal while the absence of only one is not.

Synthetic lethality lies within the scope of this thesis. To formalize the concept, let there be two genes A and B and the negative impact (measured as viability loss) of a missing gene be given as $I(A)$ and $I(B)$ respectively, the negative impact of both genes missing is called additive if $I(A, B) = I(A) + I(B)$ or synergistic (synthetically lethal) if $I(A, B) \geq I(A) \cdot I(B)$. The impact of a single, non-critical mutation is generally observable as the organism attempts to compensate for the function loss. For two non-critical (if taken individually) mutations the impact is even stronger, as the organism must now compensate for two missing genes. If the impact of a double mutation is significantly stronger than expected (immediate death is the exceptional case) then synthetic lethality is present.

Synthetic lethality opens a whole new avenue in research and particularly in cancer as tumors can now be investigated for synlet interactions and targeted accordingly.

Large-scale screenings of synthetic lethal interaction pairs have already been done in yeast [8]. Synlet interaction detection in human is not feasible as it is much more difficult to perform: different tissue types, cell immortalization, etc. Translation of pairs from yeast to human is done through gene homology.

Protein interaction networks

Protein interaction networks describe a broad group of networks the members of which are involved in direct interactions. Two important representatives of this group are metabolic networks and direct binding networks. The former describe entire sequences of chemical reactions taking place in the cell while the latter describe binary relationships between actors which form e.g. dimers.

Metabolic networks are used among others for quantitative modeling of processes using e.g. flux balance analysis and petri nets to be described as systems of linear equations and solved using linear programming. In general, protein interaction networks are used for qualitative modeling of biological processes such as A transforms B in C which leads to the expression of D.

The qualitative approach to protein interaction networks lies within the scope of this thesis. Such networks have a balanced rate of true positives to true negatives (an in-depth analysis of such aspects is provided in section 3.1).

Two relevant initiatives for curating protein interaction data are IntAct [9] and BioGRID [10]. The former is an open-source, open data molecular interaction database populated by

data either curated from the literature or from direct data depositions. IntAct contains approximately 275,000 curated binary interaction evidences from over 5,000 publications. The latter, i.e. the Biological General Repository for Interaction Datasets (BioGRID) is a public database that archives and disseminates genetic and protein interaction data from model organisms and humans. BioGRID currently holds over 500,000 interactions curated from both high-throughput datasets and individual focused studies, as derived from over 30,000 publications in the primary literature.

Molecular pathways

Molecular pathways are the epitome of curated knowledge with a distinctive focus on high-quality understanding of cellular processes. They model sequences of events occurring in the cell e.g. in response to an external stimulus. Apoptosis, for example, describes the steps taken by the cell during programmed cell death. Pathways have a low false positive rate due to the stringent review process, while at the same time exhibiting a higher false negative rate.

Pathways are usually classified in generic and disease-specific as processes may vary depending on the cell state.

Several significant pathway databases exist to date. These include the Kyoto Encyclopedia of Genes and Genomes (KEGG) [11] modeling generic and metabolic pathways as well as pathways of disease, PANTHER [12] consisting primarily of signaling pathways and Reactome [13], an expert-authored, peer-reviewed knowledge base of human reactions and pathways that functions as a data mining resource and electronic textbook.

Omics integration approaches

While networks are indeed a core concept in Systems Biology, they gain more expressive power through the integration of additional Omics data.

With an academic background, STRING [14, 15] and its related project STITCH [16] from the European Molecular Biology Laboratory (EMBL) aim at building comprehensive protein interaction networks based on genomic context data, high-throughput experiments, coexpression and literature knowledge as well as protein-small molecule networks based on information on metabolic pathways, crystal structures, binding experiments and drug-target relationships.

In the commercial setting, the Ingenuity Pathway Analysis (IPA) [17] tool is a software that helps researchers to model, analyze, and understand complex biological and chemical systems by integrating data on mRNA, miRNA, proteins, single nucleotide polymorphisms (SNP), metabolites, interactions and pathways as well as providing workflow implementations for biomarker discovery and toxicity screenings.

On a small scale, often networks are constructed and annotated with Omics data based on very specific needs.

In a work by He et al. [18] the authors attempt to identify dysfunctional modules in congenital heart disease by using a protein-protein interaction (PPI) network weighed with gene co-expression scores. They compute shortest paths and use a network flow algorithm to determine genes of interest.

Sengupta et al. [19] extend a PPI network with additional coexpression edges, i.e. they link genes with a Pearson correlation coefficient above a certain cutoff and compute clusters based on node connectivity in an attempt to investigate type 2 diabetes and associated risks.

Alexeyenko et al. [20], delineate a network constructed by inferring edge probabilities using diverse high-throughput data from different organisms with the goal of consolidating and analyzing different interactomes in one network.

1.3 Systems Medicine: are we there yet?

Systems Medicine is a novel concept referring to the use of Systems Biology tools in the context of disease in order to obtain a cell's "wiring diagram", the breakdown of which is ultimately responsible for the emergence of a particular disease phenotype. Relating to networks as a core concept in Systems Biology, Barabasi et al. [21] envision diseases-associated modules and speak of the human *diseasome*. In light of these considerations, the concept of units presented in this thesis is perfectly in-line with the scope of Systems Medicine.

Furthermore, establishing new links between genes, biological functions and a wide range of human diseases has become possible with the availability of increasingly powerful high-throughput technologies, computational tools and integrated knowledge bases, providing signatures of pathological biology and links to clinical research [22].

In a report of the Directorate of Health of the European Commission from June 2010 [23], the following potential actions in Systems Medicine are identified (excerpt):

- Clinical trials - Systems Biology approaches could guide clinical trial design, shortening times and costs.
- Redefinition of clinical phenotypes based on molecular and dynamic parameters.
- Discovery of effective biomarkers of multiple nature for disease progression (clinically useful for risk, prognosis, diagnosis).
- Combinatorial therapy, an approach that would be useful to find out a combination and lower doses of effective drugs, in particular in the case of co-morbidity, where more than one disease is affecting the patient.
- Improvement of drug development (optimized drug efficacy, safety and delivery, timing and dosage of therapy).

Additionally, the concept of Systems Medicine is summarized as follows: "The major challenge is for Systems Biology to contribute to a change in the medical paradigm in order to build the foundation for a prospective medicine that will be predictive, personalized, preventive and participatory."

In Systems Medicine the patient is placed at the center of the system and Bosquet et al. [24] go as far as to suggest patient telemonitoring and the use of predictive markers. Additionally, workflows to determine the most suitable therapeutic strategy should be developed. Initiatives such as these are met with privacy concerns and other challenges in their development as e.g. a

predictive biomarker, depending on its usage, would have to be developed with initially healthy test subjects monitored over a long period of time.

While many aspects of Systems Medicine still need to be investigated, approaches already exist today aiming to use integrative methods for learning about the connection between cellular processes and the manifestation of disease [25]. An immediate consequence of such methods is their connection to clinical pharmacology and drug repositioning, the latter referring to drugs already on the market being demonstrated as effective in single or multiple therapies of other diseases.

To summarize, we can say that Systems Medicine is a promising new concept with the potential to further our understanding of human diseases and with strong implications in the clinical field, that is, however, still in its infancy.

1.4 Thesis scope and goals

This thesis is dedicated to demonstrating that high-quality network integration of Omics supports interpretative causality. As such, this thesis is situated at the confluence of three disciplines: informatics, biology and mathematics.

The scientific output of this thesis contributes to our understanding of human disease mechanisms and intervention modalities, in particular using synthetic lethality and biological network analysis as a stepping stone towards personalized disease treatment and Systems Medicine.

Thesis outline

In the introduction to this thesis we delineated the close relationship between cellular processes as the direct cause of phenotype, Systems Biology as an integrative scientific approach, and networks as a powerful modeling paradigm.

The thesis consists of two main building blocks. The first one is an application of network concepts in the context of synthetic lethality and cancer. Here we will show how network hubs can be identified and used for tackling chemoresistance in tumor.

The second building block introduces a novel network data integration method which we then use to extend biological pathways for investigating diabetic nephropathy. Finally we introduce the concept of units for developing multi-marker profiles towards patient stratification.

The thesis is concluded by a short discussion section.

CHAPTER 2

Genetic interaction networks: synthetic lethality

2.1 Why synthetic lethality?[1]

To better put the subject at hand into context, a short introduction to cancer is necessary. Cancer is among the leading causes of mortality in the industrialized parts of the world. Some 2.9 million cases are diagnosed annually in Europe, with 1.7 million deaths attributed to this class of diseases. The efficacy of anti-cancer therapies is still limited, in a first place due to heterogeneity of cancer limiting efficacy of drugs per se, and second if a drug shows initial efficacy by the emergence of drug resistance.

Improvements in experimental molecular biology in the realm of the Omics revolution have significantly contributed to the understanding of molecular processes involved in cancer [27]. From all these advancements the available data basis for describing molecular processes specific for cancer cells has enormously increased. However, what unfortunately became clear is the significant molecular variance of cancer when comparing different organs affected, but also when analyzing a particular type of cancer, or even speculating that each cancer is to some degree patient specific. Apparently, cancer cells have a multitude of strategies for escaping natural clearance mechanisms. Aberrant cells should inherently follow apoptosis, but take routes to halt this process [28, 29]; cancer cells are at least to some extent identified as such by the patient's immune system, but exhibit immunomodulatory capabilities and other methods for escape [30]; tumors modulate their local environment e.g. triggering nutrition supply via angiogenesis [31]; cancer cells when attacked with e.g. chemotherapeutics develop resistance mechanisms [32]. Obviously, cancer has to be seen as a highly complex system in a dynamic interplay of cell

[1] This section is based largely on the patent application *Critical gene targets for cytotoxic therapy* with the publication no. WO/2011/144738, application no. PCT/EP2011/058259 by Arno Lukas, Johannes Söllner, Andreas Bernthaler, Bernd Mayer, Raul Fechete, Paul Perco, and Andreas Heinzel submitted to the World Intellectual Property Office (WIPO), 2011 [26].

internal survival routes under the selective pressure of cell internal and external death pathways, under the overall control of the body's immune system.

Complex systems in a broad definition involve a large number of elements whose interaction and respective dynamics generates unexpected (emergent) properties. Cancer is a prominent representative of this class: A well-balanced machinery of cell survival and cell death looses corrective measures, resulting in growth of tumor mass and eventually the development of resistance to therapy. Major advancements have been made in managing this disease, and an impressive body of knowledge on key molecular processes characteristic for cancer has been assembled so far. But still, putting the pieces together for reaching an understanding of cancer, being the prerequisite for rational and targeted therapy, remains a major challenge.

Malignant neoplasms, also commonly called cancers, form a group of diseases characterized by uncontrolled growth of cells in higher organisms. Division of cells, invasion of adjacent tissues and metastasis are thereby the three prime indicators for prognosis of disease development.

While many cell types can give rise to cancerous growth, malignant neoplasms can be crudely classified in a number of ways. Based on phenotype, distinction in solid neoplasms (tumours) and non-solid cancers such as leukemias is possible. Other than that, the source tissue or cell-type giving rise to carcinogenesis or the organ where growth occurs in (in case of metastasis) is used to classify cancers.

- Carcinoma: Malignant tumors derived from epithelial cells. This group represents the most common cancers, including the common forms of breast, prostate, lung and colon cancer.
- Sarcoma: Malignant tumors derived from connective tissue, or mesenchymal cells.
- Lymphoma and leukemia: Malignancies derived from hematopoietic (blood-forming) cells.
- Germ cell tumor: Tumors derived from totipotent cells. In adults most often found in the testicle and ovary.
- Blastic tumor or blastoma: A tumor (usually malignant) which resembles an immature or embryonic tissue. These tumors are mostly seen in children.

Aside of cancers also other malignancies suffer from declining efficacy of therapies through development of biological resistance. Among those infectious diseases are paramount as they contribute significantly to mortality and particularly morbidity rates, particularly in developing or non-industrialised parts of the world. In this regard, Malaria, Tuberculosis and HIV are also known as the three big killers. There is a great need for robust cytotoxic therapies avoiding or reducing development of resistances.

Cytotoxic therapy usually is directed against a specific target. Among the targeted immunotherapies there is for instance, passive immunotherapy, e.g. using monoclonal antibodies, or active immunotherapy employing vaccines. Effectivity or applicability of any particular therapy can vary greatly between types of cancer or based on presentation in individual patients. As a consequence differentiation between patients may be beneficial in many cases; however molecular biomarkers to make adequate decisions are often not available yet. One of the exceptions is

the Her2/neu monoclonal antibody therapy (marketed as Herceptin and Trastuzumab) which can greatly benefit from preemptive diagnostic tests for the expression status of the HER-2 protein as the factor is found in only approximately 25-30% of breast cancer patients at sufficient levels. Also, within a particular type of therapy differentiation by cancer type is usually necessary. For example, the choice of chemotherapeutic agent can be influenced by gender of the patient, principal effectivity on certain cancerous cells and, related to that, rate of degradation of drugs in cancerous tissues [33]. These factors may vary between cancer types as well as between patients afflicted with the same class of cancer.

Targeted chemotherapy often relies on cytostatic (cell proliferation inhibiting) chemicals which are toxic or detrimental for dividing or metabolically highly active cells and may therefore exhibit severe side-effects on tissues exhibiting naturally high proliferation rates. For drug examples, plant alcaloids, such as the Vinca alkaloids Vincristine and Vinblastine, inhibit microtubule function and therefore prevent chromosomal segregation leading to mitotic arrest. These drugs are effective or otherwise indicated only in a limited number of neoplastic disorders, however, specifically neuroblastoma, lymphomas and leukemias in the case of Vincristine. Similarly, immunotherapies such as the activation of innate and adaptive immune responses by interferon therapy are applicable only against a limited spectrum of neoplastic disorders, always additionally including a potential patient bias. Many more chemotherapeutics are in use and implicated for certain cancers in individual or combination therapies.

In contrast to classical chemotherapy, rational drug design and screening procedures have enabled certain targeted therapies by influencing activity of factors of specific relevance in cancers, sometimes allowing improved efficacy and side effect profiles as in the case of Imatinib in chronic myelogenous leukemia and a number of other cancers. Targeted therapies are thereby defined by a specific, defined molecular target commonly found in a class of cancer which is amenable to drugs. In contrast to standard chemotherapies it is possible to focus on molecular changes elementary to cancerogenesis or progression which is the basis for the mentioned improvement in effect / side effect profiles (whereas e.g. inference with microtubule formation by vincristine affects all cells of the body). While these developments are cause to hope, they suffer from some of the same problems as standard chemotherapy, namely the development of resistance against the drug stemming from mutations/alterations after onset of therapy or alternatively from genetic/epigenetic heterogeneity within treated tumors to begin with. Alternatively, only a certain subset of cancers may initially feature the drug target in adequate concentrations, as described for Herceptin before. Heterogeneity within cancerous populations of one individual, including cancer stem-cells, and between patients is the biological basis for reduced drug efficacy in the first place followed by resistance. To add robustness (avoidance of disease recurrence) to therapy therefore has to be a central goal in drug screening.

A major shortcoming of cancer therapies and certain other areas of therapy including but not limited to infectious diseases and potentially disorders related to somatic mutations is the development of resistance to those therapies. In the case of cancer and infectious diseases such resistance is mediated by development of pathogen or cancer cell populations generally resulting in diseased cells, tissues or individuals not or less affected by therapy when compared to the original population of cells, tissues or patients. This variability and potential to experience deterioration of therapy efficacy seemingly inherent to many drugged biological systems

can be seen both between individuals (for example when comparing between neoplastic tissues/growths/isolates stemming from at least two individuals) as well as within individuals (for example genetically/epigenetically/phenotypically different cell populations within one neoplastic tissue/growth/isolate). Therapies are therefore only of limited value if they cannot avoid recurrence of a disease in a form resistant to the original therapy or provide only limited efficacy in a substantial fraction of patients. Unfortunately this is a common situation when treating infectious diseases and cancer and potentially other diseases.

The concepts of synthetic lethality and synthetic sickness refer to the lethal or sickening effect of elimination of function of two specific genes or molecules encoded by these genes (for example by genetic knock-outs, natural or induced genetic or epigenetic mutation or alteration, small interfering RNAs or interfering drug effects) on a single or multicellular organism, if elimination of such function of either of the two genes is not lethal [34].

Recently, comprehensive high-throughput data of synthetically lethal genetic double knockouts have been published for the yeast Saccharomyces cerevisiae [35].

Synthetic lethality could be a promising new avenue in the attempt to overcome chemoresistance in the setting where one of the two partners is missing in tumor and the second one can be down-knocked by means of drug-intervention. In theory, this would render the cancerous cells, and only these, unviable. The purpose of the work presented in this chapter is to investigate this new approach in the specific setting of chemoresistant neuroblastoma. Subsequently, this method will be extended and presented in a more generic context of tumor.

2.2 A synlet-based approach to tackling vincristine resistant neuroblastoma[2]

Concept outline and goals

Neuroblastoma is a tumor of the peripheral nervous system and the most common extracranial solid tumor of childhood [37, 38]. This disease is one of the few malignancies where spontaneous regression is reported, however, depending on stage and biological characteristics (e.g. MYCN amplification) chemotherapy is indicated [39, 40]. Drugs shown to have substantial efficacy include anthracyclines (doxorubicin), platinum compounds (carboplatin), etoposide, vincristine, and alkylating agents as cyclophosphamide [37]. One of the major clinical issues in chemotherapy is drug resistance [41]. Resistance to a specific drug is either intrinsic [42], where the patient does not respond to therapy in the first place, or acquired, where the patient becomes non-responsive in the course of treatment [43]. Molecular mechanisms responsible for resistance are manifold including drug transport, compensatory processes leading to increased drug metabolism, or elimination/modification of the drug target, among others [44, 45].

Vinca alkaloids such as vincristine act as mitotic inhibitors by binding to tubulin and halting polymerization, consequently hampering microtubule assembly resulting in cell cycle arrest

[2]This section is based largely on the scientific article *Synthetic lethal hubs associated with vincristine resistant neuroblastoma* by Raul Fechete, Susanne Barth, Tsviya Olender, Andreea Munteanu, Andreas Bernthaler, Aron Inger, Paul Perco, Arno Lukas, Doron Lancet, Jindrich Cinatl, Martin Michaelis and Bernd Mayer published in Molecular Biosystems, January 2011 [36].

in the mitosis metaphase [46]. One phenomenological representation of vincristine resistance is changes in the cytoskeleton structure [47, 48]. Neuroblastoma is susceptible to developing drug resistance involving various specific mechanisms such as loss of p53 function [49], down-regulation of ERK1/2 phosphorylation [50], or up-regulation of p-glycoprotein (ABCB1) [51]. The clinical implications of drug resistance are further augmented by an increased invasiveness of the chemoresistant tumor cells. The particular aggressiveness of chemoresistant tumor cells has been described both in the context of neuroblastoma and vincristine [52], as well as in a more general context [53].

Mechanisms triggering intrinsic as well as acquired vincristine resistance are coupled to the genomic state of the parent tumor cell, and as for other tumors, neuroblastoma shows considerable heterogeneity impacting clinical outcome. Factors of relevance include copy number variation, chromosome gains, and changes in differential gene expression [54–56]. Options for tackling chemoresistant neuroblastoma in clinical practice are limited; various drugs as bortezomib [32], artesunate [57] or 131I-metaiodobenzylguanidine [58] have been evaluated in vitro as well as in vivo. One strategy is usage of drugs whose mode of action differs fundamentally from vincristine, thereby increasing the change for circumventing given vincristine resistance mechanisms.

Heterogeneity in solid tumors regarding genomic status, associated expression profile and eventual routes towards developing resistance under selective pressure of chemotherapeutics certainly impose a general problem for virtually any targeted therapy in cancer. In this context, the application of the concept of synthetic lethality has been proposed. Synthetic lethality describes a cellular condition in which two (or more) non-allelic and non-essential mutations being not lethal on their own, become lethal when present within the same cell [59]. However, each single mutation may show synthetic sickness, but the effect of the double mutation is synergistic and not merely additive [60, 61]. Systematic analysis has been performed in yeast and C.elegans to determine synthetically lethal gene pairs [8, 62–65].

Synthetic lethality has been pursued substantially in cancer research [34, 66–68], as this approach promises increased sensitivity for cancer cells while at the same time providing a broader therapeutic window. Only cells showing a particular mutation are specifically affected when applying a drug addressing a single lethal partner to this mutated entity. One of the most prominent examples resting on synthetic lethal interactions is PARP (poly (adenosine diphosphate-ribose) polymerase) inhibitors coupled with BRCA1 or BRCA2 mutation, all acting compensatory in repair of ssDNA breaks. Clinical trials utilizing PARP inhibitors in BRCA associated cancers as e.g. done for ovarian cancer indeed reported decreased adverse effects and at the same time anti-tumoral activity [69]. On the basis of population-based case-control studies, the estimated risk of being diagnosed with ovarian cancer increases by a factor of about 15 when carrying a BRCA1 mutation [70]. However, the prevalence of BRCA1 mutations in subjects diagnosed with ovarian cancer is estimated to be only about 5% [71], indicating a successful therapy based on BRCA1-associated synthetic lethality only for a fraction of patients We delineate from this example that drugs implicating a synthetic lethal mode of action may offer less side effect prone therapy with improved specificity, but still being hampered in broad efficacy (which we in the following denote as “coverage”): only patients whose tumor is characterized by a specific mutation (as for BRCA) are amenable by a drug addressing a synthetic lethal partner, all other tumors

are not covered.

To tackle the issue of coverage of e.g. neuroblastoma in general independent of the genomic state found in individual patients, the concept of protein hubs may be of value [72]. Protein interaction networks show a scale free topology, i.e. the degree distribution (where the degree indicates the number of interactions per protein node) follows a power law. This general principle appears also to be true for networks of synthetic lethal interactions [73]: Most genes may have none or a single synthetic lethal partner, whereas a small set of hubs may have many such interactions. If a synthetic lethal hub is hampered in its function, all cells where at least one out of the n synthetic lethal partners of a specific hub is mutated will be affected. Utilizing this principle may increase the coverage of cancerous genotypes, as schematically depicted in Fig. 2.1.

A general setup for this concept includes a reference cell population holding as phenotype e.g. “healthy” or “chemosensitive”, and n tumor cell populations holding as phenotype “cancerous” or “chemoresistant”. This setting refers to the assumption of a stable reference genotype, but numerous different cancer genotypes. We assume that each of these tumor cell populations holds at least one mutation of a non-essential gene (or generally feature), A1, A2, An. If we now identify and functionally interfer with a non-essential feature B being synthetic lethal for all combinations of mutated features Ai (B-A1, B-A2, B-An) synthetic lethality will result for all cancerous (or chemoresistant) cell populations, but at the same time leaving the reference cell population unharmed (as functionally hampered B on its own is non-essential). We denote such a feature B as a synthetic lethal hub.

This concept may now be expanded to a clinical setting with a reference gene expression profile characterizing features for the healthy (or chemosensitive) state, and individual gene expression profiles of tumors from individual patients. Having such profiles in hand allows identification of synthetic lethal hubs B according to the procedure given in Fig. 2.1. Features B are to be not differentially regulated when comparing reference and tumor samples, and have to be in a synthetic lethal interaction with partners A, where within the total set of features A at least one feature A’ is down-regulated in each single patient expression profile. If such a feature B could be identified, maximum coverage (efficacy with respect to the heterogeneous tumor cell populations) of a drug directed against the feature B would be achieved, as synthetic lethality would be induced for each individual cancer entity (patient specific tumor).

We in the following exemplify this concept for vincristine resistant neuroblastoma utilizing gene expression profiles from vincristine sensitive and resistant cell lines, and expand our findings of synthetic lethal hub candidates B to gene expression profiles derived from human neuroblastoma tissue samples.

Materials and methods

Gene expression data sets

Transcriptomics data sets characterizing vincristine sensitive and resistant cell lines, and data sets derived from neuroblastoma patient samples were used for analyzing vincristine resistance mechanisms and for delineating synthetic lethal hub proteins. Cell line data were retrieved for UKF-NB2 (isolated by the Interdisciplinary Laboratory for Tumour and Virus Re-

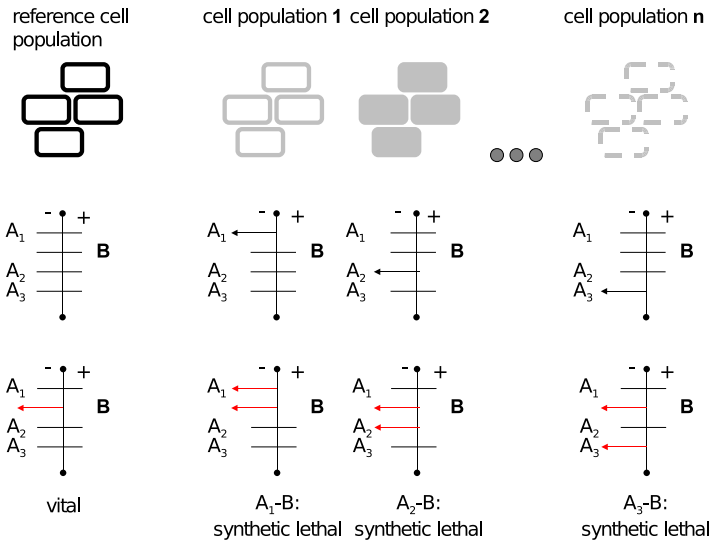

Figure 2.1: Scheme for exemplifying the concept of profile specific synthetic lethality in target identification. Top layer: Start situation is a reference cell population (normal, drug sensitive) and various cell populations (1, 2, .. , n) being malignant (or drug resistant). Middle layer: For each cell population specific gene expression is given for the features A1, A2, B, .. , An. Bottom layer: Down-regulation of features in the different cell populations. Down-regulation of (non-essential) B only has no effect on cell viability of the reference cell population. Joint down-regulation of [B, A1], [B,A2] or [B,An] results in synthetic lethality for the cell populations 1, 2 and n.

search [52, 74, 75]), UKF-NB3 (isolated by the Interdisciplinary Laboratory for Tumour and Virus Research [32, 52, 76]), UKF-NB6 (isolated at the Interdisciplinary Laboratory for Tumor and Virus Research, Klinikum der J.W. Goethe-Universität Frankfurt), SMS-KAN (received from CP Reynolds, Children's Hospital Los Angeles [77, 78]), SMS-KCN (received from CP Reynolds, Children's Hospital Los Angeles [77–79]), and IMR-32 (received from the American Type Culture Collection ATCC [80]). For each cell line, a vincristine sensitive and a vincristine resistant phenotype was available for expression profiling. Adaption of the vincristine sensitive cell lines towards rendering them resistant was done as described in [31].

The expression profile for each sensitive and resistant cell line was determined in triplicate using Applied Biosystems 1700 chemoluminescence microarrays (Applied Biosystems, Foster City, USA) holding 32,878 probes, with 25,123 probes associated to 21,012 gene symbols. Data preprocessing included removal of flagged features according to the manufacturer specifications, triplicates were averaged, and quantile-quantile normalization was performed. Identification of differentially regulated features was done using the ABarray package (Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microarray (AB1700) gene expression data) from R/bioconductor applying a FDR of 5%.

Two sets of features were derived and considered as relevant for characterizing resistance on

the level of cell lines. The first set included genes being significantly differentially regulated in a group comparison of the six sensitive and six resistant cell lines, resulting in 43 genes (21 up- and 22 down-regulated). The second set included features showing a fold change of at least five in at least one particular cell line when comparing sensitive and resistant phenotype for each cell line separately. The number of features retrieved by this second procedure was 834, with 437 genes being down-regulated and 397 features being up-regulated in at least one out of the six cell lines.

Expression data on neuroblastoma retrieved from patient tumor tissue was drawn from the ArrayExpress database (link [81]) [82] selecting the study E-TABM-38 [83]. This data set included 251 neuroblastoma patient profiles generated by using a customized array (A-MEXP-255) holding 10,707 probes corresponding to 9,878 genes. Next to the expression data, ArrayExpress provided clinical data on outcome for the 251 patients, allowing an assignment of 180 samples to "good outcome" defined as "no relapse, no death" in a follow of in the mean 60 months, and 71 samples to "bad outcome" with either relapse or death in a follow up time period of in the mean 40 months. Based on this group definition, significance of differential regulation was derived using a Student's t-test with Benjamini and Hochberg correction for multiple testing, yielding in total 2,001 genes (471 up-regulated and 1530 down-regulated) as differentially expressed when comparing "good outcome" and "bad outcome".

Synthetic lethal interactions

Synthetic lethal interaction data in *S.cerevisiae* were derived from DryGin [65]. DryGin provides data on genetic interactions for 1,712 query genes tested against 3,885 array genes resulting in interaction data for more than 6 million pairs. For each pair, an ϵ-value describing the difference between the measured and the expected double mutant population fitness is provided. A negative ϵ-value indicates synthetic sickness/lethality whereas a positive value indicates synthetic enhancement. DryGin also provide thresholds for statistical significance of sickness / enhancement based on the false positives / false negative rates, suggesting $|\epsilon| > 0.8$ at a given significance level of $p < 0.05$ for relevant interactions, among which interactions with negative ϵ-value were in the following considered as synthetic lethal.

Applying these thresholds provided 121,391 pairs showing significant synthetic lethality in *S.cerevisiae*, involving in total 4,358 unique yeast genes. For this set of unique yeast genes, a mapping to human genes was performed for deriving orthologs. This was achieved using the GeneCards (link [84]) list of human-yeast orthologs [85] as generated by integrating SGD (link [86]) and homologene (link [87]). Applying this procedure allowed a mapping of 2,052 yeast genes out of the 4,358 genes represented in the synthetic lethality data set to 3,130 human orthologs (one-to-many cardinality of the yeast to human matching for a certain number of genes). For this set of human genes a total of 150,877 synthetic lethal interactions were derived and subsequently used.

Data on gene essentiality was obtained from the Mouse Genome Database (MGD [88]) by querying the phenotype text codes of the genes possessing a human ortholog for the word "lethal". This procedure yielded a set of 2,214 human genes considered as single lethal.

Function-specific data sets

For analyzing the involvement of proteins associated with actin, and more specifically with cytoskeleton and transport in vincristine mode of action and resistance mechanisms the totality of genes associated with these categories was derived. The GeneCards database was queried for "actin" in the "pathways", "function" and "summaries" section of the database resulting in 724 genes holding NCBI gene symbols. Of these actin associated features, 119 were specifically associated with cytoskeleton and 57 with transport.

Results

Features and processes involved in vincristine resistance

Analysis of the gene expression data sets available for vincristine sensitive and resistant neuroblastoma cell lines provided 43 features being significantly differentially regulated when comparing the six cell lines in their vincristine sensitive and resistant phenotype (in the following denoted as "cell line consensus"), and 834 features showing a fold change of at least five in at least one out of the six cell lines when comparing the sensitive and resistant status for each individual cell line (in the following denoted as "cell line fold change"). Furthermore, 2,001 features were identified in human tumor tissue when comparing samples from patients with disease free survival and samples from patients experiencing tumor recurrence (in the following refered to as "human tissue feature set"). Genes associated with actin biology are present in all feature sets, specifically 3 features were found in the cell line consensus, 35 in the cell line fold change, and 112 in the human tissue feature set.

Certainly, both the in-vitro as well as in-vivo case-control studies resemble only approximations of vincristine resistance in the clinical setting. The cell lines explicitly address changes associated with adaption to vincristine; however, the changes do not necessarily mirror resistance mechanisms also found in the in-vivo situation. In-vivo data on the other hand do not explicitly reflect the development of resistance as endpoint, but the finding of disease progression is used as a surrogate marker for the potential development of resistance in these patients. In light of these considerations the validity of the given expression profiles with respect to chemoresistance, as well as the relevance of the cell line models with respect to neuroblastoma in the human (clinical) setting needs to be addressed in the first place.

For assessing the comparability of cell line and human tissue data the overlap of identified features on the level of NCBI gene symbols was retrieved as presented in Tab. 2.1 and 2.2.

data sets	up	down	total
C	21	22	43
B	437	397	834
A	471	1530	2001

Table 2.1: Number of features being up- or down-regulated in the cell line consensus (C), fold change (B), and human tissue (A).

	B,C	A,C	A,B	A,B,C
feature overlap, up-regulation	3	0	12	0
feature overlap, down-regulation	8	3	50	1

Table 2.2: Number of overlapping features when comparing the three feature lists A, B, C.

One feature, namely PTBP2 (polypyrimidine tract binding protein 2) was identified in all feature lists. PTBP2 plays a role in pre-mRNA splicing and in the regulation of translation, and is primarily expressed in brain. Various reports indicate the relevance of PTBP2 in mesothelioma, glioblastoma and other tumors of the central nervous system, also linked to tumor cell dissemination, metastasis and chemoresistance. Additional 62 features were found differentially regulated in both, patient and cell line data.

Various meta-studies aimed at comparing the direct overlap of features identified in independent gene expression studies for a given cancer were published, frequently identifying only minor overlap [30, 89], but still the same molecular processes and pathways appeared as populated by identified features. Analysis of significant enrichment of features in particular pathways (Gene Set Enrichment Analysis, GSEA [90]) using cell line and human tissue data was assessed by using a Chi^2 test comparing the number of differentially regulated features located in a specific pathway and the total number of genes assigned to this pathway. Enrichment analysis utilizing cell line and human tissue data on the level of KEGG pathways provided two cancer pathways and axon guidance as significantly enriched by features from both sample sources. Furthermore, distinct sets of pathways were identified as enriched by either only the human tissue (including cell cycle and replication) and the cell line features (like adhesion processes), respectively. The complete list of significantly affected pathways is presented in Tab. 2.3.

For the consensus expression set holding features being differentially regulated when comparing all six sensitive with the six resistant counterparts only 43 features were identified. In GSEA only one pathway came close to significance using this consensus data set, namely “ABC transporters” with a p-value of 0.07.

Also on the level of pathways the involvement of actin associated mechanisms became evident. Fig. 2.2 provides a schematic view of the intersection of the KEGG pathways “focal adhesion” (KEGG identifier hsa04510) and “pathways in cancer” (KEGG identifier hsa05200). Features identified as differentially regulated for sensitive and resistant cell lines affected ECM- and chemokine-chemokine receptor interactions leading, for both pathways, to the activation of the protein tyrosine kinase 2 (FAK) as well as the phosphoinositide-3-kinase (PI3K). Via paxilin the former can lead to the regulation of the actin cytoskeleton, while the latter leads to cell survival and proliferation via the p21 protein (Cdc42/Rac)-activated kinase 4 (PAK). Strong indications for the involvement of differentially regulated features in cell survival are provided by v-akt murine thymoma viral oncogene homolog 2 (PKB/AKT2) which together with the cyclin-dependent kinase inhibitor 1A and B (P21/P27), as well as p53 were found affected in their expression levels.

Next to identifying the overlap of cell line and human tissue data on the level of individual features and molecular pathways, the specific involvement of given features in chemoresistance and neuroblastoma may support the relevance of the data sets in the context of neuroblastoma

vincristine resistance in the clinical setting.

data set	KEGG id	pathway name	total	found	p-val
A,B	hsa05200	Pathways in cancer	329	92/27	<0.0001/0.0010
A,B	hsa05212	Pancreatic cancer	72	22/8	0.0049/0.0296
A,B	hsa04360	Axon guidance	129	31/14	0.0282/0.0026
A	hsa04110	Cell cycle	113	44	<0,001
A	hsa03030	DNA replication	36	18	<0,001
A	hsa05222	Small cell lung cancer	86	30	<0,001
A	hsa05016	Huntington's disease	181	51	<0,001
A	hsa03420	Nucleotide excision repair	44	19	<0,001
B	hsa04512	ECM-receptor interaction	84	14	<0,001
B	hsa04510	Focal adhesion	201	17	0,009
B	hsa05218	Melanoma	71	8	0,028
B	hsa04514	Cell adh. molecules (CAMs)	131	11	0,046
B	hsa00565	Ether lipid metabolism	33	5	0,047

Table 2.3: KEGG pathways being significantly affected in vincristine resistance. Given is the data set (A for human tissue, B for the fold change list derived from cell lines), the KEGG pathway id and name, the total number of features on the pathway (total), the number of differentially regulated features on this pathway (found), and the p-value of a Chi^2 test comparing the number of present and the number of affected features for a specific pathway.

To address this issue, each feature was queried in NCBI PubMed together with the MeSH terms "drug resistance, neoplasm" and "neuroblastoma" for deriving already reported association of the features with these terms. The number of citations of a certain gene in the context of drug resistance and neuroblastoma was retrieved by invoking the ESearch tool provided by the NCBI Entrez Programming Utilities tool set (link [91]). Results are presented in Tab. 2.4.

Joint occurrence of resistance-associated features involved in both cell line and human tissue is given among others for JUN, KIT, BID, RET, CDKN2A, GAL and ABCC4. Features well known in resistance such as ABCB1 (N-glycoprotein, described as a major active transport mechanism for drug resistance [92,93]) or MYC (where patients with amplified MYCN are known to have a poorer prognosis [94]) were only identified for either cell line or human system.

For further evaluating the significance of these findings, the PubMed database was queried for drug resistance and neuroblastoma association for all 15,547 genes assigned in PubMed, resulting in 2,004 genes with at least one reference to "drug resistance" and 1,987 genes with at least one reference to "neuroblastoma".

From the double standard deviation retrieved from the distribution of feature-to-term assignments as computed for the reference genes, all features from the cell line and human tissue lists with at least 36 references for neuroblastoma or at least 53 references for drug resistance can be considered significant in the respective category (Tab. 2.4).

To evaluate the enrichment of the terms "neuroblastoma" and "resistance" for the cell line and human data feature sets a Fisher's exact test was performed. Each of the three feature lists

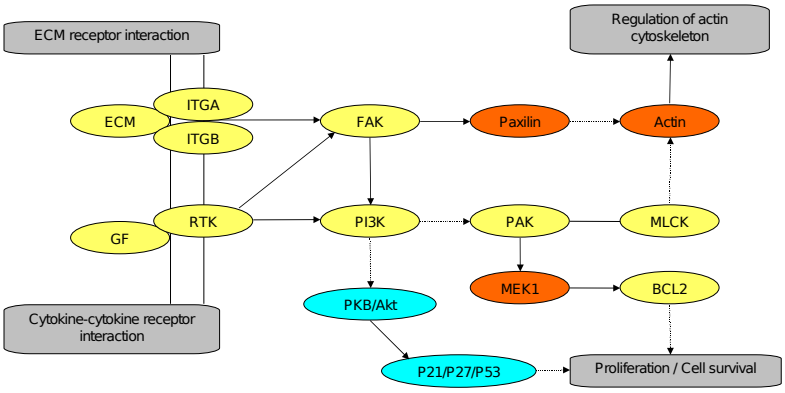

Figure 2.2: Reconstruction of pathway segments (utilizing the KEGG pathways "focal adhesion" and "pathways in cancer") including features identified in the gene expression experiments comparing sensitive and resistant cell lines. Features being differentially regulated and actin associated are marked in yellow, actin associated features not differentially regulated are marked in orange, features being differentially regulated but not actin associated are marked in blue.

was split into groups holding either genes with no reference or genes with at least one reference in the respective MeSH category. The number of genes assigned to these groups was tested against the respective number of features per group derived for the reference set of 15,547 genes. All three gene sets showed statistically significant enrichment ($p < 0.05$) in both neuroblastoma and drug resistance.

Analysis results from feature comparison, gene set enrichment and literature association indicated the validity of the expression data sets in the context of neuroblastoma chemoresistance. Nevertheless, the substantial heterogeneity and complexity of the cell line and human profiles also became evident. This is true on both levels, namely genotype (associated expression profiles of analyzed tumor cells and tissue samples) as well as processes responsible for embedding vincristine resistance. This aspect also becomes clear when analyzing the cell line expression profiles. Fig. 2.3 provides the result of hierarchical clustering of the expression profiles derived from the six sensitive and resistant cell lines.

The cell lines, although on the phenotype level clearly assignable to being vincristine sensitive or resistant, cluster by cell type. The difference in expression status of the individual sensitive cell lines exceeded changes introduced by rendering the cell resistant to vincristine. This finding is in line with the comparably low number of features (43) identified as differentially regulated in the cell line consensus compared to the 834 features being significantly affected in at least one cell line when comparing sensitive and resistant phenotype. For these 834 features only 15% were found affected in two, and only 1% in three cell lines. One single feature was found up-regulated in all six cell lines (ABCB1), consequently also being a member of the consensus feature list.

These results support the arguments in favor of a synthetic lethality approach as an attempt to target different features in different samples for reflecting this heterogeneity rather than targeting specific features resting on a putative "consensus" over all samples.

cell lines				human tumor tissue			
gene sym.	**DR&NB**	**DR**	**NB**	**gene sym.**	**DR&NB**	**DR**	**NB**
ABCB1	27	87*	2505*	MYC	20	1152*	191*
JUN	1	132*	149*	MYCN	23	755*	23
KIT	1	52	201*	CA2	0	697*	39*
CDKN1A	3	57*	163*	TNF	3	131*	417*
TNFRSF10B	2	23	147*	BAX	5	156*	379*
VIP	0	96*	6	ERBB2	1	23	436*
BID	1	21	78*	ACHE	1	271*	146*
HTR3A	0	96*	0	CASP8	8	118*	269*
RET	0	67*	8	CASP9	8	107*	197*
CDKN2A	0	31	38*	**JUN**	1	132*	149*
PDGFRA	0	9	60*	BCL2L1	2	54*	205*
ANXA5	0	28	38*	KITLG	2	99*	158*
NMB	0	42*	9	**KIT**	1	52	201*
GEM	0	4	38*	SRC	0	88*	160*
GAL	0	31	10	DES	0	135*	96*
ATM	0	11	29	**BID**	1	21	78*
CXCR4	1	20	19	RET	0	67*	8
LYN	0	3	35	**CDKN2A**	0	31	38*
ABCC4	1	1	32	**GAL**	0	31	10
ALK	0	24	8	**ABCC4**	1	1	32

Table 2.4: Features (only the top hits are listed) being differentially regulated in the cell lines (left) and in human tumor tissue (right) and assigned to drug resistance and/or neuroblastoma. Given are the gene symbol and the number of reports in PubMed associated with a given gene based on the search terms "drug resistance, neoplasm and neuroblastoma" (DR&NB), "drug resistance, neoplasm" (DR), and "neuroblastoma" (NB). Features identified in both data sets are given in bold. Numbers holding an asterisk are computed as statistically significant for a given category.

Identification of synthetic lethal hubs

Gene expression profiles for six vincristine sensitive cell lines and the corresponding resistant phenotypes including 25,123 features associated to 21,012 genes were available as start point. Each expression value was given in triplicate, and the average was calculated for each feature and each cell line with a given resistance phenotype. For each feature and cell line the fold change comparing the sensitive and resistant phenotype was computed. Fold change values x found below 1.0 were re-computed as -1/x. Genes being represented on the array by more than one probe were consolidated on this fold change level by calculating the median fold change over all probes.

One important criterion for identifying synthetic lethal hubs was their validity in all six cell lines (see Fig. 2.1). To be valid, the fold changes of a potential synthetic lethal hub (B) had to

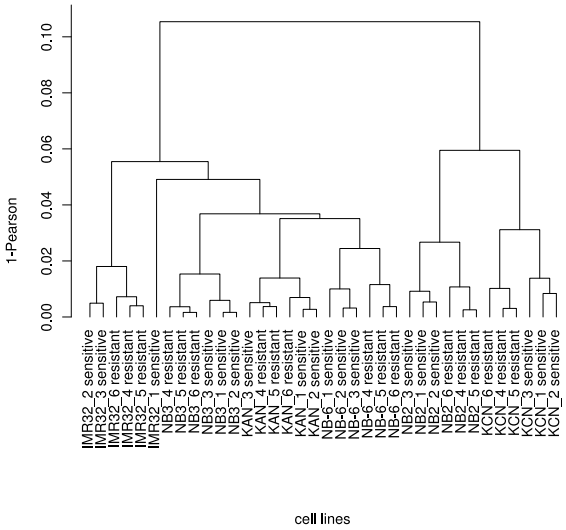

Figure 2.3: Hierarchical clustering of cell line expression profiles applying complete linkage and Pearson correlation as distance measure. The expression profiles for six sensitive and six resistant cell lines, each in triplicate, are shown.

be above a specific fold change threshold in each of the six cell lines. Ideally, this fold change should be positive for all cell lines (i.e. a feature B is up-regulated in the resistant cell line). We selected a minimum fold change of -2.0 as valid cutoff for respecting some variability of the array signals as identified by the triplicate measurements, but still assuring that a given B is present (a natural requirement for enabling a subsequent knock-down of a feature B).

As second criterion, a partner for a given B was considered as valid from the fold change perspective if this partner A showed a fold change below -3.0 in at least one resistant cell line, i.e. being substantially down-regulated in the resistant situation for at least one cell line. Next to these specific fold change requirements, a third constraint was imposed on the selection of A and B features, namely A and B had to be a synthetic lethal pair as determined on the basis of the yeast-to-human matching of synthetic lethal partners identified in *S.cerevisiae*. Applying these criteria on the six cell lines resulted in the selection of 83 features (synthetic lethal hub genes B) having down-regulated synthetic lethal partners A assigned in all six cell lines, consequently providing 100% cell line coverage. For each selected B, there was at least one A for each cell line which i) was down-regulated in the resistant phenotype in at least one (or more) cell lines and ii) B and A were assigned as synthetic lethal partners following the yeast data.

Due to the evident importance of actin associated features in the context of vincristine resistance, only synthetic lethal hubs functionally assigned to this category were further considered. Among the 83 features 13 were also present in the actin feature list (Tab. 2.5).

For these 13 hub features, 94 synthetic lethal partners (A's) were identified, the majority of which was found to be associated with several hub features. Further details for the synthetic lethal partners A are listed in Tab. 2.6. Eight of the 94 genes are also found to be actin associated. For further characterizing the set of 94 partners, GSEA was performed on the level of Gene

gene sym.	name	min FC	max FC	avg FC	# of A's
ACTA1	Actin, alpha 1, skeletal muscle	-1,71	1,96	0,07	37
ACTB	Actin, beta	-1,73	1,96	0,09	37
ACTG1	Actin, gamma 1	-1,72	1,24	-0,24	37
ACTR1B	Arp1 actin-related protein 1 homolog b, centractin beta	-1,15	1,47	-0,26	18
ARPC2	Actin related protein 2/3 complex, subunit 2, 34kda	-1,78	1,22	-0,07	21
ARPC5	Actin related protein 2/3 complex, subunit 5, 16kda	-1,44	1,39	0,39	19
ARPC5L	Actin related protein 2/3 complex, subunit 5-like	-1,22	1,45	-0,31	19
FMN1	Formin 1	-1,38	1,16	-0,04	12
FMN2	Formin 2	-1,72	2,25	0,45	12
MAP2K1	Mitogen-activated protein kinase kinase 1	-1,28	1,52	0,46	19
MAP2K2	Mitogen-activated protein kinase kinase 2	-1,42	2,06	0,08	19
PKN1	Protein kinase n1	-1,59	2,57	0,17	20
PPP1CA	Protein phosphatase 1, catalytic subunit, alpha isoform	-1,34	1,66	0,48	19

Table 2.5: The list of actin associated synthetic lethal hubs covering all six cell lines. Given are gene symbol and name, the minimum (min FC), maximum (max FC) and average fold change (avg FC) when comparing sensitive and resistant cell lines, as well as the number of synthetic lethal partners A (# of A's).

Ontology (GO) terms [95] as depicted in Tab. 2.7.

In this procedure, each GO term was analyzed for the ratio of the number of A's found in the term with respect to the total number of genes assigned to the term. A Fisher's exact test was used to calculate the significance of enrichment using Benjamini and Hochberg correction for multiple testing, further eliminating all terms with a corrected p-value above 0.05. Next, all enriched terms holding only one single hub feature were removed leading to the selection of 34 significant ontology concepts as listed in Tab. 2.7. Next to actin associated terms (which were expected due to the biased selection of B features), an enrichment of terms related to cell motion, metabolism and regulation of MAP kinase activity was found.

For further analyzing the relation between synthetic lethal hubs and synthetic lethal partners, the connection between the two GO concept lists, one for the hubs and one for their partners, was studied. The first group held terms enriched by the hub gene list for which the uncorrected p-value of Fisher's exact test was below 0.05 and which had at least 3 hub genes. The second group refered to terms enriched by the partner gene list for which the uncorrected p-value of Fisher's exact test was below 0.05 and which held at least 5 partner genes.

gene sym.	max FC	avg FC	#CL	#B's	gene sym.	max FC	avg FC	#CL	#B's
ABCC4	-3,98	-1,83	1	1	OSBPL1A	-3,59	0,12	1	3
ACSS1	-20,20	-3,96	1	3	PCCA	-4,11	-0,83	1	3
ACSS2	-4,93	-0,48	1	3	PEX5	-3,59	-1,24	1	3
ADA	-4,23	-0,86	1	3	PEX5L	-3,13	-0,80	1	3
AGPAT3	-4,73	-0,49	1	1	PLEKHG2	-5,22	-0,21	1	1
AP1S3	-3,56	-0,99	1	1	PPAN	-3,14	-0,37	1	4
AQP10	-12,03	-2,34	1	5	PPEF1	-3,97	1,71	1	3
ARHGAP29	-7,00	-2,45	3	11	PSAT1	-11,00	-2,23	2	1
ARHGEF4	-3,43	-1,99	2	1	PSPH	-4,14	0,21	1	1
C12ORF5	-12,40	-1,54	1	5	RAB25	-26,11	-4,33	1	1
C8ORF30A	-4,65	-2,00	1	1	RAB35	-3,76	-0,78	1	1
CLPB	-4,38	-0,46	1	1	RAB6B	-3,42	-1,36	1	3
CORO1C	-4,43	-0,06	1	7	RER1	-3,25	-0,05	1	5
CORO6	-10,71	-1,99	1	7	RPS4Y1	-4,27	36,88	1	3
CPVL	-3,72	-0,91	1	3	SC5DL	-3,34	-0,85	1	5
CSTF2T	-3,46	0,31	1	2	SCPEP1	-4,39	-0,54	1	3
DAPK1	-26,50	-4,82	1	1	SDHB	-3,07	0,05	1	3
DBNL	-6,56	-0,59	1	7	SELI	-3,59	-1,64	1	2
DCTN1	-4,47	-0,08	1	5	SLC25A29	-3,52	-0,76	1	1
DMXL1	-5,62	0,45	1	1	SLC26A8	-3,64	0,08	1	2
DNM1L	-3,30	-1,47	1	5	SLC2A10	-5,20	-1,28	1	1
DPH2	-3,38	-0,57	1	2	SNX10	-9,47	1,33	2	1
EPS15	-4,09	0,08	1	2	SNX12	-4,77	-0,97	1	1
EPS15L1	-3,11	0,25	1	2	SNX13	-3,26	-0,13	1	1
FOXB1	-3,20	0,18	1	1	SORCS1	-3,59	0,32	1	2
FOXJ1	-3,91	-0,93	1	3	SPAG1	-3,05	-0,96	1	2
FOXK2	-3,78	-0,25	1	1	SRCAP	-3,49	0,42	1	4
MTOR	-3,01	0,41	1	5	STX7	-26,21	-4,15	1	2
FYCO1	-3,12	-0,69	1	1	TCEA1	-3,16	-0,30	1	3
GPT2	-4,11	-0,01	1	6	TCEA3	-3,74	-0,90	1	3
GRAMD1A	-3,02	-0,66	1	3	TEAD2	-43,23	-10,57	2	5
HCLS1	-72,05	-11,43	1	7	TKTL1	-21,91	2,60	1	5
HHAT	-3,64	-0,55	1	3	TMLHE	-3,64	-1,34	1	1
HIST1H2AI	-3,38	-1,90	1	3	**TPM2**	-22,79	-3,21	1	4
HK2	-3,41	-0,52	1	2	TSSK4	-3,25	-1,14	1	5
HSPH1	-3,64	-1,08	1	3	TTL	-4,26	-1,11	1	2
IER3IP1	-3,64	-0,20	1	2	UBB	-731,34	-120,92	1	6
LOC147804	-3,15	-0,17	1	4	UBE4B	-3,11	0,25	1	2
LY6G5B	-5,98	-0,28	1	6	ULK2	-10,31	-2,01	1	4
MAPK4	-9,20	-0,52	1	9	USP27X	-3,81	-1,16	1	4
MAPRE2	-3,16	-0,04	2	2	VAC14	-3,13	-0,17	1	3
MYO1C	-3,17	-0,60	1	6	VAMP5	-7,55	0,04	1	6
MYO5A	-4,66	-1,42	1	3	VAMP8	-3,44	-1,18	1	6
NOTCH2	-8,45	-1,29	1	2	VPS13D	-4,19	-0,63	1	1
NUAK1	-7,03	-2,73	2	1	WIPI1	-41,69	-6,88	1	1
NUAK2	-9,46	-2,46	1	1	YARS	-3,09	0,29	1	1
ONECUT1	-5,80	-0,94	1	2	ZNF22	-66,09	-10,51	1	6

Table 2.6: Synthetic lethal partners (A's), with maximum (max FC) and average (avg FC) down-regulation fold change, number of resistant cell lines with down-regulation (#CL), and the number of synthetic lethal hubs B (#B's) associated. Actin relevance is marked in bold.

term ID	term name	size	hubs	p-val.	c. p-val.
GO:0016043	cellular component organization	2229	7	0,008	0,039
GO:0006996	organelle organization	1184	6	0,002	0,019
GO:0030036	actin cytoskeleton organization	217	5	<0.001	0,001
GO:0030029	actin filament-based process	232	5	<0.001	0,001
GO:0007010	cytoskeleton organization	386	5	<0.001	0,002
GO:0006928	cell component movement	420	5	<0.001	0,002
GO:0032268	regulation of cellular protein metabolic process	458	4	0,001	0,017
GO:0051246	regulation of protein metabolic process	531	4	0,002	0,020
GO:0006796	phosphate metabolic process	848	4	0,010	0,039
GO:0006793	phosphorus metabolic process	848	4	0,010	0,039
GO:0030833	regulation of actin filament polymerization	49	3	<0.001	0,001
GO:0008064	regulation of actin polymerization or depolymeriz.	56	3	<0.001	0,001
GO:0032271	regulation of protein polymerization	57	3	<0.001	0,001
GO:0030832	regulation of actin filament length	58	3	<0.001	0,001
GO:0043254	regulation of protein complex assembly	78	3	<0.001	0,002
GO:0032535	regulation of cellular component size	79	3	<0.001	0,002
GO:0032956	regulation of actin cytoskeleton organization	81	3	<0.001	0,002
GO:0032970	regulation of actin filament-based process	84	3	<0.001	0,002
GO:0090066	regulation of anatomical structure size	117	3	<0.001	0,004
GO:0051493	regulation of cytoskeleton organization	118	3	<0.001	0,004
GO:0044087	regulation of cellular component biogenesis	123	3	<0.001	0,004
GO:0033043	regulation of organelle organization	200	3	0,001	0,016
GO:0051128	regulation of cellular component organization	429	3	0,008	0,039
GO:0051656	establishment of organelle localization	42	2	0,001	0,013
GO:0000187	activation of MAPK activity	77	2	0,002	0,020
GO:0043406	positive regulation of MAP kinase activity	96	2	0,004	0,025
GO:0007265	Ras protein signal transduction	99	2	0,004	0,026
GO:0048812	neuron projection morphogenesis	103	2	0,004	0,028
GO:0048858	cell projection morphogenesis	110	2	0,005	0,030
GO:0032990	cell part morphogenesis	119	2	0,006	0,033
GO:0010038	response to metal ion	126	2	0,006	0,035
GO:0043405	regulation of MAP kinase activity	134	2	0,007	0,037
GO:0010035	response to inorganic substance	145	2	0,008	0,039
GO:0048545	response to steroid hormone stimulus	179	2	0,012	0,047
GO:0032989	cellular component morphogenesis	182	2	0,013	0,047

Table 2.7: GO terms associated with the synthetic lethal hubs. For each significant term the GO identifier (term id), name, and the total number of genes in that particular term as well as the number of hubs for the particular term are given. Additionally the raw p-value of the Fisher's exact test, as well as the values after correction for multiple testing (Benjamini & Hochberg) are given for comparing total number of genes per term and genes affected in a specific term.

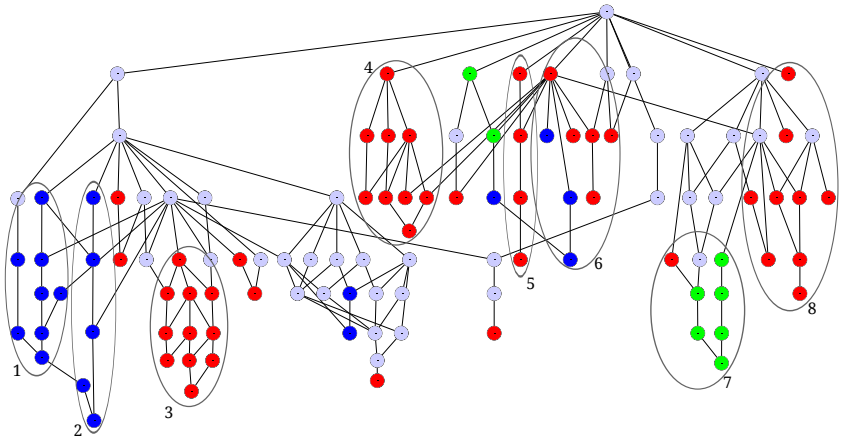

Figure 2.4: Section of the Gene Ontology populated by synthetic lethal hubs B (blue), synthetic lethal partners A (red), and terms populated by both (green). The main term groups are: 1: regulation of actin cytoskeleton organization, 2: regulation of actin filament polymerization, 3: positive and negative regulation of apoptosis, 4: establishment of localization and (vesicle mediated) transport, 5: apoptosis, 6: various cellular processes (cycle, development, etc.), 7: post translational modifications / phosphorylation and 8: various metabolic processes (lipid, organic acid, ketone, alcohol, etc.).

Term selection was done by considering the ontology as a directed acyclic graph and walking its structure bottom-up (in the direction of the "is a" relationship up to the root concept) beginning in each of the 27 (hub B) and 54 (partner A) significant terms. All concepts reached and edges used during this walk in the acyclic graph were selected. The resulting graph is presented in a hierarchical layout in Fig. 2.4.

This approach allowed identification of terms being strongly affected by the synthetic lethal pairs including, among others, cytoskeleton organization and actin filament polymerization for the synthetic lethal hubs (B), apoptosis and the regulation thereof, transport and various metabolic processes for the partners (A), and post-translational modifications and phosphorylation for both (B and A features).

This ontology analysis clearly showed disjoint areas regarding hubs and partners. Synthetic lethal interactions were mostly found between different ontology terms rather than term-internal. Relevant between-term interactions and their weight (i.e. number of synthetic lethal pairs associated with the terms) are shown in Fig. 2.5.

Particularly strong interactions were found between hub-associated terms and the terms transport, organelle organization, post translational modifications and regulation of cell communication.

Relevance of hubs in tumor tissue

For assessing the relevance of synthetic lethal hubs B delineated from the cell line data, a respective coverage analysis was performed using the expression data characterizing human tissue. The

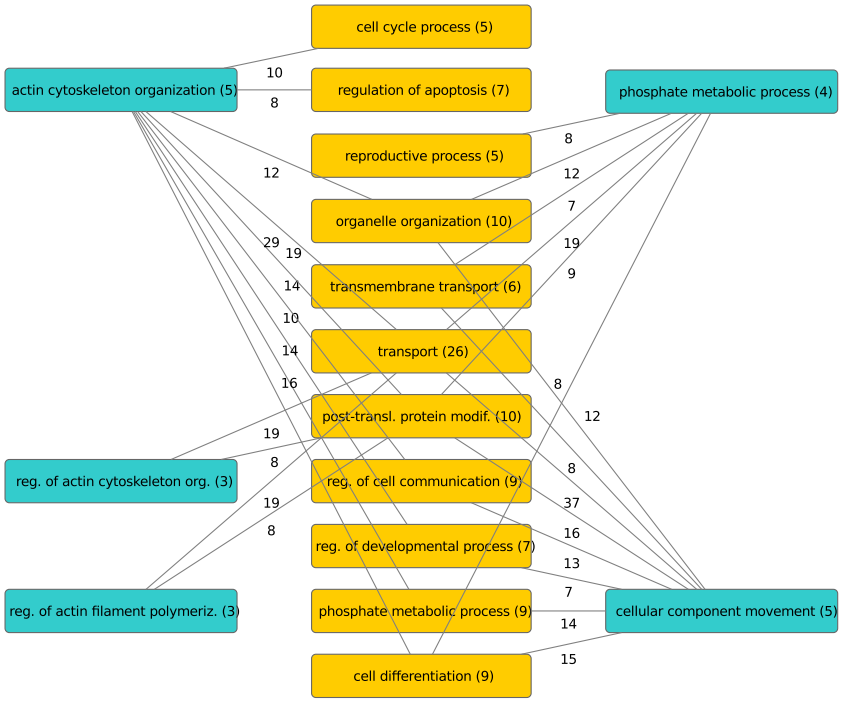

Figure 2.5: Reconstruction of dependencies between lethal hubs B (blue) and their partners A (orange) on the level of GO processes. The number of synthetic lethal links between terms associated with synthetic lethal hubs and partners are given.

study by Oberthür et al. [83] provided 251 gene expression profiles. Based on the given clinical data on outcome, 180 arrays were assigned as "good outcome" controls and 71 arrays as "bad outcome" cases (34 samples with tumor relapse and 37 samples with death).

Equivalently to the procedure applied for cell line data, the expression values of genes being represented on the arrays by more than one probe were consolidated computing the median over all probes. This reduced the 10,163 probes (with 10,156 features holding a gene symbol) to 9,878 genes. In contrast to the cell line data, the human data set was composed of two groups of samples based on clinical outcome instead of having a sensitive and a resistant phenotype for each individual sample. For respecting this situation the median expression value for each gene including the 180 controls ("good outcome") was computed, and for each individual case ("bad outcome") the fold change for each gene was computed with respect to the median value derived from the 180 control expression values. Again the rationale is that a hub feature B ideally had to be present in all cases and controls, whereas a particular partner A had to be be down-regulated in one (or many, ideally all) cases. In this setting, down-regulation of a feature A was expressed as fold change of A in a particular case with respect to the median of A as determined from the controls. This setup allowed the computation of the coverage of hub features B in analogy to the cell line procedure presented above.

Fold change threshold values were adjusted for allowing comparability to the cell line se-

lection procedures. Specifically, cutoff values were selected such that the same percentage of presence (B features) and down-regulation (A features) were available in both, cell line and human tissue experiments: A fold change cutoff of -3.0 (down-regulation for A features) for the cell line data resulted in selecting 2.285% of down-regulated features on the cell line level.

For reaching the same number of features for the human tissue samples a corresponding fold change of -1.7696 was computed. A fold change constraint of at least -2.0 (presence) on the cell line data, selected as minimum fold change for B features, yielded 92.5% features, corresponding to a cutoff of -1.3667 in the patient data. Based on these cutoff values, the coverage of the 13 hub features B identified on the cell line level was tested on the human tissue data by first identifying partners A fulfilling the fold change criterion, and then selecting B - A pairs reported as being synthetic lethal as derived from the yeast data set.

This procedure could be applied for only 6 out of the 13 hub genes, as the remaining seven features were not represented on the arrays used for analyzing the human samples. Characteristics on coverage, minimum, maximum and average fold changes, and the number of associated synthetically lethal partners for the six hubs are presented in Tab. 2.8.

gene sym.	coverage	min FC	max FC	avg FC	# of A's
ACTA1	59,15%	-1,53	6,63	-0,43	25
ACTG1	57,75%	-2,37	1,9	0,34	25
ACTB	56,34%	-1,55	1,46	-0,07	25
MAP2K2	52,11%	-1,74	1,45	0,5	21
ACTR1B	38,03%	-1,74	1,61	-0,12	15
ARPC5	35,21%	-1,44	1,77	-0,07	15

Table 2.8: List of actin associated synthetic lethal hubs B also found in the human tumor tissue data set. The table lists gene symbol, coverage of human samples considered, minimum (min FC), maximum (max FC) and average fold change (avg FC) when comparing the case expression values to the median control expression value, as well as the number of synthetic lethal partners A. Genes identified as statistically significant regarding their coverage are given in bold.

For the six synthetic lethal hubs the best coverage is reached by ACTA1 (59.15%) and the lowest by ARPC5 (35.21%). For evaluating the significance of these findings, a reference distribution of coverage was computed for all 733 genes from the yeast synthetic lethality data set which were successfully mapped from yeast to human and which were also present on the custom array used. Based on this distribution of coverage, a mean, single and double standard deviation was computed, and the coverage reached by the six hub features derived from the cell line data was analyzed in the context of these distribution descriptors. The result is shown in Fig. 2.6.

Four of the lethal hubs (namely ACTA1, ACTG1, ACTB and MAP2K2) were found to be highly significant with respect to coverage, i.e. outside the double standard deviation of the underlying reference distribution. The remaining two (ACTR1B and ARPC5) were significant being outside the single standard deviation.

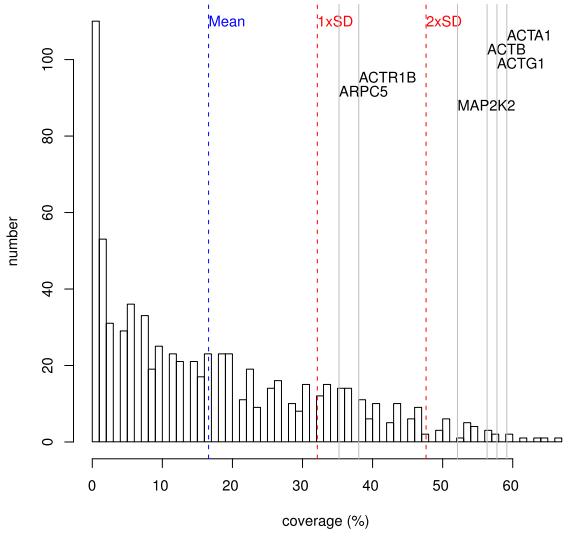

Figure 2.6: Synthetic lethal hub validation in human tumor tissue data. The histogram of the case coverage of all available synthetic lethal features as derived from the yeast data set forms the distribution given by the bars, showing the number of features (number) with a specific coverage. Characteristics of the distribution are provided (mean, single and double standard deviation). Coverage of selected synthetic lethal hubs as derived from the cell lines and their gene symbol are provided.

Discussion and conclusion

Models of resistance in neuroblastoma

We explored two model systems for characterizing vincristine resistance in neuroblastoma. The cell line model explicitly addressed vincristine resistance by comparing six sensitive (parental) cell lines and their counterparts rendered resistant. The human tissue samples on the other hand only serve as surrogate for drug resistant neuroblastoma in the in vivo situation. Disease progression certainly rests on various reasons, with resistance to chemotherapy just being one (also highly prevalent) cause.

Analysis of differentially regulated features and affected molecular processes shed some light on the relevance of the two model systems used. Direct overlap of features being differentially regulated when comparing cell line and human data with respect to resistance was found minor (Tab. 2.1 and 2.2), including 62 features. Of further notice in this context is the low number of features found consistently differentially regulated on the cell line level when comparing sensitive and resistant phenotype, i.e. only 43 features. However, gene expression meta-studies done for other tumor entities also reported little feature overlap [30, 89], and it was postulated that on a functional level (molecular processes and pathways) the factual overlap becomes more evident.

Identified pathways significantly populated by both differentially regulated cell line and human tissue features included pathways in cancer as well as a process linked to the tissue type

under analysis (axon guidance) (Tab. 2.3). Pathways prominently enriched in the in vivo situation included cell cycle, DNA replication and repair, in line with mechanisms linked to expanding tumor mass. The in-vitro data on the other hand predominantly showed extracellular matrix and adhesion processes as being affected, presumably indicating the specific situation of the cell cultures.

Analyzing the 62 features identified as differentially regulated in both cell line and human tissue again provided "pathways in cancer" as significantly affected, next to "endocytosis". The latter functional entity has been associated with vincristine since its introduction to the field [96, 97], and recently has been associated with increase in lysosomal volume and leakage and sensitization of cells to lysosomal membrane permeabilization [98]. Screening for vincristine resistance associated features e.g. in gastric cancer also revealed the involvement of endocytosis [99].

Clear reference to actin-associated features is found for both model systems, in line with the assumption of the pivotal role of cytoskeleton and transport in vincristine resistance. For features derived from cell lines 35 genes, and for the human tissue 112 genes were represented in the actin list holding in total 724 genes. Specific features were also identified in ECM receptor interaction leading to regulation of actin cytoskeleton as well as cell proliferation (Fig. 2.2). Cytoskeletal involvement was broadly discussed in the context of invasive migration of neoplastic cells [100–103] clearly linking the development of resistance and the clinical finding of increased aggressiveness. Interestingly, these reports pinpoint the relevance of focal adhesion, cell adhesion molecules and extracellular matrix/ECM interactions, being in line with pathways found to be enriched by features differentiating drug sensitive and resistant cell lines.

Further evidence supporting the relevance of the given feature lists with respect to drug resistance mechanisms was derived from mining literature (Tab. 2.4). Association analysis of the approx. 15,000 genes assigned in PubMed resulted in a significant enrichment of drug resistance associated features for both, cell line as well as human data. Interestingly, ABCB1 (p-glycoprotein), a well known transporter associated with drug resistance [104], was found differentially regulated only in the cell lines and not in patient profiles. Another gene highly ranked in the literature annotation but not differentially regulated in either the patient or the cell lines is caspase 3 (CASP3). Contrary to ABCB1, CASP3 is not a direct target for triggering drug resistance but a downstream element, as the activation of caspases plays a central role in apoptosis.

In summary, the given feature lists as well as enriched pathways populated by these grasp major elements associated with drug resistance. However, also substantial heterogeneity becomes evident, not only when comparing cell line and human tissue data, but also when analyzing the cell line data as such. Clustering the expression profiles characterizing the six chemosensitive and resistant cell lines provided a closer linkage on the level of cell type than on the level of sensitive versus resistant phenotype (Fig. 2.3). Apparently no set of features (or single feature) could be identified as target for tackling vincristine resistance. ABCB1 has been considered as such a target based on its diverse involvement in drug resistance [105], and L-type calcium channel blockers such as verapamil were shown to increase chemosensitivity [106, 107]. A recent study reported the reversal of docetaxel- and vincristine-induced multidrug resistance also independent of ABCB1 expression in human lung cancer cell lines [108].

For addressing the heterogeneity of drug escape routes which appear fundamentally grounded on the heterogeneity of primary neuroblastoma, we utilized the concept of synthetic lethality. In theory, this approach promises improved specificity for targeting cancer cells, exhibiting at the same time an improved safety profile.

Synthetic lethal hubs and vincristine resistant neuroblastoma

Extended data regarding genetic interactions was published recently on *S.cerevisiae*, showing a dense network of interactions involving both enhancement as well as synthetic sickness/lethality. Conservation of such genetic interactions, and consequently the validity of traversing data on synthetic lethality between species is heavily discussed [109] with first comparative studies ongoing involving *S.cerevisiae*, *S.pombe*, and *C.elegans* [110]. In the presented example, about 150,000 synthetic lethal interactions were derived for human genes utilizing available data from *S.cerevisiae*.

This interaction data set was then integrated with expression data given for the six individual cell lines. Primary selection rationale was the identification of a feature B which is present (i.e. expressed) in all sensitive as well as resistant cell lines, and interacts with one or more features A which are absent (weakly expressed or absent) in resistant cell lines. A coverage of 100% is reached by a feature B if there is at least one particular A in each individual resistant cell line. If B is non-essential, then a knock-down / knock-out of B in all cell lines would result in no (or minor) change of viability for sensitive, but in synthetic lethality for resistant cells (Fig. 2.1). In line with this approach a mutation (as found in the definition of the concept of synthetic lethality) is set equivalent to down-regulation of a feature as both, mutation as well as down-regulation are assumed to result in decreased effective activity of a feature.

Thresholds of fold changes have to be introduced for implementing this approach for defining at what level a feature is considered as "present" and as "absent". Instead of using fixed cutoff values and imposing the constraint of presence for all B features (as applied in this work) a dynamic procedure could be applied by searching optimal presence of B and maximum down-regulation of respective features A. Such alternative approach would also allow selection of suboptimal B features with respect to presence in all samples, but still providing better coverage by selecting alternative A features.

For our data set on vincristine resistance, 83 synthetic lethal hubs could be identified interacting with partners A in all six resistant cell lines (thus providing 100% coverage). 13 out of these 83 features are found to be actin associated (Tab. 2.5), and GSEA of these 83 hubs identified "regulation of actin cytoskeleton" as significantly enriched. This pathway bias is solely coming from the gene expression data, again indicating the important role of actin associated processes in vincristine resistance. The hubs showed an average fold change of zero when comparing sensitive and resistant phenotype, and a substantial number of synthetic lethal partners A, ranging from 12 to 37. The 13 hub features are in total associated with 94 A features, of which 87 are down-regulated (mutated) in only one cell line (Tab. 2.6), six in two, and one in three cell lines (Rho-type GTPase-activating protein, ARHGAP29). Of importance is the redundancy in interactions, where each A is in most cases linked to more than one B, for features as ARHGAP29 even 11 out of the 13 hubs interact. Of further interest is the mean fold change of features A, being in many cases around zero (as also given for B features, see Tab. 2.5),

but showing substantial down-regulation for individual cell lines. Apparently, gain in resistance severely impacts specific features depending on the genotype of the parental, sensitive cell.

One mechanism causing synthetic lethal association between two features is (functional) paralogs, i.e. one feature compensates the loss of the second feature. PARP and BRCA may serve as example. However, redundancy is frequently encoded by more complex mechanisms spanning functional categories [62]. Next to actin-associated terms coming from the synthetic lethal hubs (Tab. 2.7) transport, apoptosis, cell cycle and development, and metabolic processes are found for the combined set of hubs and interacting partners (Fig. 2.4). The five major functional terms given for the synthetic lethal hubs couple with 11 major functional terms found for partner features A. The functional interaction scheme given in Fig. 2.5 clearly indicates degeneracy, i.e. distributed robustness [111] as major mechanism for proposed synthetic lethal interactions and not functional compensation on the basis of redundancy.

Of the 13 synthetic lethal hub features discussed above, 6 were present on a custom array used for profiling of neuroblastoma tissue from 251 patients. Among these, 180 showed stable disease or regression, whereas 71 experienced relapse or died in a mean follow up of 40 months. Based on this group specification we performed a verification of the six valid synthetic lethal hubs as identified on the level of cell lines. Equivalent criteria as selected for the cell lines were applied for the human tissue data. All six synthetic lethal hub candidates showed in the mean a fold change around zero, and utilizing the synthetic lethal partners for the given hub proteins showed a coverage between 35% and 59% (Tab. 2.8). Specifically, the increased coverage realized by the hubs ACTA1, ACTG1, ACTB and MAP2K2 is significant with respect to a reference distribution of coverage (Fig. 2.6). ACTA1 (actin alpha 1), specifically expressed in skeletal muscle, is one of the six different actin isoforms also including ACTG1 (cytoplasmic actin found in non-muscle cells) and beta-actin (ACTB). MAP2K2 is a protein kinase known to play a critical role in mitogen growth factor signal transduction, also involved in stress induced apoptosis and actin reorganization.

Certainly, synthetic lethal hub candidates as derived by the screening procedure described in this work may fail in experimental verification due to various reasons: one fundamental issue is the assumption of stable genetic association across species. To what extent traversing genetic interactions from yeast to human provides valid interactions still needs to be clarified. Here, the specificity of mapping yeast sequences to the human genome has also to be considered. Next, for genes successfully mapped a bias towards essentiality may be given, and naturally synthetic lethal hubs must not be essential as a knock-down of these would also impact the viability of cells not carrying the mutated (down-regulated) partner. Based on a data set on single lethality, five of the 13 actin-associated hubs, namely ACTA1, ACTB, FMN1, MAP2K1 and MAP2K2 are indicated as essential, presumably not being valid candidates. However, cross-species evaluation of essentiality carries the same restrictions as given for mapping genetic interactions. For other hub candidates derived by the screening routine explicit disease annotation is provided (e.g. for mutation in ACTG1 [112]).

Significant redundancy is suspected in genetic interaction networks, to a large extent based on more complex interactions than explicit implementation of redundancy [62]. The procedure proposed in this work is centrally focusing on coverage, resulting in hubs exhibiting a large number of interaction partners, as this fact per se increases the likelihood of identifying a set of

down-regulated partners in given expression profiles. The node degree distribution of the entire synthetic lethality network showed that genes with the highest coverage given for the validation utilizing patient tissue expression data, namely ACTA1, ACTB and ACTG1, are situated well outside the double standard deviation of the degree distribution. Here, the risk of side effects increases, as tissue specific gene expression may show absence of one (or many) of such partners, resulting in significant off-target effects. On the other hand a fraction of synthetic lethal pairs may be embedded in a higher order redundancy, i.e. the effect of a double mutant may still be buffered by a third entity or alternative mechanism.

From these considerations various aspects impacting false positive as well as false negative findings on synthetic lethal hubs become clear, alluding the relevance of extended experimental verification. Here, the development of high throughput RNA interference screening [113] has become a perfect complement, allowing large scale verification of synthetic lethal hub candidates as derived by screening procedures exemplarily outlined in this work.

2.3 A generic synlet-based approach to tackling chemoresistance in tumor[3]

Concept outline and goals

In the previous sections we described and exemplified the concept of synthetic lethality in chemoresistant neuroblastoma. Knock-down hubs were identified in resistant cell-lines and were then validated in human tissue samples showing in both cases promising results.

The approach, however, is not at all restricted to this particular tumor type or, by that matter, to cancer. Theoretically, any cells exhibiting a particular phenotype in which one synthetic lethal partner is missing (as opposed to all other tissues where both are present) can be targeted by these means. Cancer is an especially interesting target as few alternative therapies exist.

In this section we will generalize the concept of synthetic lethal hubs and their detection and further exemplify the approach using curated human data on breast, lung and colon cancer.

To formalize this concept, an individual data set referring to an organism or biological sample presents a set of genes A1, A2, A3, .. , An which are down-modulated or functionally compromised and are part of a set of genes S together with a functionally non-compromised gene B, where compromising the function of B should result in a lethal effect or should lead to substantial morbidity due to pairwise given synthetic lethal relationships to genes A1, A2, A3, .. , An.

Classical synthetic lethality or synthetic sickness screens have previously been used to identify potential drug candidates specific to certain diseased cells. These cells were considered amenable to drugs if a synthetically lethal partner A of a direct or indirect drug target B is functionally inactivated in those cells, for example by down-regulation, mutation, deletion or other mechanisms, however, present or otherwise functionally available in healthy/uninfected

[3]This section is based largely on the patent application *Critical gene targets for cytotoxic therapy* with the publication no. WO/2011/144738, application no. PCT/EP2011/058259 by Arno Lukas, Johannes Söllner, Andreas Bernthaler, Bernd Mayer, Raul Fechete, Paul Perco, and Andreas Heinzel submitted to the World Intellectual Property Office (WIPO), 2011 [26].

cells. Drug target candidates identified in such a way are not less likely to be subject to the development of resistance than drug target candidates identified by other means, for example high-throughput small-molecule screens. Multiple mono-lethal relationships to different genes A might be effective in a patient sample targeting a selected gene B. Therefore it is rather likely to preserve sensitivity to the drug even if single mono-lethal relationships to a gene partner A are bypassed by resistance mechanisms in a patient. In the same sense the efficacy can be increased in terms of patient coverage. Variability of down-modulated genes A in different patient samples is taken into consideration by the approach.

Based on the concept of synthetic multi-lethality a workflow has been developed to select drug targets characterized by increased probability of therapy efficacy across individuals and reduced risk of development of resistance within individuals as well as reduced risk of disease recurrence in the case at least of, but not limited to, neoplastic disorders, including solid tumor disease, and infectious disease.

Specifically in a first step down-regulated, mutated or otherwise functionally compromised genes A or gene products are identified. Such identification can be conducted by several means including, but not limited to, transcriptomics experiments indicating significantly down-regulated genes, nucleotide sequencing indicating mutated genes or a multitude of other experimental approaches leading to the identification of gene function partially or completely missing or otherwise compromised in particular diseased individuals. Specifically this step is performed sequentially and repeatedly for multiple individuals or biological samples, for example by analyzing transcriptomics data from multiple individuals or cell-lines associated with or suffering from a particular disease or disease condition. Synthetic multi-lethality genes A1..An can be envisioned where at least one B common to all these A1..An has to be present in as many individuals/biological samples as possible. In each iteration new combinations of genes A1..An, (primary, functionally compromised) and B (secondary, functionally non-compromised gene) can be identified. Based on this iterated sampling technique genes B (gene products not functionally inactivated or down-regulated) can be identified which are associated by synthetic multi-lethality with one or several different genes A (gene products functionally inactivated or down-regulated). More precisely, a specific gene taken from set A1..An which is associated with a specific gene B may or may not be down-regulated in a specific patient, tissue or otherwise sample of organic matter associated with a particular disease or disease state as long as it is down-regulated or functionally inactivated in at least one of these samples. Elements from A1..An may but need not be shared between individual samples or individuals while B must be present in at least the substantial majority of analyzed samples. A preferred aim is to identify sets of multiple genes, where those sets contain at least three elements (a gene triple) where at least one is a potential drug target or biomarker B functionally present in as many analyzed individuals/tissue samples as feasibly possible. This iterative aspect for enriching genes B optimally generally available in diseased organisms/cells associated with a set of genes A1..An of size of at least two highlights the potential to identify drug targets or biomarkers, which will be characterized by enhanced patient coverage or therapy stability where therapy stability, is defined as the reduced likelihood or delayed onset of resistance against therapy compared to standard therapies.

The described approach is mainly geared towards identification of drug targets and ultimately drug therapies or targets of diagnostic methods applicable in a multitude of patients with

improved therapy stability and reduced likelihood of disease recurrence after end of drug therapy. The same approach is well suited as a basis for individualized medicine as well. Individual patients can be profiled for the presence of genes B in a synthetic multi-lethal relationship with sets of down-modulated genes A. While such combinations may be infrequent in most patients, certain individuals, particularly but not exclusively in neoplastic disorders can show particular B and A1..An combinations which make this particular patient suitable for a particular therapy based on drug targets encoded by B. While this approach is related to testing for the presence of Her2/neu for the targeted therapy with herceptin our approach adds stability to therapy for a particular patient, which would otherwise not be present or subject to chance or luck. Other patients not featuring as many missing functionalities in potential genes A should be more likely to experience remissions or inefficacy of therapy. On the other hand they may exhibit other B and A1..An combinations allowing selection of other accepted therapies for efficient therapy based on molecular profiling of available functions of potential genes B and missing functions of potential genes A.

The identified hubs can be addressed in specific therapeutic settings. Drug targets can be drugged directly and indirectly, the essential part is that function is removed or altered.

In many cases drugs affecting the presence of gene B encoded functionality will already be known, allowing for repositioning of drugs for new indications. In certain cases, such as anti-neoplastic agents, the applicability for certain tumors may be new or untested and the presented approach can be used to propose the most suitable therapy from a set of drugs. Drug databanks such as DrugBank [114] and STITCH 2 [16] play important roles in the identification of drugs for repositioning as they allow identification of drugs associated with molecular drug targets identical or associated with identified products of genes B.

Materials and methods

The examples presented here describe the application of the method for identifying critical gene targets to predict the response to a targeted cytotoxic therapy in a patient for three different neoplastic disorders of high relevance in terms of incidence and number of associated deaths. These are ductal breast carcinoma, non small-cell lung adenocarcinoma and colon adenocarcinoma. For each of these cancer types the method was applied separately. To identify critical gene targets B specific for a cancer type the following stepwise procedure was applied.

Synthetic lethal interactions

In the absence of a comprehensive set of experimentally determined synthetic lethal relationships between human genes an alternative way was chosen, taking respective available information for yeast genes and mapping these genes to ortholog human genes.

Therefore the publicly available data set of synthetic lethality relationships between yeast genes [35] was filtered for significant synthetic lethal interactions according to specifications proposed by the authors. This set of genes was mapped to ortholog (homology between organisms) human genes. The aim of this step was gaining information of synthetic lethal relations between genes in homo sapiens (human) based on the data given for yeast. To do so mapping information was obtained from five providers: Homologene [87, 115] (content downloaded on

2010/02/03), Inparanoid [116, 117] (content downloaded on 2010/02/03), RoundUp [118, 119] (content downloaded on 2010/02/03), OMA [120,121] (content downloaded on 2010/02/03) and ENSEMBL biomart [122, 123] (content downloaded on 2010/02/22).

While this process entailed several technical steps primarily because of different sequence ID spaces both the raw synthetic lethality dataset provided for yeast and the generated orthology dataset could be mapped to ENSEMBL gene IDs, thus also removing ambiguities due to protein isoforms mapped to in component databases. The mapping was done assigning human ENSEMBL gene IDs to yeast ENSEMBL gene IDs leading to a human gene set for which synthetic lethal relations are extrapolated as given in Tab. 2.9.

Minimum number of sources supporting ortholog mapping	**1**	**2**	**3**	**4**	**5**
Number of ortholog human genes	3714	1631	1055	662	340
Number of synthetic lethal, ortholog human gene pairs	204124	31068	13144	5408	1478

Table 2.9: Overview on numbers of human genes and their synthetic lethal relations as resulting from ortholog mapping of yeast to human data sets. Given are the numbers of distinct human genes ortholog to at least one yeast gene which takes place in a synthetic sick interaction with orthology reported in minimum N databases and the numbers of distinct extrapolated, synthetic lethal human gene pairs (mapped by orthology with support of minimum N data sources).

The set union of the human genes mappable from yeast, corresponding genes supported by at least one orthology source were used for further analysis.

Gene expression data sets

To determine expression of genes B and down-modulation of genes A in patient samples as well as expression of genes A in healthy tissue samples of a particular cancer type, publicly available transcriptomics raw data from primary tumor tissue and of samples from healthy tissue of the same tissue type were identified in a literature and database search. The sources for these data are listed in Tab. 2.10, 2.11 and 2.12. Transcriptomics raw data from studies as listed in the aforementioned tables were preprocessed to assess data quality. Therefore outlier detection was done based on hierarchical clustering and detection call distribution of the quantile-quantile normalized data set. Sample duplicates and outliers were excluded to receive the data pools subsequently used for the drawing procedure.

In the given example, downmodulation is determined (using data from gene expression micro-arrays) by measuring lowered concentration of gene expression products (messenger RNA). For each patient's sample each gene is statistically tested for being downmodulated in the sense of observing a significantly lowered concentration of the expression product in the patient's sample compared to the normal tissue sample.

Study	No. of cases	No. of controls
Landi et al. 2008	58	49
Su et al. 2007	27	29
Yap et al. 2005	49	9
Shedden et al. 2008	321	0
Total	455	87

Table 2.10: Selected studies providing transcriptomics raw data gained using HG-U133A microarrays (Affymetrix, CA, USA) from human samples comprising primary tumor tissue and respective normal tissue. Number of samples from non small-cell lung adenocarcinoma (cases) and healthy tissue (controls) used for further analysis are given (outlier adjusted). References are: [124–127]

Study	No. of cases	No. of controls
Richardson et al. 2006	40	7
Chen et al. 2010	42	143
Total	82	150

Table 2.11: Selected studies providing transcriptomics raw data gained using HG-U133 Plus 2.0 microarrays (Affymetrix, CA, USA) from human samples comprising primary tumor tissue and respective normal tissue. Number of samples from ductal breast carcinoma (cases) and healthy tissue (controls) used for further analysis are given (outlier adjusted). References are: [128, 129]

Study	No. of cases	No. of controls
Sabates-Bellver et al. 2007	32	32
Jorissen et al. 2009	290	0
Gyorffy et al. 2009	30	8
Bjerrum et al. 2009	0	9
Total	349	49

Table 2.12: Selected studies providing transcriptomics raw data gained using HG-U133 Plus 2.0 microarrays (Affymetrix, CA, USA) from human samples comprising primary tumor tissue and respective normal tissue. Number of samples from colon adenocarcinoma (cases) and healthy tissue (controls) used for further analysis are given (outlier adjusted). References are: [130–133]

Bootstrapping and feature selection

While described transcriptomics raw data pre-processing led to reasonably homogeneous data sets an iterative drawing process from these cancer type specific data sets as given in Tab. 2.10, 2.11 and 2.12 was applied to be able to estimate stability of results, consider varying numbers of case and control samples for the different cancer types and reduce biases. Each single drawing step was done by randomly selecting 30 tumor tissue samples for the cases group and randomly

selecting 30 healthy samples for the control group from the cancer type specific sample set. This drawing is iterated 20 times for each cancer type leading to the respective result statistics. For each stratified random sampling, respective raw data were separately quantile-quantile normalized. To estimate non specific contributions from the analysis procedure and the data sets, separate runs with inverted sample assignments to the analysis groups were performed. This means 30 tumor samples were drawn for the control group and 30 healthy samples were drawn for the case group in each iterative run.

In each of these drawing steps a ranked list of critical genes B is derived based on a defined procedure. To consider a gene B expressed in a patient's sample or a gene A expressed in a healthy tissue sample, the MAS5 detection calls for features on Affymetix microarray platforms were used. This means, a valid intensity value from a microarray experiment exists for a feature called "present". However, no valid intensity value could be measured due to high mismatch intensities or zero concentration of respective RNA for features called "absent". A gene might be represented on a microarray by multiple features. It was defined to consider a gene as expressed in case all of its features are called present in a sample measurement. Otherwise the gene is called absent. An intensity value for a gene is derived as mean from respective feature intensities in a sample measurement.

In a single drawing step defining 30 cases and 30 controls each gene of the ortholog human gene set is checked for each case to be

1. downmodulated compared to the 30 controls and / or
2. expressed.

Based on this criteria for each gene of the ortholog human gene set a coverage value is calculated applying the information of synthetic lethal relations between genes. The coverage represents the fraction of cases where a synthetic lethal effect is expected when repressing the respective gene B. A case is considered as affected if the gene B is expressed and at least one gene A exists for this case. For a gene A additionally the condition to be present in controls is essential. By repressing a gene B, different cases can be affected due to the same or different genes A. A gene might be a gene B and / or contribute as a gene A to the coverage of another gene B for a particular case depending on the synthetic lethal relationships to other genes.

Targeting of gene B or gene B products shall only lead to reduced viability in tumor cells due to the multi-synthetic lethal effect. Therefore if function encoded by gene B is depleted or missing (due to therapeutic intervention) only cells with missing or depleted functions encoded by genes A in multi-synthetic lethal relation with gene B (diseased cells) are affected. This specificity of therapeutic intervention to diseased cells is secured by the need of genes A to be expressed in healthy tissue.

Performing 20 drawing steps standard deviation for the coverage value can be calculated. The coverage of each gene was corrected by subtraction of the background coverage value gained by inverting cases and controls. The corrected coverage is used for sorting where highest coverage is preferred. As result of this analysis process a ranked list of critical genes B is derived. Additionally to each gene B a list of genes A was derived indicating the percentage the gene A contributes to the coverage of gene B.

Potential gene targets B were ranked by the coverage of cases (maximum coverage of patient samples indicating optimum expected stability of therapy). This ranking scheme has been applied for generation of target lists specific for a particular cancer and enriched for encoded functions putatively less likely to be lost during cancer development.

Drug data

These data were further enriched by drugability data present in the STITCH 2 database [16] allowing identification of already known drugs targeting identified, putative drug targets B and thus readily paving the way for drug re-positioning in context of our newly identified genes B and their products.

As an important note, selection of drug targets and ranking of these for individual patients is done differently, basically by selecting those genes B or combinations of genes B covered by a substantial multitude of genes A. Genes B with hundreds of A's may also be discouraged, however, as this multitude may indicate increased levels of toxicity upon removal of function. Manual analysis of gene-associated function and, for example presence and dimension of paralogous clusters among genes A can help substantially in selecting targets. Independent whether drug targets for individuals or populations of patients are to be selected, rationalization of gene function has been proven useful in many cases.

Results and discussion

Tab. 2.13, 2.14 and 2.15 are extracts of gene B lists specific for the three cancer types ranked by coverage. Other genes B, which are still unknown with regard to their targetability of cytotoxic therapy have not been included in the extracts. Genes B are listed with descending coverage values from top (rank 1) to the bottom (rank n) where the rank indicates the position of the gene in the list relative to other genes. Only genes B are extracted from these lists having an entry for inhibiting chemicals in STITCH 2 database investigated or discussed for the treatment of the specific cancer type. The tables hold the following information indicated by the column headers: Gene Symbol, ENSG: ENSEMBL Gene ID, Rank: position in the list sorted by coverage, Cvrg %: coverage value in % indicting the fraction of cases affected when targeting gene B as mean value derived from sample drawing, Stdv Cvrg %: standard deviation of the coverage value in % derived from sample drawing. Expr. change %: change of medium expression intensity of gene B in Patient's samples in % relative to medium expression intensity in controls, Drug: drugs targeting the expression product of gene B according to the STITCH 2 database.

Dihydrofolate reductase (gene symbol: DHFR) is known as a target for the approved drugs Pemetrexed, Trimetrexate and Methotrexate. These where clinically investigated in colon cancer patients as antineoplastic agents showing activity in particular in combination therapy. These drugs are antifolates, which impair the function of folic acids. Antifolates are used in cancer chemotherapy.

Aldehyde dehydrogenase 1 family, member A1 (gene symbol: ALDH1A1) is known as a target for Tretinoin, a naturally occurring derivative of vitamin A (retinol). The group of retinoids such as tretinoin shows antineoplastic activity and is used in the treatment of acute

Gene Sym.	ENSG	Rank	Cvrg %	Stdev Cvrg %	Expr Change %	Drug
DHFR	ENSG00000228716	26	80	10	30	Pemetrexed, Methotrexate, Trimetrexate
ALDH1A1	ENSG00000165092	198	50	15	-92	Tretinoin
HDAC1	ENSG00000116478	248	45	12	-21	Virinostat
HDAC3	ENSG00000171720	250	45	12	4	Virinostat
HDAC8	ENSG00000147099	293	41	14	-10	Virinostat
TOP1	ENSG00000198900	419	32	16	11	Irinotecan

Table 2.13: Extract of genes B from the ranked result list derived from processing data specific for colon adenocarcinoma samples.

promyelocytic leukemia. However in-vitro studies have shown modulating affects also for colon cancer cells.

Histone deacetylase 1, 3 and 8 (gene symbol: HDAC1, HDAC3 and HDAC8) are known as targets for Vorinostat, which is an antineoplastic agent approved for cutaneous T-cell lymphoma. It inhibits the enzymatic activity of histone deacetylases Class I and II. It has shown anti cancer activity in colon tumor cells and is currently under clinical investigation for colorectal cancer patients.

Topoisomerase (DNA) I (gene symbol: TOP1) is known as target for Irinotecan which is an antineoplastic enzyme inhibitor primarily used as part of the front line treatment of metastatic colorectal cancer. Irinotecan prevents religation of the DNA strand by binding to topoisomerase I-DNA complex, and causes double-strand DNA breakage and cell death.

Gene Sym.	ENSG	Rank	Cvrg %	Stdev Cvrg %	Expr Change %	Drug
TUBA1C	ENSG00000167553	66	51	25	-8	Epothilone B
POLA2	ENSG00000014138	181	34	9	12	Decarbacine

Table 2.14: Extract of genes B from the ranked result list derived from processing data specific for non small-cell lung adenocarcinoma samples.

Tubulin, alpha 1c (gene symbol: TUBA1C) is known as target for Epothilone B which is a 16-membered macrolide with antineoplastic effects. It inhibits microtubule function and is investigated for use/treatment in lung cancer and other neoplasms. Polymerase (DNA directed), alpha 2 (70kD subunit) (gene symbol: POLA2) is known as target for the non-classical alkylating agent Dacarbazine. It is usually used in conjunction with other drugs as an antineoplastic second-line therapy. Dacarbazine when used with other chemotherapeutic agents has shown activity in treatment of non-small cell lung cancer.

Gene Sym.	ENSG	Rank	Cvrg %	Stdev Cvrg %	Expr Change %	Drug
CYP51A1	ENSG00000001630	1	77	8	42	Letrozole
CCNA2	ENSG00000145386	18	55	9	68	LY293111, DENSPM
SLC25A5	ENSG00000005022	196	18	12	35	Clodronate
SLC25A6	ENSG00000169100	209	17	11	-24	Clodronate

Table 2.15: Extract of genes B from the ranked result list derived from processing data specific for ductal breast carcinoma samples.

Cytochrome P450, family 51, subfamily A, polypeptide 1 (gene symbol: CYP51A1) is known to be affected in activity by Letrozole which is a non-steroidal aromatoase inhibitor. CYP51A1 is involved in cholesterol biosynthesis. Letrozole is used as adjuvant treatment of hormonally-responsive breast cancer.

Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5 (gene symbol: SLC25A5) and solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6 (gene symbol: SLC25A6) are known targets for Clodronate which is a (non-nitrogenous) bisphosphonate affecting calcium metabolism. Antineoplastic activity of clodronate was clinically investigated as adjuvant treatment in metastatic breast cancer patients.

Cycline A2 (gene symbol: CCNA2) is known to be repressed by LY293111 and DENSPM. LY293111 is known to be a leukotriene B4 antagonist, a 5-lipoxygenase inhibitor and a peroxisome proliferator-activated receptor (PPAR)-gamma agonist with cytotoxic properties in cell lines. It shows synergistic activity with the active metabolite of capecitabine in two breast cancer cell lines. DENSPM (N1,N11-Diethylnorspermine) induce programmed cell death in breast cancer models and it was clinically investigated for treating metastatic breast cancer patients.

Selection of drugs based on proposed target molecule lists

The described example results in the selection and ranking of molecules the functional activity of which should be removed or impaired to produce a synthetic multi-lethal effect in a specific diseased tissue. For many of these small molecule drugs have been experimentally implicated to directly (for example through irreversible binding to the active site of the gene B encoded molecule) or indirectly (through effective down-regulation of gene B transcription) effect impairment of gene B encoded function. The discussed example has therefore been extended to include small molecule data contained in the publicly available database STITCH 2. Chemicals or drugs indicated by the database to have repressive effect on particular genes B were associated with these based on ENSG.

2.4 Summary of genetic interaction networks

Cancer is one of the leading causes of mortality in the industrialized world with only a very limited therapy efficacy. This is due on the one side to strong variability in the malignant cell population, making a targeted therapy difficult to perform. On the other side it is due to acquired chemoresistance, making an initially successful therapy eventually fail due to relapse of disease.

Synthetic lethality promises to address both issues by widening the therapeutic window through the targeting of synthetic lethal hubs. Assuming tumors exhibit in their array of mutations responsible for malignancy knock-downs of genes normally present in healthy cells, it is conceivable that an induced knock-down the missing genes' synlet partners would lead to the death of the tumor cells, while having a minimum impact on the healthy ones.

In a first approach we investigated the feasibility of this method for chemoresistant neuroblastoma in selected cell lines and showed that we can indeed validate the results in human tissue samples by maintaining good patient coverage with the selected targets. In a second approach we generalized the method and applied it to lung, colon and breast cancer and extracted significant synlet hubs including eligible drugs.

A potentially vulnerable spot of this approach is the homology mapping of synthetic lethal pairs from yeast to human. The results showed multiple matchings of yeast to human genes, leading to an increased number of synlet pairs in human. This is a fact that remains to be investigated further through experimental means.

Overall, the results of our work proved promising and with improving synthetic lethality screenings it enables us to hope that this approach will provide a key mechanism for large-scale treatment of cancer in the medium term.

CHAPTER 3

Protein interaction networks

3.1 A new model for protein interaction networks: omicsNET[1]

Concept outline and goals

Molecular interaction networks representing the human protein coding gene universe - together with the various types of molecular interactions and relations - have become widely used for analyzing Omics profiles. Here the challenge is to translate the descriptive set of features identified as (from a statistical perspective) relevant in an Omics experiment into pathways and processes, being effectively amenable for results interpretation and hypothesis generation. This need is even more pronounced for cross-Omics data interpretation, where the challenge is the combined analysis of heterogeneous feature types essentially spanning from the genome to the metabolome level [20]. Such integrated analysis strategies rest on expanding knowledge regarding molecular feature catalogs, on their biological role and interaction specifics, altogether resembling the core of Systems Biology approaches [134], in the clinical context leading to Network Medicine [21] or Systems Pharmacology [134, 135].

Repositories for molecular catalogues are maintained by major institutions such as the National Center for Biotechnology Information (NCBI) or ENSEMBL operated by the European Bioinformatics Institute together with the Sanger Institute. The same is true for specific molecular interaction networks with the Kyoto Encyclopedia of Genes and Genomes [11] (KEGG), PANTHER [12], or BioCarta [136] as prominent representatives. Further databases focus explicitly on protein-protein interaction data, as the Ontario Population Health Index of Databases [137] (OPHID), IntAct [9], or BioGRID [10]. Each such repository exhibits specific characteristics regarding type of interaction represented, coverage of molecular catalogs, as well as evidence and relevance of interactions in the biological context. KEGG for example offers various types of interactions ranging from protein complex formation to enzyme-substrate interactions

[1] This section is based largely on the scientific article *Expanded human protein coding gene interaction networks for serving Omics profile interpretation* by Raul Fechete, Andreas Heinzel, Johannes Söllner, Paul Perco, Arno Lukas and Bernd Mayer submitted to Molecular Biosystems pending peer review, August 2012.

at high level of evidence, but falls short on completeness regarding the human molecular catalog, as 6,198 (version as of January 2012) human protein coding genes (compared to the total set of 19,980 reported in ENSEMBL [138]) are represented. OPHID on the other hand provides a considerable set of protein interactions covering 14,612 UniProt/SwissProt identifiers which can be translated to a roughly similar number of protein coding genes. Protein interactions represented in such repositories span heterogeneous types like physical interactions or procedural dependencies [139] detected via different experimental methods such as affinity chromatography, yeast-2-hybrid screens, or being predicted based on cross-species analogies. Consequently, the data holds false-positives [140–142], as well as an undetermined false negative rate [143–145], with further uncertainly insofar as e.g. identified binding affinities exhibit relevance in the biological context. On top the total number of protein-protein interactions in the human proteome is not known, with estimates ranging from 130,000 to 650,000 [146, 147].

Obviously, the type of specific interaction network used for Omics profile interpretation influences hypothesis generation, driven by mere quantity of represented features, and biological nature of encoded interactions. For respecting these facts hybrid interaction networks have been developed, aiming at integrating diverse sources for obtaining networks showing improved coverage on the level of molecular feature catalogs, and at the same time providing a more complete representation of biologically relevant interactions. Alexeyenko et al. [20], for example, delineated a network constructed by inferring edge probabilities using diverse high-throughput data that achieves high gene coverage through orthology-based integration of different model organisms. Tyagi et al. [141] described a framework for delineating a human protein interactome including experimental details on complex structures and their binding interfaces together with evolutionary conservation. Kuchaiev et al. [144] presented a technique using geometric graphs to assess the confidence levels of interactions in PPI networks obtained from experimental studies in order to predict new interactions.

In this work we attempt to expand the set of human protein-protein interactions as provided by IntAct, BioGRID and Reactome [148] with inferred interactions being computed based on pathway (Reactome, PANTHER), ontology (Gene Ontology [95]) and protein domain data (InterPro [149]). The goal of this approach is to improve coverage of the human protein coding gene set represented in such a combined network, and at the same time extend it with putative interactions. Furthermore, we aim at categorizing interactions into the types “procedural” and “functional”. The semantics of these two categories are similar to those in Gene Ontology, i.e. a biological process is a series of events accomplished by one or more ordered assemblies of molecular interactions, while the function refers to a gene’s own properties. A procedural relationship between two entities exists in our context if both entities are involved in the same process, e.g. A phosphorylates B, whereas a functional relationship hints at an increased level of similarity, e.g. both A and B are ion transporters, but need not be involved in the same process.

Having such an expanded network in hand promises an improved representation of features identified as relevant in Omics profiling, additionally providing expanded information on interactions as well as type of interaction, together leading to improved hypothesis generation.

Materials and methods

Gene and protein identifier cross-referencing

For gene and protein cross-referencing the BioMart interface of ENSEMBL [138] was used (version as of October 2011). Identifiers considered as relevant included the ENSEMBL gene (19,980) and protein (86,934) IDs, NCBI gene symbols (18,981) and IDs (18,994), NCBI protein identifiers (31,628), TrEMBL [150] (42,399) and SwissProt IDs and accessions (37,864). The NCBI gene symbols and summaries were imported directly from the NCBI FTP site [151], and the lists of deprecated identifiers were imported from NCBI and UniProt, respectively. IDs referring to the same biological entity (gene or protein) were interlinked using abstract hyperstructures. Each such structure resembles a protein sequence as retrieved from ENSEMBL. All gene identifiers and symbols of genes coding for this sequence, and all protein identifiers were linked to such protein sequence. These hyperstructures encode the nodes of the interaction network.

Data sources on experimentally identified and literature curated protein interactions

Human protein-protein interaction data was retrieved from IntAct, BioGRID and Reactome, all in their versions as of October 2011. IntAct provided 35,634 raw interactions, BioGRID 41,496 and Reactome 91,002, respectively. Reactome offers two types of interactions, namely physical associations for all interactors occurring in the same complex, polymer or in the same reaction, as well as associations for interactors involved in neighboring reactions or being associated via (positive or negative) regulation. Consolidation of these interaction data sources provided the core interaction network. All interactions are retrieved and stored with their characterization provided as Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) Molecular Interaction (MI) [152] ontology terms. The PSI-MI ontology is rooted in the term molecular interaction (MI:0000) and the two branches relevant for this work are the interaction detection method (MI:0001) and the interaction type (MI:0190).

Gene and protein parameterization

For biological annotation of genes and proteins three categories of biological data was explored: biological pathway assignment, ontology assignment, and protein domain information. Pathway information was retrieved from Reactome and PANTHER. Reactome provided 1,128 pathways covering 5,292 genes, while PANTHER contributed 140 pathways covering 2,120 genes. Ontology information was retrieved from the Gene Ontology Consortium for the branches biological process and molecular function. The process category contributed 21,433 terms covering 14,110 genes, while the function category contributed 9,087 terms covering 14,693 genes. Protein domain information was retrieved from InterPro, which provided 6,041 specific domains covering 38,275 protein sequences. This annotation data was subsequently used for expanding interaction information.

Information consolidation

For each gene/protein entity the respective hyperstructures were populated with the retrieved annotation information. This information was subsequently used for computing relation scores (edge weights) between entities. A major implication of this information inheritance approach was that for each gene information became redundantly associated to all its splice variants. Considering the large number of splice variants, i.e. 86,934 sequences for 19,980 genes, but not having splice variant specific annotation available, only one representative protein sequence per protein coding gene was further considered. For this, the hyperstructure representing the protein with the longest sequence of each ENSEMBL gene was marked as the canonical sequence. Only such canonical hyperstructures were then used as nodes in the interaction network. Protein domain information for non-canonical sequences (if available) was also linked to the canonical one.

Interaction scoring

Based on the consolidated information, relation scores were computed for all pair-wise combinations of canonical hyperstructures. For each pair, six individual scores were computed, one for each of the available annotation types: GO biological process, GO molecular function, Reactome pathways, PANTHER pathways, protein domains and given interactions as retrieved from the interaction databases. These were subsequently combined into an overall relation score, expressed as edge weight.

All parameters (with the exception of the given interactions) were computed based on the information content (IC) of the terms, pathways and domains shared by two hyperstructures. For pathways and domains, the information content was computed as depicted in equation 1.

$$IC(E) = -ln(P(E)) = -ln(\frac{n_E}{N}) \quad (1)$$

E is the entity (pathway or domain) the information content is being computed for. N is the total number of nodes with this type of annotation. n_E is the number of nodes on this particular pathway or having this particular domain.

For computing the information content of the GO terms, this general calculation was extended to accommodate for the hierarchical structure of the ontology: each node was not only associated to the terms explicitly mentioned by the GO but also to all parent terms in the hierarchy following the "is-a" relationship. Further, the information content of a term is not only a function of the number of genes associated to it and its children [153–155] but also a function of its specificity given by the relative position in the ontology (number of its child terms) [156]. Based on these considerations two separate IC functions needed to be computed: one for the genes and one for the terms (equation 2). We added a value of 1 to the number of genes and child terms to enable computation of the information content values also for terms which were either empty or which had no children.

$$IC_{terms}(E) = -ln(\frac{n_{terms,E}+1}{N_{terms}}) \qquad IC_{genes}(E) = -ln(\frac{n_{genes,E}+1}{N_{genes}}) \quad (2)$$

Subsequently, the overall information content of a term was computed by using the root mean square (RMS) function (equation 3).

$$IC(E) = \sqrt{\frac{IC_{terms}(E)^2 + IC_{genes}(E)^2}{2}} \quad (3)$$

For a given node pair (X,Y), each of the individual parameter scores was computed as the sum of the information content values of the entities E (domains, pathways and terms) common to both nodes (equation 4).

$$f(X,Y) = \sum_{E \in X \cap Y} IC(E) \quad (4)$$

By construction, all parameter values were equal to or greater than zero. However, the individual parameter distributions were found to be strongly right-skewed. A high number of zero values resulted due to a limited set of parameter overlaps. In order to ensure a similar impact on the final edge weight delineated by utilizing distance functions on the individual parameter level for node pairs, we performed score adjustment. Due to the semantic contrast between the zero and the non-zero values, i.e. definitively inexistent versus given level of relation, we ignored the zero values during the normalization process. Each parameter distribution was then scaled with an individual factor alpha and the logarithm was computed. The alpha values were chosen such as to ensure a maximum curve overlap after the log-transformation and rescaling on the interval [0,1]. The 0.5 quantiles of the resulting distributions were situated around 0.35.

For the computation of the final edge weight as a proxy for the aggregate relation between two gene/protein nodes, we distinguished between procedural (composed of GO process, Reactome and PANTHER) and functional (composed of protein domain and GO molecular function) parameters. Procedural parameters were considered as describing a dependency (B follows A and leading to C) while functional parameters address similarity.

For each node pair, the average relation score of each category (procedural, functional) was computed, and the effective weight of a relation between two nodes (i.e. the edge weight) was given by the maximum of the procedural and the functional parameter. Additionally, the sources on experimentally described interactions (INT1) were included (equation 5).

$$F(X,Y) = max \left\{ \begin{array}{c} avg(f_{Reactome}, f_{PANTHER}, f_{GOProcess}) \\ avg(f_{Domains}, f_{GOFunction}) \\ 1 | (X,Y) \in INT1 \end{array} \right\} \quad (5)$$

Individual parameters missing due to incomplete node annotation were omitted from the computation, with the constraint of having at least one shared parameter for effectively computing a score. Consequently, node pairs not sharing a single parameter could not be further evaluated.

Results

Data sets and parameter characteristics

The number of gene/protein nodes taken into consideration for the protein interaction graph construction corresponded to the total number of protein coding genes represented in ENSEMBL, being 19,980. Of these, 14,212 had GO process and 14,773 GO function annotation, 5,340

were present in Reactome, 2,138 in PANTHER and 13,681 in InterPro. Characteristics of data sources included for annotation on experimentally determined and / or manually curated interactions (edges), given in distinct entities and interactions per source is depicted in Tab. 3.1. BioGRID had with close to 9,000 genes the best node coverage, while Reactome showed with close to 91,000 interactions the best edge coverage.

No. of nodes				
	IntAct	**BioGRID**	**Reactome**	**All**
IntAct	8,419	5,348	1,248	930
BioGRID		8,988	1,952	
Reactome			4,458	
No. of edges				
	IntAct	**BioGRID**	**Reactome**	**All**
IntAct	35,634	13,928	2,258	1,406
BioGRID		41,496	3,987	
Reactome			91,002	

Table 3.1: The table holds the number of unique gene identifiers and interactions, as well as overlap, for the data sources included in gene/protein node annotation.

Consolidating the available annotation information for each gene provided 17,022 of the 19,980 nodes with information from at least one data source. This set of nodes was further taken into consideration for delineating the interaction graph, neglecting the 2,958 protein coding genes not holding any annotation. The level of annotation, given by the number of data sources per node is depicted in Fig. 3.1 (left). 1,199 node pairs showed complete annotation on the level of all five data sources, 6,682 nodes held annotation from three sources.

Computing all-to-all edges between these nodes led to an edge evidence distribution as depicted in Fig. 3.1 (right). Based on the 17,022 nodes annotated with at least one source a total of 140.4 million edges could theoretically be computed. Of these, 5 million edges had no evidence (i.e. both nodes had not a single shared annotation), and were therefore omitted from further processing.

Of the about 140 million viable edges 145,391 rested on experimentally derived / manually curated interactions as provided by IntAct, BioGRID or Reactome. Such edges are subsequently denoted as INT1 (in contrast to INT0 edges not having such background on interaction).

General graph characteristics

The cumulative edge weight distribution of the 140 million edges, together with completeness, i.e. the number of nodes holding at least one edge with a weight above a certain cutoff value, is depicted in Fig. 3.2 (left). Weighing all 145,391 INT1 interactions with a value of 1.0 resulted in 11,162 nodes already present at this maximum weight cutoff. The maximum number of nodes, on the other hand, is reached at a cutoff value of 0.0. This is due to the about 23 million edges holding a weight of zero, as becoming evident in the weight dependent number of edges also depicted in Fig. 3.2 (left).

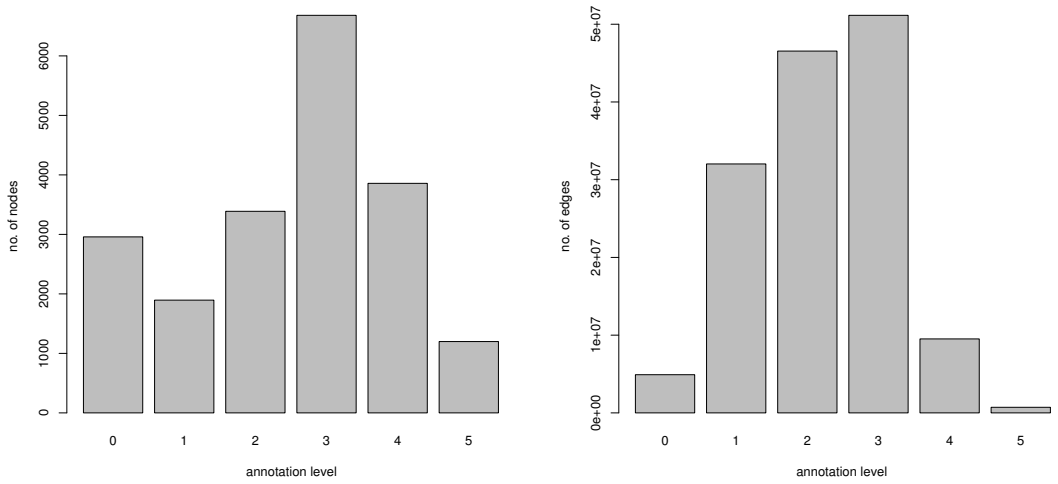

Figure 3.1: Overview of the number of data sources per node (left) and shared between two nodes of an inferred edge (right), with a maximum of five according to the annotation sources.

Plotting the number of nodes and edges in dependence of edge weight provides for a given number of nodes the number of edges being necessary to cover the nodes as depicted in Fig. 3.2 (right). Covering e.g. about 14,500 nodes results in a significant increase in the number of additionally required edges, i.e. new edges reached by lowering the weight cutoff tend to link nodes already being part of the graph rather than adding additional nodes.

In order to investigate potential literature bias in deriving edge weights, the association to the gene characterization index [157] (GCI) was calculated. The Pearson correlation coefficient between each node's strongest edge weight and its GCI was 0.43, thus a weak positive correlation between the two parameters could be observed. No correlation (Pearson R = 0.15) could be determined for comparing each node's strongest edge weight to the number of papers associated to the respective gene based on NCBI's gene2pubmed [158] links. A third analysis regarding eventual bias in node or edge annotation focused on the node annotation level (Fig. 3.1, left). An increase in the maximum edge weight per node with rising annotation level was identified (Pearson correlation score of 0.53), mainly because the edge weight was computed as the maximum of the procedural and the functional category. To estimate the magnitude of this bias, we performed a complementary analysis involving the edge annotation basis (Fig. 3.1, right). The Pearson correlation coefficient between edge weights and edge evidence levels was -0.2.

Topological graph characteristics

Already at the highest cutoff of 1.0, the Index of Aggregation (IoA) was close to 1.0, i.e. paths between virtually all 11,162 nodes represented in IntAct, BioGRID or Reactome are available. Nodes with computed relations are added to this given sub-graph when lowering the edge weight cutoff levels. The node degree distribution of the graph at different cutoff levels is provided in

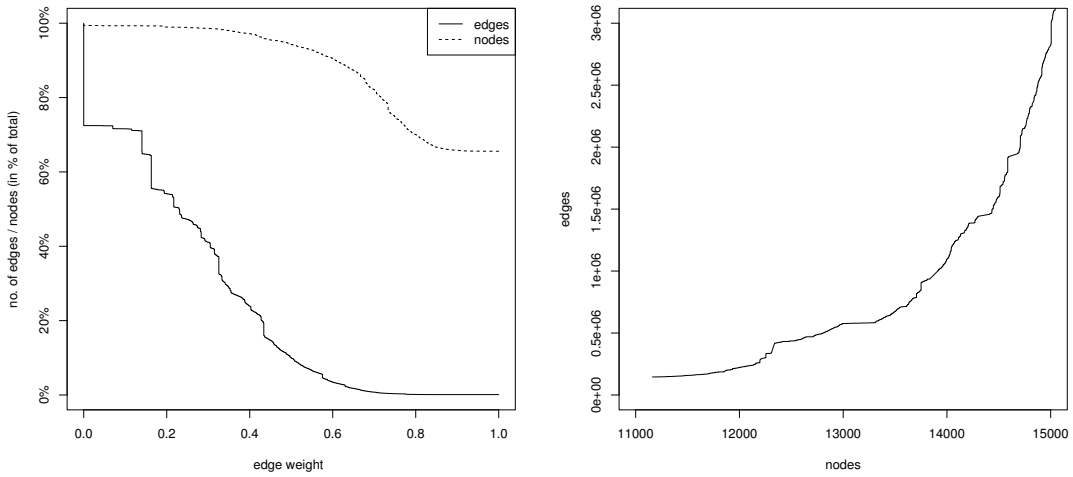

Figure 3.2: Edge weight distribution and node completeness as a function of edge weight (left). Node and edge count are presented for each weight cutoff in %, with the maximum number of nodes being 17,022, and maximum number of edges being 140 million. Relation of node and edge counts as derived at the different edge weight cutoff values (right).

Fig. 3.3. The median number of neighbors remains below 1% of included nodes for the weight interval [0.7, 1.0], the connectivity outliers at each cutoff provide an explanation for the high IoA. To further investigate this finding the five nodes with the highest degrees were exemplarily extracted at various cutoffs. At high cutoff levels (0.8-1.0), where the graph is mainly composed of INT1 edges, genes such as UBC (ubiquitin C) and the MYC oncogene [159] were identified as strongly connected. This finding may be explained in the light of the biological function of e.g. ubiquitin [160], or reflecting the significant literature association of e.g. MYC in a wide range of molecular processes [161]. At lower edge weight cutoff levels (0.5-0.7) mainly genes annotated by highly specific ontology terms, as e.g. members of the sirtuin [162] family or genes such as the P2RX4 [163] purinergic receptor P2X, ligand-gated ion channel 4 are present. Following the logics for computation of edge weights, highly specific ontology terms result in high edge weights for all genes sharing at least one of such specific term.

The global clustering coefficient (GCC) in relation to edge weights in the interval [0.5, 1.0] is depicted in Fig. 3.4 (left). The graph provided on the basis of INT1 alone serves as "seed" network, with an initial GCC of 0.33. Decreasing the weight cutoff, i.e. adding computed relations, resulted in an increase of the GCC, reaching a plateau at edge weights of 0.6. To further characterize the graph in comparison to a random network, two additional analyses starting from the same INT1 seed network were performed. In the first approach edges were added to the given INT1 graph in a random manner in the same pace as for the computed edges. In the second approach the very same procedure was performed but additionally respecting the node degree distribution. The results showed a strong GCC divergence for the three networks with increasing number of edges, with the two reference networks generated by adding edges in a

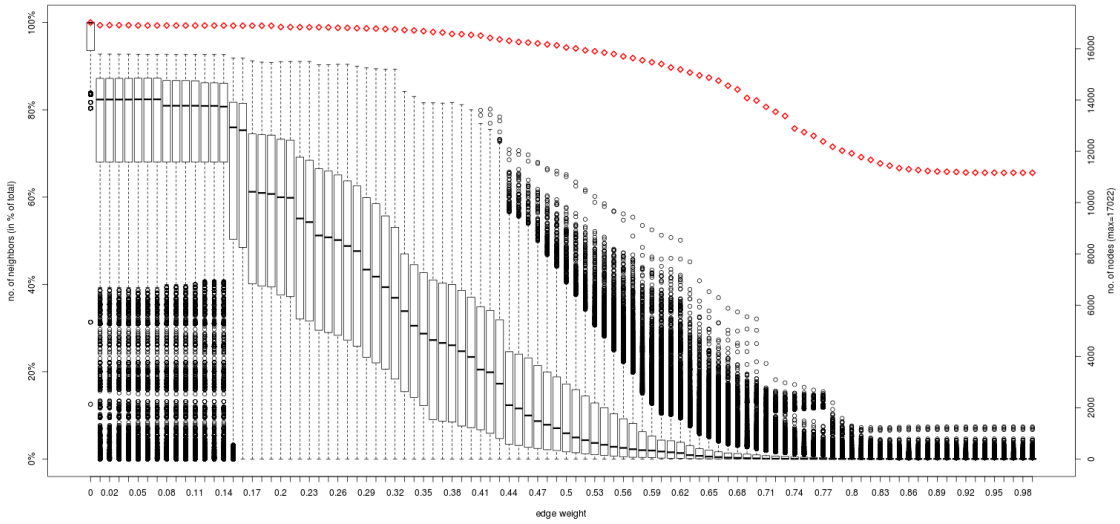

Figure 3.3: Degree centrality distribution as a function of edge weight. The Box Plots show the node degree distributions for all edges in the graph above the indicated edge weight cutoff. Additionally, the number of nodes included at each weight interval is provided (red curve).

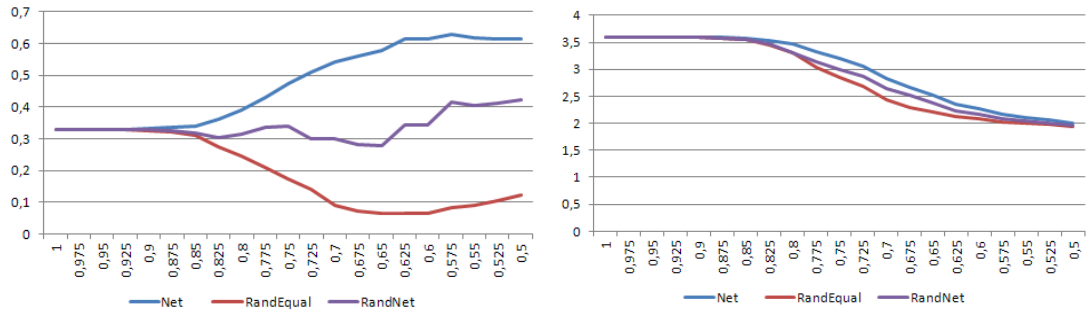

Figure 3.4: Global clustering coefficient (GCC) (left) and characteristic path length (CPL) (right) in relation to edge weight. The INT1 graph is extended by computed edges (Net), the same number of edges being randomly distributed (RandEqual), and the same number of edges as for Net taking into account the computed graph's degree distribution (RandNet).

random fashion reaching GCC values of about 0.4 (when reflecting the node degree) and 0.1 (adding edges entirely random). Clearly, the computed graph differs on the level of the GCC significantly from networks populated by edges instantiated randomly between nodes. For the three graphs also the characteristic path length (CPL) was computed. The results are depicted in Fig. 3.4 (right).

The initial characteristic path length of the INT1 network was 3.5, decreasing with lower edge weight cutoffs to a length of 2.0 at an edge weight of 0.5. The divergence of the computed

network in comparison to adding random effects also became clear on the level of CPL, with smallest CPL values reached for the graph complemented with entirely random edges. At an edge weight of about 0.5 the three graphs converge on the level of CPL.

Computed edge weights as surrogate for experimentally determined interactions

To investigate the suitability of the computed edge weights for predicting interactions as provided in databases resting on experimental evidence, the scores were investigated with respect to their prediction performance when inferring curated or experimentally determined interactions (INT1). All edges were split into two groups, namely with (INT1) and without (INT0) experimental backup, subsequently comparing their computed edge weights. Weight distributions of the two edge sets showed a distinctive right shift for INT1, indicating higher procedural / functional scores than for edges only holding computed scores. To assess the impact of this shift in terms of reflecting the nature of procedural and functional edges the receiver operator characteristic (ROC) curves for the two groups of edges with INT1 as the prediction target were computed. INT1 edges above a specific computed edge weight cutoff were therefore interpreted as true positives, while edges above the same value but not holding experimental evidence were interpreted as false positives (Fig. 3.5).

Fig. 3.5 (left) included the entire INT1 set of 145,391 edges, i.e. all interactions annotated with the PSI-MI term MI:0190. However, Reactome was used both for deriving interactions contributing to INT1 but also for category annotation utilized in computing edge weights. To identify a potential bias resting on this assignment, the computation of the ROC curves was also performed on an INT1 data set omitting contributions from Reactome (Fig. 3.5, right) including 3,872 interactions annotated with the PSI-MI term MI:0407 direct interaction. No bias could be identified as the initial AUC of 0.82 for the data set also including Reactome only dropped to 0.79 in the latter case.

The precision (the percentage of true positives from all positives) and the recall rate (the percentage of correctly predicted INT1 from all INT1) were investigated next. The precision starts off at 100% and drops to 13.5% at a cutoff of 0.92, being mainly due to noise as only a small number of edges is included. It peaks again at 63% at a cutoff of 0.89 and then decreases continuously with decreasing cutoff. For estimating the test accuracy the F1 score, computed as the harmonic mean of the precision and recall rates, was investigated to identify the best precision / recall ratio, being at a cutoff of 0.74.

A further independent analysis was performed using 31,996 interactions from KEGG as the set of interactions for the prediction analysis. For each cutoff in the edge weight interval [0,1] the true positive rate (the number of KEGG edges correctly identified as such, relative to the total number of KEGG edges) and the false positive rate (the number of non-KEGG edges above that cutoff, relative to the total number of non-KEGG edges) was computed. The resulting ROC curve is depicted in Fig. 3.6, showing an AUC of 0.79. Similarly to evaluation of INT1 edges the procedural parameter performed equally well as the effective weight (AUC of 0.77), while the functional parameter showed a lower AUC of 0.69.

In a further analysis we investigated the relative contribution of the parameters used for computing relations (Reactome, PANTHER, protein domains, GO process, GO function) to the assessment of specific interaction types, namely functional and procedural. The consolidated

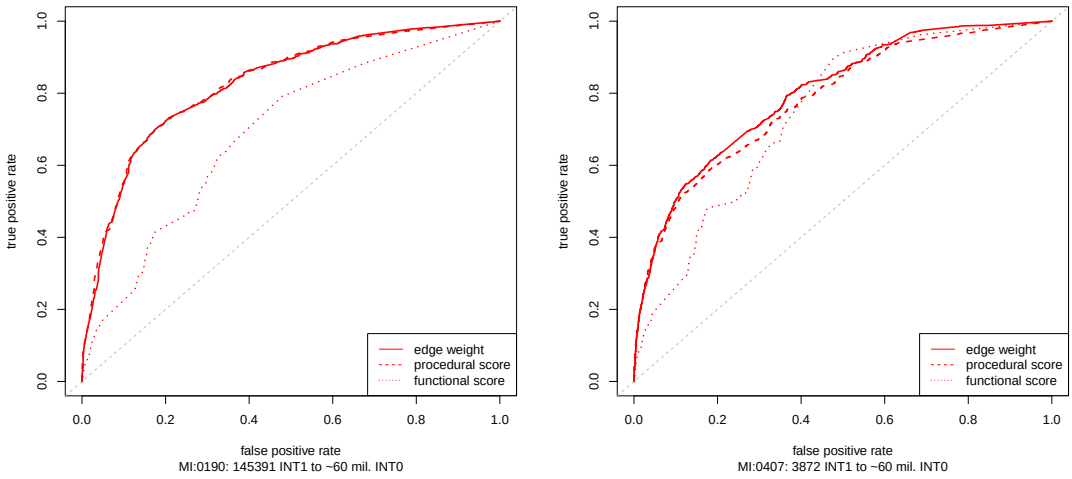

Figure 3.5: Receiver operator characteristic (ROC) curve for the entire set of INT1 (left) and for MI:0407 direct interactions only (right). For each weight cutoff in [0,1] the true positive rate (percentage of INT1 correctly identified from total INT1) is plotted against the false positive rate (percentage of computed interactions falsely assigned as interactions). ROC curves are plotted for the entire set of interactions, as well as separately for procedural and functional interactions.

INT1 edge set appeared to be mainly of type procedural, showing a significantly higher computed score for this interaction type compared to the scores achieved for the functional type. Considering computed scores above a value of 0.7 for edges not supported by database evidence, the interaction type functional was the dominant form. Interesting to note is that the relative distance between the two specific scores for functional and procedural relation remained constant for the whole edge weight interval.

Evaluation of graph characteristics done up to this point rested on the absolute edge weights of all edges (and nodes respectively) above a certain cutoff. Consequently, nodes holding edges with low weights were neglected, hampering the completeness of the node set effectively represented in the graph derived at a specific edge weight cutoff. As alternative edge selection strategy a relative order relationship rather than an absolute one may be applied by picking the top ranked edges of each node irrespective of the absolute weight. Validation of such edge selection strategy was performed again using the set of INT1 edges, as well as KEGG nodes and edges. Prediction performance with respect to INT1 decreased only marginally (AUC of 0.77 compared to an AUC of 0.82 when using the absolute cutoff criterion), while predicting KEGG edges was basically identical to the former approach (data not shown). While this alternative edge selection approach raised the node coverage at apparently little cost on accuracy, other factors need to be considered: Highly weighted edges were omitted if they do not fit in the top x% of their adjacent nodes, obviously increasing the false negative count.

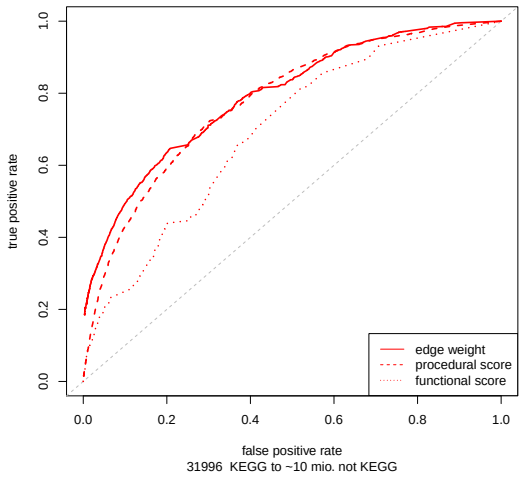

Figure 3.6: Receiver operator characteristic (ROC) curve for the KEGG classification. For each weight cutoff in [0,1] the true positive rate (percentage of KEGG edges correctly identified from total KEGG) is plotted against the false positive rate (percentage of edges not represented in KEGG but predicted as being present.

Discussion and conclusion

Construction of molecular networks represents a major leap in understanding cellular processes. Various types of networks exist, including direct protein-protein interaction networks [164], metabolic [165] or regulatory networks as well as RNA networks [166]. A major shortcoming of available interaction data is their lack of completeness on both molecular feature level as well as interaction information, often coupled with a significant number of false positive interactions in the context of biological relevance. In a previous work [89] we presented an approach to construct a dependency network based on adding biological information next to protein-protein interaction datasets. These additional data sources included pathways and ontologies, gene expression profiles and WPSORT-predicted subcellular localization information [167]. In the current work we present a refined version of this dependency graph based on a core network embedding interactions backed by experimental evidence or literature curation (145,391 edges), and then expanding this core network with computed interactions (relations) to increase coverage both on a node (gene/protein) as well as on the edge (interaction) level. Datasets used for delineating such interactions include pathway information from Reactome and PANTHER, ontology information from Gene Ontology and protein domain data from InterPro. The data is mapped to all presently known human protein sequences in ENSEMBL (86,934 entries) and subsequently consolidated on the canonical sequence of each gene/protein (17,022 entries).

Cutoff values for optimal coverage and accuracy

Inferring relations between these 17,022 nodes resulted in a complete graph with relation values for edges between 0.0 and 1.0, leading to the need to identify an edge weight cutoff for optimal accuracy and node coverage. While a higher edge weight implies a higher probability regarding a true positive relation, such a cutoff also leads to a drop in the number of edges included, in turn reducing the number of nodes represented in the graph. Accordingly, lowering the cutoff increases node coverage going in hand with loss of precision on the edge level. The process of choosing an edge weight cutoff implies finding a sensible balance between node coverage and edge accuracy.

We evaluated two methods of defining a cutoff, one where a single cutoff value was defined for all edges (absolute ranking method) and a second where for each node the top x% ranked edges were considered as positive dependencies (relative ranking method). Following the absolute ranking identified an edge weight of 0.65 as intrinsic lower boundary as the network differs above 0.65 from random on the level of global clustering coefficient (GCC) and characteristic path length (CPL). This provides approximately 2.5 million edges for 14,872 nodes. As upper boundary regarding the edge weight 0.8 may be perceived. At this cutoff the network holds 204,128 edges and 11,921 nodes, being only a minor increase with respect to the core (INT1) network holding 145,391 edges and 11,162 nodes. As further criterion the characteristics of the degree centrality may be considered, being at steady values in the interval [0.71, 1.0], and seeing a significant increase below an edge weight cutoff of 0.71 (at this point providing 830,470 edges for 13,730 nodes). Further supportive factor for setting an edge weight cutoff in this range is the computed precision F1, seeing a maximum at an edge weight of 0.74, at this point providing 12,891 nodes and 533,020 edges.

Result graph characteristics

Stumpf et at. [146] estimated the size of the human interactome to hold about 650,000 interactions. In this context, the size of our resulting network at an edge weight cutoff of 0.71 is roughly the same order of magnitude. Hart et al. [168] estimated the number of human protein-protein interactions to be situated somewhat lower between 154,000 and 369,000, while Venkatesan and colleagues [147] speculated on approximately 130,000 interactions. Certainly, with respect to node coverage there is still a gap of 6,250 protein coding genes; However, utilizing the annotation approach presented in this work excludes 2,958 nodes due to lack of any annotation data, leaving 3,292 protein coding genes not exhibiting a single relation scoring at least with an edge weight of 0.71. Contrasting the hybrid network with given reference networks provides additional 2,568 nodes when compared to the consolidated data from IntAct, Reactome, and BioGRID, naturally showing significantly increased coverage when e.g. compared to high evidence networks as KEGG (6,198 nodes).

Notably, the hypothesis that the computed edges exhibit a strong literature bias proved unfounded (Pearson R for comparing edge evidence level and edge weight of -0.2). Positive correlation, however, was identified for each node's strongest edge and the node's Gene Characterization Index (Pearson R = 0.43) and level of node annotation (Pearson R = 0.53) respectively. The network at an edge weight cutoff of 0.71 is found to be more compact than the INT1 network,

with a clustering coefficient of 0.51 as compared to 0.32 for the INT1 graph. The characteristic path length is found at about 3.0 compared to 3.5 for the INT1 graph. New edges added to the graph tend to link already included nodes in contrast to adding further nodes.

Graph validation

Validating the constructed network is an essential step towards assessing the quality of the presented method. This is, however, in practice difficult to perform. One of the main challenges to address is finding an appropriate dataset to validate against. As the method aims at extending the presently accepted set of protein-protein interactions (INT1, both experimentally determined and literature-curated), it is sensible to address the quality of the INT1 prediction. This is achieved by assessing the discriminative power of the method with respect to the INT0/INT1 classification. An inherent shortcoming of the INT1 dataset, however, is that it is by itself a heterogeneous collection of interactions ranging from direct binding to genetic interactions. Predicting INT1 members is therefore highly unspecific. For the complete INT1 dataset, the receiver operating characteristic curve showed an AUC of 82% for different cutoff values in the interval between 0 and 1. While this value is indicative for good prediction quality with respect to INT1 detection, this is still subjected to debate as the goal of the method lies not in the sole prediction but in the extension of the INT1 dataset. In this context, it can be argued that a high prediction value adds little novelty to the existing network, whereas a low prediction value may add novelty, but also a significant amount of false positives.

Additional difficulty is added to the validation by the strongly asymmetric character of the two sets' sizes, i.e. 145,391 edges for INT1 and 60 million for INT0 (only the set of INT0 edges between nodes present in INT1 was taken into consideration). While an AUC of 82% stands both for good precision and significant novelty, the disproportionately large number of INT0 edges adds a great absolute number of INT0 edges already at a small false positive rate. At a cutoff at 0.71 the true positive and false positive rates were 11% and 0.7%, respectively. This stands for approx. 16,000 INT1 edges and 685,000 INT0 edges. By comparison, a 50% true positive rate with an 8% false positive rate was reached at a cutoff of 0.55.

To reduce the heterogeneity of the prediction target, the classification performance was subsequently investigated for a highly specific set of 3,872 direct interaction edges only. This also yielded a high AUC of 0.79. This finding shows that the described method has good performance irrespective of the targeted INT1 edge type.

The performance of the KEGG interaction set prediction (31,996 KEGG edges vs. approximately 10 million inferred edges) was performed to assess prediction quality for an independent source in spite of increased edge heterogeneity. The AUC of 0.79 showed again good results, especially when considering the variability of the KEGG edge types.

Notably, in the performance investigation using INT1 as target, the ROCs are computed using the inferred edge weights for the INT1 network. In our final setup, however, all INT1 edges have a preset value of 1.0 and therefore, all 145,391 INT1 interactions are present within the resulting 830,470 edge set.

Conclusions

Molecular networks have become a central ingredient in Omics profile interpretation and hypothesis generation, consequently demanding networks with significant coverage of molecular entities combined with a comprehensive representation of interactions. The latter see various types of interactions, together with different levels of evidence regarding biological relevance.

Hybrid networks aiming at integrating diverse data sources are a straightforward approach for expanding both, node and edge count, and offering a single reference network for Omics data mapping and interpretation. Although information on interactions of protein coding genes expands on a continuous basis, computational inference of interactions on top of database information, as introduced in this work, adds to a more complete representation of the interactome, expanding opportunities for Omics profile-based hypothesis generation.

3.2 A new platform for visualizing molecular information: BIO

Concept outline and goals

The purpose of the work described in this section is the design and implementation of a new platform, called BIO [169], for the visualization of molecular information. Envisioned is a tool with which the user can query and navigate biological data on molecular entities like proteins and genes. This includes both relations such as protein-protein interactions and entity-centric functional annotation such as pathway membership and ontology term association.

A series of platforms addressing related aspects exists to date. STRING [14], for example, concentrates on molecular relationships providing comprehensive information on biological dependencies between genes and proteins. Fig. 3.7 shows the immediate neighborhood of BCL2 with the relation type encoded in the edge color. The figure also shows a typical information card describing the evidence supporting the edge. STRING is directed mainly towards academia.

By contrast, an approach directed strongly towards the industry is presented by NextBio [170]. The NextBio platform integrates different organisms and data types (drugs, diseases, clinical trials, etc.) and was designed with a clear focus on the clinical aspect of research.

When comparing NextBio to STRING the distinction in their target audience becomes evident. While STRING focuses on relationships -with less node annotation- NextBio focuses on the node characterization from a clinical point of view, lacking on the other hand the neighborhood aspect.

The goal of BIO is to unite the strengths of both systems by simultaneously providing neighborhood views and annotation information of clinical relevance for selected genes. The development of BIO also has a synergistic effect with omicsNET as BIO becomes a front-end for the former, while the latter gains a valuable dependency neighborhood.

Architecture and implementation

Considering, among others, the large amount of preprocessed data and the need to be easily accessible from anywhere, BIO was designed and implemented as a three-tier application. A diagram of the project landscape is depicted in Fig. 3.8.

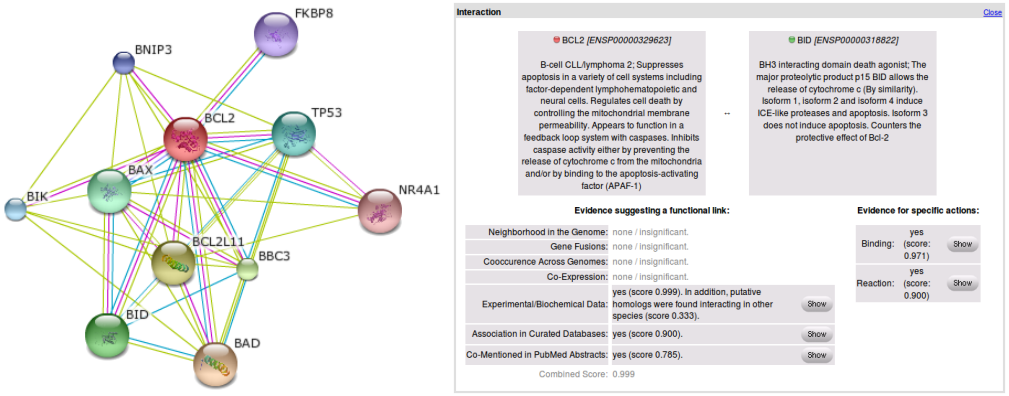

Figure 3.7: Typical STRING view with network structure and edge card

The direction of the information flow is represented in the figure from left to right. Biological data sources for the gene and protein annotation, such as drugs, patents and clinical trials are retrieved from the Internet and are preprocessed within Metaverse. Metaverse is a distinct project being developed independently from BIO, with the explicit goal to create a comprehensive repository of biological data. A selection of this data is then prepared and handed over to BIO over a predefined interface.

Biological data sources necessary for identifier cross-referencing and for the construction of the molecular dependency network are retrieved from the Internet and preprocessed within the omicsNET project. omicsNET is also a distinct project from BIO with the explicit goal to create a high confidence gene/protein dependency graph. Details on the exact data sources and the computation procedure, as well as an analysis of the results can be found in section 3.1. The identifier cross-reference data and the top-ranked dependency edges are handed over to BIO over a predefined interface.

The selected information from Metaverse and omicsNET is subsequently imported and consolidated in a database optimized for speed. This constitutes the first tier of the BIO project.

The second tier is represented by the main server application. It is responsible for processing requests coming from the user, authentication, logging, statistics, etc.

The third and last tier is represented by the client application running on the user's machine. Presently, this is a Microsoft Silverlight application running in the user's browser. Also part of this tier is a Mozilla Firefox plugin highlighting and displaying information on gene symbols found on visited websites. Furthermore, a Joomla Content Management System (CMS) is responsible for user registration, buying modules, displaying news and help items and generally providing an entry point into the system.

omicsNET and the server were developed by me and therefore lie within the scope of this thesis. They are also the ones to be outlined in the following. Metaverse and the third tier of the application consisting of the viewer, the Firefox extension and the Joomla CMS are independent projects and are not within the scope of this thesis.

BIO can be reached using the following link [169].

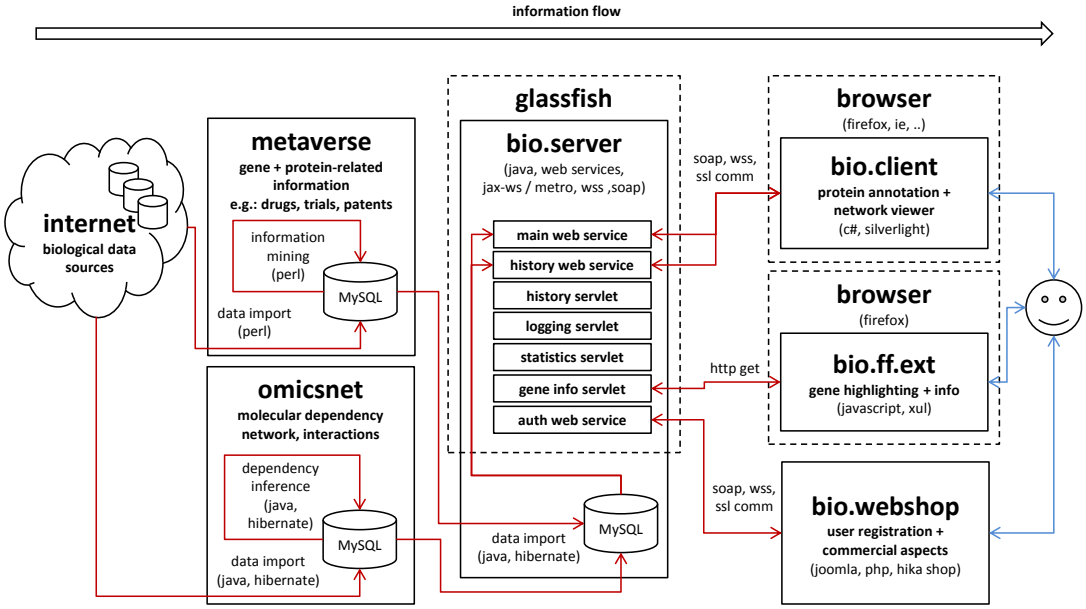

Figure 3.8: The BIO project landscape

Source: omicsNET

As was previously mentioned, the molecular dependency graph is computed in omicsNET. Method description and quality assessment were provided in section 3.1. Here we will give a quick technical overview of the project. Early development of omicsNET was done until 2009 by my colleague Andreas Bernthaler and me and was described in his doctor's thesis [171] and my master's thesis [172]. During the following years I continued the development and refinement of the project.

The omicsNET software is written completely in Java, employing Hibernate [173] as persistence layer. The Object Relational Modeling (ORM) facilities of Hibernate are used to store Java objects into a relational database, in this particular case MySQL.

omicsNET consists of two main building blocks:

1. The Framework is mainly declarative, and it holds the data model along with other definitions necessary to interface with the omicsNET data. It is implemented as a software library to be used by all omicsNET and custom applications.

2. Apport holds the imperative logic necessary to download, import, process and analyze the

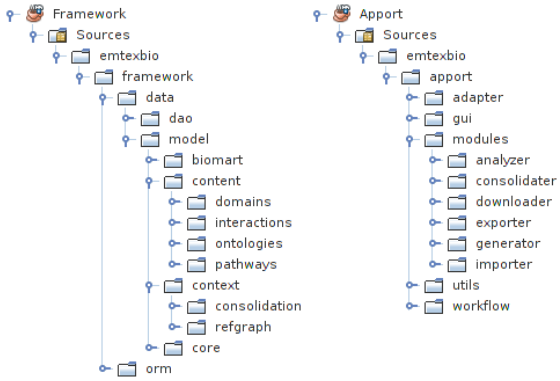

Figure 3.9: Framework and Apport project structure

information. Apport is an executable application holding a series of modules which do the actual work. It is used to generate the database from scratch, compute the graph as well as run various analyses. It uses the Framework library to access the data.

The goal of the Framework is to provide an interface towards the omicsNET data. In omicsNET, the database is accessed only through the intermediate Hibernate (ORM) layer. This bears portability advantages, as the DBMS can be replaced at any time with another, such as ORACLE or Postgre, without having to modify the Java code. The entity classes for data storage are annotated using JPA (the Java Persistence API) and are then persisted at runtime.

The structure of the Framework is depicted in Fig. 3.9. The Framework comprises the data model, which holds the actual class definitions for saving into the database and the Data Access Objects (DAO), which is a collection of methods to facilitate access to the stored entities. Furthermore, it holds ORM definitions which tweak the way Hibernate saves the data. Such changes are, e.g. adding prefixes to the table names for easier reference. Since database handling is done transparently, this is not a mandatory step. It is, however, a good practice to use table name prefixes similar to the java package names of the persisted classes in order to facilitate maintenance and error handling.

Within the model section, the core holds classes for the biological hyperstructure and all other gene and protein identifiers. These classes represent the unification of the different namespaces, both on the biological level (genomic and proteomic) as well as between different databases and organizations (NCBI, UniProt, etc.). They hold entity names, descriptions, summaries, deprecation relationships, protein sequences, etc. The identifier cross-referencing and biological information mapping is outlined in Fig. 3.10. A hyperstructure (e.g. HS1) can refer to one or more genes, while a gene can have one or more protein sequences. Biological information is retrieved on both levels. The content holds raw data imported from external databases such as domains, pathways, ontologies, etc. The context represents the further processing of the raw data from the content area. This includes consolidating it for each hyperstructure and subsequently computing the reference dependency graph using the hyperstructures as nodes. The biomart models the interface between omicsNET and the BIO project and the data stored here will be imported into BIO.

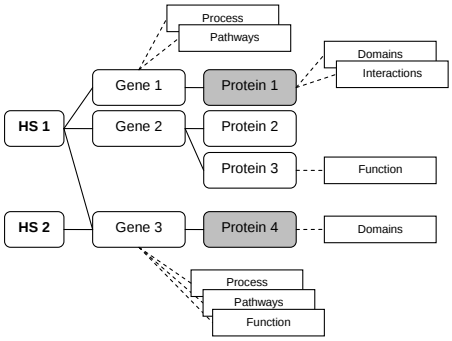

Figure 3.10: Identifier cross-referencing and biological information mapping

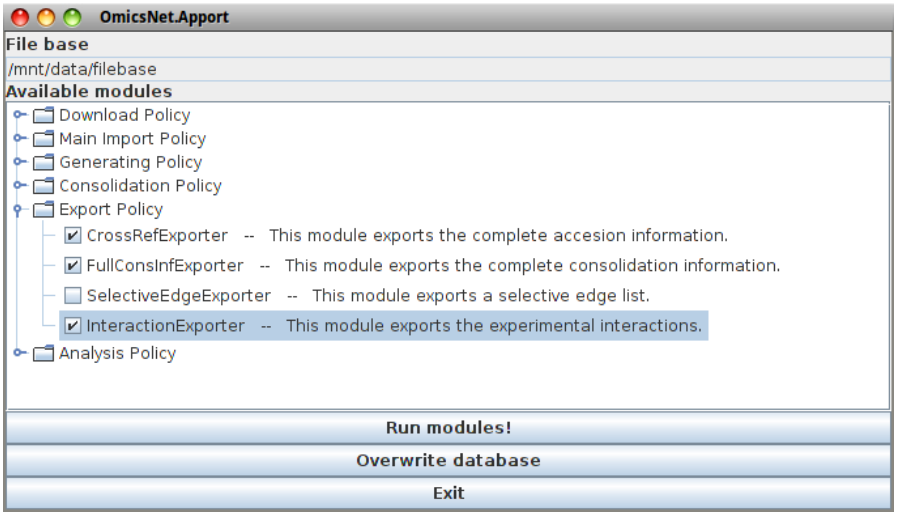

Figure 3.11: A screenshot of Apport with expanded exporter list and selected modules

Apport is an executable application implementing the graphical user interface (GUI) and the construction of the omicsNET by organizing all necessary steps in a modular workflow. The modules contain the work units grouped in downloaders, importers, consolidaters, generators and analyzers. This is also the precise data processing sequence. A major advantage of a modular workflow is that adding a new data source to omicsNET implies adding at most one module of each type and not changing any of the existing code. This leads to improved maintainability. Apport also allows the selective execution of modules. A screenshot of the program is depicted in Fig. 3.11.

Server: Data

Data retrieved from omicsNET and Metaverse is consolidated for usage by the server in a database optimized for speed, the biomart, depicted in Fig. 3.8 at the bottom of the BIO server component. While the data itself is partially redundant, this drawback is outweighed by the

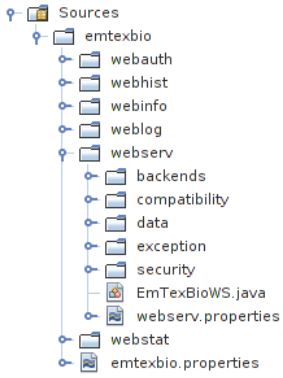

Figure 3.12: eBioWs project structure

reduced number of necessary table joins granting the system a significant performance boost at high workloads.

The design details are similar to omicsNET, i.e. a library provides access to a relational database over a Hibernate ORM layer while an application using this library implements a modular workflow to create the database. The database and access tools were designed and implemented by me.

Server: eBioWs

The second tier of the BIO project consists of the BIO server, eBioWs, depicted in the middle block of Fig. 3.8. The server was also designed and implemented by me. It is a Java web application exposing both SOAP web services and standard HTTP interfaces over the Secure Socket Layer (SSL). eBioWs is deployed in a Glassfish Java Application Server [174].

The main components of the server, depicted in Fig. 3.12, are listed below:

1. WebServ is the main component. It is a SOAP-based web service with an external interface. It handles incoming calls from the client.

2. WebAuth is also a SOAP-based web service with an external interface. Its role is to handle authentication requests coming from Joomla. Authentication information is stored in a local registry and provided, as needed during credentials checks, to WebServ. The exact authentication mechanism is described in below.

3. WebStat is a HTTP servlet handling GET requests. It returns information on the currently authenticated (logged in) users.

4. WebLog is a HTTP servlet handling POST requests. It can be used by an external entity (e.g. by Joomla or the client) to send messages to the server.

5. WebHist contains both a SOAP-based web service and a servlet. It is used to generate a report out of the session history. The client submits the session information to the server over SOAP and retrieves a PDF document over the HTTP servlet.

6. WebInfo is a HTTP servlet handling GET requests. It returns information on queried genes, such as gene name and available content types. It is used mainly by the Firefox extension.

The structure of the main service of eBioWs, the WebServ, is depicted in Fig. 3.12:

1. EmTexBioWS is the web service class. It holds the methods available for remote calling.

2. Backends holds the classes that are responsible for the information retrieval from the database. To ensure maximum maintainability, the EmTexBioWS class does not interface with the data source, i.e. not even with the Hibernate ORM layer directly, but passes all calls it receives to a proxy. The proxy class first checks the user credentials (if the user is authorized to access this type of data), and if everything is ok, it then forwards the call to a back-end class, the sole purpose of which is to retrieve data. By using this mechanism, it is possible to handle credentials checking and licensing issues separately from the web service - if the licensing policies change, the service remains unaffected. It is also possible to reroute the information retrieval separately from the web service and the licensing issues - the back-end can be a dummy data generator, a database interface, etc.

3. Compatibility holds classes necessary to ensure Java-Silverlight compatibility. If a method throws an exception in the web service, the HTTP response of the message returned to the client will have a response code different than 200 (OK). Due to the architecture of Silverlight, the browser will be the first to intercept this response and since the code is not 200, the message will be discarded. Silverlight will only receive a generic error. To ensure that messages are forwarded to Silverlight even if they contain exceptions, we need to manually set the response code to 200 prior to sending it back. The Compatibility package holds a HTTP filter which does exactly that.

4. Exception holds exceptions that can be thrown during remote method calls (e.g. licensing exceptions if the user is not authorized to access certain information).

5. Security holds classes responsible for user authentication. It holds, among others, the user validator class used during basic authentication (with user name and password). This step of the authentication is performed automatically by Glassfish, which uses this class to determine whether the message needs to be forwarded to the service or not. A more detailed explanation of the authentication mechanism can be found later in this section.

6. Data holds the data objects to be marshaled and sent to the client. It revolves mainly around a graph and various content annotations for its nodes and edges.

   The client retrieves data by querying specific nodes and edges. The only possibility to retrieve data without providing a valid node id is during a search query using a generic string. This results in a list of hits, from which the client / user picks one. A node is defined by its unique id like “ENSPxxxxxxxxxx”, while an edge is defined by two node ids.

All graphs, nodes and edges have types. When retrieving a graph, the client needs to specify which node is being inspected (the graph will be its neighborhood) and the desired graph type. A literature co-citation graph, for example, will have literature co-citation edges.

For each node and edge, various contents are available. They are specified in an inheritance hierarchy rooted in a generic Content class. This class also has two flags: licensed and available. The former is set to true if the user requesting this type of information has the right to retrieve it and false otherwise. The later is set to true if there is any content of this type available for this node. If a user without authorization requests this type of content, he will receive an empty object with the licensed flag set to false and the available flag set accordingly.

Two major groups of contents exist:

- The general node content describes the gene and proteins that are associated to this node. Information such as gene name, function, splice variants, etc. is provided
- The specific node contents hold information such as which papers are associated to this gene, which diseases and drugs are known to be linked to this protein, which clinical studies address this protein, etc. These contents are mostly enumerations. It is possible to retrieve all content types at once or to query only one specific content type. In both cases, the first X entries of each one will be returned. Additionally, WebServ also provides methods for pagination, i.e. the client can retrieve the entries X through Y of a specific content. The edge contents hold information such as which papers are associated to both nodes, or how strong the interaction of the two proteins is.

The second component of eBioWs, the WebAuth, has a similar structure to WebServ. The main difference is the presence of a user registry used to keep track of the registered users presently logged into the system, as well as their respective rights (which licenses they hold). Additionally an IP filter ensures that only log-in request from an authorized machine are accepted.

WebStat, WebLog and WebInfo are simple HTTP servlets without any special architectural considerations.

WebHist comprises a SOAP-based web service which is used by the client to submit a history log containing the user's session activity. Based on this logging information the server generates a PDF history file with additional information on the visited nodes and edges, which can be subsequently downloaded over the HTTP servlet of WebHist.

The user authentication mechanism of eBioWs is depicted as sequence diagram in Fig. 3.13. This process is used in conjunction by WebAuth and WebServ to determine if a user exists, and if the user is allowed to perform a certain operation.

The first authentication aspect is the registration of a user with the service upon login onto the website. This step is executed by Joomla. Clients wishing to perform calls (on WebAuth and WebServ) need to authenticate all their messages using basic authentication (user name and password). This happens implicitly, before the call is executed. Services wishing to check the

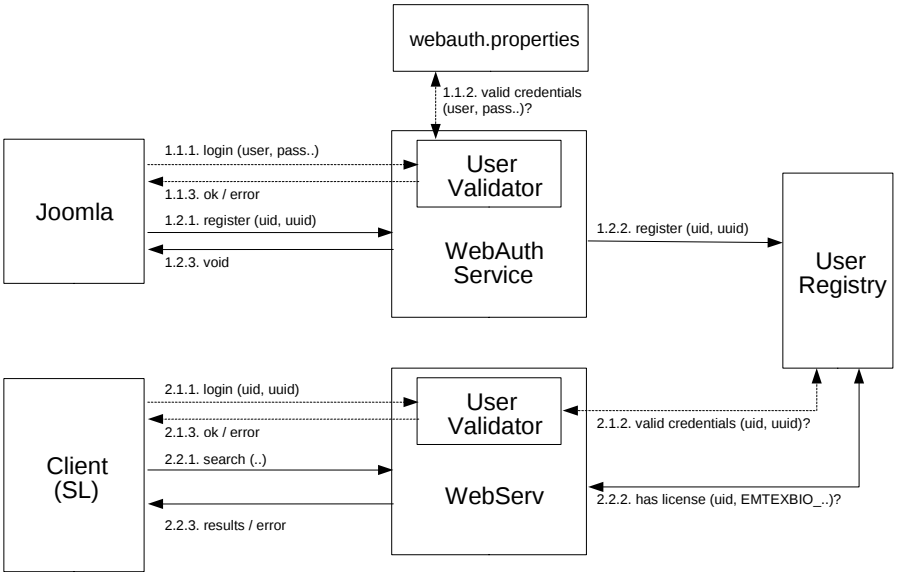

Figure 3.13: eBioWs authentication mechanism

credentials of the incoming messages need to register a UserValidator class with Glassfish which will be called every time a message arrives. This initial credentials check is represented by the login message. This is not a remote call per se, just the first stage of the message processing. It is only depicted as such to enable a more intuitive representation of the process. The UserValidator class in WebAuth is called and handed the credentials (1.1.1.). In case of WebAuth, there is only one user (Joomla authenticates itself always with the same credentials) which are specified in the webauth.properties file. If the credentials are ok, the message is accepted (1.1.3.) and processed further, i.e. forwarded to the service core. The register call itself (1.2.1.) holds user information (name, current UUID, licenses, etc). This data is entered in the user registry (1.2.2.), which is also available to WebServ. This type of call is performed by Joomla when a user logs in on the web page, buys a new module, or proceeds to the navigator.

The second authentication aspect is necessary for each attempt by the user to retrieve data from the server. Such calls are performed by the Silverlight client. Again, the message needs to be checked with basic authentication (2.1.1.-2.1.3.). This time, WebServ checks if this combination of user id and UUID (the UUID for this session) are present in the registry. This is done for all messages regardless of licensing issues. If the user and UUID exist, then the message is processed further, if not, the message is dropped. This procedure is triggered automatically by Glassfish. Before the call itself is executed, another check is performed (2.2.2.) by the back-end proxy, to see if the user is allowed to perform this operation (if a corresponding license is available). If the license is ok, the call is forwarded to the back-end. This, in turn, processes the search query and returns the results (or an error).

All components of the BIO web service are only accessible over HTTPS. Further, the service provides SOAP 1.1 interfaces and it uses two extensions, namely WS-Security (WSS) and WS-Addressing (WSA). The service requires clients to use Message Authentication over SSL to

identify themselves and if using WSA the client must include a WSA message id in the SOAP header. To use these security mechanisms, the METRO 2.0 libraries are necessary.

Summary and outlook

In the previous sections we introduced a platform for the visualization of biological information. Other similar approaches such as STRING or NextBio exist, but they do not address the neighborhood and the annotation aspects simultaneously. The application presented here attempts to address these shortcomings and unite the strengths of these approaches by providing a comprehensive information repository on both levels.

BIO is a three tier application with the information traveling from the omicsNET and Metaverse projects over the BIO server to the Navigator and Firefox extension, with a Joomla website providing the entry point.

A shortcoming of this approach is the usage of Microsoft specific technology for the viewer, restricting its usage to only Windows and MacOS. With an increase in the market share for Linux-based distributions, addressing this shortcoming is a next step in the evolution of BIO. A possible approach is the usage of a cross-platform technology like HTML5. Additional interesting next steps are the optimization of BIO for mobile devices, especially Android and iOS-based systems and a continuous quality assessment and improvement workflow for the provided information.

3.3 A network-based in-silico analysis of diabetic nephropathy[2]

For diseases with complex phenotype such as diabetic nephropathy large scale data integration approaches such as omicsNET promise an improved description of the disease pathophysiology, being the basis for novel diagnostics and therapy. In light of these considerations, omicsNET and BIO constitute essential pillars in investigating the mechanisms of disease.

The work presented in this section is meant to shed light on the disease-associated processes involved in diabetic nephropathy at different Omics levels by using the omicsNET graph as a starting point in the analysis of the molecular features underlying the disease phenotype.

The omicsNET version used in this analysis corresponds to a previous version of the one described in section 3.1. Here, gene expression information across 32 human tissue types as obtained from the Gene Expression Omnibus dataset GSE7905 [175] was included in the edge weight computation. Additionally, information on the sub-cellular localization of proteins as imported from SwissProt or computed with the WPSORT [167] algorithm was included. A more detailed description of the omicsNET version used in this section can be found here [172].

[2]This section is based largely on the scientific article *Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy* by Raul Fechete, Andreas Heinzel, Paul Perco, Konrad Mönks, Johannes Söllner, Gil Stelzer, Susanne Eder, Doron Lancet, Rainer Oberbauer, Gert Mayer and Bernd Mayer published in Proteomics - Clinical Applications, June 2011 [134].

Concept outline and goals

Diabetic nephropathy occurs in both, type 1 and 2 diabetes mellitus. In patients with type 1 disease approximately 20 to 30 percent of affected individuals will develop microalbuminuria as the first clinical sign of renal disease after a median diabetes duration of about 15 years [176]. Early detection of these individuals is crucial as it has been shown that less than half of the patients will further progress to overt nephropathy (defined as urinary albumin excretion > 300 mg/day). This finding is mainly due to better metabolic control, more aggressive blood pressure reduction, and the use of agents blocking the renin-angiotensin system [177,178]. In type 2 diabetes robust data on the incidence of nephropathy were derived from the United Kingdom Prospective Diabetes Study: Among the 5,100 patients with newly diagnosed diabetes the prevalence of microalbuminuria, macroalbuminuria and either an elevated plasma creatinine concentration or requirement of renal replacement therapy after 10 years was 25%, 5% and 0.8%, respectively [179]. However, as the prevalence of type 2 diabetes for all age groups worldwide is expected to rise from 2.8% seen in the year 2000 to 4.4% in 2030, which corresponds to an increase in absolute patient numbers from 171 million to 366 million, it is not surprising that already at present diabetic nephropathy is by far the leading cause of end stage renal disease [180]. Of particular importance is the fact that many patients die because of cardiovascular disease before reaching dialysis [179].

The early diagnosis of diabetic nephropathy rests on the measurement of protein and/or albumin excretion in urine, and microalbuminuria (MIA, excretion of 30-300 mg albumin per day) is currently considered as the diagnostic gold standard. In patients with type I diabetes MIA has an excellent sensitivity as well as specificity to identify patients at risk for the development of more severe nephropathy, and a reduction of urinary albumin excretion is associated with the preservation of the glomerular filtration rate. In patients with type 2 diabetes, which form the majority of subjects in clinical practice, the specificity of MIA for diabetic nephropathy is much lower even though it is an established ominous sign for an adverse cardiovascular prognosis [176, 181]. This lack of specificity however leads to treatment problems. Whereas blockade of the renin-angiotensin system still is very effective in preventing progression of renal disease in patients where MIA is a sign of early nephropathy, the same therapeutic intervention did not preserve glomerular filtration rate in the ONTARGET study, which included subjects where MIA was more an indicator of increased cardiovascular risk [182]. Accordingly, in the ACCOMPLISH trial the combination therapy of benazepril and hydrochlorothiazide reduced albuminuria in hypertensive subjects but resulted in an almost doubled rate of chronic kidney disease compared to benazepril/amlodipine therapy [183]. These data led the participants of a NKF and FDA (National Kidney Foundation and US Food and Drug Administration) sponsored meeting to the conclusion that proteinuria in fact is only a surrogate outcome in kidney disease progression [184]. Multiple risk factors for the development of diabetic nephropathy have been identified. These include amongst others genetic susceptibility, age, race, obesity, smoking, blood pressure, glomerular hypertrophy and hyperfiltration, hyperglycemia, and the formation of advanced glycation end products as well as cytokine activation [185]. Some of these risk factors can be modified by therapy and thus current treatment regimen are directed against systemic and especially intraglomerular hypertension (blood pressure lowering medication and blockade of the renin-angiotensin-aldosterone system), as well as strict glycemic control [186–189]. How-

ever, as cardiovascular mortality is also a major focus of treatment a multifactorial approach, which additionally includes smoking cessation, dietary and behavior modification [190], lipid lowering therapy and administration of aspirin is recommended [191, 192].

Numerous Omics technologies have been applied for deciphering biological processes linked to DN, including genome-wide association studies [193], transcriptomics [194], proteomics [195] and metabolomics [196]. Traversing Omics screening results towards identification of biomarkers is seen with proteomics: Rossing and colleagues applied proteomics profiling in a cohort of over 300 patients for studying diabetic nephropathy and non-diabetic kidney diseases [197], and the method was recently expanded in a multicenter validation study for identification of subjects with DN resting on detection and quantification of urinary collagen fragments [198].

Further integrating such data (cross-Omics) leading to multi-marker models to be used for diagnosis or disease monitoring is seen as future perspective, as recently outlined by Fox et al. [199]. More generally, DN may be considered as ideal case for utilizing the concept of Systems Biology [200], as complex pathophysiology is seen for the kidney in the realm of diabetes mellitus and hypertension, in turn closely linking to bone metabolism and cardiovascular implications denoted as the cardiorenal syndrome. Cross-omics integration of Omics data from kidney transcriptomics and proteomics, together with literature annotation of molecular features relevant in the cardiovascular context identified the coagulation pathway as important cardiorenal process [201, 202]. In this work protein-protein interaction networks (PPIs) were used for studying causative as well as associative dependencies of feature profiles and clinical indication. In the last years the analysis of Omics data on gene or protein interaction networks has been established as de-facto standard for multidimensional data mapping and interpretation [203], and concepts for linking molecular data and clinical phenotype space have emerged. Hidalgo et al. [204] introduced the Phenotypic Disease Network (PDN) as a map summarizing phenotypic connections between diseases. Barrenas et al. employed disease-gene networks [205] with specific focus on topological characteristics of such graphs. Shortest path-based algorithms (among others) are frequently applied for linking multilevel networks aimed at identifying key regulatory genes and proteins [206–209], but also for identification of functional modules and pathways affected in the diseased state [210]. Furthermore, concepts for building disease-specific drug-protein connectivity maps have been introduced [211], altogether aimed at building relations between molecular feature space, clinical data space, and associated markers, drug targets and associated drugs.

In the present work we integrated data sources characterizing DN on a molecular level by utilizing a full human proteome interaction network as common denominator, specifically including data coming from DN case-control Omics studies. We further annotated the network with DN-associated data retrieved from scientific literature, patent text and clinical trial references, and then used this annotated PPI for linking the pathophysiological description of DN with biomarker candidates and therapy targets discussed for diagnosis, prognosis and therapy of DN.

Materials and methods

Data: omics studies

A search in PubMed (link [212], database status as of June 2010) was performed in order to identify publications utilizing Omics in the context of DN. Next to the clinical phenotype "diabetic nephropathy" the keywords "microarray", "transcriptomics", "proteomics", "metabolomics", "metabonomics" or "SNP" were applied. This retrieval of Omics studies resulted in 146 publications, of which 31 specifically reported Omics screening results and features deregulated in DN as identified in a case-control study design. Of these, 17 were done with human specimens further considered, comprised of 4 transcriptomics, 7 proteomics, 2 metabolomics and 4 SNP studies, as referenced in Tab. 3.2.

SNPs found in open reading frames of protein coding genes were directly mapped to the respective gene ID, or otherwise assigned to the gene ID of the next transcription start site of a protein coding reading frame, all according to the ENSEMBL human genome representation. Metabolites were mapped to associated enzymes using the Human Metabolome Database (HMDB) [213] (status as of September 2010). Across all four Omics domains 851 unique gene symbols could be retrieved.

Data: literature mining

Features associated with DN were furthermore extracted via literature mining. A PubMed search using the MeSH heading "diabetic nephropathy" was applied in conjunction with the gene2pubmed association file (link [214], status as of July 2010) maintained by NCBI enabling a gene-to-publication assignment. For determining the relevance of an identified gene-to-DN association, i.e. if a gene was reported significantly more frequent with DN as by chance, a statistics procedure was applied: Contingency tables were computed by splitting the full set of PubMed IDs first with respect to the occurrence in the context of DN and secondly with respect to the occurrence of each specific gene. 265 unique gene symbols identified as having enriched association with diabetic nephropathy could be retrieved. The same approach was applied for the headings "diabetic nephropathy" AND "biomarker". In this restricted setting 108 unique gene symbols were identified.

Data: clinical trials mining

Information on clinical trials focusing on DN was retrieved from ClinicalTrials (link [215]) maintained by the NIH. 260 studies could be identified by using the keyword "diabetic nephropathy" in the search field provided at the clinical trials web site. Among these studies 238 specifically focusing on type 2 diabetes were further considered. 111 interventional drugs were extracted from these 238 studies, and were subsequently queried in the DrugBank database [114]. 64 drugs could be identified in DrugBank holding 176 drug-gene relations, linking to 144 unique gene symbols.

Study	Omics	Reference
1	mRNA	Baelde, H.J. et al., Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. Am J Kidney Dis 2004, 43, 636-650.
2	mRNA	Berthier, C.C. et al., Enhanced expression of Janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. Diabetes 2009, 58, 469-477.
3	mRNA	Cohen, C.D. et al., Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis. PLoS One 2008, 3, e2937.
4	mRNA	Rudnicki, M. et al., Gene expression profiles of human proximal tubular epithelial cells in proteinuric nephropathies. Kidney Int 2007, 71, 325-335.
5	prot.	Dihazi, H. et al., Characterization of diabetic nephropathy by urinary proteomic analysis: identification of a processed ubiquitin form as a differentially excreted protein in diabetic nephropathy patients. Clin Chem 2007, 53, 1636-1645.
6	prot.	Jain, S. et al., Proteomic analysis of urinary protein markers for accurate prediction of diabetic kidney disorder. J Assoc Physicians India 2005, 53, 513-520.
7	prot.	Kim, P.K. et al., Proteome analysis of the rat hepatic stellate cells under high concentrations of glucose. Proteomics 2007, 7, 2184-2188.
8	prot.	Mischak, H. et al., Proteomic analysis for the assessment of diabetic renal damage in humans. Clin Sci (London) 2004, 107, 485-495.
9	prot.	Otu, H.H. et al., Prediction of diabetic nephropathy using urine proteomic profiling 10 years prior to development of nephropathy. Diabetes Care 2007, 30, 638-643.
10	prot.	Rossing, K. et al., Urinary proteomics in diabetes and CKD. J Am Soc Nephrol 2008, 19, 1283-1290.
11	prot.	Sharma, K. et al., Two-dimensional fluorescence difference gel electrophoresis analysis of the urine proteome in human diabetic nephropathy. Proteomics 2005, 5, 2648-2655.
12	metabo.	Xia, J.F. et al., Ultraviolet and tandem mass spectrometry for simultaneous quantification of 21 pivotal metabolites in plasma from patients with diabetic nephropathy. J Chromatogr B Analyt Technol Biomed Life Sci 2009, 877, 1930-1936.
13	metabo.	Zhang, J. et al., Metabonomics research of diabetic nephropathy and type 2 diabetes mellitus based on UPLC-oaTOF-MS system. Anal Chim Acta 2009, 650, 16-22.
14	SNP	Chambers, J.C. et al., Genetic loci influencing kidney function and chronic kidney disease. Nat Genet 2010, 42, 373-375.
15	SNP	Köttgen, A. et al., Multiple loci associated with indices of renal function and chronic kidney disease. Nat Genet 2009, 41, 712-717.
16	SNP	Köttgen, A. et al., New loci associated with kidney function and chronic kidney disease. Nat Genet 2010, 42, 376-384.
17	SNP	Lim, S.C. et al., Microarray analysis of multiple candidate genes and associated plasma proteins for nephropathy secondary to type 2 diabetes among Chinese individuals. Diabetologia. 2009, 52, 1343-1351.

Table 3.2: Omics type and scientific reference of public domain Omics sources used in this work.

Data: patent mining

Patent Lens (link [216]) was used to identify molecular features provided in patents and patent applications associated with DN. Patent Lens offers the full text on granted patents from the US, Europe and Australia, as well as patent applications from the US, WIPO/PCT and Australia. A Patent Lens search was performed using the full-text keyword query "diabetic AND nephropathy", providing 30,604 hits. Only patents linked to patent classes 424 (Drug, bio-affecting and body treating), 436 (Chemistry: Analytical and immunological testing), and 514 (extension to category 424) were retained. For retrieving only patents with evident association with DN a search on each patent was done for extracting patents where either "diabetic nephropathy" was present in the patent title or in the patent claims, or where the search string was found at least five times in the full-text, or the string was found in full text together with specific tags in the claims ("diabetic", "diabetes" or "diab" AND "nephropathy", "kidney" or "nephro"). This process resulted in a list of 1,561 patents with evident link to DN. From these the claims were kept for identification of genes and proteins associated with DN. The present list of valid NCBI symbols and full gene names was retrieved from the current version of GenBank, holding 19,066 genes, further expanded by associated 72,559 deprecated symbols and aliases. These were filtered by an English thesaurus (link [217]) containing 62,109 words, resulting in 18,905 valid symbols and full names, and 71,054 aliases. This query set was then applied for a case insensitive search among text tokens extracted from patent claims, resulting in 901 gene symbols.

Protein interaction network

All data sets generated on DN were mapped to gene symbols as common name space. These were subsequently mapped on a full human proteome interaction network (omicsNET, [89]). This network aims at a complete representation of protein coding genes, furthermore integrating the various types of protein interactions (being physical interactions, procedural interactions, or paralogs). Our network holds 26,419 nodes representing canonical proteins (as provided by SwissProt in combination with ENSEMBL). The starting point in the construction of omicsNET was the set of experimentally determined PPIs as obtained through the unification of available PPI sources (IntAct, OPHID, BioGRID, KEGG, PANTER and Reactome). On top, data on tissue specific gene expression, sub-cellular localization (retrieved from SwissProt and computed by WPSORT), ontology (GO molecular function and biological process) and pathway (KEGG, PANTHER) annotation and protein domains (PDB) were included. This procedure resulted in a maximum of seven data sources per node. On the basis of this parameterization we inferred the probability of a relation between node pairs by weighting the strength of the arguments in favor and against such a relation as encoded in the given parameters for a specific pair, technically represented by a metafunction. This procedure resulted in a matrix of pair relation weights being in the interval [-1,2], where negative values indicate lack of relation, and positive values indicate a relation for a given pair. The accuracy for estimating a relation between nodes certainly depends on the number of available parameters. E.g. tissue specific gene expression (represented in the metafunction as pair-wise correlation coefficient) is available only for 20,282 of the in total 26,419 nodes. The assumption therefore is that a higher number of valid parameters for a given pair (maximum seven parameters) increases the validity of a given edge weight, subsequently

termed as evidence level of the relation. For 19,236 nodes at least four out of the seven parameters were found valid for computing an edge weight, and by selecting an edge weight cutoff of >= 0.58 a graph holding 18,948 nodes resulted, each specified by a unique molecular identifier. This node set and respective interaction network was then annotated by the diverse data sets, i.e. for each node information was added regarding identification in Omics studies, literature, clinical trials or patent text.

PPI network layout

For allowing an interpretation of the annotated omicsNET graph a functional layout in scope of KEGG pathways was performed utilizing the KEGG status as of October 2010. From the in total 214 pathways provided by KEGG all entries specifically focusing on disease phenotypes (as "pathways in cancer") were removed, resulting in a set of 171 "generic" pathways. For each pathway, the KEGG identifiers of the assigned genes were retrieved and mapped to ENSEMBL protein identifiers via Entrez gene identifiers. The cross-referencing used in the process was obtained from the current version of the ENSEMBL database, resulting in 20,462 ENSEMBL protein identifiers. The set was then restricted to proteins present in our PPI network as described above, yielding 7,009 items. 2,924 of these were present in more than one KEGG pathway, and for assuring uniqueness of assignment each of these objects was assigned only to the KEGG pathway already holding the node which showed the strongest weight (as computed in omicsNET) to the non-uniquely assigned node. 2,644 nodes could be assigned this way, resulting in 151 populated pathways further considered. The same procedure was then applied for all nodes from the omicsNET node set not assigned at all to a KEGG pathway. This procedure allowed assigning the remaining nodes to a single pathway of the in total 151 pathways, where the allocation either rested on the given assignment by KEGG as such, or by the edge weights available in omicsNET. Of the 18,948 nodes provided in omicsNET in total 17,995 nodes could be assigned. This approach allowed the clustering of omicsNET in KEGG pathways, subsequently easing interpretation of the DN data sets. For analyzing this KEGG-based clustering of omicsNET a treemap [218] was computed. The inter-pathway similarity between two pathways was calculated as the average over all omicsNET edge weights between the two pathways. This similarity was then transformed into inter-pathway distance. On the basis of these distances a hierarchical clustering using Ward's linkage was computed and represented as a tree with merge steps as dummy nodes and pathways as leaves. From the resulting tree, a treemap was constructed using the Treemap program (link [219]) enabling an inspection of pathway distances. For evaluating if a pathway showed a significant enrichment of annotation with respect to a specific data category (Omics, literature, clinical trials, patents) a Fisher's Exact Test was applied using a p-value < 0.05 as significance level.

Results

Linking disease processes with markers and targets

Due to the prevalence of DN a solid body of data is available, spanning various Omics screening levels and scientific literature, but also patent text and clinical trial descriptions. We ex-

tracted relevant data files and documents on a keyword search basis with the aim of identifying molecular features associated with the disease, and linking these into a canonical gene/protein interaction network (omicsNET).

All primary sources were processed for gene-disease associations, being directly provided from Omics screening results, via MeSH-disease-to-pubmed assignment coupled with gene-to-pubmed data for MEDLINE sources, via drug-gene assignment for clinical trials associated with DN, and via identification of gene symbols and names in DN-relevant patent text. For each extracted association the respective data structure of the molecular feature was expanded for mirroring the identified association, also providing the type of data source. In this approach no specific level of detail regarding the type of association was taken into consideration (e.g. direction and amplitude of differential regulation/abundance as derived in Omics profiling), but solely the fact of a qualified association between a molecular feature and the disease was retrieved. Next to this disease-specific annotation each molecular feature holds a second layer of annotation specifically used for computing a relation score between genes/proteins (molecular context). Applying this procedure to the full set of human protein coding genes provided a complete interaction network for the human proteome. Functional grouping of the molecular features in line with core KEGG pathways allows identification of pathways specifically enriched by one (or more) feature sets holding a disease annotation. Based on the sources used for feature retrieval qualitative layers can be defined holding i) a description of the pathophysiology of the disease (represented by features coming from Omics profiling and literature mining), and ii) holding features reported in the realm of pre-clinical or clinical application, i.e. identified explicitly in the context of biomarkers and therapy targets found in scientific publications, patent text and clinical trial descriptions (in the following denoted as "clinical context"). As a result of this integration strategy each affected pathway becomes amenable for analyzing its association to the pathophysiology as well as to biomarkers and therapy targets.

Feature overlap

First question to be addressed when analyzing heterogeneous feature lists is their direct overlap on the level of molecular identifiers. Tab. 3.3 provides an overview on the molecular feature sets derived from the various data and literature sources.

The source providing most hits is patent text with 901 gene symbols, followed by transcriptomics studies with 708 symbols. Mining of scientific literature provided 287 symbols linked with DN, and 108 symbols associated specifically with biomarkers in the context of DN. On the level of a direct NCBI gene symbol overlap 44 symbols are identified in both, literature and Omics profiles. The gene symbol overlap of combining Omics and literature (the pathophysiology set), and individually comparing this set with the other sources classified under clinical context is shown in Fig. 3.14.

Of the in total 1,094 unique symbols assigned to pathophysiology 102 are also found in the restricted literature mining focused on biomarkers, 144 are represented in patents, and 39 in clinical trial descriptions. Specific comparison of Omics tracks on a feature level is provided in Fig. 3.14. From the in total 708 gene symbols reported from transcriptomics 5 are also reflected on the SNP, 5 on the proteome, and 6 on the metabolome level. Weak overlap on the level of individual features, however, has been reported in various meta-studies of Omics profiles

Level	Source	Classification	# features
Pathophysiology	Literature	“Diabetic Nephropathy”	287
	Omics	Transcriptomics	708
		Proteomics	30
		SNP	36
		Metabolomics	94
	Total unique features		**1094**
Clinical context	Literature	“Diabetic Nephropathy” AND “Biomarker”	108
	Trials	“Diabetic Nephropathy”	144
	Patents	“Diabetic Nephropathy”	901
	Total unique features		**1059**

Table 3.3: Mined sources assigned to either describing the pathophysiology of DN or to the clinical context of DN, source classification and keywords, and number of features (NCBI gene symbols) extracted.

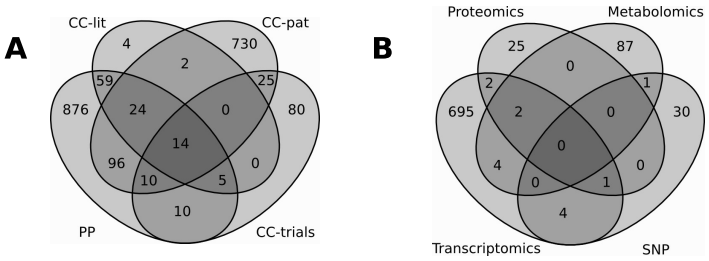

Figure 3.14: Venn diagrams comparing A: features assigned to pathophysiology (PP, Omics, literature) and their overlap with features extracted from a biomarker-biased literature search (CC-lit), patent text (CC-pat), and clinical trials (CC-trials) text. B: features assigned to the individual Omics tracks (SNP, transcriptomics, proteomics, metabolomics).

[30, 89]. This alleged heterogeneity changes when going to the level of pathways populated by identified features instead of comparing individual features as such.

GSEA: KEGG pathway level

A gene set enrichment analysis (GSEA) [90] utilizing KEGG pathways provided a more coherent picture when comparing affected pathways grounded on data sets assigned to pathophysiology and to clinical context, as shown in Tab. 3.4.

Two pathways, namely the renin-angiotensin system (hsa04614) as well as the complement and coagulation cascades (hsa04610) appeared as significantly affected in both, the pathophysiology as well as the clinical context layer. Although the total number of features for both layers

KEGG term	# features	CC, hits	CC, p-value	PP, hits	PP, p-value
hsa04614:Renin-angiotensin system	17	10	0.01	10	0
hsa04610:Complement and coagulation cascades	69	31	0	19	0.01
hsa00760:Nicotinate and nicotinamide metabolism	24	-	-	10	0.04
hsa03320:PPAR signaling pathway	69	-	-	19	0.01
hsa00563:Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	25	-	-	22	0
hsa04060:Cytokine-cytokine receptor interaction	262	93	0	-	-
hsa04010:MAPK signaling pathway	267	77	0	-	-
hsa04660:T cell receptor signaling pathway	108	43	0	-	-
hsa04620:Toll-like receptor signaling pathway	101	41	0	-	-
hsa04340:Hedgehog signaling pathway	56	29	0	-	-
hsa04930:Type II diabetes mellitus	47	25	0	-	-
hsa04621:NOD-like receptor signaling pathway	62	29	0	-	-
hsa04350:TGF-beta signaling pathway	87	35	0	-	-
hsa04722:Neurotrophin signaling pathway	124	42	0	-	-
hsa04062:Chemokine signaling pathway	187	54	0	-	-
hsa04672:Intestinal immune network for IgA production	49	24	0	-	-
hsa04916:Melanogenesis	99	36	0	-	-
hsa04012:ErbB signaling pathway	87	32	0	-	-
hsa04920:Adipocytokine signaling pathway	67	26	0	-	-
hsa04630:Jak-STAT signaling pathway	155	44	0	-	-
hsa04664:Fc epsilon RI signaling pathway	78	28	0	-	-
hsa04662:B cell receptor signaling pathway	75	27	0	-	-
hsa04910:Insulin signaling pathway	135	39	0	-	-
hsa04912:GnRH signaling pathway	98	30	0	-	-
hsa04020:Calcium signaling pathway	176	43	0	-	-
hsa04914:Progesterone-mediated oocyte maturation	86	26	0.01	-	-
hsa04520:Adherens junction	77	24	0.01	-	-
hsa04370:VEGF signaling pathway	75	23	0.01	-	-
hsa04080: Neuroactive ligand-receptor interaction	256	54	0.02	-	-
hsa04110: Cell cycle	125	32	0.02	-	-
hsa04150: mTor signaling pathway	52	18	0.02	-	-

Table 3.4: KEGG identifier and term name, the total number of features assigned to this pathway, the number of features found to be affected utilizing 'pathophysiology' (PP) and 'clinical context' (CC) data sets, as well as associated p-values indicating the significance of association.

is in the very same range only five pathways were found to be significantly affected when mapping the pathophysiology feature set, but 28 pathways were found significantly affected when supplying the clinical context list. Among these are pathways frequently reported in the context of kidney disease, as MAPK signaling [19], VEGF signaling [220, 221], or TGF-beta signaling [222]. For completeness and as positive control we explicitly included the type II diabetes mellitus pathway also provided in KEGG, indeed showing enrichment for the clinical context data set.

GSEA: Extended KEGG pathway level

For overcoming the limitation in coverage of genes in KEGG, but also for going beyond the procedural interactions provided in KEGG, we generated an extended KEGG pathway set holding in total 17,995 proteins. We mapped the five feature data sets (literature, combined Omics, literature with focus on biomarker, patents, and clinical trials) individually to the extended pathways and searched for pathways showing significant enrichment in affected features (i.e. being members of the DN data sets) utilizing a treemap representation for reflecting pathway neighborhood.

Treemaps align pathways with respect to their relatedness, but only a minor number of affected pathways were found in proximity for any of the five data sets. Furthermore, only three pathways were coherently found as enriched in all five maps, namely i) complement and coagulation cascade, ii) the renin-angiotensin system, and iii) the PPAR signaling pathway, where the first two were also identified on the KEGG level, and PPAR was in a first place only found on the level of pathophysiology. Only one additional pathway is jointly affected for Omics data and general literature data, namely "focal adhesion". For the literature data set "TGF-beta signaling" is an affected and direct neighbor of the renin-angiotensin pathway, whereas for the Omics data set "vascular smooth muscle contraction" is significantly affected, apparently reflecting Ca2+ signaling. Two pathways were coherently identified on the clinical context level also being direct neighbors in the treemap representation, namely i) neuroactive receptor-ligand interactions and ii) cytokine-cytokine receptor interaction. The total number of pathways found as significantly enriched on the individual data set level, as well as the comparative display of affected pathways is provided in Tab. 3.5.

Source	O	L	LB	P	CT
Omics (O)	19	6	3	5	6
Literature (L)		17	7	7	4
Literature, biomarker (LB)			10	7	5
Patents (P)				20	7
Clinical trials (CT)					14

Table 3.5: The feature source, the number of pathways enriched on the level of an individual source, and the number of pathways jointly affected in a pair-wise comparison.

Among the 151 pathways represented in the treemaps at maximum 20 are affected by a single source (patent feature list), and only three to seven pathways are jointly found as enriched when

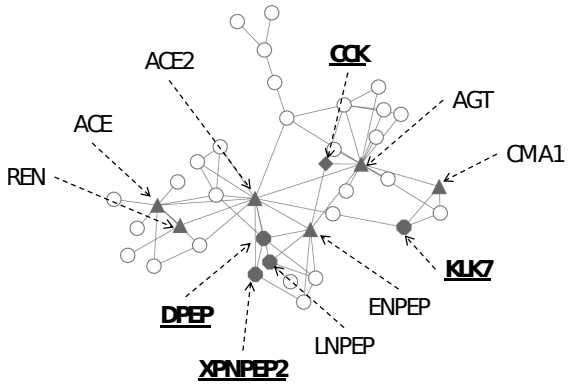

Figure 3.15: Network view on the renin-angiotensin system (hsa04614). Nodes are encoded according to category assignment as triangles (multiple sources), octagons (pathophysiology-omics), diamonds (patents), and circles (not identified as affected in any of the given data sets). Edges represent relations as provided from omicsNET. Underlined gene symbols were assigned to this pathway based on the extension procedure applied.

comparing two data sets. In our pathway representation each pathway is populated by members assigned by KEGG, and members derived from the extended assignment of gene symbols based on omicsNET edge weights. On the basis of omicsNET all members of a specific pathway have relations assigned, enabling the extraction of pathway-specific subgraphs. The subgraph of the extended renin-angiotensin pathway is shown in Fig. 3.15.

Of the 37 members (level of gene symbols) represented in the extended KEGG "renin-angiotensin pathway" (compared to 17 members of the original KEGG pathway hsa04614) 11 are found in at least one data set, with multiple data source annotation for the proteins ACE (angiotensin 1 converting enzyme 1), ACE2 (angiotensin 1 converting enzyme 2), AGT (angiotensin), CMA1 (chymase 1), ENPEP (glutamyl aminopeptidase), and REN (renin). Further features identified as relevant but not being members of the original KEGG pathway are CCK (Cholecystokinin), XPNPEP2 (X-prolyl aminopeptidase (aminopeptidase P) 2, DPEP (a renal dipeptidase), and KLK7 (kallikrein 7). Three pathways were found individually enriched by all five extracted data sets. Biomarkers (derived from the focused literature mining) and therapy targets (as derived from clinical trials text) assigned to these pathways are listed in Tab. 3.6.

Biomarkers and targets associated with hsa04614 (renin-angiotensin system) obviously hold ACE and AGT on both, biomarker and target level. A further consensus finding were the pathways has04080 (neuroactive receptor-ligand interactions) and hsa04060 (cytokine-cytokine receptor interaction), being enriched in the clinical context setting resting on the focused biomarker search in the scientific literature, patent information, as well as clinical trial text. In total 16 specific drugs are linked to hsa04080 (mainly sartans acting as angiotensin II receptor antagonists), and two to hsa04060 (statins for lowering cholesterol levels via inhibiting HMG-CoA reductase) according to the gene-drug assignments from DrugBank. Their main associated disease indications as assigned in MeSH are given in Tab. 3.7, clearly reflecting cardiovascular medication for drugs linked to hsa04080 and cholesterol reduction, and lipid balance and diabetes for hsa04060.

ext. KEGG id	biomarker	target	drug
hsa04610	CD59	SERPINE1	Atorvastatin
	F3	F10, F2, SERPINC1	Enoxaparin
	FGB	C1QA, C1QB, C1QC, C1R, C1S	Rituximab
	HP	F2, MMP3, SERPINE1	Simvastatin
	KLKB1	SERPINE1, SERPINC1, SERPIND1	Sulodexide
	LPA		
	PTX3		
hsa03320	ALB	APOB, PON1, PPARA, PPARG	Atorvastatin
	APOB	ALB	Captopril, Insulin-Glargine
	FABP1	VDR	Paricalcitol
	HPR	PPARG	Pioglitazone, Rosiglitazone
	PON1	PPARA	Simvastatin
hsa04614	ACE	AGT, REN	Aliskiren
	ACE2	AGT	Atorvastatin, Simvastatin, Lisinopril, Irbesartan
	AGT	ACE	Benazepril, Captopril, Enalapril, Fosinopril, Lisinopril, Ramipril, Trandolapril

Table 3.6: Biomarkers and targets including assigned drugs as represented in the extended KEGG pathways hsa04610 (complement and coagulation cascade), hsa03320 (PPAR signaling pathway) and hsa04614 (renin-angiotensin system).

MeSH terms associated with PubMed hits found by individually searching these 16 drugs together with a provided MeSH modifier for obtaining systematic reviews of value to clinicians ("systematic[sb]") allowed us to count and rank literature occurrence of diseases associated with these drugs, as listed in Tab. 3.8.

Discussion and conclusion

Omics technologies have significantly broadened our experimental capabilities for describing the molecular status of a given biological matrix (tissue, blood or urine level). Improvements in experimental workup of samples, SOPs for Omics screening, and standardization in reporting including both, experimental description as well as execution have provided the ground for utilizing Omics also in the clinical context. Here individual Omics tracks, as proteomics in the field of diabetic nephropathy, have already entered validation in disease diagnosis and prognosis.

KEGG id	drug	pharmacology
hsa04080	ATENOLOL	Adrenergic beta-Antagonists, Anti-Arrhythmia Agents, Antihypertensive Agents, Sympatholytics
	BOSENTAN	Antihypertensive Agents
	CANDESARTAN	Angiotensin II Type 1 Receptor Blockers, Antihypertensive Agents
	CLONIDINE	Adrenergic alpha-Agonists, Analgesics, Antihypertensive Agents, Sympatholytics
	CORTICOTROPIN	Hormones
	DOXAZOSIN	Adrenergic alpha-Antagonists, Antihypertensive Agents
	EXENATIDE	Hypoglycemic Agents
	FOLIC ACID	Hematinics, Vitamin B Complex
	IRBESARTAN	Angiotensin II Type 1 Receptor Blockers, Antihypertensive Agents
	LOSARTAN	Angiotensin II Type 1 Receptor Blockers, Anti-Arrhythmia Agents, Antihypertensive Agents
	OLMESARTAN	Angiotensin II Type 1 Receptor Blockers
	PERINDOPRIL	Angiotensin-Converting Enzyme Inhibitors, Antihypertensive Agents
	SERTRALINE	Antidepressive Agents, Serotonin Uptake Inhibitors
	SPIRONOLACTONE	Aldosterone Antagonists, Diuretics
	TELMISARTAN	Angiotensin II Type 1 Receptor Blockers, Angiotensin-Converting Enzyme Inhibitors
	VALSARTAN	Angiotensin II Type 1 Receptor Blockers, Antihypertensive Agents
hsa04060	ATORVASTATIN	Anticholesteremic Agents, Hydroxymethylglutaryl-CoA Reductase Inhibitors
	SIMVASTATIN	Hydroxymethylglutaryl-CoA Reductase Inhibitors, Hypolipidemic Agents

Table 3.7: Drugs assigned to the pathway hsa04080 (neuroactive receptor-ligand interactions) and hsa04060 (cytokine-cytokine receptor interaction), and pharmacological action.

Workup of microdissected kidney tissue forwarded to transcriptomics has resulted in the discovery of major pathways in the tubular compartments seen with chronic kidney disease [221]. However, most of these procedures have been implemented "within the Omics domain", still centrally driven by statistics procedures aimed at identifying differentially regulated transcripts or differentially abundant proteins. This domain separation also becomes evident by the highly valuable Omics profile consolidation efforts e.g. seen for transcriptomics at ArrayExpress [223], or for proteomics with PRIDE [224], or more general gene-centric data consolidation as with GeneCards [225]. Disease-specific Omics repositories are unfortunately still sparse and valuable examples as Oncomine (Oncomine^{TM} - Compendia Bioscience, Ann Arbor, MI; focus on

hsa04080	count	hsa0460	count
HYPERTENSION	68	HYPERCHOLESTEROLEMIA	38
COLORECTAL NEOPLASMS	56	CARDIOVASCULAR DISEASES	39
NEURAL TUBE DEFECTS	41	HYPERTENSION	5
CARDIOVASCULAR DISEASES	64	DIABETES MELLITUS	11
HEART FAILURE	24	RHABDOMYOLYSIS	4
FOLIC ACID DEFICIENCY/ HYPERHOMOCYSTEINE-MIA	26	DEMENTIA	2
PREGNANCY COMPLICATIONS	13	STROKE	2
STROKE	13		
OPIOID-RELATED DISORDERS	13		
DIABETES MELLITUS, TYPE 2	12		

Table 3.8: Name and frequency (count) of indications assigned to drugs found as associated with the pathway hsa04080 (neuroactive receptor-ligand interactions) and hsa04060 (cytokine-cytokine receptor interaction).

neoplasm) or Nephromine (link [226], focus on renal expression profiles) are transcriptomics centered. Another initiative, SysKid (link [227]), takes a different route, namely limiting the clinical phenotype under analysis, but in turn broadening the Omics disease annotation beyond transcriptomics by also including GWAS, proteomics and metabolomics. The present paper follows this concept for characterizing diabetic nephropathy, furthermore including text mining results from scientific literature, patent text as well as clinical trial descriptions yielding 1,094 features characterizing the pathophysiology, and 1,059 features associated with the clinical context of the disease.

A first finding in cross-Omics data comparison is on the level of direct feature comparison, where meta-analysis within an Omics domain [89] as well as cross Omics domains [201] shows sparse overlap. Utilizing this procedure for the given Omics data sets on DN provides the same conclusion, and also when broadening the features included beyond Omics does not substantially change this picture. In terms of biological causality with respect to abundance levels from Omics (the central result from a statistics-driven profile analysis), however, this finding is not surprising: SNP for instance may affect efficacy of a protein's function but this fact is eventually not mirrored on the transcript level with respect to different concentration as determined in a microarray experiment. Furthermore, different Omics profiling done in different sample types (e.g. tissue, plasma, urine) characterize joint, but also up- and downstream processes. Based on this background cross-Omics analysis on the level of networks and pathways became the method of choice with the underlying assumption that the heterogeneity of identified individual features

consolidates on the level of functional units (pathways), as these provide a causal link between concerted molecular processes being responsible for a given disease phenotype. For pathways also interdependencies can be described (e.g. encoded as pathway distance in treemaps), which might support elucidation of sequences of processes (as inherently given when e.g. integrating tissue transcriptomics and urinary proteomics profiles). Data on functional units, represented as interaction networks (graphs), is available for a number of generic cellular processes with KEGG as the most prominent example. Mapping the data sets on DN to KEGG and performing a GSEA identified five pathways as affected when using combined Omics and literature feature sets. 80 out of the in total 1,094 features are assigned to these five pathways, all other features are assigned to other pathways where the statistical analysis of the number of total features given in a pathway and the number of respective features also found in either the Omics or the literature data set did not show significant enrichment. In any case, identification of affected pathways is significantly improved when combining the different Omics domains. The number of enriched pathways is substantially expanded when using the clinical context lists, resulting in 28 pathways resting on in total 406 features from the 1,059 features assigned to clinical context. Two pathways are found as affected for both, pathophysiology as well as clinical context, namely the renin-angiotensin system as well as the complement and coagulation cascade. The renin-angiotensin system depicts an important mechanism in the regulation of blood pressure with an increased production of renin in the kidney leading to a constriction of blood vessels and an increased blood pressure. Next to the direct inhibition of renin in order to lower blood pressure in patients with diabetic nephropathy, therapeutic options targeting the angiotensin-converting enzyme or the angiotensin II receptor are in clinical use. Next to the blood pressure lowering capabilities of angiotensin-converting enzyme inhibitors (ACEIs) and angiotensin II type 1 receptor blockers (ARBs), their independent anti-proteinuric effect made them a major factor in treatment of chronic kidney disease. The complement cascade is an essential part of the innate immune system, and the complement components C5b-9 have also been detected in urine of diabetic nephropathy patients [228]. The link between complement and the coagulation cascade and the contribution to inflammation and other diseases has been outlined by Markiewski and colleagues [229].

However, available pathway data as encoded in KEGG are far from being complete with respect to the number of protein coding genes, and fall short in integrating the various types of interactions. For expanding both, comprehensiveness as well as coverage of interaction types we utilized omicsNET aimed at a complete representation of protein coding genes in a network of relations encapsulating various types of interactions. However, omicsNET does not provide functional units, and recalling the above said such units are deemed necessary for consolidating and interpreting the diverse Omics data. Therefore nodes not represented in KEGG were assigned to KEGG pathways using the omicsNET relations weight as assignment criterion: Each node in the first place not represented in KEGG was assigned to a given KEGG node showing strongest relation with the non-assigned node. Result of this procedure is a clustering of omicsNET in KEGG pathway categories.

The diverse data sets retrieved on DN were mapped on this extended KEGG category set, and GSEA was again performed resulting in 10 to 20 enriched pathways for each data source. The two pathways already seen in KEGG, namely the renin-angiotensin as well as the comple-

ment and coagulation cascade, were again identified, but in this setting significant enrichment was given individually for all five data sets. Due to the fact that omicsNET is underlying the pathways, extraction of the pathway specific relations network became possible. The functional unit presents as a single connected component, and short paths are seen for major players of this pathway including ACE, REN and AGT, as well as of members which were assigned to this pathway as KLK7, linking the kallikrein-kinin system. Due to mapping all data sources on a common name space the assignment of features to biomarkers and therapy targets is straight forward. Cholesterol lowering drugs, insulin sensitizers, and ACE inhibitors are prominently linked in this category together with sulodexide, a glycosaminoglycan mixture actively tested in DN patients with urinary albumin excretion [230]. Additionally the PPAR signaling pathway, in the first setup only seen for the combined Omics and literature data set, became significant for all five data sets. PPARs are involved in modulating insulin resistance, hypertension, dyslipidemia, obesity, and inflammation. PPARs depict promising alternative therapeutic targets next to the above mentioned molecules of the renin-angiotensin pathway for diseases like type 2 diabetes, obesity, hypertension, hyperlipidemia, or atherosclerosis. A number of clinical trials suggest the renoprotective effects of PPAR agonists [231].

From this unbiased data selection and analysis three pathways became evident allowing a conclusive link of biomarkers and therapy targets on the basis of a molecular description of the pathophysiology of the disease. However, two additional pathways, functionally being in close proximity, were identified congruently for the clinical context data sets, but not being significantly affected on the pathophysiology level, namely “neuroactive receptor-ligand interactions” and “cytokine-cytokine receptor interaction”. Next to statins found for the cytokine-cytokine receptor interaction pathway the class of sartans (angiotensin II receptor antagonists) is prominently linked to the neuroactive receptor-ligand interactions. Cross-evaluating the drugs found in these two pathways with assigned disease names according to NCBI MeSH clearly reflects hypertension and diseases afflicted with lipid metabolism.

The molecular pathways retrieved by the multi-source consolidation perfectly match the clinical view regarding risk factors for developing DN, involving among others hyperglycemia and cytokine activation, and associated therapy regimes. Data integration also demonstrates the intricate entanglement of decreased kidney function and hypertension in the realm of diabetes mellitus. Certainly the various data sets utilized in this work carry different levels of evidence and specific biases. For explorative (bias-free) Omics studies the study design, and here specifically proper case and control definition as well as appropriate statistical power, define evidence of features termed as relevant in the context of DN. Automated data retrieval from literature and patents on the other hand results in less evident and also biased feature extraction. Integration of the clinical relevance data space with Omics profiles nevertheless is supportive for interpretation of explorative data in the realm of known associations next to identification of novel features and respective pathways. Consolidation of molecular features on interaction networks provides furthermore the basis for linking molecular profiles with associated biomarkers and therapy targets. Expanding the concept presented here towards also linking detailed clinical data for each individual Omics profile as yet another layer would generate patient (cohort) specific molecular pathology landscapes, which in a next step could be linked with cohort specific diagnostics and therapy regimes. An apparent shortcoming of the approach presented here is the on a clinical

level still broad spectrum of DN (level of albuminuria, stage of the disease), which was not specifically addressed in the course of data set retrieval. For implementing a true Systems Biology approach, however, specific clinical data specifying the context of Omics profiles, but also specifying the context of individual molecular features are necessary. Feeding a multilayered reference network as delineated in this work with patient specific Omics profiles/marker profiles, and analyzing affected pathways in the realm of specific clinical data might then offer a route towards individualized therapy strategies.

Multilevel data consolidation and integration for diabetic nephropathy, involving next to Omics profiles also literature, patent and clinical trial text mining provides a heterogeneous spectrum of molecular features. This alleged heterogeneity vanishes when interpreting the features in the context of pathways and protein interaction networks, where distinct sets of disease associated processes become evident. Identification of such disease associated pathways furthermore allows to link pathway-specific biomarkers and drug targets. This integration concept proves valid in representing the intricate interplay of diabetic nephropathy, cardiovascular disease and diabetes on a molecular network level, in turn offering a platform for analysis of biomarkers and drugs in the context of specific Omics profiles characterizing diabetic nephropathy.

3.4 A units-based in-silico approach exemplified for diabetic nephropathy

In the previous section, an approach to disease characterization based on network and gene set enrichment analysis (GSEA) was delineated for diabetic nephropathy. Information on related patents and drugs was additionally cross-linked and interpreted. As is the case with most approaches, heterogeneity on the molecular level is a major issue hampering results quality.

The variance issue has already been addressed in the course of this thesis by introducing the concept of synthetic (multi-)lethality. However, this was done in the context of cancer, in which coverage was interpreted as the percentage of tumors that could be driven into apoptosis by targeting a single, or otherwise small set of genes. Synthetic lethality is therefore reasonable for targeting a wide range of heterogeneous patients when there is no need to differentiate between the mechanisms underlying the disease phenotype.

For diseases where treatment does not involve removing the affected cells, but instead attempting to stimulate protection or even damage reversal mechanisms, a detailed knowledge of the involved processes is necessary. Ideally, each patient would be screened down to the molecular level and a corresponding treatment would be prescribed. Such approaches, however, would be prohibitively expensive and complex, not to mention technically difficult at the present moment so they are not an option for clinical routine. An alternative is patient stratification based on biomarker panels, i.e. patients exhibiting the same disease phenotype are screened more in-depth with a predefined set of clinical parameters and provided treatment depending on the outcome.

The following section will introduce the concept of units, an approach to disease investigation and patient stratification based on identifying disease-specific gene groups.

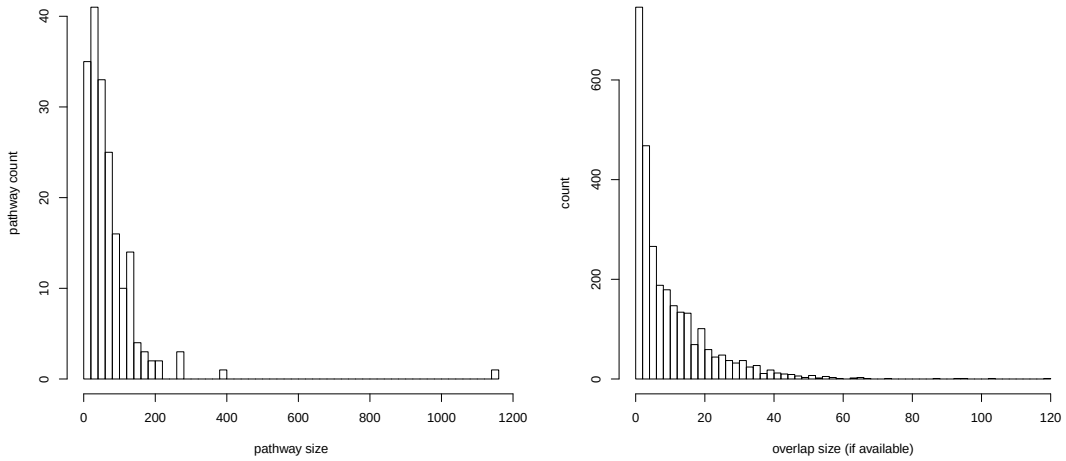

Figure 3.16: Distribution of KEGG pathway sizes (left) and distribution of KEGG pathway pairwise overlaps (right) when available.

Concept outline and goals

Our concept is based on the hypothesis that small sets of genes (units) which perform well-defined tasks exist, and that processes in the broader sense, like pathways, consist of one or more such units.

To better outline the concept of units, a quick comparison to pathways will be provided next. Fig. 3.16 (left) depicts the distribution of KEGG pathway sizes. Aside from pathway hsa01100 with 1,150 genes, which is an integrated map of all metabolic networks stored in KEGG, as well as several other outliers with approximately 200 members each, the vast majority of pathways holds less than 150 genes. The median value lies at 50 genes per pathway. This provides a rough estimate for the upper bound of the envisioned unit size. To further characterize this value, an exhaustive binary pathway overlap was computed. Fig. 3.16 (right) depicts the sizes of the pathway-pathway intersections. Of the 17,955 possible binary combinations, 84% were empty with no overlap between the investigated pathways. These values were not depicted in the figure. The remaining 16% had a median overlap value of 6.

The median pathway size of 50 and the median overlap size of 6 characterize the approximate number of genes in the envisioned units. In the subprocess interpretation of units, the pathway overlap size alone, however, does not provide any information on the number of units in the intersection. It is conceivable, e.g. that in an overlap of size 10, one or maybe two units may be present. Considering that KEGG is a highly curated data source, we can expect that these values lie actually higher than the ones determined here. This is even more the case when considering that KEGG only holds high confidence information implicitly containing false negatives. The median of the term size distribution for the GO process ontology lies at 4 genes per term. This further supports the hypothesis of a relatively small unit size.

A further question to be addressed is the unit computation procedure. Plainly using ontology terms would be a valid approach with the additional advantage of good literature curation. A drawback is that the ontology would still need to be artificially pruned. Additionally, the graph topology would be ignored in this approach. Barabasi et al. [21] specifically emphasize the role of topology in disease modules. Functionally related genes tend to be stronger connected to each other than unrelated ones. This finding leads the investigation of unit computation mechanisms towards network clustering methods.

Topology-based approaches on hybrid networks such as omicsNET make sense as they integrate both information on graph characteristics and background biological data. The latter includes not only the Gene Ontology but also additional sources to improve completeness. In the course of this work, several approaches to building clusters were investigated, such as walktrap [232], MCL [233], CAST [234] and hierarchical clustering, most leading to a single giant connected component. Better results, in terms of smaller units, were obtained with the MCODE [235] algorithm for computing k-core components.

Assuming units can be computed with a given algorithm, the question to be addressed next is the biological function embodied by the resulting sets. In a hybrid network, functional information is merged, making a subsequent dissection of units cumbersome. To obtain a rough description of the retrieved subprocesses, functional annotation of the groups needs to be performed. A sensible approach is to link units to related ontology terms and pathways. This functional annotation step is reasonable when expert analysis of the results is the next intended step. Alternatively, when units are computed based on background data specific to different phenotypes, a black-box approach is conceivable, in which patient assignment to disease-specific units is the actual goal.

Based on the units interpretation as subprocesses, it would be sensible to compute such groups on the whole graph. It is known, however, that genes play more than one biological role, which is not only reflected in units being part of different processes but also in genes being involved in different units. This fact is prohibitive for a meaningful whole-graph segmentation, as processes and units overlap, leading to large and densely interlinked components. This fact became particularly obvious during clustering attempts on the full graph.

A different method for handling units can be derived by putting a disease in relation to the normal cell state. It can be assumed that on the molecular level a disease does not consist of single transcriptional deregulations with no relation in-between, but instead it consists of entire subprocesses that are affected and are therefore acting abnormally. The reasons why during omics methods only some genes are identified as differentially regulated are manifold, and include down-stream effects, measurement errors, etc. In our context, this would translate into entire units being affected. This means further, that, at least theoretically, when comparing healthy vs. diseased tissue, precisely the genes constituting the units underlying the disease phenotype would be identified. It can also be expected that during disease the majority of the cell functions remain intact, so attempting to determine units based on affected genes should prove easier that on the complete graph, as less functional overlap is present in the input set.

The units - disease phenotype relation is bidirectional. Mechanistically, deregulated processes lead to changed phenotype. When investigating the disease, however, identifying a clear phenotype is the first step to obtaining a clear picture of the deregulated processes.

Two distinct application scenarios for units are conceivable:

- **Explorative investigation of disease.** In this approach, phenotype-specific gene sets are used to compute individual unit groups. Relation strengths between units (both intra- and inter-group) as well as between units and GO terms / pathways / drugs / etc. are subsequently computed to support functional annotation. This method is ideal for supporting hypothesis generation and drug repositioning. This is also the method that will be exemplified in the following.

- **Patient stratification.** In this second approach, the focus lies mainly on classifying patients by disease phenotype and less on understanding the mechanisms of disease.

  As previously mentioned, patient disease profiles are heterogeneous depending on the affected processes. As treatment is tightly coupled with the causes of disease, an accurate patient classification is crucial towards effective treatment. Classical clinical approaches using single biomarkers as common denominators for testing generally have poor results due to strong variance in the data. A solution to this shortcoming is using panels of biomarkers characterizing the different processes commonly affected by the disease. Depending on which of the biomarkers are found a classification of the patients for therapy is possible.

  To obtain the biomarker panel, the relevant disease processes need to be identified first. For this, units are computed based on background data, ideally with a clear phenotype definition. Each heterogeneous patient profile constitutes -assuming our hypothesis is correct- henceforth a sub-selection from these units, i.e. the genes of aberrant subprocesses in this particular patient should provide a selection from the ones found in the background data. In practice this matching is not expected to be perfect due to experimental limitations. Furthermore, a genome-wide patient screening is again not feasible due to complexity. To address this shortcoming, biomarkers can be chosen from each unit, which together constitute the actual screening panel.

  This second method is currently under assessment and is not within the scope of this thesis.

Materials and methods

Workflow overview

The workflow for deriving units is depicted in Fig. 3.17. Initially, a phenotype-specific gene set is retrieved. The members are then connected using omicsNET edges. Only direct connections between the genes are taken into consideration. On the resulting graph, strongly connected components are derived using the MCODE algorithm. Subsequently, unit-unit dependencies as well as unit-term and unit-pathway relations are computed. Finally, sub-cellular location and drug-target relationships are determined for the members of each unit.

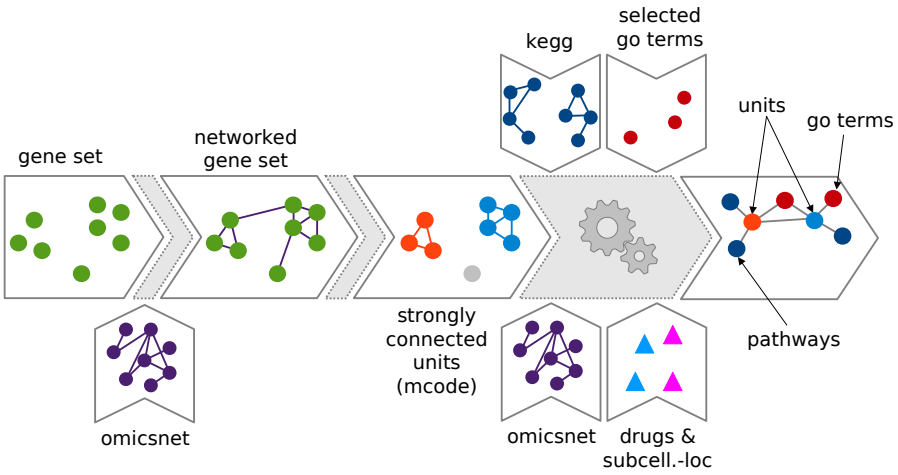

Figure 3.17: Overview of the units workflow.

Omics studies

For demonstrating the principle of units in the context of diabetic nephropathy, a set of nine scientific articles with clear phenotype definition for early and late disease stage respectively was selected. The list of papers is given in Tab. 3.9 and it represents a subset of the literature used for the analysis performed in section 3.3.

The feature count obtained from the selected literature is shown in Tab. 3.10. Noticeably the gene sets are significantly larger for transcriptomics than for proteomics. The original data sets were retrieved with NCBI gene ids, gene symbols and SwissProt identifiers and were subsequently mapped to ENSEMBL ids. The set consolidation led to 187 features specific for the early and 450 features specific for the late disease stage. 67 features were common to both groups.

Protein interaction network

For connecting the features, the omicsNET graph at a cutoff value of 0.73, holding 616,816 edges was employed. A description of the omicsNET graph can be found in section 3.1.

Units computation

The MCODE [235] algorithm aims at finding strongly connected components and was initially developed to identify protein complexes in PPI networks. In our case, the base hypothesis is that subprocesses are tighter connected than their surrounding neighborhood.

The algorithm encompasses three steps:

1. **Node scoring.** Each node is assigned a score based on how strongly its immediate neighbors are interconnected.

Study	Omics	Reference
1	mRNA	Baelde, H.J., Eikmans, M., Doran, P.P., Lappin, D.W., et al., Gene expression profiling in glomeruli from human kidneys with diabetic nephropathy. Am J Kidney Dis 2004, 43, 636-650.
2	mRNA	Berthier, C.C., Zhang, H., Schin, M., Henger, A., et al., Enhanced expression of Janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. Diabetes 2009, 58, 469-477.
3	mRNA	Cohen, C.D., Lindenmeyer, M.T., Eichinger, F., Hahn, A., et al. Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis. PLoS One 2008, 3, e2937.
4	mRNA	Woroniecka KI, Park AS, Mohtat D, Thomas DB, Pullman JM, Susztak K., Transcriptome analysis of human diabetic kidney disease., Diabetes. 2011 Sep;60(9):2354-69. Epub 2011 Jul 13.
5	prot.	Jain, S., Rajput, A., Kumar, Y., Uppuluri, N., et al., Proteomic analysis of urinary protein markers for accurate prediction of diabetic kidney disorder. J Assoc Physicians India 2005, 53, 513-520.
6	prot.	Kim, P.K., Kim, M.R., Kim, H.J., Yoo, H.S., et al., Proteome analysis of the rat hepatic stellate cells under high concentrations of glucose. Proteomics 2007, 7, 2184-2188.
7	prot.	Alkhalaf A, Zürbig P, Bakker SJ, Bilo HJ, Cerna M, Fischer C, Fuchs S, Janssen B, Medek K, Mischak H, Roob JM, Rossing K, Rossing P, Rychlík I, Sourij H, Tiran B, Winklhofer-Roob BM, Navis GJ; PREDICTIONS Group., Multicentric validation of proteomic biomarkers in urine specific for diabetic nephropathy., PLoS One. 2010 Oct 20;5(10):e13421.
8	prot.	Rossing K, Mischak H, Dakna M, Zürbig P, Novak J, Julian BA, Good DM, Coon JJ, Tarnow L, Rossing P; PREDICTIONS Network., Urinary proteomics in diabetes and CKD., J Am Soc Nephrol. 2008 Jul;19(7):1283-90. Epub 2008 Apr 30.
9	prot.	Sharma, K., Lee, S., Han, S., Lee, S., et al., Two-dimensional fluorescence difference gel electrophoresis analysis of the urine proteome in human diabetic nephropathy. Proteomics 2005, 5, 2648-2655.

Table 3.9: Literature list for the diabetic nephropathy unit analysis.

2. **Cluster finding.** Build clusters by starting from the highest score nodes and moving outwards. Adjacent nodes are included if their score is above a given threshold. This step is repeated until no more changes can be performed. Clusters with less than k-cores are filtered out.

3. **Post-processing.** Optional *haircut* (removal of poorly connected nodes from the clusters) and *fluff* (next-neighbor expansion of clusters) can be performed.

For the units computation, the MCODE algorithm was executed with the standard configuration parameters (node degree cutoff = 2, no loops, node score cutoff = 0.2, k-core = 2, maximum depth = 100) for each of the early and late stage gene sets.

Fig. 3.18 (bottom) shows the early disease stage network loaded in Cytoscape and being processed with the MCODE implementation provided by the AllegroMCODE plugin. The nodes highlighted in yellow show the unit with the highest score (also highlighted on the right side).

Stage	Omics	Glomeruli	Tubuli	Urine	Blood
early	mRNA	49 / 128	-	-	-
	prot.	-	-	3 / 0	3 / 10
late	mRNA	53 / 178	262 / 29	-	-
	prot.	-	-	12 / 13	3 / 10

Table 3.10: Feature set sizes for early and late disease stages grouped by tissue. The two numbers in each cell indicate how many features are up- and down-regulated respectively.

Ontology terms

Considering that the process section of the Gene Ontology holds 21,433 concepts with any number of genes between 0 and 14,110 a selection mechanism for the terms had to be developed to allow meaningful unit annotation. To achieve this, a method similar to the one presented by Alterovitz et al. [236] was developed.

The procedure encompasses the following steps:

1. Associate each term t an information content value $ic(t) = -log_2(\frac{n}{N})$ based on its size n relative to the total number of genes N in the ontology.

2. For a preselected information content value x and a maximum deviation k, choose all terms t_i with $ic(t_i) \in [x - k, x + k]$ for further processing.

3. While keeping the selected terms in a sorted queue based on their increasing distance from x, choose the one nearest to x, mark it as valid and invalidate all its ancestors and descendants in the ontology by removing them from the queue.

4. Repeat the previous step until the queue is empty.

The values x and k are used to limit the set of relevant terms to a certain range in order to avoid too specific or too general ones. Removing the ancestors and descendants ensures that the terms in the output set cover mostly distinct processes. This latter step prevents over-annotation of units with similar terms.

For the processing of the early / late dataset we chose $x = 9$ and $k = 1$ leading to a selection of 999 terms holding between 15 and 60 genes each. This set of terms annotates 7,403 genes.

Pathways

To facilitate KEGG annotation of units, the pathway-gene associations for all KEGG pathways (version as of January 2012) was retrieved. The data included 248 pathways and 6,198 genes. From these, pathways with distinctive disease focus were removed, reducing the original dataset to 190 pathway and 5,781 genes.

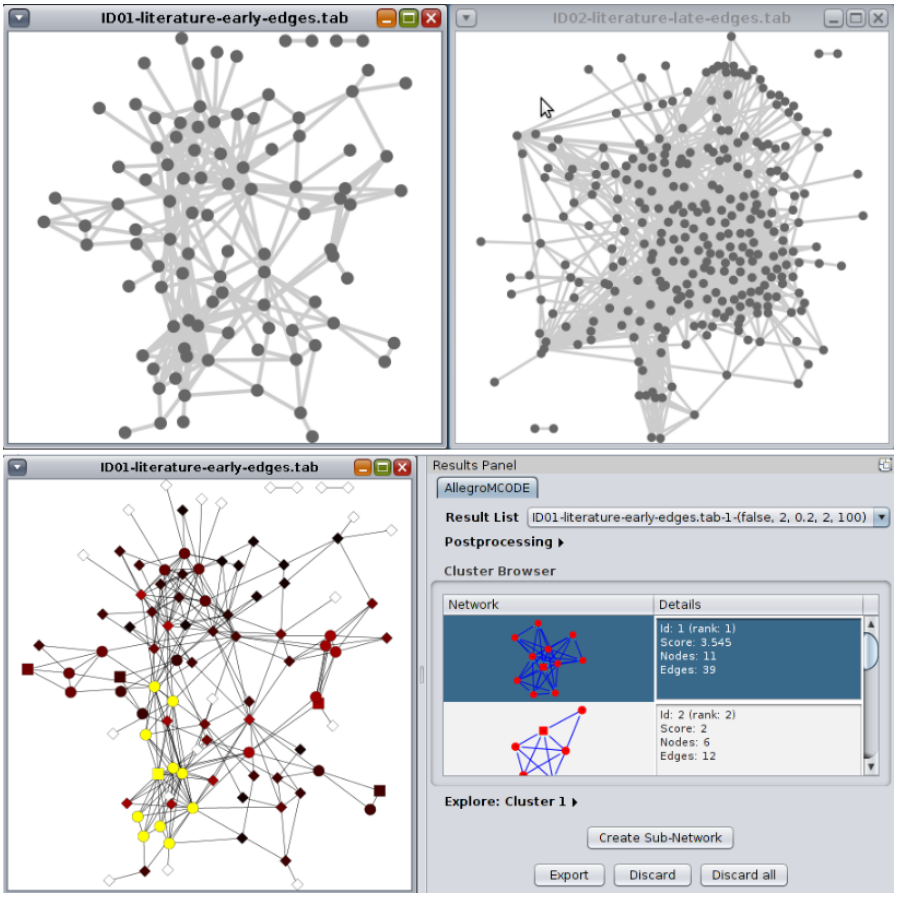

Figure 3.18: Phenotypic networks for early and late disease stages (top). Early disease network with units computed by the AllegroMCODE Cytoscape plugin (bottom). The nodes in yellow correspond to the highlighted unit on the right side.

Relation inference

To obtain information on the processes embodied by the units, the workflow puts them into context by inferring relations (similarity, dependence) both between units and between units and pathways / terms.

In standard GSEA only set overlap is taken into account. A shortcoming of this method is that it ignores dependencies between members of the one set and the other. Such relations are biologically relevant and therefore need to be taken into consideration.

The procedure used for relation inference is defined as follows:

- For all pairs (x, y) where x is a unit and y is a unit/pathway/term,
- the relation strength between x and y, given the omicsNET graph G(V,E) is:

$$w(x,y) = \frac{2 \cdot |x \cap y|}{|x|+|y|} + \frac{|\{(a,b) \ for \ (a,b) \in E \wedge a \in x \wedge b \in y\}|}{|x| \cdot |y| - |x \cap y|}$$

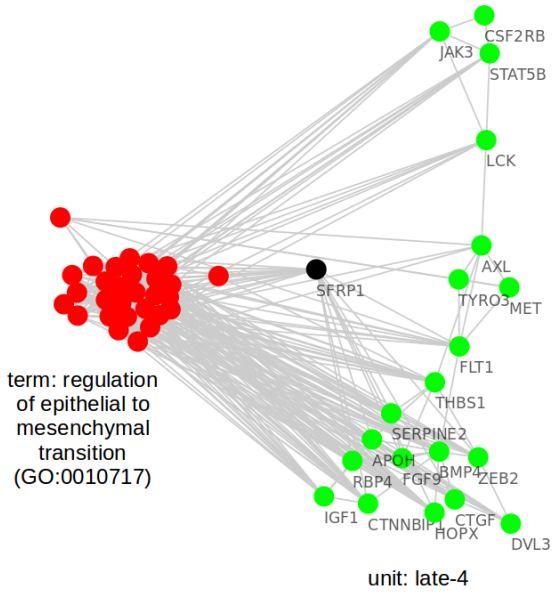

Figure 3.19: omicsNET edges linking members of a unit (late-4, given in green) to members of an ontology term (GO:0010717, given in red). One gene (SFRP1, given in black) is common to both groups.

The first part of the procedure ensures that only edges between units or between units and pathways / terms are computed. Calculating dependencies between pathways and terms is beyond the scope of this approach. In the relation strength function the first part of the formula is the Dice coefficient for the set overlap, while the second one determines which percentage of all theoretically possible edges between the two sets actually is present. The reason for subtracting the size of the common set of genes from the total number of possible interactions is that self-loops are not present in omicsNET.

In empirical testing, this function showed with a Pearson score of -0.1 no correlation between the size of the inspected modules and the strength of their relation.

Fig. 3.19 is especially representative for the rationale behind using a dependency function based both on overlap and connectivity. In the depicted case, only one gene is common to both sets making an enrichment analysis return a low score. On the edge level, however, the strong connectivity between the nodes of the two sets become evident.

Drug-target relations

Drug-target information was retrieved from the DrugBank (version as of May 2012) specifically for approved drugs. The dataset included 1,331 drugs addressing 1,461 targets.

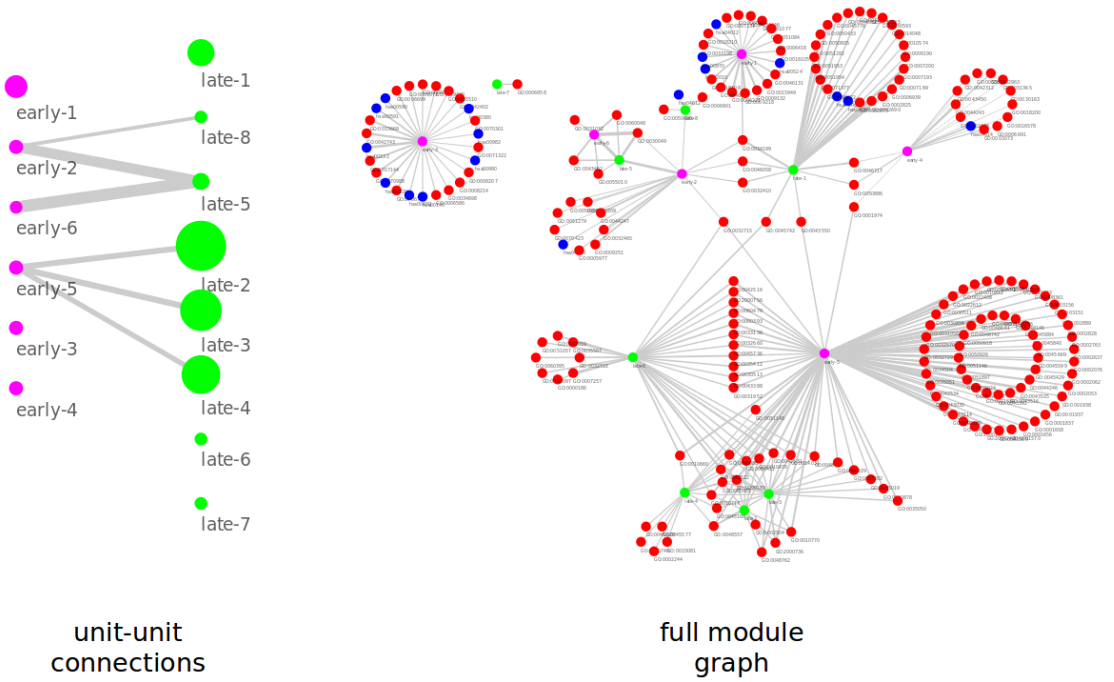

Figure 3.20: Unit-unit network with inter-unit edges (left). Units discovered in the early disease stage are given in purple, units specific to the late disease stage are given in green. The edge width and node size are indicative for the edge strength and the number of genes in the unit, respectively. Full module graph (right) with connections between units (given in purple/green) and ontology terms (given in red) and pathways (given in blue).

Sub-cellular location

Information on the sub-cellular location of proteins was extracted from SwissProt (version as of May 2012) for 20,246 entities. The free-text blocks containing the information were queried for 10 reference locations, including nucleus, cytoplasm, etc. The information was retrieved with the goal of visualizing it in Cytoscape using the Cerebral plugin. Cerebral organizes the network in layers, with each node occurring once on the layer corresponding to its sub-cellular location. To accommodate genes present in two adjacent locations, composite layers were introduced, such as nucleus-cytoplasm and cytoplasm-membrane. Proteins present in more than two or non-adjacent locations were annotated with *multiple*. For proteins without a known location, the term *other* was used.

Results and discussion

During preprocessing of the disease-specific gene sets, from the original 187 features of the early stage set, 98 could be connected using 267 omicsNET edges. From the original 450 features of the late stage set, 291 could be connected using 1,561 edges.

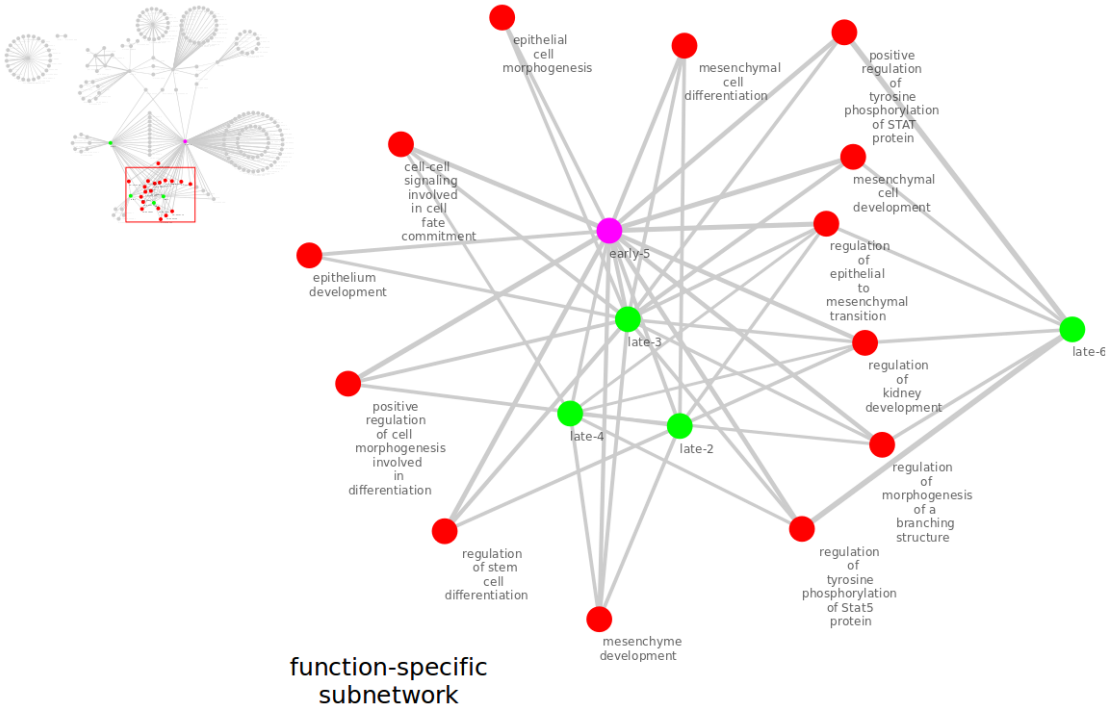

Figure 3.21: Subnetwork with distinct functional convergence towards mesenchyme related processes consisting of early (given in purple) and late (given in green) units and their related GO annotation (given in red).

The MCODE computation yielded 6 units for the early gene set with sizes in the interval [3,10] while for the late gene set 8 units could be computed with sizes in the interval [3,29].

When inferring the unit - pathway / term dependencies, a complete graph is computed. To reduce the size of the resulting network, two ranking methods were used for the selection of edges: an absolute one (for the whole graph) and a relative one (for each node individually). All edges above the average plus two standard deviations were considered to be significant. The reason for using a relative cutoff in addition to the absolute one was due to the fact that some units had no edge above the absolute cutoff value and were therefore discarded.

Fig. 3.20 (left) depicts the 6 + 8 units produced by the MCODE algorithm and their connecting edges. On the right side of the figure, the complete module graph consisting of units, pathways and terms is shown. Interestingly, it is often the case that units that are not directly connected, share a large number of common ontology terms, as is the case e.g. for early-5 and late-6.

In Fig. 3.21 a function-specific subgraph is depicted, in which the units have a distinctive focus on mesenchyme development and epithelial to mesenchymal transition. The involvement of these processes in diabetic nephropathy has been discussed in literature [237–239] and the strong presence of this association in the results represents a first validation of the method.

Fig. 3.22 shows an in-depth view of two units (early-5 and late-4), with their members or-

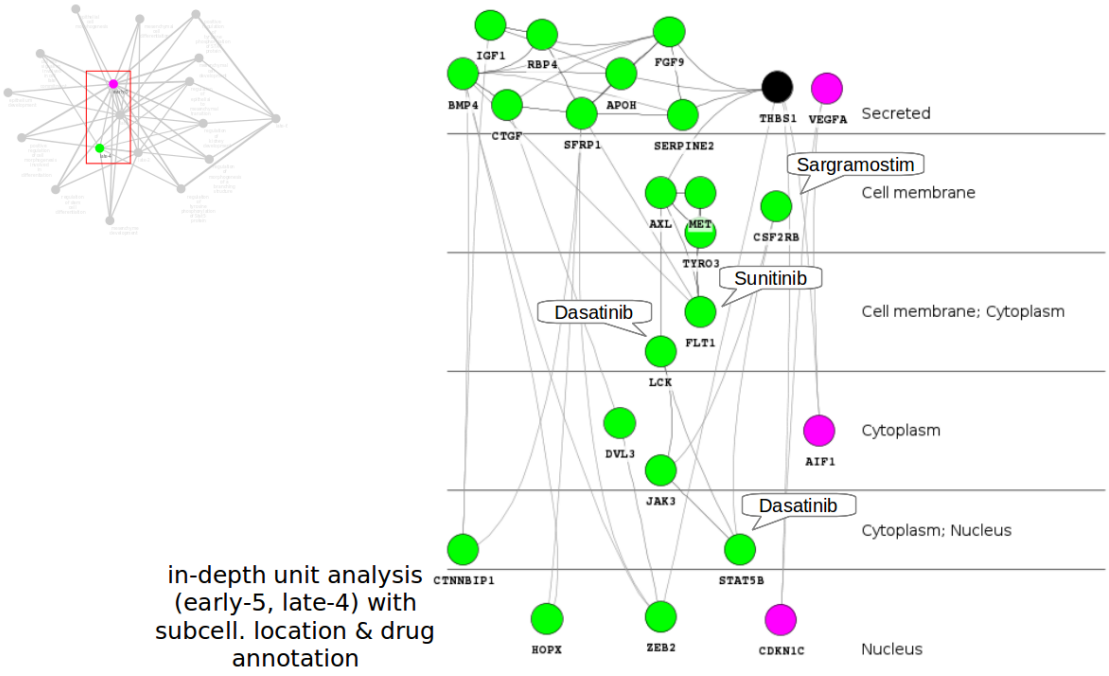

Figure 3.22: In-depth unit analysis showing the members of two units (early in purple and late in green) with omicsNET edges, ordered by sub-cellular location from nucleus to secreted. One gene (THBS1, given in black) is common to both units. Depicted with call-outs are also the drugs targeting specific genes in the units.

dered in cellular compartments as well as the omicsNET edges connecting them. Also depicted are the drugs targeting member genes. This view is particularly useful after an initial screening of the units, when a detailed analysis for hypothesis generation is necessary.

After a positive first evaluation, the method and results presented in this section have been submitted for expert review and are pending experimental validation.

3.5 Summary of protein interaction networks

Networks are a core paradigm in Systems Biology and a powerful tool for representing and analyzing biological processes. They can be used to model a wide array of information such as gene expression correlation, transcriptional regulation or direct protein binding together with relation direction information, interaction strength, sequence of events and much more.

Protein interaction networks represent a subclass of biological networks with a distinctive focus on direct relations between partners. In this chapter we relax this definition to a certain extent to also allow inferred dependency edges, where two genes or proteins are assumed to share an unspecified relation based on the available biological annotation information.

In the introduction to this thesis we provided an overview of existing approaches to network

information integration, including large academic and commercial projects such as STRING, STITCH and IPA as well as specific methods delineated in various scientific articles.

In this chapter we focused on a new network model, omicsNET, a concept resting on the extension of existing PPI networks with inferred edges based on high-quality pathway, ontology and protein domain information.

After presenting the graph construction method and testing it using interaction and KEGG edges, we then described the design and implementation details of a platform, BIO, for the visualization of molecular information. This platform is also a front-end for omicsNET.

Subsequently we performed an in-silico analysis of diabetic nephropathy using the omicsNET graph to extend KEGG pathways and identify mechanisms of disease addressed in literature, patents and clinical trials.

In the final section, a novel concept, units, was introduced to tackle the limited performance of single biomarkers in the clinical context. In this approach, subprocesses are identified in sets of disease-associated genes by linking them with omicsNET edges and subsequently determining clusters. By choosing biomarkers from different units it is expected to reach better performance in patient stratification. This method is currently pending experimental validation.

The results proved again to be promising both in the verification of the omicsNET construction and in its application for analyzing diabetic nephropathy. This allows us to hope that the units concept will indeed also prove valuable in the field of Systems Medicine.

CHAPTER 4

Discussion and conclusion

With the advent of the Omics revolution, the scientific community was presented with a whole new array of possibilities and challenges. High-throughput experiments and large-scale genomics, transcriptomics and proteomics as well as related methods produced in a relatively short time period vast amounts of data and enabled researchers to gain new insights in cell biology. On the downside, high false positive rates, measurement errors, methodological limitations, suboptimal experiment setup and a data bulk increasingly difficult to handle were the result. Information extraction became the search for the needle in the proverbial haystack.

To tackle these issues new disciplines were co-opted. Mathematics, informatics and even sociology, the latter through analogies discovered between molecular and social networks, provided tools for the analysis of the new data.

Increasing knowledge of the connection between pathways and phenotypes led to the desire to influence them in case of disease. With the shift of the scientific perspective towards processes, the theoretical model best suited for describing dependencies and complex events was identified to be networks. These became in turn a central concept in Systems Biology, a branch of biology focused on integrating the Omics knowledge now readily available to provide a holistic view of the cell functioning.

Biological networks also constitute the scope of this thesis. Within these boundaries, two novel concepts were investigated.

Genetic interaction networks, more specifically, synthetic lethality networks were used to demonstrate that by combining a synlet map with patient data, it is possible to find target hubs which could resolve the tumor variability issue. Having published the results in a scientific article, the concept was generalized for additional tumor types and a patent application was submitted.

In a different scope where the death of the affected cells is not the desired outcome, but instead strengthening protection mechanisms and reversing damage is required, a network concept was proposed aimed at overcoming the completeness issue of today's networks while maintaining a significant level of accuracy. The method was tested with promising results by investigating its prediction performance for existing PPI and KEGG interactions and was submitted as a

scientific article for review and publication.

For manual analysis, a software for viewing the proposed network was presented including architecture and implementation details. For semi-automatic analysis, an approach to diabetic nephropathy investigation using KEGG pathways extended with inferred network edges was delineated and the results were published in a scientific article. For the automatic isolation of disease-associated processes towards finding multi-marker panels for patient stratification, a novel concept (units) and its associated workflow were presented and are currently pending experimental validation.

In the introduction to this thesis we spoke about an emerging branch of Systems Biology, the Systems Medicine. The latter focuses on the integration of Omics with the clear purpose of addressing diseases in the clinical context. The work presented in this thesis, whether it is cancer treatment by synthetic lethality means, network-based analysis of diabetic nephropahty or units for clinical patient segmentation is perfectly in-line with the goals of Systems Medicine.

Overall, this thesis showed that high-quality network integration of Omics indeed supports interpretative causality and its scientific output furthers our understanding of human disease mechanisms and intervention modalities.

Bibliography

[1] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. j.d. watson and f.h.c. crick. published in nature, number 4356 april 25, 1953. *Nature*, 248(5451):765, Apr 1974.

[2] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

[3] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, Mar 2011.

[4] Carlos Prieto, Alberto Risueño, Celia Fontanillo, and Javier De las Rivas. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One*, 3(12):e3911, 2008.

[5] Huan Yang, Chao Cheng, and Wei Zhang. Average rank-based score to measure deregulation of molecular pathway gene sets. *PLoS One*, 6(11):e27579, 2011.

[6] Barabasi and Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[7] Konrad Mönks, Irmgard Mühlberger, Andreas Bernthaler, Raul Fechete, Paul Perco, Rudolf Freund, Arno Lukas, and Bernd Mayer. *Computational Reconstruction of Protein Interaction Networks*, pages 155–180. Wiley-VCH Verlag GmbH & Co. KGaA, 2011.

[8] S. J. Dixon, M. Costanzo, A. Baryshnikova, B. Andrews, and C. Boone. Systematic mapping of genetic interaction networks. *Annu Rev Genet*, 43:601–25, 2009.

[9] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The intact molecular interaction database in 2012. *Nucleic Acids Res*, 40(Database issue):D841–D846, Jan 2012.

[10] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M Rust, Andrew Winter, Kara Dolinski, and Mike Tyers. The biogrid interaction database: 2011 update. *Nucleic Acids Res*, 39(Database issue):D698–D704, Jan 2011.

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

[12] Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, Sep 2003.

[13] Peter D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol*, 694:49–61, 2011.

[14] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–D568, Jan 2011.

[15] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–D416, Jan 2009.

[16] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Monica Campillos, Christian von Mering, Lars Juhl Jensen, Andreas Beyer, and Peer Bork. Stitch 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res*, 38(Database issue):D552–D556, Jan 2010.

[17] Ingenuity Systems. `www.ingenuity.com`.

[18] Danning He, Zhi-Ping Liu, and Luonan Chen. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics*, 12:592, 2011.

[19] U. Sengupta, S. Ukil, N. Dimitrova, and S. Agrawal. Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. *PLoS One*, 4(12):e8100, 2009.

[20] Andrey Alexeyenko, Thomas Schmitt, Andreas Tjärnberg, Dmitri Guala, Oliver Frings, and Erik L L Sonnhammer. Comparative interactomics with funcoup 2.0. *Nucleic Acids Res*, 40(Database issue):D821–D828, Jan 2012.

[21] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, Jan 2011.

[22] Charles Auffray, Zhu Chen, and Leroy Hood. Systems medicine: the future of medical genomics and healthcare. *Genome Med*, 1(1):2, 2009.

[23] European Commision, DG Research, Directorate of Health. From Systems Biology to Systems Medicine. ftp://ftp.cordis.europa.eu/pub/fp7/health/docs/final-report-systems-medicine-workshop_en.pdf, June 2010.

[24] Jean Bousquet, Josep M Anto, Peter J Sterk, Ian M Adcock, Kian Fan Chung, Josep Roca, Alvar Agusti, Chris Brightling, Anne Cambon-Thomsen, Alfredo Cesario, Sonia Abdelhak, Stylianos E Antonarakis, Antoine Avignon, Andrea Ballabio, Eugenio Baraldi, Alexander Baranov, Thomas Bieber, Joël Bockaert, Samir Brahmachari, Christian Brambilla, Jacques Bringer, Michel Dauzat, Ingemar Ernberg, Leonardo Fabbri, Philippe Froguel, David Galas, Takashi Gojobori, Peter Hunter, Christian Jorgensen, Francine Kauffmann, Philippe Kourilsky, Marek L Kowalski, Doron Lancet, Claude Le Pen, Jacques Mallet, Bongani Mayosi, Jacques Mercier, Andres Metspalu, Joseph H Nadeau, Grégory Ninot, Denis Noble, Mehmet Oztürk, Susanna Palkonen, Christian Préfaut, Klaus Rabe, Eric Renard, Richard G Roberts, Boleslav Samolinski, Holger J Schünemann, Hans-Uwe Simon, Marcelo Bento Soares, Giulio Superti-Furga, Jesper Tegner, Sergio Verjovski-Almeida, Peter Wellstead, Olaf Wolkenhauer, Emiel Wouters, Rudi Balling, Anthony J Brookes, Dominique Charron, Christophe Pison, Zhu Chen, Leroy Hood, and Charles Auffray. Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med*, 3(7):43, 2011.

[25] John Cijiang He, Peter Y Chuang, Avi Ma'ayan, and Ravi Iyengar. Systems biology of kidney diseases. *Kidney Int*, 81(1):22–39, Jan 2012.

[26] Arno Lukas, Johannes Soellner, Andreas Bernthaler, Bernd Mayer, Raul Fechete, Paul Perco, and Andreas Heinzel. Critical gene targets for cytotoxic therapy, November 2011. App.No.: EP2011/058259, Pub.No.: WO/2011/144738.

[27] S. W. Lowe and A. W. Lin. Apoptosis in cancer. *Carcinogenesis*, 21(3):485–495, Mar 2000.

[28] William C.S. Cho, editor. *An Omics Perspective on Cancer Research*. Springer, NY, 2010.

[29] I. G. Khalil and C. Hill. Systems biology for cancer. *Curr Opin Oncol*, 17(1):44–48, Jan 2005.

[30] Ronald Rapberger, Paul Perco, Cornelia Sax, Thomas Pangerl, Christian Siehs, Dietmar Pils, Andreas Bernthaler, Arno Lukas, Bernd Mayer, and Michael Krainer. Linking the ovarian cancer transcriptome and immunome. *BMC Syst Biol*, 2:2, 2008.

[31] Martin Michaelis, Denise Klassert, Susanne Barth, Tatyana Suhan, Rainer Breitling, Bernd Mayer, Nora Hinsch, Hans W Doerr, Jaroslav Cinatl, and Jindrich Cinatl. Chemoresistance acquisition induces a global shift of expression of aniogenesis-associated genes and increased pro-angogenic activity in neuroblastoma cells. *Mol Cancer*, 8:80, 2009.

[32] M. Michaelis, I. Fichtner, D. Behrens, W. Haider, F. Rothweiler, A. Mack, J. Cinatl, H. W. Doerr, and Jr. Cinatl, J. Anti-cancer effects of bortezomib against chemoresistant neuroblastoma cell lines in vitro and in vivo. *Int J Oncol*, 28(2):439–46, 2006.

[33] Masashi Takemura, Harushi Osugi, Shigeru Lee, Masahiro Kaneko, Yoshinori Tanaka, Yushi Fujiwara, Satoshi Nishizawa, and Hiroshi Iwasaki. [choice of chemotherapeutic drugs for colorectal cancers by dpd and oprt activities in cancer tissues]. *Gan To Kagaku Ryoho*, 31(7):1053–1056, Jul 2004.

[34] Nuria Conde-Pueyo, Andreea Munteanu, Ricard V Solé, and Carlos Rodríguez-Caso. Human synthetic lethal inference as potential anti-cancer target gene detection. *BMC Syst Biol*, 3:116, 2009.

[35] Judice L Y Koh, Huiming Ding, Michael Costanzo, Anastasia Baryshnikova, Kiana Toufighi, Gary D Bader, Chad L Myers, Brenda J Andrews, and Charles Boone. Dryygin: a database of quantitative genetic interaction networks in yeast. *Nucleic Acids Res*, 38(Database issue):D502–D507, Jan 2010.

[36] Raul Fechete, Susanne Barth, Tsviya Olender, Andreea Munteanu, Andreas Bernthaler, Aron Inger, Paul Perco, Arno Lukas, Doron Lancet, Jindrich Cinatl, Martin Michaelis, and Bernd Mayer. Synthetic lethal hubs associated with vincristine resistant neuroblastoma. *Mol Biosyst*, 7(1):200–214, Jan 2011.

[37] N. F. Schor. New approaches to pharmacotherapy of tumors of the nervous system during childhood and adolescence. *Pharmacol Ther*, 122(1):44–55, 2009.

[38] S. Ootsuka, S. Asami, T. Sasaki, Y. Yoshida, N. Nemoto, H. Shichino, M. Chin, H. Mugishima, and T. Suzuki. Analyses of novel prognostic factors in neuroblastoma patients. *Biol Pharm Bull*, 30(12):2294–9, 2007.

[39] J. Hirsch. An anniversary for cancer chemotherapy. *Jama*, 296(12):1518–20, 2006.

[40] H. Joensuu. Systemic chemotherapy for cancer: from weapon to treatment. *Lancet Oncol*, 9(3):304, 2008.

[41] Aung Naing and Razelle Kurzrock. Chemotherapy resistance and retreatment: a dogma revisited. *Clin Colorectal Cancer*, 9(2):E1–E4, Apr 2010.

[42] X. Li, M. T. Lewis, J. Huang, C. Gutierrez, C. K. Osborne, M. F. Wu, S. G. Hilsenbeck, A. Pavlick, X. Zhang, G. C. Chamness, H. Wong, J. Rosen, and J. C. Chang. Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *J Natl Cancer Inst*, 100(9):672–9, 2008.

[43] W. C. Huang and M. C. Hung. Induction of akt activity by chemotherapy confers acquired resistance. *J Formos Med Assoc*, 108(3):180–94, 2009.

[44] M. M. Gottesman. Mechanisms of cancer drug resistance. *Annu Rev Med*, 53:615–27, 2002.

[45] Daniel S-W Tan, Marco Gerlinger, Bin-Tean Teh, and Charles Swanton. Anti-cancer drug resistance: understanding the mechanisms through the use of integrative genomics and functional rna interference. *Eur J Cancer*, 46(12):2166–2177, Aug 2010.

[46] G. Kuruvilla, S. Perry, B. Wilson, and H. El-Hakim. The natural history of vincristine-induced laryngeal paralysis in children. *Arch Otolaryngol Head Neck Surg*, 135(1):101–5, 2009.

[47] M. W. Chan, C. D. Chiang, E. J. Song, and V. C. Yang. Effects of cytoskeletal inhibitors on the accumulation of vincristine in a resistant human lung cancer cell line with high level of polymerized tubulin. *Cancer Biochem Biophys*, 16(4):347–63, 1998.

[48] V. Ong, N. L. Liem, M. A. Schmid, N. M. Verrills, R. A. Papa, G. M. Marshall, K. L. Mackenzie, M. Kavallaris, and R. B. Lock. A role for altered microtubule polymer levels in vincristine resistance of childhood acute lymphoblastic leukemia xenografts. *J Pharmacol Exp Ther*, 324(2):434–42, 2008.

[49] N. Keshelava, E. Davicioni, Z. Wan, L. Ji, R. Sposto, T. J. Triche, and C. P. Reynolds. Histone deacetylase 1 gene expression and sensitization of multidrug-resistant neuroblastoma cell lines to cytotoxic agents by depsipeptide. *J Natl Cancer Inst*, 99(14):1107–19, 2007.

[50] R. A. Blaheta, M. Michaelis, I. Natsheh, C. Hasenberg, E. Weich, B. Relja, D. Jonas, H. W. Doerr, and Jr. Cinatl, J. Valproic acid inhibits adhesion of vincristine- and cisplatin-resistant neuroblastoma tumour cells to endothelium. *Br J Cancer*, 96(11):1699–706, 2007.

[51] R. Huang, D. J. Murry, D. Kolwankar, S. D. Hall, and D. R. Foster. Vincristine transcriptional regulation of efflux drug transporters in carcinoma cell lines. *Biochem Pharmacol*, 71(12):1695–704, 2006.

[52] R. A. Blaheta, F. H. Daher, M. Michaelis, C. Hasenberg, E. M. Weich, D. Jonas, R. Kotchetkov, H. W. Doerr, and Jr. Cinatl, J. Chemoresistance induces enhanced adhesion and transendothelial penetration of neuroblastoma cells by down-regulating ncam surface expression. *BMC Cancer*, 6:294, 2006.

[53] Y. Liang, S. McDonnell, and M. Clynes. Examining the relationship between cancer invasion/metastasis and drug resistance. *Curr Cancer Drug Targets*, 2(3):257–77, 2002.

[54] E. Hiyama, K. Hiyama, M. Nishiyama, C. P. Reynolds, J. W. Shay, and T. Yokoyama. Differential gene expression profiles between neuroblastomas with high telomerase activity and low telomerase activity. *J Pediatr Surg*, 38(12):1730–4, 2003.

[55] L. McArdle, M. McDermott, R. Purcell, D. Grehan, A. O'Meara, F. Breatnach, D. Catchpoole, A. C. Culhane, I. Jeffery, W. M. Gallagher, and R. L. Stallings. Oligonucleotide microarray analysis of gene expression in neuroblastoma displaying loss of chromosome 11q. *Carcinogenesis*, 25(9):1599–609, 2004.

[56] A. Oberthuer, P. Warnat, Y. Kahlert, F. Westermann, R. Spitz, B. Brors, B. Hero, R. Eils, M. Schwab, F. Berthold, and M. Fischer. Classification of neuroblastoma patients by published gene-expression markers reveals a low sensitivity for unfavorable courses of mycn non-amplified disease. *Cancer Lett*, 250(2):250–67, 2007.

[57] M. Michaelis, M. C. Kleinschmidt, S. Barth, F. Rothweiler, J. Geiler, R. Breitling, B. Mayer, H. Deubzer, O. Witt, J. Kreuter, H. W. Doerr, J. Cinatl, and Jr. Cinatl, J. Anti-cancer effects of artesunate in a panel of chemoresistant neuroblastoma cell lines. *Biochem Pharmacol*, 79(2):130–6, 2010.

[58] A. Garaventa, O. Bellagamba, M. S. Lo Piccolo, C. Milanaccio, E. Lanino, L. Bertolazzi, G. P. Villavecchia, M. Cabria, G. Scopinaro, F. Claudiani, and B. De Bernardi. 131i-metaiodobenzylguanidine (131i-mibg) therapy for residual neuroblastoma: a monoinstitutional experience with 43 patients. *Br J Cancer*, 81(8):1378–84, 1999.

[59] D. Canaani. Methodological approaches in application of synthetic lethality screening towards anticancer therapy. *Br J Cancer*, 100(8):1213–8, 2009.

[60] N. Le Meur and R. Gentleman. Modeling synthetic lethality. *Genome Biol*, 9(9):R135, 2008.

[61] P. F. Suthers, A. Zomorrodi, and C. D. Maranas. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol Syst Biol*, 5:301, 2009.

[62] A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Menard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A. M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–13, 2004.

[63] C. Boone, H. Bussey, and B. J. Andrews. Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, 8(6):437–49, 2007.

[64] S. J. Dixon, B. J. Andrews, and C. Boone. Exploring the conservation of synthetic lethal genetic interaction networks. *Commun Integr Biol*, 2(2):78–81, 2009.

[65] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P St Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M Wallace, Joseph A Whitney, Matthew T Weirauch, Guoqing Zhong, Hongwei

Zhu, Walid A Houry, Michael Brudno, Sasan Ragibizadeh, Balazs Papp, Csaba Pal, Frederick P Roth, Guri Giaever, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Anne-Claude Gingras, Quaid D Morris, Philip M Kim, Chris A Kaiser, Chad L Myers, Brenda J Andrews, and Charles Boone. The genetic landscape of a cell. *Science*, 327(5964):425–431, Jan 2010.

[66] D. A. Chan and A. J. Giaccia. Targeting cancer cells by synthetic lethality: autophagy and vhl in cancer therapeutics. *Cell Cycle*, 7(19):2987–90, 2008.

[67] H. C. Reinhardt, H. Jiang, M. T. Hemann, and M. B. Yaffe. Exploiting synthetic lethal interactions for targeted cancer therapy. *Cell Cycle*, 8(19):3112–9, 2009.

[68] J. D. Iglehart and D. P. Silver. Synthetic lethality–a new direction in cancer-drug development. *N Engl J Med*, 361(2):189–91, 2009.

[69] P. C. Fong, D. S. Boss, T. A. Yap, A. Tutt, P. Wu, M. Mergui-Roelvink, P. Mortimer, H. Swaisland, A. Lau, M. J. O'Connor, A. Ashworth, J. Carmichael, S. B. Kaye, J. H. Schellens, and J. S. de Bono. Inhibition of poly(adp-ribose) polymerase in tumors from brca mutation carriers. *N Engl J Med*, 361(2):123–34, 2009.

[70] A. S. Whittemore, G. Gong, and J. Itnyre. Prevalence and contribution of brca1 mutations in breast cancer and ovarian cancer: results from three u.s. population-based case-control studies of ovarian cancer. *Am J Hum Genet*, 60(3):496–504, 1997.

[71] J. F. Stratton, S. A. Gayther, P. Russell, J. Dearden, M. Gore, P. Blake, D. Easton, and B. A. Ponder. Contribution of brca1 mutations to ovarian cancer. *N Engl J Med*, 336(16):1125–30, 1997.

[72] A. Platzer, P. Perco, A. Lukas, and B. Mayer. Characterization of protein-interaction networks in tumors. *BMC Bioinformatics*, 8:224, 2007.

[73] R. Albert. Scale-free networks in cell biology. *J Cell Sci*, 118(Pt 21):4947–57, 2005.

[74] R. Kotchetkov, J. Cinatl, R. Blaheta, J. U. Vogel, J. Karaskova, J. Squire, P. Hernaiz Driever, T. Klingebiel, and Jr. Cinatl, J. Development of resistance to vincristine and doxorubicin in neuroblastoma alters malignant properties and induces additional karyotype changes: a preclinical model. *Int J Cancer*, 104(1):36–43, 2003.

[75] M. Michaelis, J. Cinatl, P. Anand, F. Rothweiler, R. Kotchetkov, A. von Deimling, H. W. Doerr, K. Shogen, and Jr. Cinatl, J. Onconase induces caspase-independent cell death in chemoresistant neuroblastoma cells. *Cancer Lett*, 250(1):107–16, 2007.

[76] R. Kotchetkov, P. H. Driever, J. Cinatl, M. Michaelis, J. Karaskova, R. Blaheta, J. A. Squire, A. Von Deimling, J. Moog, and Jr. Cinatl, J. Increased malignant behavior in neuroblastoma cells with acquired multi-drug resistance does not depend on p-gp expression. *Int J Oncol*, 27(4):1029–37, 2005.

[77] N. Keshelava, R. C. Seeger, S. Groshen, and C. P. Reynolds. Drug resistance patterns of human neuroblastoma cell lines derived from patients at different phases of therapy. *Cancer Res*, 58(23):5396–405, 1998.

[78] N. Keshelava, D. Tsao-Wei, and C. P. Reynolds. Pyrazoloacridine is active in multidrug-resistant neuroblastoma cell lines with nonfunctional p53. *Clin Cancer Res*, 9(9):3492–502, 2003.

[79] N. Rosen, C. P. Reynolds, C. J. Thiele, J. L. Biedler, and M. A. Israel. Increased n-myc expression following progressive growth of human neuroblastoma. *Cancer Res*, 46(8):4139–42, 1986.

[80] J. J. Tumilowicz, W. W. Nichols, J. J. Cholon, and A. E. Greene. Definition of a continuous human cell line derived from neuroblastoma. *Cancer Res*, 30(8):2110–8, 1970.

[81] The ArrayExpress. `www.ebi.ac.uk/microarray-as/ae`.

[82] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. Arrayexpress–a public repository for microarray gene expression data at the ebi. *Nucleic Acids Res*, 33(Database issue):D553–5, 2005.

[83] A. Oberthuer, F. Berthold, P. Warnat, B. Hero, Y. Kahlert, R. Spitz, K. Ernestus, R. Konig, S. Haas, R. Eils, M. Schwab, B. Brors, F. Westermann, and M. Fischer. Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *J Clin Oncol*, 24(31):5070–8, 2006.

[84] The GeneCards database. `www.genecards.org`.

[85] M. Safran, V. Chalifa-Caspi, O. Shmueli, T. Olender, M. Lapidot, N. Rosen, M. Shmoish, Y. Peter, G. Glusman, E. Feldmesser, A. Adato, I. Peter, M. Khen, T. Atarot, Y. Groner, and D. Lancet. Human gene-centric databases at the weizmann institute of science: Genecards, udb, crow 21 and horde. *Nucleic Acids Res*, 31(1):142–6, 2003.

[86] The YeastGenome project. `www.yeastgenome.org`.

[87] The HomoloGene database. `www.ncbi.nlm.nih.gov/homologene`.

[88] C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and J. A. Blake. The mouse genome database (mgd): mouse biology and model systems. *Nucleic Acids Res*, 36(Database issue):D724–8, 2008.

[89] A. Bernthaler, I. Muhlberger, R. Fechete, P. Perco, A. Lukas, and B. Mayer. A dependency graph approach for the analysis of differential gene expression profiles. *Mol Biosyst*, 5(12):1720–31, 2009.

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

[91] The Entrez Programming Utilities. `eutils.ncbi.nlm.nih.gov/corehtml/query/static/esearch_help.html`.

[92] Emily Crowley and Richard Callaghan. Multidrug efflux pumps: drug binding–gates or cavity? *FEBS J*, 277(3):530–539, Feb 2010.

[93] M. D. Hall, M. D. Handley, and M. M. Gottesman. Is resistance useless? multidrug resistance and collateral sensitivity. *Trends Pharmacol Sci*, 30(10):546–56, 2009.

[94] Emma Bell, Lindi Chen, Tao Liu, Glenn M Marshall, John Lunec, and Deborah A Tweddle. Mycn oncoprotein targets and their therapeutic potential. *Cancer Lett*, 293(2):144–157, Jul 2010.

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

[96] M. Sehested, T. Skovsgaard, B. van Deurs, and H. Winther-Nielsen. Increase in nonspecific adsorptive endocytosis in anthracycline- and vinca alkaloid-resistant ehrlich ascites tumor cell lines. *J Natl Cancer Inst*, 78(1):171–9, 1987.

[97] G. M. Galbraith and R. M. Galbraith. Metabolic and cytoskeletal modulation of transferrin receptor mobility in mitogen-activated human lymphocytes. *Clin Exp Immunol*, 42(2):285–93, 1980.

[98] L. Groth-Pedersen, M. S. Ostenfeld, M. Hoyer-Hansen, J. Nylandsted, and M. Jaattela. Vincristine induces dramatic lysosomal changes and sensitizes cancer cells to lysosome-destabilizing siramesine. *Cancer Res*, 67(5):2217–25, 2007.

[99] X. Wang, M. Lan, Y. Q. Shi, J. Lu, Y. X. Zhong, H. P. Wu, H. H. Zai, J. Ding, K. C. Wu, B. R. Pan, J. P. Jin, and D. M. Fan. Differential display of vincristine-resistance-related genes in gastric cancer sgc7901 cell. *World J Gastroenterol*, 8(1):54–9, 2002.

[100] D. H. Geho, R. W. Bandle, T. Clair, and L. A. Liotta. Physiological mechanisms of tumor-cell invasion and migration. *Physiology (Bethesda)*, 20:194–200, 2005.

[101] J. M. Vasiliev. Cytoskeletal mechanisms responsible for invasive migration of neoplastic cells. *Int J Dev Biol*, 48(5-6):425–39, 2004.

[102] P. Friedl and K. Wolf. Tumour-cell invasion and migration: diversity and escape mechanisms. *Nat Rev Cancer*, 3(5):362–74, 2003.

[103] Giuseppina Bozzuto, Paola Ruggieri, and Agnese Molinari. Molecular aspects of tumor cell migration and invasion. *Ann Ist Super Sanita*, 46(1):66–80, 2010.

[104] Helen M Coley. Overcoming multidrug resistance in cancer: clinical studies of p-glycoprotein inhibitors. *Methods Mol Biol*, 596:341–358, 2010.

[105] K. Reed, S. L. Hembruff, J. A. Sprowl, and A. M. Parissenti. The temporal relationship between abcb1 promoter hypomethylation, abcb1 expression and acquisition of drug resistance. *Pharmacogenomics J*, 10(6):489–504, Dec 2010.

[106] W. G. Simpson. The calcium channel blocker verapamil and cancer chemotherapy. *Cell Calcium*, 6(6):449–67, 1985.

[107] A. M. Rogan, T. C. Hamilton, R. C. Young, Jr. Klecker, R. W., and R. F. Ozols. Reversal of adriamycin resistance by verapamil in human ovarian cancer. *Science*, 224(4652):994–6, 1984.

[108] Ling-Yen Chiu, Jiunn-Liang Ko, Yi-Ju Lee, Tsung-Ying Yang, Yi-Torng Tee, and Gwo-Tarng Sheu. L-type calcium channel blockers reverse docetaxel and vincristine-induced multidrug resistance independent of abcb1 expression in human lung cancer cell lines. *Toxicol Lett*, 192(3):408–418, Feb 2010.

[109] S. J. Dixon, Y. Fedyshyn, J. L. Koh, T. S. Prasad, C. Chahwan, G. Chua, K. Toufighi, A. Baryshnikova, J. Hayles, K. L. Hoe, D. U. Kim, H. O. Park, C. L. Myers, A. Pandey, D. Durocher, B. J. Andrews, and C. Boone. Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A*, 105(43):16653–8, 2008.

[110] B. Lehner, C. Crombie, J. Tischler, A. Fortunato, and A. G. Fraser. Systematic mapping of genetic interactions in caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet*, 38(8):896–903, 2006.

[111] J. Macia and R. V. Sole. Distributed robustness in cellular networks: insights from synthetic evolved circuits. *J R Soc Interface*, 6(33):393–400, 2009.

[112] E. van Wijk, E. Krieger, M. H. Kemperman, E. M. De Leenheer, P. L. Huygen, C. W. Cremers, F. P. Cremers, and H. Kremer. A mutation in the gamma actin 1 (actg1) gene causes autosomal dominant hearing loss (dfna20/26). *J Med Genet*, 40(12):879–84, 2003.

[113] J. Mullenders, A. W. Fabius, M. Madiredjo, R. Bernards, and R. L. Beijersbergen. A large scale shrna barcode screen identifies the circadian clock component arntl as putative regulator of the p53 tumor suppressor pathway. *PLoS One*, 4(3):e4798, 2009.

[114] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36(Database issue):D901–6, 2008.

[115] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, Oleg Khovayko, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, Vadim Miller, James Ostell, Kim D Pruitt, Gregory D Schuler, Martin Shumway, Edwin Sequeira, Steven T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 36(Database issue):D13–D21, Jan 2008.

[116] Ann-Charlotte Berglund, Erik Sjölund, Gabriel Ostlund, and Erik L L Sonnhammer. Inparanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, 36(Database issue):D263–D266, Jan 2008.

[117] The InParanoid database. `inparanoid.sbc.su.se/cgi-bin/index.cgi`.

[118] The Roundup database. `roundup.hms.harvard.edu/site/index.php`.

[119] Todd F Deluca, I-Hsien Wu, Jian Pu, Thomas Monaghan, Leonid Peshkin, Saurav Singh, and Dennis P Wall. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, 22(16):2044–2046, Aug 2006.

[120] The OMA database. `www.omabrowser.org/cgi-bin/gateway.pl`.

[121] Alexander C J Roth, Gaston H Gonnet, and Christophe Dessimoz. Algorithm of oma for large-scale orthology inference. *BMC Bioinformatics*, 9:518, 2008.

[122] The ENSEMBL BioMart. `www.ensembl.org/index.html`.

[123] T. J P Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617, Jan 2007.

[124] Maria Teresa Landi, Tatiana Dracheva, Melissa Rotunno, Jonine D Figueroa, Huaitian Liu, Abhijit Dasgupta, Felecia E Mann, Junya Fukuoka, Megan Hames, Andrew W Bergen, Sharon E Murphy, Ping Yang, Angela C Pesatori, Dario Consonni, Pier Alberto Bertazzi, Sholom Wacholder, Joanna H Shih, Neil E Caporaso, and Jin Jen. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*, 3(2):e1651, 2008.

[125] Li-Jen Su, Ching-Wei Chang, Yu-Chung Wu, Kuang-Chi Chen, Chien-Ju Lin, Shu-Ching Liang, Chi-Hung Lin, Jacqueline Whang-Peng, Shih-Lan Hsu, Chen-Hsin Chen, and Chi-Ying F Huang. Selection of ddx5 as a novel internal control for q-rt-pcr from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*, 8:140, 2007.

[126] Yee Leng Yap, David C L Lam, Girard Luc, Xue Wu Zhang, David Hernandez, Robin Gras, Elaine Wang, S. W. Chiu, Lap Ping Chung, W. K. Lam, David K Smith, John D Minna, Antoine Danchin, and Maria P Wong. Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. *Nucleic Acids Res*, 33(1):409–421, 2005.

[127] Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Kerby Shedden, Jeremy M G Taylor, Steven A Enkemann, Ming-Sound Tsao, Timothy J Yeatman, William L Gerald, Steven Eschrich, Igor Jurisica, Thomas J Giordano, David E Misek, Andrew C Chang, Chang Qi Zhu, Daniel Strumpf, Samir Hanash, Frances A Shepherd, Keyue Ding, Lesley Seymour, Katsuhiko Naoki, Nathan Pennell, Barbara Weir, Roel Verhaak, Christine Ladd-Acosta, Todd Golub, Michael Gruidl, Anupama Sharma, Janos Szoke, Maureen Zakowski, Valerie Rusch, Mark Kris, Agnes Viale, Noriko Motoi, William Travis, Barbara Conley, Venkatraman E Seshan, Matthew Meyerson, Rork Kuick, Kevin K Dobbin, Tracy Lively, James W Jacobson, and David G Beer. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*, 14(8):822–827, Aug 2008.

[128] Andrea L Richardson, Zhigang C Wang, Arcangela De Nicolo, Xin Lu, Myles Brown, Alexander Miron, Xiaodong Liao, J. Dirk Iglehart, David M Livingston, and Shridar Ganesan. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, 9(2):121–132, Feb 2006.

[129] Dung-Tsa Chen, Aejaz Nasir, Aedin Culhane, Chinnambally Venkataramu, William Fulp, Renee Rubio, Tao Wang, Deepak Agrawal, Susan M McCarthy, Mike Gruidl, Gregory Bloom, Tove Anderson, Joe White, John Quackenbush, and Timothy Yeatman. Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res Treat*, 119(2):335–346, Jan 2010.

[130] Jacob Sabates-Bellver, Laurens G Van der Flier, Mariagrazia de Palo, Elisa Cattaneo, Caroline Maake, Hubert Rehrauer, Endre Laczko, Michal A Kurowski, Janusz M Bujnicki, Mirco Menigatti, Judith Luz, Teresa V Ranalli, Vito Gomes, Alfredo Pastorelli, Roberto Faggiani, Marcello Anti, Josef Jiricny, Hans Clevers, and Giancarlo Marra. Transcriptome profile of human colorectal adenomas. *Mol Cancer Res*, 5(12):1263–1275, Dec 2007.

[131] Robert N Jorissen, Peter Gibbs, Michael Christie, Saurabh Prakash, Lara Lipton, Jayesh Desai, David Kerr, Lauri A Aaltonen, Diego Arango, Mogens Kruhøffer, Torben F Orntoft, Claus Lindbjerg Andersen, Mike Gruidl, Vidya P Kamath, Steven Eschrich, Timothy J Yeatman, and Oliver M Sieber. Metastasis-associated gene expression changes

predict poor outcomes in patients with dukes stage b and c colorectal cancer. *Clin Cancer Res*, 15(24):7642–7651, Dec 2009.

[132] Balazs Gyorffy, Bela Molnar, Hermann Lage, Zoltan Szallasi, and Aron C Eklund. Evaluation of microarray preprocessing algorithms based on concordance with rt-pcr in clinical samples. *PLoS One*, 4(5):e5645, 2009.

[133] Jacob Tveiten Bjerrum, Morten Hansen, Jørgen Olsen, and Ole Haagen Nielsen. Genome-wide gene expression analysis of mucosal colonic biopsies and isolated colonocytes suggests a continuous inflammatory state in the lamina propria of patients with quiescent ulcerative colitis. *Inflamm Bowel Dis*, 16(6):999–1007, Jun 2010.

[134] Raul Fechete, Andreas Heinzel, Paul Perco, Konrad Mönks, Johannes Söllner, Gil Stelzer, Susanne Eder, Doron Lancet, Rainer Oberbauer, Gert Mayer, and Bernd Mayer. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics Clin Appl*, 5(5-6):354–366, Jun 2011.

[135] Andreas Heinzel, Raul Fechete, Johannes Söllner, Paul Perco, Georg Heinze, Rainer Oberbauer, Gert Mayer, Arno Lukas, and Bernd Mayer. Data graphs for linking clinical phenotype and molecular feature space. *International Journal of Systems Biology and Biomedical Technologies (IJSBBT)*, 1(1):11–25, 2012.

[136] BioCarta. `www.biocarta.com`.

[137] K. R. Brown and I. Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*, 8(5):R95, 2007.

[138] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kahari, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Harrow, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. Searle. Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–90, 2012.

[139] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, G. Cesareni, A. Chatr-aryamontri, E. Chautard, C. Chen, M. Dumousseau, J. Goll, R. E. Hancock, L. I. Hannick, I. Jurisica, J. Khadake, D. J. Lynn, U. Mahadevan, L. Perfetto, A. Raghunath, S. Ricard-Blum, B. Roechert, L. Salwinski, V. Stumpflen, M. Tyers, P. Uetz, I. Xenarios, and H. Hermjakob. Protein interaction data curation: the international molecular exchange (imex) consortium. *Nat Methods*, 9(4):345–50, 2012.

[140] G. Liu, J. Li, and L. Wong. Assessing and predicting protein interactions using both local and global network topological metrics. *Genome Inform*, 21:138–49, 2008.

[141] M. Tyagi, K. Hashimoto, B. A. Shoemaker, S. Wuchty, and A. R. Panchenko. Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep*, 13(3):266–71, 2012.

[142] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85, 2004.

[143] H. Bjorkelund, L. Gedda, and K. Andersson. Avoiding false negative results in specificity analysis of protein-protein interactions. *J Mol Recognit*, 24(1):81–9, 2011.

[144] O. Kuchaiev, M. Rasajski, D. J. Higham, and N. Przulj. Geometric de-noising of protein-protein interaction networks. *PLoS Comput Biol*, 5(8):e1000454, 2009.

[145] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–23, 2003.

[146] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A*, 105(19):6959–64, 2008.

[147] K. Venkatesan, J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi, and M. Vidal. An empirical framework for binary interactome mapping. *Nat Methods*, 6(1):83–90, 2009.

[148] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37(Database issue):D619–22, 2009.

[149] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, and S. Y. Yong. Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*, 40(Database issue):D306–12, 2011.

[150] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue):D115–9, 2004.

[151] The NCBI FTP site. `ftp.ncbi.nlm.nih.gov/gene/DATA/`.

[152] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler. The hupo psi's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–83, 2004.

[153] M. Mistry and P. Pavlidis. Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327, 2008.

[154] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, pages 601–12, 2003.

[155] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–81, 2007.

[156] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal Of Artificial Intelligence Research*, 11:95–130, 1999.

[157] D. Kemmer, R. M. Podowski, D. Yusuf, J. Brumm, W. Cheung, C. Wahlestedt, B. Lenhard, and W. W. Wasserman. Gene characterization index: assessing the depth of gene annotation. *PLoS One*, 3(1):e1440, 2008.

[158] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 39(Database issue):D52–7, 2011.

[159] I. Paul, S. F. Ahmed, A. Bhowmik, S. Deb, and M. K. Ghosh. The ubiquitin ligase chip regulates c-myc stability and transcriptional activity. *Oncogene*, 2012.

[160] D. Mukhopadhyay and H. Riezman. Proteasome-independent functions of ubiquitin in endocytosis and signaling. *Science*, 315(5809):201–5, 2007.

[161] L. Soucek, J. Whitfield, C. P. Martins, A. J. Finch, D. J. Murphy, N. M. Sodir, A. N. Karnezis, L. B. Swigart, S. Nasi, and G. I. Evan. Modelling myc inhibition as a cancer therapy. *Nature*, 455(7213):679–83, 2008.

[162] B. J. North and E. Verdin. Sirtuins: Sir2-related nad-dependent protein deacetylases. *Genome Biol*, 5(5):224, 2004.

[163] R. A. North. Molecular physiology of p2x receptors. *Physiol Rev*, 82(4):1013–67, 2002.

[164] J. Hou and X. Chi. Predicting protein functions from ppi networks using functional aggregation. *Math Biosci*, 2012.

[165] D. De Martino, M. Figliuzzi, A. De Martino, and E. Marinari. A scalable algorithm to explore the gibbs energy landscape of genome-scale metabolic networks. *PLoS Comput Biol*, 8(6):e1002562, 2012.

[166] P. Sumazin, X. Yang, H. S. Chiu, W. J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva, and A. Califano. An extensive microrna-mediated network of rna-rna interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–81, 2011.

[167] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, C. J. Adams-Collier, and Kenta Nakai. Wolf psort: protein localization predictor. *Nucleic Acids Res*, 35(Web Server issue):W585–W587, Jul 2007.

[168] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006.

[169] emergentec.BIO. `bio.emergentec.com`.

[170] NextBio. `www.nextbio.com`.

[171] A. Bernthaler. *Disease specific gene expression profiles in the context of molecular networks.* Phd thesis, Technical University of Vienna, 2009.

[172] R. Fechete. Data models, graph analysis, and information retrieval from biological data. Master's thesis, Technical University of Vienna, 2009.

[173] The Hibernate framework. `www.hibernate.org`.

[174] The GlassFish Java Application Server. `glassfish.java.net`.

[175] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue):D760–D765, Jan 2007.

[176] D. J. Newman, M. B. Mattock, A. B. Dawnay, S. Kerry, A. McGuire, M. Yaqoob, G. A. Hitman, and C. Hawke. Systematic review on urine albumin testing for early detection of diabetic complications. *Health Technol Assess*, 9(30):iii–vi, xiii–163, 2005.

[177] M. Bojestig, H. J. Arnqvist, G. Hermansson, B. E. Karlberg, and J. Ludvigsson. Declining incidence of nephropathy in insulin-dependent diabetes mellitus. *N Engl J Med*, 330(1):15–8, 1994.

[178] P. Finne, A. Reunanen, S. Stenman, P. H. Groop, and C. Gronhagen-Riska. Incidence of end-stage renal disease in patients with type 1 diabetes. *Jama*, 294(14):1782–7, 2005.

[179] A. I. Adler, R. J. Stevens, S. E. Manley, R. W. Bilous, C. A. Cull, and R. R. Holman. Development and progression of nephropathy in type 2 diabetes: the united kingdom prospective diabetes study (ukpds 64). *Kidney Int*, 63(1):225–32, 2003.

[180] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care*, 27(5):1047–53, 2004.

[181] Akin Özyilmaz, Stephan J L Bakker, Dick de Zeeuw, Paul E de Jong, Ronald T Gansevoort, and P. R. E. V. E. N. D. Study Group. Selection on albuminuria enhances the efficacy of screening for cardiovascular risk factors. *Nephrol Dial Transplant*, 25(11):3560–3568, Nov 2010.

[182] J. F. Mann, R. E. Schmieder, M. McQueen, L. Dyal, H. Schumacher, J. Pogue, X. Wang, A. Maggioni, A. Budaj, S. Chaithiraphan, K. Dickstein, M. Keltai, K. Metsarinne, A. Oto, A. Parkhomenko, L. S. Piegas, T. L. Svendsen, K. K. Teo, and S. Yusuf. Renal outcomes with telmisartan, ramipril, or both, in people at high vascular risk (the ontarget study): a multicentre, randomised, double-blind, controlled trial. *Lancet*, 372(9638):547–53, 2008.

[183] George L Bakris, Pantelis A Sarafidis, Matthew R Weir, Björn Dahlöf, Bertram Pitt, Kenneth Jamerson, Eric J Velazquez, Linda Staikos-Byrne, Roxzana Y Kelly, Victor Shi, Yann-Tong Chiang, Michael A Weber, and A. C. C. O. M. P. L. I. S. H. Trial investigators. Renal outcomes with different fixed-dose combination therapies in patients with hypertension at high risk for cardiovascular events (accomplish): a prespecified secondary analysis of a randomised controlled trial. *Lancet*, 375(9721):1173–1181, Apr 2010.

[184] A. S. Levey, D. Cattran, A. Friedman, W. G. Miller, J. Sedor, K. Tuttle, B. Kasiske, and T. Hostetter. Proteinuria as a surrogate outcome in ckd: report of a scientific workshop sponsored by the national kidney foundation and the us food and drug administration. *Am J Kidney Dis*, 54(2):205–26, 2009.

[185] Ana M Blázquez-Medela, José M López-Novoa, and Carlos Martínez-Salgado. Mechanisms involved in the genesis of diabetic nephropathy. *Curr Diabetes Rev*, 6(2):68–87, Mar 2010.

[186] A. C. C. O. R. D. Study Group, William C Cushman, Gregory W Evans, Robert P Byington, David C Goff, Richard H Grimm, Jeffrey A Cutler, Denise G Simons-Morton, Jan N Basile, Marshall A Corson, Jeffrey L Probstfield, Lois Katz, Kevin A Peterson, William T Friedewald, John B Buse, J. Thomas Bigger, Hertzel C Gerstein, and Faramarz Ismail-Beigi. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med*, 362(17):1575–1585, Apr 2010.

[187] P. Fioretto, M. W. Steffes, D. E. Sutherland, F. C. Goetz, and M. Mauer. Reversal of lesions of diabetic nephropathy after pancreas transplantation. *N Engl J Med*, 339(2):69–75, 1998.

[188] R. R. Holman, S. K. Paul, M. A. Bethel, H. A. Neil, and D. R. Matthews. Long-term follow-up after tight control of blood pressure in type 2 diabetes. *N Engl J Med*, 359(15):1565–76, 2008.

[189] KDOQI. Clinical practice guidelines and clinical practice recommendations for diabetes and chronic kidney disease.

[190] Y. Pan, L. L. Guo, and H. M. Jin. Low-protein diet for diabetic nephropathy: a meta-analysis of randomized controlled trials. *Am J Clin Nutr*, 88(3):660–6, 2008.

[191] P. Gaede, P. Vedel, N. Larsen, G. V. Jensen, H. H. Parving, and O. Pedersen. Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. *N Engl J Med*, 348(5):383–93, 2003.

[192] B. V. Howard, M. J. Roman, R. B. Devereux, J. L. Fleg, J. M. Galloway, J. A. Henderson, W. J. Howard, E. T. Lee, M. Mete, B. Poolaw, R. E. Ratner, M. Russell, A. Silverman, M. Stylianou, J. G. Umans, W. Wang, M. R. Weir, N. J. Weissman, C. Wilson, F. Yeh, and J. Zhu. Effect of lower targets for blood pressure and ldl cholesterol on atherosclerosis in diabetes: the sands randomized trial. *Jama*, 299(14):1678–89, 2008.

[193] Anna Köttgen, Cristian Pattaro, Carsten A Böger, Christian Fuchsberger, Matthias Olden, Nicole L Glazer, Afshin Parsa, Xiaoyi Gao, Qiong Yang, Albert V Smith, Jeffrey R O'Connell, Man Li, Helena Schmidt, Toshiko Tanaka, Aaron Isaacs, Shamika Ketkar, Shih-Jen Hwang, Andrew D Johnson, Abbas Dehghan, Alexander Teumer, Guillaume Paré, Elizabeth J Atkinson, Tanja Zeller, Kurt Lohman, Marilyn C Cornelis, Nicole M Probst-Hensch, Florian Kronenberg, Anke Tönjes, Caroline Hayward, Thor Aspelund, Gudny Eiriksdottir, Lenore J Launer, Tamara B Harris, Evadnie Rampersaud, Braxton D Mitchell, Dan E Arking, Eric Boerwinkle, Maksim Struchalin, Margherita Cavalieri, Andrew Singleton, Francesco Giallauria, Jeffrey Metter, Ian H de Boer, Talin Haritunians, Thomas Lumley, David Siscovick, Bruce M Psaty, M. Carola Zillikens, Ben A Oostra, Mary Feitosa, Michael Province, Mariza de Andrade, Stephen T Turner, Arne Schillert, Andreas Ziegler, Philipp S Wild, Renate B Schnabel, Sandra Wilde, Thomas F Munzel, Tennille S Leak, Thomas Illig, Norman Klopp, Christa Meisinger, H-Erich Wichmann, Wolfgang Koenig, Lina Zgaga, Tatijana Zemunik, Ivana Kolcic, Cosetta Minelli, Frank B Hu, Asa Johansson, Wilmar Igl, Ghazal Zaboli, Sarah H Wild, Alan F Wright, Harry Campbell, David Ellinghaus, Stefan Schreiber, Yurii S Aulchenko, Janine F Felix, Fernando Rivadeneira, Andre G Uitterlinden, Albert Hofman, Medea Imboden, Dorothea Nitsch, Anita Brandstätter, Barbara Kollerits, Lyudmyla Kedenko, Reedik Mägi, Michael Stumvoll, Peter Kovacs, Mladen Boban, Susan Campbell, Karlhans Endlich, Henry Völzke, Heyo K Kroemer, Matthias Nauck, Uwe Völker, Ozren Polasek, Veronique Vitart, Sunita Badola, Alexander N Parker, Paul M Ridker, Sharon L R Kardia, Stefan Blankenberg, Yongmei Liu, Gary C Curhan, Andre Franke, Thierry Rochat, Bernhard

Paulweber, Inga Prokopenko, Wei Wang, Vilmundur Gudnason, Alan R Shuldiner, Josef Coresh, Reinhold Schmidt, Luigi Ferrucci, Michael G Shlipak, Cornelia M van Duijn, Ingrid Borecki, Bernhard K Krämer, Igor Rudan, Ulf Gyllensten, James F Wilson, Jacqueline C Witteman, Peter P Pramstaller, Rainer Rettig, Nick Hastie, Daniel I Chasman, W. H. Kao, Iris M Heid, and Caroline S Fox. New loci associated with kidney function and chronic kidney disease. *Nat Genet*, 42(5):376–384, May 2010.

[194] C. D. Cohen, M. T. Lindenmeyer, F. Eichinger, A. Hahn, M. Seifert, A. G. Moll, H. Schmid, E. Kiss, E. Grone, H. J. Grone, M. Kretzler, T. Werner, and P. J. Nelson. Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis. *PLoS One*, 3(8):e2937, 2008.

[195] M. L. Merchant and J. B. Klein. Proteomics and diabetic nephropathy. *Curr Diab Rep*, 5(6):464–9, 2005.

[196] J. L. Sebedio, E. Pujos-Guillot, and M. Ferrara. Metabolomics in evaluation of glucose disorders. *Curr Opin Clin Nutr Metab Care*, 12(4):412–8, 2009.

[197] K. Rossing, H. Mischak, M. Dakna, P. Zurbig, J. Novak, B. A. Julian, D. M. Good, J. J. Coon, L. Tarnow, and P. Rossing. Urinary proteomics in diabetes and ckd. *J Am Soc Nephrol*, 19(7):1283–90, 2008.

[198] Alaa Alkhalaf, Petra Zürbig, Stephan J L Bakker, Henk J G Bilo, Marie Cerna, Christine Fischer, Sebastian Fuchs, Bart Janssen, Karel Medek, Harald Mischak, Johannes M Roob, Kasper Rossing, Peter Rossing, Ivan Rychlík, Harald Sourij, Beate Tiran, Brigitte M Winklhofer-Roob, Gerjan J Navis, and P. R. E. D. I. C. T. I. O. N. S. Group. Multicentric validation of proteomic biomarkers in urine specific for diabetic nephropathy. *PLoS One*, 5(10):e13421, 2010.

[199] Caroline S Fox, Philimon Gona, Martin G Larson, Jacob Selhub, Geoffrey Tofler, Shih-Jen Hwang, James B Meigs, Daniel Levy, Thomas J Wang, Paul F Jacques, Emelia J Benjamin, and Ramachandran S Vasan. A multi-marker approach to predict incident ckd and microalbuminuria. *J Am Soc Nephrol*, 21(12):2143–2149, Dec 2010.

[200] Franck Molina, Matthias Dehmer, Paul Perco, Armin Graber, Mark Girolami, Goce Spasovski, Joost P Schanstra, and Antonia Vlahou. Systems biology: opening new avenues in clinical research. *Nephrol Dial Transplant*, 25(4):1015–1018, Apr 2010.

[201] I. Mühlberger, K. Mönks, A. Bernthaler, C. Jandrasits, B. Mayer, G. Mayer, R. Oberbauer, and Perco P. Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome. *International Journal of Nephrology*, 2011:10, 2011.

[202] P. Perco, J. Wilflingseder, A. Bernthaler, M. Wiesinger, M. Rudnicki, B. Wimmer, B. Mayer, and R. Oberbauer. Biomarker candidates for cardiovascular disease and bone metabolism disorders in chronic kidney disease: a systems biology perspective. *J Cell Mol Med*, 12(4):1177–87, 2008.

[203] Aldons J Lusis and James N Weiss. Cardiovascular networks: systems-based approaches to cardiovascular disease. *Circulation*, 121(1):157–170, Jan 2010.

[204] C. A. Hidalgo, N. Blumm, A. L. Barabasi, and N. A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*, 5(4):e1000353, 2009.

[205] F. Barrenas, S. Chavali, P. Holme, R. Mobini, and M. Benson. Network properties of complex human disease genes identified through genome-wide association studies. *PLoS One*, 4(11):e8090, 2009.

[206] Z. Dezso, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, and A. Bugrim. Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol*, 3:36, 2009.

[207] Joel T Dudley and Atul J Butte. Identification of discriminating biomarkers for human disease using integrative network biology. In *Pac Symp Biocomput*, pages 27–38, 2009.

[208] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi. The human disease network. *Proc Natl Acad Sci U S A*, 104(21):8685–90, 2007.

[209] A. Ozgur, T. Vu, G. Erkan, and D. R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–85, 2008.

[210] M. Xu, M. C. Kao, J. Nunez-Iglesias, J. R. Nevins, M. West, and X. J. Zhou. An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics*, 9 Suppl 1:S12, 2008.

[211] J. Li, X. Zhu, and J. Y. Chen. Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. *PLoS Comput Biol*, 5(7):e1000450, 2009.

[212] The PubMed database. `www.ncbi.nlm.nih.gov/pubmed`.

[213] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, and I. Forsythe. Hmdb: a knowledgebase for the human metabolome. *Nucleic Acids Res*, 37(Database issue):D603–10, 2009.

[214] The NCBI Gene2PubMed file. `ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz`.

[215] The ClinicalTrials.gov. `www.clinicaltrials.gov`.

[216] The PatentLens. `www.patentlens.net`.

[217] The Mozilla language tools. addons.mozilla.org/en-US/firefox/language-tools.

[218] Brian Johnson and Ben Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *IEEE Conference on Visualization*, 1991.

[219] The TreeMap project. www.cs.umd.edu/hcil/treemap-history.

[220] Maria Mironidou-Tzouveleki, Stergios Tsartsalis, and Constantinos Tomos. Vascular endothelial growth factor (vegf) in the pathogenesis of diabetic nephropathy of type 1 diabetes mellitus. *Curr Drug Targets*, 12(1):107–114, Jan 2011.

[221] M. Rudnicki, P. Perco, J. Enrich, S. Eder, D. Heininger, A. Bernthaler, M. Wiesinger, R. Sarkozi, S. J. Noppert, H. Schramek, B. Mayer, R. Oberbauer, and G. Mayer. Hypoxia response and vegf-a expression in human proximal tubular epithelial cells in stable and progressive renal disease. *Lab Invest*, 89(3):337–46, 2009.

[222] Weina Jiang, Yan Zhang, Huijuan Wu, Xin Zhang, Hualei Gan, Jianyong Sun, Qi Chen, Muyi Guo, and Zhigang Zhang. Role of cross-talk between the smad2 and mapk pathways in tgf-beta1-induced collagen iv expression in mesangial cells. *Int J Mol Med*, 26(4):571–576, Oct 2010.

[223] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S. A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–72, 2009.

[224] J. A. Vizcaino, R. Cote, F. Reisinger, J. M. Foster, M. Mueller, J. Rameseder, H. Hermjakob, and L. Martens. A guide to the proteomics identifications database proteomics data repository. *Proteomics*, 9(18):4276–83, 2009.

[225] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, Alexandra Sirota-Madi, Tsviya Olender, Yaron Golan, Gil Stelzer, Arye Harel, and Doron Lancet. Genecards version 3: the human gene integrator. *Database (Oxford)*, 2010:baq020, 2010.

[226] The NephroMine. www.nephromine.org.

[227] The SysKid project. www.syskid.eu.

[228] Y. Morita, H. Ikeguchi, J. Nakamura, N. Hotta, Y. Yuzawa, and S. Matsuo. Complement activation products in the urine from proteinuric patients. *J Am Soc Nephrol*, 11(4):700–7, 2000.

[229] M. M. Markiewski, B. Nilsson, K. N. Ekdahl, T. E. Mollnes, and J. D. Lambris. Complement and coagulation: strangers or partners in crime? *Trends Immunol*, 28(4):184–92, 2007.

[230] H. J. Lambers Heerspink, M. J. Fowler, J. Volgi, A. T. Reutens, I. Klein, T. A. Herskovits, D. K. Packham, I. R. Fraser, S. L. Schwartz, C. Abaterusso, and J. Lewis. Rationale for and study design of the sulodexide trials in type 2 diabetic, hypertensive patients with microalbuminuria or overt nephropathy. *Diabet Med*, 24(11):1290–5, 2007.

[231] S. Kume, T. Uzu, K. Isshiki, and D. Koya. Peroxisome proliferator-activated receptors in diabetic nephropathy. *PPAR Res*, 2008:879523, 2008.

[232] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10:284–293, 2004.

[233] Stijn van Dongen and Cei Abreu-Goodger. Using mcl to extract clusters from networks. *Methods Mol Biol*, 804:281–295, 2012.

[234] Ai Li and Steve Horvath. Network module detection: Affinity search technique with the multi-node topological overlap measure. *BMC Res Notes*, 2:142, 2009.

[235] Gary D Bader and Christopher W V Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, Jan 2003.

[236] Gil Alterovitz, Michael Xiang, Mamta Mohan, and Marco F Ramoni. Go pad: the gene ontology partition database. *Nucleic Acids Res*, 35(Database issue):D322–D327, Jan 2007.

[237] Min Zheng, Lin-Li Lv, Yu-Han Cao, Jian-Dong Zhang, Min Wu, Kun-Ling Ma, Aled O Phillips, and Bi-Cheng Liu. Urinary mrna markers of epithelial-mesenchymal transition correlate with progression of diabetic nephropathy. *Clin Endocrinol (Oxf)*, 76(5):657–664, May 2012.

[238] Claire E Hills and Paul E Squires. The role of tgf-beta and epithelial-to mesenchymal transition in diabetic nephropathy. *Cytokine Growth Factor Rev*, 22(3):131–139, Jun 2011.

[239] Claire E Hills and Paul E Squires. Tgf-beta1-induced epithelial-to-mesenchymal transition and therapeutic intervention in diabetic nephropathy. *Am J Nephrol*, 31(1):68–74, 2010.

APPENDIX A

Curriculum vitae

Raul Fechete

Higher education

Mar 2009 – present	PhD thesis in Computer Science at the Technical University of Vienna in collaboration with the University of Vienna and emergentec biodevelopment GmbH on *Omics profile interpretation on molecular interaction graphs*.
Mar 2006 – Mar 2009	Master of Science in Software Engineering at the Technical University of Vienna in collaboration with emergentec biodevelopment GmbH, with thesis on *Data Models, Graph Analysis, and Information Retrieval from Biological Data*, graduated with distinction.
Oct 2002 – Jan 2006	Bachelor of Science in Software and Information Engineering at the Technical University of Vienna with thesis on *Control-Flow/Busy-Waiting Analysis*.

Career-related activities

Feb 2008 – present	Research and development with focus on biological network analysis at emergentec biodevelopment GmbH.
Oct 2010 – Jun 2011	First year of the Austrian dance instructor certification training graduated June 2011.
Oct 2010 – Jun 2011	Dance instructor at the Bierbach Dance School, Vienna.
Oct 2007 – Sept 2010	Assistant dance instructor for ballroom dance at the Bierbach Dance School, Vienna.
Sept 2004 – Feb 2008	Teaching Assistant at the Algorithms and Data Structures Group TU Vienna, responsible for courses, tutor coordination, exam preparation and grading.

Oct 2006 – Nov 2007	Project on abstract syntax tree to control flow graph transformation at the at the Institute of Computer Aided Automation with technical report on A Framework for CFG-Based Static Program Analysis of Ada Programs, presented at the Ada Europe 2008 conference on reliable software technologies.
Feb, Jul – Sept 2005-2007	Multiple Siemens internships. Application development for the Siemens Public WLAN Project with C/C++ on Linux with focus on mobile client authentication with EAP.
Jul – Sept 2004	Siemens internship. Software development for the Session Initiated Protocol with VBA.
Feb 2004	Siemens internship. Application development using J2EE, EJB, JBoss and Oracle.

Miscellaneous

Languages	Fluent in German, English and Romanian.
Other interests	Dancing, horse riding.

APPENDIX B

Publication list

- Raul Fechete, Andreas Heinzel, Johannes Söllner, Paul Perco, Arno Lukas, and Bernd Mayer. **Expanded human protein coding gene interaction networks for serving Omics profile interpretation.** Mol Biosyst, in review, 2012.
- Andreas Heinzel, Raul Fechete, Johannes Söllner, Paul Perco, Georg Heinze, Rainer Oberbauer, Gert Mayer, Arno Lukas, and Bernd Mayer. **Data graphs for linking clinical phenotype and molecular feature space.** International Journal of Systems Biology and Biomedical Technologies (IJSBBT), 1(1):11–25, 2012.
- Raul Fechete, Andreas Heinzel, Bernd Mayer, Paul Perco and the SysKid team. **Integration towards a SysKid biomarker panel.**, SysKid 2nd Annual Meeting, Bergamo, Italy, 2012.
- Irmgard Mühlberger, Konrad Mönks, Raul Fechete, Gert Mayer, Rainer Oberbauer, Bernd Mayer, and Paul Perco. **Molecular pathways and crosstalk characterizing the cardiorenal syndrome.** OMICS, 16(3):105–112, Mar 2012.
- Arno Lukas, Johannes Soellner, Andreas Bernthaler, Bernd Mayer, Raul Fechete, Paul Perco, and Andreas Heinzel. **Critical gene targets for cytotoxic therapy.**, November 2011. Patent Application No.: EP2011/058259, Publication Np.: WO/2011/144738.
- Raul Fechete, Andreas Heinzel, Paul Perco, Konrad Mönks, Johannes Söllner, Gil Stelzer, Susanne Eder, Doron Lancet, Rainer Oberbauer, Gert Mayer, and Bernd Mayer. **Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy.** Proteomics Clin Appl, 5(5-6):354–366, Jun 2011.
- Raul Fechete, Susanne Barth, Tsviya Olender, Andreea Munteanu, Andreas Bernthaler, Aron Inger, Paul Perco, Arno Lukas, Doron Lancet, Jindrich Cinatl, Martin Michaelis, and Bernd Mayer. **Synthetic lethal hubs associated with vincristine resistant neuroblastoma.** Mol Biosyst, 7(1):200–214, Jan 2011.

- Konrad Mönks, Irmgard Mühlberger, Andreas Bernthaler, Raul Fechete, Paul Perco, Rudolf Freund, Arno Lukas, and Bernd Mayer. **Computational Reconstruction of Protein Interaction Networks.**, pages 155–180. Wiley-VCH Verlag GmbH & Co. KGaA, 2011.

- Irmgard Mühlberger, Julia Wilflingseder, Andreas Bernthaler, Raul Fechete, Arno Lukas, and Paul Perco. **Computational analysis workflows for omics data interpretation.** Methods Mol Biol, 719:379–397, 2011.

- Johannes Söllner, Andreas Heinzel, Georg Summer, Raul Fechete, Laszlo Stipkovits, Susan Szathmary, and Bernd Mayer. **Concept and application of a computational vaccinology workflow.** Immunome Res, 6 Suppl 2:S7, 2010.

- Andreas Bernthaler, Irmgard Muhlberger, Raul Fechete, Paul Perco, Arno Lukas, and Bernd Mayer. **A dependency graph approach for the analysis of differential gene expression profiles.** Mol Biosyst, 5(12):1720–31, 2009.

- Irmgard Mühlberger, Paul Perco, Raul Fechete, Bernd Mayer, and Rainer Oberbauer. **Biomarkers in renal transplantation ischemia reperfusion injury.** Transplantation, 88(3 Suppl):S14–S19, Aug 2009.

- Raul Fechete, Georg Kienesberger, and Johann Blieberger. **A framework for cfg-based static program analysis of ada programs.** Reliable Software Technologies – Ada-Europe 2008, volume 5026 of Lecture Notes in Computer Science, pages 130–143. Springer Berlin Heidelberg, 2008.