Systematic analysis of time resolved high-throughput data using stochastic network inference methods

Inaugural - Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany

for the degree of
Doctor of Natural Sciences

presented by Dipl.-Bioinf. Christian Bender

born in Bamberg, Germany

Oral examination: 23rd May, 2011

The presented work was conducted between March 2008 and March 2011 in the division of Molecular Genome Analysis at the German Cancer Research Center, Heidelberg.

Referees:

Prof. Dr. Roland Eils
University of Heidelberg, Department of Bioinformatics/Functional Genomics

Prof. Dr. Tim Beißbarth
University of Göttingen, Department of Medical Statistics

Danksagung

Ich danke Herrn Prof. Dr. Roland Eils für die freundliche Übernahme des Erstgutachtens.

Herrn Prof. Dr. Tim Beißbarth gilt mein besonderer Dank für die intensive und hervorragende Betreuung meiner Arbeit, sowie für die vielen Gelegenheiten, an denen er es mir ermöglichte, meine Arbeit aktiv in meinem Forschungsumfeld zu präsentieren und mich dadurch weiterzuentwickeln.

Herrn PD Dr. Stefan Wiemann, in dessen Abteilung Molekulare Genomanalyse ich meine Promotion durchführen konnte, danke ich für seine Ratschläge und Korrekturen während und für meine Doktorarbeit.

Frau Dr. Frauke Henjes möchte ich für die Durchführung der biologischen Experimente, die ich in meiner Arbeit verwendet habe danken, sowie für die Diskussionen über die biologischen Inhalte meiner Arbeit. Außerdem möchte ich Ihr für Ihr Korrekturlesen meiner Publikationen und meiner Doktorarbeit danken.

Ein weiterer Dank gilt der Bioinformatik-Gruppe der Molekularen Genomanalyse und der Abteilung Molecular Genetics am Nationalen Zentrum für Tumorerkrankungen. Insbesondere gilt mein Dank Anika Joecker, Maria Fälth, Marc Johannes und Stephan Gade für fruchtbare Diskussionen und eine tolle Arbeitsatmosphäre.

Außerdem möchte ich mich bei allen anderen ehemaligen Kollegen und Projektpartnern bedanken, durch die ich in Gesprächen oder durch deren Hinweise für meine schriftlichen Beiträge meinen Horizont erweitern konnte.

Schließlich möchte ich ganz herzlich meiner Freundin Nina und meinen Eltern danken, die mich immer auf meinem Weg unterstützt haben, mit Rat zur Seite standen und für mich da waren.

Abstract

Breast Cancer is the most common cancer in women and is characterised by various deregulations in signalling processes, leading to abnormal proliferation, differentiation or apoptosis. Several treatments for breast cancer exist, including the human monoclonal antibody Trastuzumab and the small molecule erlotinib, which both target and inhibit receptors of the ERBB receptor network. However, signalling processes in cancers, especially under drug treatment are not yet completely understood, and methods that learn treatment specific regulation and signalling patterns on a system-wide view from experimental data are needed. One approach is the reconstruction of interaction networks for genes or proteins under external perturbation, and many different algorithms have been proposed in the past. These include Boolean networks, Bayesian Networks, Dynamic Bayesian networks and differential equation systems, all describing the system on a different level of accuracy and complexity. However, if external perturbation is applied, the targets of the perturbations either have to be known, or only the targets of a single perturbation can be learned directly from data in current algorithms. And in general, dependencies of signalling events at different time points should be included into the modelling frameworks, too. This work proposes a novel approach to learn networks from longitudinal and externally perturbed data, called 'Dynamic Deterministic Effects Propagation Networks (*DDEPN*)'. Nodes in the network correspond to genes or proteins, selected from a particular biological system, while edges describe the interactions between the nodes. *DDEPN* models the activity of a node as boolean variable (either active or passive) and creates an activity profile of all nodes for the given time frame, depending on a given network structure. The activity profile is assessed by a likelihood score that describes the probability of the measured data given the activity profile. A network structure that fits best the measured data is identified by modifying the network such that the likelihood score is optimised. *DDEPN* is applied to a phosphoproteomic dataset from the ERBB signalling cascade, as well as to gene expression data measuring cell cycle related genes. Known signalling cascades from the ERBB and cell cycle networks could be successfully reconstructed and *DDEPN* also outperformed related network inference approaches. Further, in the ERBB data set, the combined application of the drugs erlotinib and Trastuzumab to the breast cancer cell line HCC1954 resulted in potent inhibition of growth promoting signalling effects, reflected in the down-regulation of the MAPK and AKT signalling pathways. This suggests that this combination therapy could be also a promising option for treatment of breast cancer patients.

Zusammenfassung

Brustkrebs ist die bei Frauen häufigste Krebsart, und wie auch bei anderen Krebserkrankungen ist hier die Regulation einer Vielzahl von Signalprozessen gestört, was eine verstärkte und unkontrollierte Zellproliferation zur Folge hat. Zur Bekämpfung von Krebsleiden werden derzeit verschiedenste Behandlungsmethoden angewandt, wie z.B. der monoklonale Antikörper Trastuzumab sowie das Medikament Erlotinib, die beide die Rezeptoraktivität des ERBB-Signalnetzwerkes inhibieren. Obwohl das ERBB-Netz zu einer der am Besten untersuchten Signalkaskaden gehört, sind viele Regulationsprozesse, insbesondere in Krebszellen und unter Medikamenteneinfluss, noch unbekannt. Daher sind analytische Methoden, die die behandlungsabhängige Regulation von Signalwegen aus experimentellen Daten rekonstruieren und als System beschreiben können, vielversprechende Ansätze für ein verbessertes Verständnis dieser Prozesse. Zur Rekonstruktion biologischer Netzwerke aus experimentellen Daten wurden bereits verschiedenste Methoden beschrieben, wie z.B. Boolsche, Bayes- und Dynamische Bayes-Netzwerke, sowie auch Modelle aus gekoppelten Differentialgleichungen. Sofern ein System durch externe Behandlung gestört wird, müssen bisher allerdings entweder die Zielknoten dieser Perturbationen bekannt sein, oder es kann nur der Effekt für einen einzelnen Einfluss gelernt werden. Bei zeitaufgelösten Daten ist es zudem nötig, Abhängigkeiten zwischen Messpunkten in den Signalprofilen in die Modellierung einzubinden. Im Fokus dieser Arbeit steht die Netzwerkrekonstruktion aus zeitaufgelösten Daten, die nach externer Perturbation des Systems (wie z.B. Zugabe von Inhibitoren) generiert wurden. Hierfür wird die neue Methode „Dynamic Deterministic Effects Propagation Networks (*DDEPN*)“ vorgestellt. Knoten im Netzwerk entsprechen hier den gemessenen Genen oder Proteinen und Kanten den Interaktionen zwischen diesen. Die Aktivität eines Knotens wird als boolsche Variable modelliert, und ein Aktivitätsprofil abhängig von der gegebenen Netzstruktur für den gemessenen Zeitrahmen hergeleitet. Für ein solches Profil wird mit Hilfe eines Likelihoodmodells ein Wahrscheinlichkeitsmaß berechnet, das bewertet, wie gut die gemessenen Daten die Netzwerkhypothese repräsentieren. Mit diesem Maß wird schließlich die Netzwerkstruktur im Bezug auf die Daten optimiert. Als Anwendungsbeispiele werden zwei Datensätze vorgestellt, die die Proteinphosphorylierung im ERBB-Signalnetz bzw. die Expression von Zellzyklusrelevanten Genen untersuchen. In beiden Fällen konnten nicht nur bekannte Protein- bzw. Geninteraktionen erfolgreich rekonstruiert werden, sondern auch verbesserte Ergebnisse im Vergleich zu verwandten Ansätzen erzielt werden. So wurde eine verminderte Aktivität der MAPK- und AKT-Signalkaskaden in HCC1954-Brustkrebszellen nach Behandlung mit Trastuzumab und Erlotinib identifiziert. Diese Kombinationsbehandlung könnte somit auch für Brustkrebspatienten eine vielversprechende Therapieoption darstellen.

Contents

1 Introduction

The size of data from modern biological experiments is steadily increasing, making the interpretation and visualisation of the results a challenging task. Biological networks are frequently used to represent biological knowledge in a graphical way. Despite the great variety of experimental setups, the notion of a network as representation of experimental results can be seen in a very general way. A network is a composition of nodes and edges, the latter being either directed or undirected. Nodes can correspond to biological entities such as genes or proteins, more complex structures like cells or even whole organisms. Edges describe some relationship between the entities, either in a pairwise manner, when edges are undirected and describe a mutual influence of two components, or in a parent-child relationship, pointing to a directed influence of a parental node onto its child. Examples for biological networks include undirected protein interaction networks, directed transcriptional regulation networks, protein signalling networks or metabolic networks. This work deals with two kinds of networks, in particular with protein signalling networks and transcriptional regulation networks. A novel approach for the reconstruction of networks from experimental data, called 'Dynamic Deterministic Effects Propagation Networks (*DDEPN*)', is proposed and discussed throughout this dissertation. In short, the theory behind *DDEPN* as well as testing procedures to assess its theoretical reconstruction performance are described. Further, two applications to data generated in cancer related biological systems are presented. In particular, these data are protein phosphorylation measurements in the ERBB signalling cascade, as well as gene expression measurements in cell cycle related genes. It is shown that *DDEPN* performs well from both a theoretical and practical perspective, and that it can be used successfully to generate hypotheses for the system-wide effects of external treatments that might reveal promising therapeutical options for patients carrying diseases like cancer.

1.1 Reconstruction methods of different types of biological networks

Cells are the basic building blocks of all living organisms and their complexity varies between simpler prokaryotic and higher eukaryotic cells. In prokaryotes,

cells are relatively simple structures without inner compartments, making up mostly unicellular organisms. Eukaryotic cells contain membrane enclosed compartments (like the nucleus or mitochondria) and often form complex multicellular organisms. The etymologic origin of the word 'cell' is the latin word 'cella', translated as 'small room' and indicates that cells are separated spaces in the organism that concentrate functional entities like DNA, RNA or proteins in spatial manner (Harper, 2001). Of central importance is the interplay between the various components in a cell, as, for example, binding of small molecules to proteins, micro RNAs to mRNA molecules, protein to DNA- or protein to protein-interactions. A suitable way of representing these interactions is the graphical representation as network. Herein, the nodes correspond to the interaction partners, while the edges represent the interaction between the nodes. A map of interactions can be seen as a description of the system's functioning, and comparison between these networks for different cells or under different conditions can provide insight into the regulatory processes of biological systems. Current research in molecular biology focuses more and more on the interplay of many interactors in a cell and thereby the combination of traditional experimental techniques with computational approaches for modelling and simulation of biological systems becomes increasingly important (Ideker et al., 2001; Kitano, 2002a,b). Diseases like cancer do not only affect single proteins or genes, but influence the whole system, and the investigation of system-wide abnormalities is an important step for the comprehension and also treatment of the respective disease (Vogelstein and Kinzler, 2004).

To obtain an intuitive understanding of the dynamics or the regulatory program of a biological system is difficult, due to the inherent complexity of regulatory circuits that frequently contain positive or negative feedback mechanisms (de Jong, 2002). Figure 1.1 shows schematically how experimental and computational research can work together to create and refine a holistic analysis of a biological system. Expert knowledge about a system is integrated into an initial model of the system that is used for simulation and prediction (left side of the figure). As soon as experimental data is available (right side), the system is revised and checked for adequacy in an iterative manner. There have been big efforts to set up various kinds of network models, and especially the reconstruction of biological networks from experimental data is an important subject of research. Some examples for available network inference approaches are given in the following paragraphs. In addition, comprehensive reviews can be found in de Jong (2002) or Bansal et al. (2007).

Bayesian Networks (BN) (Heckerman, 1996) have been frequently used to reconstruct gene regulatory networks from gene-expression experiments (Friedman et al., 2000; Segal et al., 2005) or to infer causal protein-protein relationships from intensity measurements quantifying protein abundance (Sachs et al.,

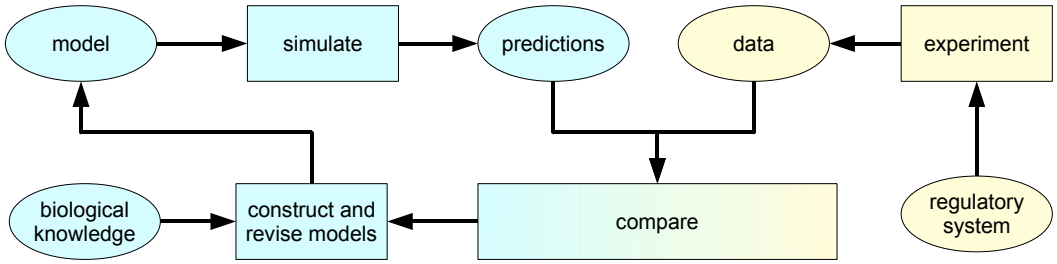

Figure 1.1 – *Combined computational and experimental analysis of biological systems. Figure adapted from de Jong (2002).*

2005). The latter is an example for directed perturbations of several measured proteins, in order to resolve the structure of the underlying interactions. Also for gene expression measurements after external perturbation, networks have been reconstructed using BNs (Pe'er et al., 2001). External interventions can be introduced by multiple means, like changing environmental conditions, applying drugs or using gene silencing methods like RNA interference (RNAi, Fire et al. (1998)). For example, Kaderali et al. (2009) utilise RNAi in a genome wide screen to reconstruct gene regulatory networks for perturbed data.

A common problem in this type of network analysis is the problem of searching for an optimal network structure depending on some evaluation criterion that maps the structure to the measured data. The problem is described to be NP-complete for BN inference (Chickering, 1996), and several heuristic approaches were proposed in the past. Early approaches include deterministic procedures like greedy search and the K2 algorithm (Heckerman, 1996; Chickering, 2003; Cooper and Herskovits, 1992). Stochastic heuristics, trying to avoid getting trapped in local optima, include simulated annealing (Heckerman, 1996), Markov Chain Monte Carlo Model Composition (MC^3, Madigan et al. (1995)), the order MCMC algorithm (Friedman and Koller, 2003), as well as evolutionary approaches (Yu et al., 2004; Spieth et al., 2006; Bevilacqua et al., 2009). Although these models have been successfully applied to various settings for BN inference, the problem of identifying unique model structures from data remains and is usually not solvable. Often, the same scoring is yielded for a number of network structures, building so called equivalence classes of network structures. A modification of the structure sampler, proposed by Castelo and Kocka (2003), is able to learn networks in equivalence classes of networks, although convergence and mixing of this structure MCMC approach seem to be unsolved. Grzegorczyk and Husmeier (2008) introduced a general edge reversal move to improve the traditional MC^3 approach and structure learning performance can be further improved by inclusion of prior

knowledge on the network structure itself. This was demonstrated in a number of publications (see Imoto et al. (2004); Gat-Viks et al. (2006); Werhli and Husmeier (2007); Mukherjee and Speed (2008); Sheridan et al. (2010)). Utilising prior knowledge seems to be most promising approach to improve reconstruction results, since the amount of publicly available information that can be used is already huge and steadily increasing (see also section 2.2).

Besides BNs, there are several related approaches to infer networks from perturbation data. Markowetz et al. (2005) derived networks from data generated after knock-out of specific genes by analysing expression patterns in the discretised gene expression measurements. Fröhlich et al. (2008a) extended this approach to perform inference on non-discretised expression levels. Perturbation effects to a system can be measured within a certain time frame. Using time resolved measurements provides insight into the dynamical behaviour of the system and does not restrict modelling to a 'snapshot' of the system's state. A suitable approach for network inference from time resolved data are Dynamic Bayesian Networks (DBN), a family of reconstruction methods including boolean network models, state-space models or regression models (Akutsu et al., 1999; Murphy and Mian, 1999; Imoto et al., 2002; Yu et al., 2004; Lébre, 2009; Rau et al., 2010).

A more concise modelling of the system is achieved by ordinary differential equation (ODE) systems, in which changes in the concentrations or intensities of the nodes are related to both external perturbations and the other nodes in the network. There are several proposals for ODE based systems. Tegner et al. (2003) suggested iterative perturbation of the system in order to reveal the underlying network structure. Perturbations were modelled as a linear combination of inputs, and weights for the pairwise node to node influences were inferred. Nelander et al. (2008) improved this idea by using non-linear perturbation effects and modelled the interaction behaviour of a number of components after several single and combinatorial perturbations. An example for a differential equation system after only stimulating perturbation can be found in Busch et al. (2008). When using steady state data, differential equations can be reduced to linear regressions for each node. Gardner et al. (2003) used this approach to infer gene regulatory networks from steady state microarray data. They solved the linear regression problem and inferred the optimal network under perturbation condition. For all of the latter approaches, the targets of the perturbations have to be known. Advances in methodology with respect to this issue include di Bernardo et al. (2005) and Bansal et al. (2006). Their approaches are able to infer the most likely targets of a single perturbation condition and additionally optimal gene regulatory networks.

As noted above, either the perturbation targets have to be known in advance, or only the effect of a single perturbation can be estimated. However, the successful usage of combinatorial perturbation was shown, for instance,

in Nelander et al. (2008); but a dynamical determination of the perturbation effects onto the network nodes is also not done. Thus, methods that explicitly include and estimate the effects of an arbitrary number of perturbations from longitudinal data are necessary. In addition, it is apparent that most of the current network reconstruction methods are tailored to the analysis of gene regulatory networks based on gene expression data from microarray experiments. Rather few studies deal with the signalling flow between proteins based on the analysis of protein activation and abundance coupled with intervention effects. In Fröhlich et al. (2009), we developed a network inference method for protein networks. Data were obtained using Reverse Phase Protein Arrays (RPPA, see section 1.2) after knockdown of the measured components at two time points. The method can handle multiple time points, but treats each time point as independent measurement and does not model the time dependent behaviour of the system explicitly. The approach is extended in this work to use multiple perturbations and to dynamically derive their effects from the data, thus filling this methodological gap (Bender et al., 2010). The approach, called *DDEPN*, will be described in section 3.1

1.2 Data generation with gene- and protein-expression-microarrays

In the previous section, two experimental techniques, DNA microarrays and RPPAs, were mentioned as options for large scale generation of data that are suitable for the different reconstruction algorithms. These two methods were used to produce the data for the application examples in this work (sections 3.6.1 and 3.7). However, these are just examples for a whole family of related technologies, all having in common the goal of high-throughput analysis of samples on various molecular levels (Hoheisel, 2006). In the following, an overview is given on the plethora of microarray techniques that are currently in use. In principle, all of them could be used either to generate input data for network reconstruction algorithms or as source for external knowledge on the networks to be reconstructed - provided, that the algorithms are properly adapted to the different data types.

To start with, the idea of a microarray goes back nearly half a century, where for the first time DNA-RNA hybrids were produced on nitrocellulose filters (Gillespie and Spiegelman, 1965). The principle idea is that a number of reference nucleic acid sequences is immobilised onto a surface, and the samples of DNA or RNA to be analysed are first labelled (nowadays usually by fluorescent dyes) and then hybridised to the arrayed DNA material. Af-

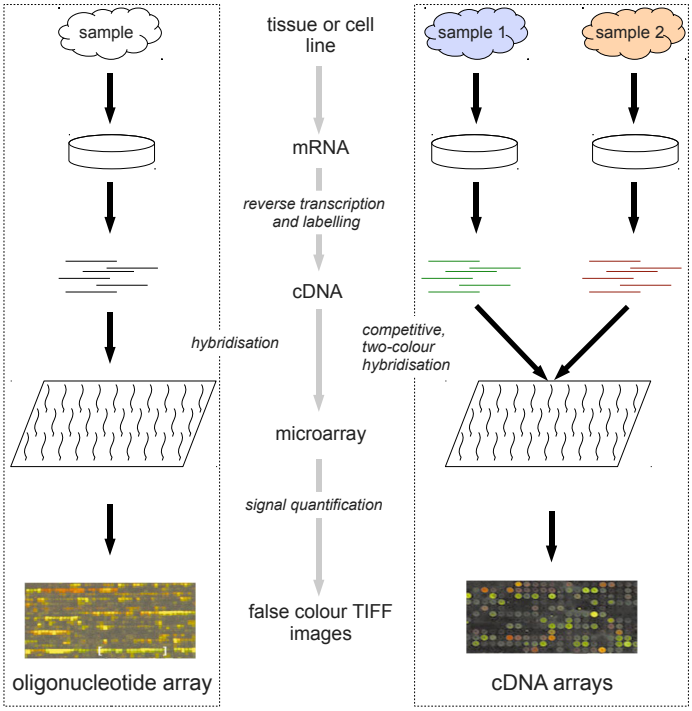

Figure 1.2 – *Analysis of gene expression using DNA-microarrays. mRNA is extracted from tissue or cell line samples. During reverse transcription labelling by fluorescent dye is performed. Afterwards, the cDNA is hybridised to a library of complementary DNA sequences that were immobilised to a surface. A read-out is generated by quantifying the intensities of the array spots after excitation of the fluorescent dyes. Left: Single sample hybridisations are performed on oligonucleotide arrays (usually multiple probes per sequence fragment) and intensities of separate arrays can be compared. Right: In two-colour arrays, competitive hybridisation is conducted and the relative amounts of sample 1 to sample 2 are quantified as expression values. Array images extracted from Lockhart and Winzeler (2000).*

ter washing away the remainder of the sample material, by excitation of the dye and subsequent quantification of the signal intensities, analysis on the relative expression strength of the contained DNA or RNA can be made (compare figure 1.2 and Brown and Botstein (1999)). Several techniques have been developed over the years, including cDNA arrays (Schena et al., 1995), oligonucleotide arrays (Southern et al., 1994) and most recently, bead arrays (Kuhn et al., 2004). So far, data from DNA-microarrays are the most frequently used data type for network reconstruction methods, and also in this work an oligonucleotide array is utilised in the CAMDA data example (see section 3.7) for data generation used as input to the *DDEPN* method.

By the transcription of DNA into RNA just one possibility of cellular control mechanisms is represented. Many events on the path from DNA to the

protein determine the fate of the cell, such as splicing, protein modifications or effects triggered by regulatory, non-coding RNAs. Microarray techniques exist to measure data on these various levels of molecular control. For example, analysing known splice variants of genes and even determining novel splice variants is possible using exon arrays, on which exonic structures are fixed on the array and probed with the RNA material from samples. The analysis of patterns of up- and down-regulated exons can then lead to the identification of active and missing splice variants (Johnson et al., 2003). Other technologies have been devised to perform genotypic profiling, used to generate high-resolution maps of genomes of various organisms (see e.g. International HapMap Consortium (2005)). Goals in genotypic profiling range from copy number variation studies, using array comparative genomic hybridisation (aCGH, Pinkel et al. (1998); Cheung et al. (2005)) to single nucleotide polymorphism (SNP) genotyping (Jobs et al., 2003; Beroukhim et al., 2010) as well as to parallel sequencing approaches (Pleasance et al., 2010). Even epigenetic analyses are possible using microarray techniques, for example by DNA methylation arrays (see Callinan and Feinberg (2006) for a review). The latter approaches are most suitable for inclusion as external knowledge into network inference procedures or for integrative analysis workflows (for example Jung et al. (2009); Li et al. (2009); Akavia et al. (2010)).

Moving away from genomic and transcriptomic profiling methods, an increasingly important field are studies on the proteome of organisms. Often the immediate response of a cell to a stimulus or change in environmental conditions is directly reflected in post-translational modifications or changes in the activity of proteins. Examples are transformational changes in protein structure, complex formation, phosphorylation or translocation of proteins in the cellular compartments and successive binding to target proteins or DNA. To measure the abundance of proteins and their modifications, Sachs et al. (2005) used flow cytometry based methodology and a BN network inference approach to reconstruct a protein signalling network. Also mass spectrometry is another widely used approach for protein signalling network investigation (Tedford et al., 2009). The focus in this work is put on measuring protein abundance by the use of protein specific antibodies, as done with the Reverse Phase Protein Array technique (Paweletz et al. (2001); Loebke et al. (2007)). Figure 1.3 shows the steps to perform automated screening for protein abundance in cellular lysates. First, after lysing the cells, the solutions are spotted on a slide, arranged in a grid. Each array is incubated with a target specific antibody that binds the protein of interest. A near-infrared dye-labelled secondary antibody is used to detect the primary antibody and to quantify the amount of protein during the scanning process afterwards. The name 'reverse phase' is derived from the fact that protein lysates are spotted on the slide, which is the reverse procedure as in antibody microarrays, which are also named 'forward arrays' (Nielsen et al., 2003; Korf et al., 2008). For this tech-

nology, the capture antibodies are immobilised on the surface and proteins in the lysate will bind to these. RPPAs were used in this work to measure the phosphorylation patterns found in a breast cancer cell line after external perturbation of the cells. The experiment, application to *DDEPN* and the obtained results are described later in section 3.6.

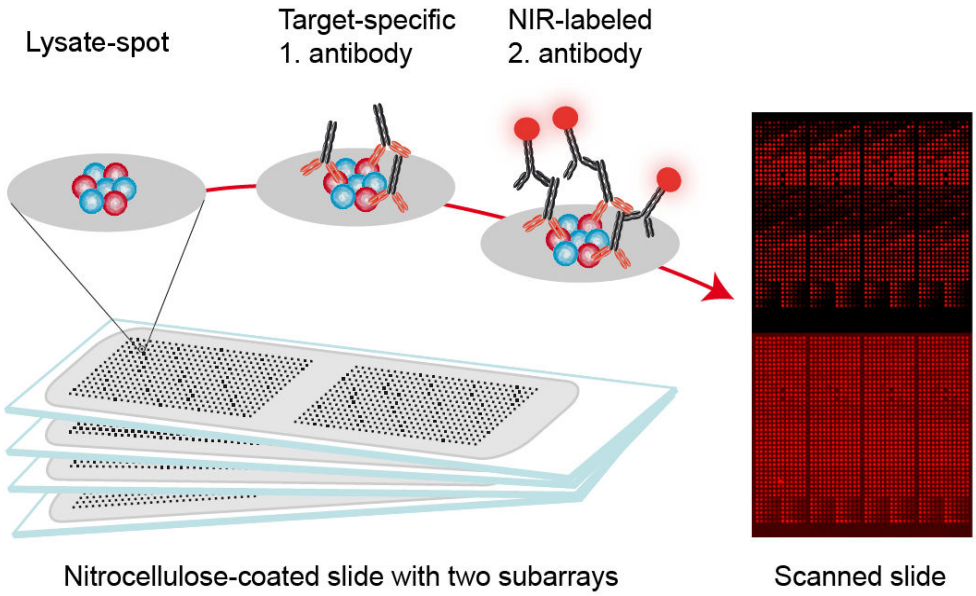

Figure 1.3 – *Measuring protein abundance using RPPAs. Cell lysate is spotted onto a nitrocellulose-coated slide. A target-specific antibody is incubated on the array, binding specifically to the protein that should be measured. Finally, incubation with a near-infrared dye-labelled secondary antibody and quantification of the signal intensities are performed. Figure adapted from Henjes (2010).*

1.3 Breast cancer is a heterogeneous disease that affects many regulatory circuits

According to the World Health Organisation databases cancer is the leading cause of deaths worldwide with around 7.6 million deaths and a number of about 12.3 million new cancer cases occurring in 2008 (Ferlay et al., 2010; World Health Organization Databank, 2010). Cancers with the highest incidence of death are lung, stomach, liver, colorectal and breast cancer. The development of cancer is characterised by multiple genetic alterations that alter the molecular circuitry and thus transform normal to malignant tumour cells. Risk factors for the development of cancers include physical, chemical and biological carcinogens like e.g. ultraviolet radiation, tobacco smoke or virus infections, respectively. In general, cancer is a genetic disease, and mul-

tiple genetic defects are required for the development of cancer (Vogelstein and Kinzler, 2004). Six physiological alterations in the cell can be defined that drive malignant tumour cell development: self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of programmed cell death, limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis (Hanahan and Weinberg, 2000). In a recent publication, the authors extend the six types of alterations by two emerging hallmarks, involved in the pathogenesis of cancer (Hanahan and Weinberg, 2011). One is the deregulation of cellular energetics, such that neoplastic proliferation is supported, and the second is avoiding immune destruction. Further, two enabling characteristics are described, in particular genome instability and mutation, as well as tumour-promoting inflammation. These two help tumours to acquire the formerly described core and emerging hallmarks. These alterations cover a wide range of regulatory processes in the cell and touch many distinct regulatory pathways.

Genetic alterations happen in three types of genes: oncogenes, tumour suppressor genes and stability genes, all contributing differently to cancer development (Vogelstein and Kinzler, 2004). Oncogene alterations lead to constitutive activity of the gene leading to a selective advantage of the cells containing the altered gene. An example is the BRAF1 gene, which frequently contains an activating mutation in its kinase domain causing constant growth stimulatory signalling (Davies et al., 2002). Tumour suppressors on the other hand are inactivated by genetic alterations and again result in a selective advantage. For instance, the TP53 gene causes inhibition of cell growth and stimulation of cell death. When an alteration renders TP53 inactive, growth control is lost (Oren, 2003). At last, stability genes (also known as 'caretakers') promote cancer development indirectly by keeping genetic alterations at a low level through DNA-repair- or chromosome-recombination and segregation mechanisms. Inactivation of these mechanisms leads to higher mutation rates which cause more frequent alterations in oncogenes or tumour suppressors. The BRCA1 gene, involved in DNA repair processes, is a prominent example for this class of cancer driving genes (Daniel, 2002). There are different types of genetic alterations, all characterised by a change in the sequence of the genome. These changes range from single point mutations to large deletions, insertions or translocation. A distinction is also made in the origin of the genetic alterations. First, mutations can occur in the germline and these are mostly point mutations or small deletions or insertions. They cause a hereditary predisposition to cancer. Second, somatic mutations in tumour cells happen spontaneously and cause sporadic cancers. As it can be seen, some type of cancer is characterised by its own composition of genetic alterations, which means that few pivotal 'causes' for a specific cancer type cannot easily be determined.

The focus in this work is set around breast cancer, the most frequent type of cancer in women with around 1.38 million new cases in 2008, both in de-

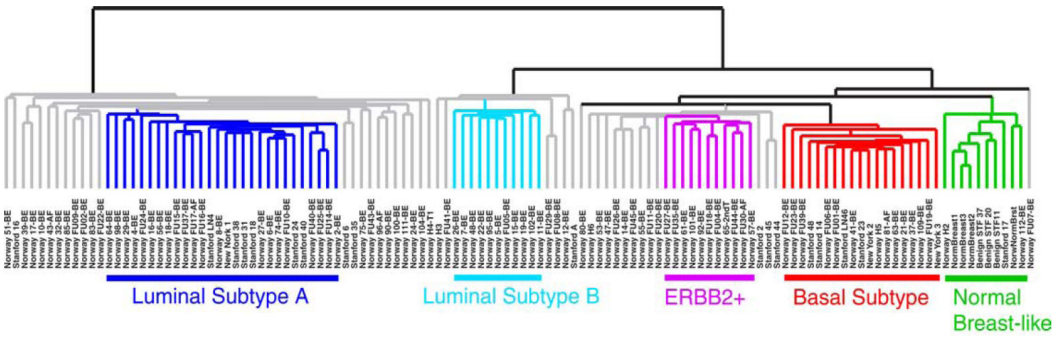

Figure 1.4 – *Molecular subtypes identified by hierarchical clustering of 115 tumour and 7 non-malignant tissues using cDNA microarrays. For the identification the 'intrinsic' gene set of 540 genes showing high variation across the microarrays was used. The colours indicate the identified subtypes. Figure from Sørlie (2004).*

veloping and developed regions (Ferlay et al., 2010). However, incidence rates vary widely across different countries, ranging from 19.3 per 100000 women in Eastern Africa to 89.7 per 100000 in Western Europe. Despite its relatively low mortality rate (6-19 per 100000), breast cancer is the most frequent cause of cancer death in women. Studies on the molecular characteristics of breast cancer patient samples using cDNA microarrays revealed five molecular subtypes with different clinical implications: the luminal A, luminal B, basal-epithelial like, ERBB2-positive and normal breast-like subtypes (Perou et al. (2000); Sørlie et al. (2001), see also figure 1.4). High oestrogen receptor (ER) levels were observed in the luminal subtypes, while the basal epithelial-like subtype showed a low ER expression and higher levels of basal epithelial molecular markers. ERBB2-positive subtype samples showed high expression of genes located in the ERBB2 amplicon at chromosome 17q22.24 and normal breast-like samples were most similar to non-epithelial cells. The different subtypes could be associated to distinct clinical outcomes. In short, the luminal subtypes showed better clinical outcome with respect to overall and relapse-free survival, while ER-positive, basal epithelial-like and ERBB2 positive samples were associated with poor overall survival (Sørlie et al., 2001). These are only some of the molecular characteristics of breast cancer, already showing its heterogeneity on the molecular level.

Vogelstein and Kinzler (2004) note that many genes are affected by cancer related alterations, but all of the genes act as part of distinct regulatory pathways in the cell. There are naturally fewer pathways than genes, and to understand a disease like cancer, a more extended view on the function of genes in the context of the pathways they are part of is necessary.

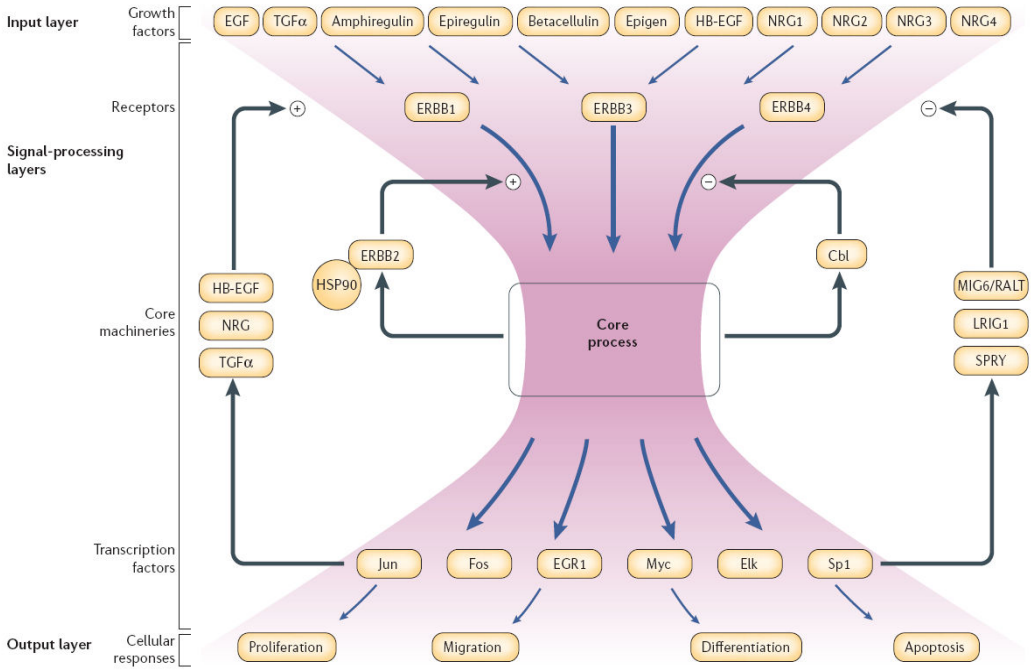

Figure 1.5 – *Abstract delineation of the ERBB signalling system. Apparent is the general bow-tie structure of the pathway. A wide range of receptor stimulating ligands in the input layer are targeting the ERBB receptors. These pass the signal through a dense core process, in order to yield a diverse set of cellular responses, shown in the output layer. These include proliferation, migration, differentiation and apoptosis. The architecture ensures redundancy, i.e. effects can be triggered in different ways, and modularity, i.e. autonomous signalling cascades are present. Additionally, several feedback mechanisms assure the system's robustness. Figure from Citri and Yarden (2006).*

1.3.1 The ERBB signalling pathway in breast cancer

The Epidermal Growth Factor (EGF)-ERBB signalling pathway is often deregulated in human cancers and a promising target for cancer therapeutics. It interweaves smaller sub-pathways like the Mitogen Activated Protein Kinase (MAPK) or PI3K/AKT pathways, and is tightly connected to growth control and cell cycle regulation (see also section 1.3.2). Genetic aberrations in the pathway lead to constitutively active signalling, and mitogenic signals triggered by activation of the ERBB pathway lead to constitutive stimulation of proliferation signals (like the phosphorylation of the pRb protein), which cause constant progression of the cells through the cell cycle. Thus, the ERBB pathway is an example for a regulatory circuit that exhibits self-sufficiency in its growth signals, one of the hallmarks of cancer mentioned earlier. It is one of the best studied biological pathways, and comprehensive reviews on existing knowledge exist, in particular reviews that relate the pathway to therapeu-

tic opportunities in the context of cancer treatment (Citri and Yarden, 2006; Hatakeyama, 2007).

The ERBB receptors (or EGF-receptors, EGFR) are a family of membrane-spanning receptor tyrosine kinases (RTK). This family comprises four receptors (ERBB1-4) which are targeted by a multitude of ligands and trigger diverse cellular processes, including proliferation, migration, differentiation and apoptosis. ERBB1 (or HER1, EGFR, Ullrich et al. (1984)) is activated through the ligands epidermal growth factor (EGF), transforming growth factor alpha (TGF-α), amphiregulin (AR), epiregulin (EPR), betacellulin (BTC), epigen and heparin-binding EGF-like growth factor (HB-EGF). It mainly activates the MAPK pathways (Seger and Krebs, 1995) by forming both homo- and heterodimers with members of the ERBB receptor family. In contrast, ERBB2 (or HER-2/neu, Yamamoto et al. (1986)) is a non-autonomous receptor without an extracellular ligand binding domain. It can only be activated by heterodimerisation with the other ERBB family members and is their preferred binding partner. ERBB2 plays an important role in many cancer related signalling processes, because its potent mitogenic signalling is often deregulated, rendering the receptor constitutively active. ERBB3 (or HER3, Kraus et al. (1989); Plowman et al. (1990)) is another non-autonomous receptor that lacks the intracellular kinase domain. It also has to interact with the other receptors to form heterodimers in order to transduce a signal into the cell. Its activation is triggered by the ligands heregulin 1 and 2 (HRG1/2, also NRG1/2) and it is a strong activator of the PI3K/AKT pathways (Manning and Cantley, 2007) when bound to ERBB2. Finally, ERBB4 (or HER4, Plowman et al. (1993)) shares some properties with ERBB1 (for example, it binds GRB2, Shc and STAT5 directly) and is activated through BTC, HB-EGF, EPR and HRG1-4 (also termed NRG1-4).

The abstract depiction of the ERBB network in figure 1.5 shows some general properties of the system. Its bow-tie architecture (Citri and Yarden, 2006) points out that signals are generated at a highly redundant input layer and passed into the network through a small number of molecular switches (the core process, i.e. common signalling cascades like MAPK and PI3K/AKT). Remarkably diverse effects can be triggered at the output layer through activation of various transcription factors. The whole system is highly modular, meaning that several functional units act more or less autonomously to trigger their effect. This also represents an important characteristic encoded in the ERBB network, namely redundancy of signalling processes. For example, various receptors of the ERBB family share the same target pathways, and effects can be triggered using up to eight receptor dimers. Both, modularity and redundancy, contribute to the robustness of the ERBB network, since activation of the core processes can happen independently over multiple separate paths through the network.

Besides, the presence of feed-forward and feed-back mechanisms further enhances the robustness of the system (Avraham and Yarden, 2011). For instance, ERBB2 serves as positive enhancer of signalling processes, once it is activated. Also secondary stimulating signalling events are triggered by activation of the MAPK cascade, such as TGFα and HB-EGF production. Furthermore, negative regulation is prevalent as control mechanism, such as receptor internalisation and degradation, or the synthesis of signal attenuators that is caused by the activation of the signalling cascade itself.

What is the role of the ERBB signalling cascade in cancer? As noted before, cancer is a disease that evolves by multiple cancer causing events like DNA amplifications or deletions, translocations or mutation accumulation. In the ERBB pathway, several mutations can be found that cause aberrant signalling behaviour and contribute to tumour development. ERBB2, for example, has been shown to contain cancer-relevant mutations in lung adenocarcinomas (Shigematsu et al., 2005). More frequently, over-expression of receptors is found to be related to several cancer types, e.g. ERBB1 over-expression in lung, pancreas and breast cancers (Nicholson et al., 2001) or ERBB2 over-expression in breast, lung, pancreas, colon, endometrium and ovarian cancer. Focusing on breast cancer, ERBB2 is found to be overexpressed in 20–30% of breast tumours and is associated with a poor prognosis and short overall survival (Slamon et al., 1987; Sørlie et al., 2001). Besides surgery, standard cancer treatment includes chemo- and radiation therapy, which target tumour cells by inhibiting cell division or DNA replication. Their use is accompanied by severe side effects, because the therapeutics do not specifically target tumour cells, but also healthy tissue. Therefore, several targeted therapies avoiding these side effects have been devised and are currently in use to treat the signalling processes leading to increased cell growth, proliferation and migration specifically. These include tyrosine kinase inhibitors, angiogenesis inhibitors, proteasome inhibitors and immunotherapeutics. Two drugs from the class of tyrosine kinase inhibitors are presented in the following that are used as perturbation treatment later in this work.

Figure 1.6 shows a simplified downstream network of the ERBB receptors and the targets of the therapeutics trastuzumab (Carter et al., 1992) and erlotinib (Hidalgo, 2003). Trastuzumab (or Herceptin) is a humanised monoclonal antibody targeting ERBB2. It was approved in patients with metastatic and ERBB2 over-expressing breast cancer and is used in combination with standard chemotherapy. A number of reviews exist on the function of trastuzumab (see e.g. Nahta and Esteva (2007); Valabrega et al. (2007)), although it should be noted that its function is not yet fully understood. Trastuzumab is reported to inhibit the MAPK and PI3K/AKT pathways and to positively regulate the p27 cell cycle inhibitor (Yakes et al. (2002), see also section 1.3.2 for an introduction to cell cycle regulation). It also induces antibody dependent cellular toxicity (ADCC, Cooley et al. (1999); Clynes et al. (2000)) and

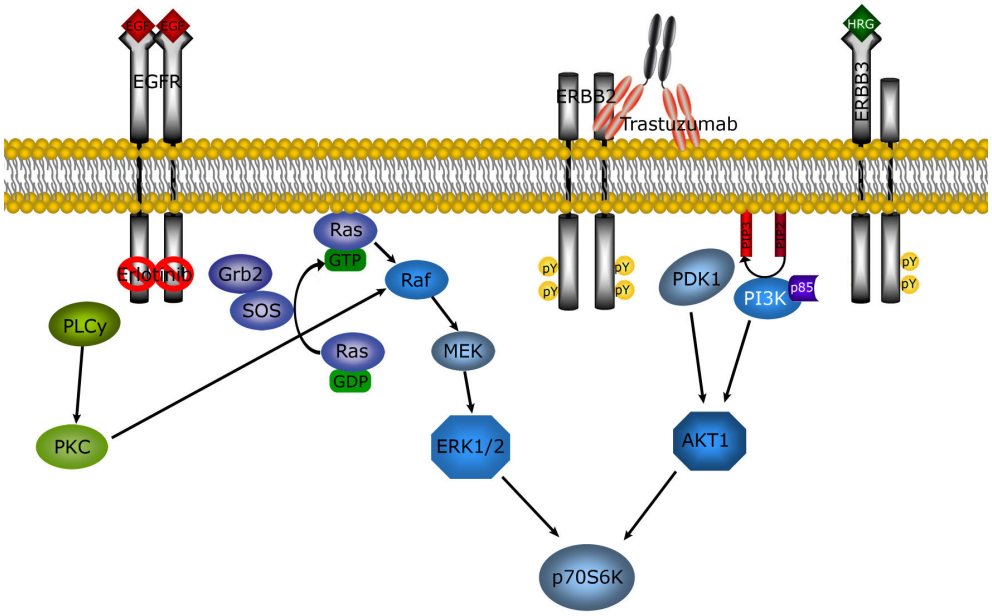

Figure 1.6 – *Basic ERBB downstream signalling including the MAPK and PI3K/AKT pathways. Inhibitors trastuzumab and erlotinib are shown, binding their target receptors. Figure adapted from Henjes (2010).*

reduces angiogenesis (Izumi et al., 2002; Klos et al., 2003). Potential impact on signalling activity might be caused by trastuzumab's ability to block extra cellular domain shedding of the ERBB2 receptor, which would leave a truncated and constitutively active form of ERBB2, named p95 (Molina et al., 2001; Christianson et al., 1998). The second drug, erlotinib, is a small molecule inhibitor that targets the intracellular tyrosine kinase domain of EGFR and, due to its lacking absolute specificity to EGFR, other kinases, too (Karaman et al., 2008). First, it was approved in 2005 for non-small cell lung cancers and afterwards for pancreatic cancers in 2007. Currently, it is tested in several phase II clinical studies in breast cancer patients (Dickler et al., 2009). Both drugs are used in section 3.6 to perturb the ERBB signalling cascade in a breast cancer cell line and the results of reconstructing the ERBB signalling pathway under various treatments by applying the *DDEPN* method are shown there.

1.3.2 ERBB signalling and its connection to the cell cycle

The ERBB signalling cascade is tightly connected to cell proliferation, migration and survival through the activation of the major downstream pathways

MAPK and PI3K/AKT. The activation of these pathways leads to activation of regulators that stimulate the master governor of the cell's fate, the cell cycle clock (Weinberg, 2007a). Here, the decision is made whether the cell will enter an active cell cycle state or a resting state, in which proliferation and growth are stopped. The cell cycle clock is composed of a network of interacting proteins that respond to extracellular signals, like those sent by RTKs or other types of receptors.

Figure 1.7 (A) depicts schematically the cell cycle with its different phases and the major proteins that are responsible for the progression through the phases (a review on the mammalian cell cycle can be found in Schafer (1998), for instance). The cycle consists of four phases, two gap phases G1 and G2 that connect the DNA-synthesis phase (S) and mitosis (M) phase. Progression through G1, S, G2 and M is mainly controlled by the expression and interaction of cyclin dependent kinases (CDKs) and cyclins. CDKs are serine/threonine protein kinases that bind to cyclins, which in turn translocate the complex to the nucleus via nuclear localisation signals. There, the CDKs can activate their targets by phosphorylation, which catalyse the progression through the various phases of the cell cycle. In G1 phase, CDK4 and CDK6 are associated with D-type cyclins and promote progression to S phase by phosphorylation of the retinoblastoma protein pRb. pRb-phosphorylation stimulates CDK2 and expression of E-type cyclins, and both proteins bind to each other and further phosphorylate the pRb protein. As soon as pRb-phosphorylation happens by CDK–Cyclin E, the cell cycle is independent from CDK4/6–Cyclin D complexes, so the first restriction point is passed. Starting at the G1/S transition, Cyclin A expression is stimulated, which induces formation of the pre-replication complex for DNA-synthesis, containing the origin recognition complex (ORC), the minichromosome maintenance proteins (MCM) and further proteins. After S phase, CDK1 binds to both Cyclin A and Cyclin B during G2 phase, which promote the entry into mitosis, named M-phase. During anaphase of the mitosis, Cyclins A and B are polyubiquitinated by the anaphase promoting complex (APC, which is activated by the Cdc20 protein) and subsequently degraded, which resets the cell's state to G1 phase for a new cell cycle. Thus, progression through the cell cycle phases occurs by precisely timed expression and degradation of cyclins and CDKs, as is depicted in figure 1.7 (B).

There are several negative regulators of the cell cycle, the p16 family and the p21 family of inhibitors, collectively termed cyclin-kinase inhibitors (CKI). The p16-family members (p16, p15, p18 and p19) inactivate the G1 CDKs (CDK4 and 6), thus preventing phosphorylation of pRb. The p21-family members p21 (or Cip1, Waf1, CDKN1A), p27 (or Kip1) and p57 (or Kip2, CDKN1C) bind and inactivate cyclins and prevent pRb-phosphorylation during G1 (Sherr and Roberts, 1999).

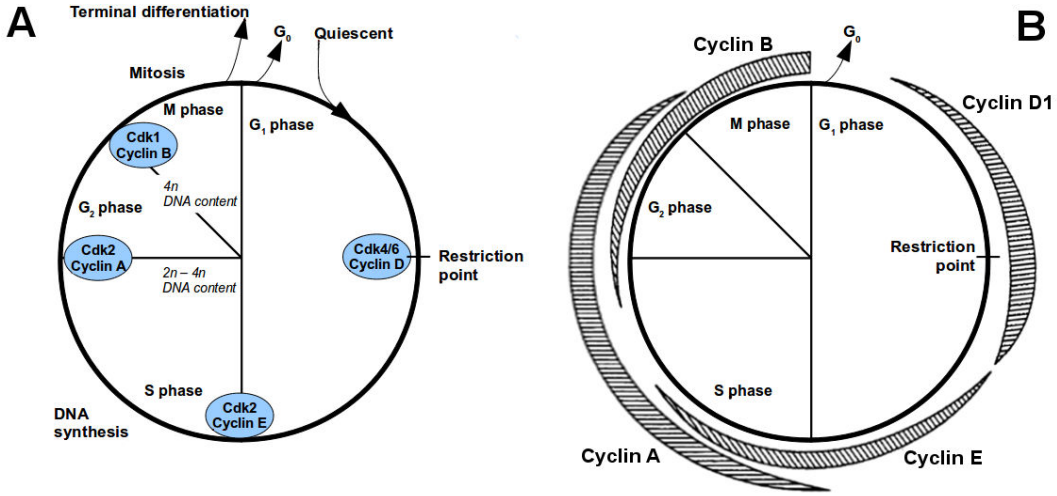

Figure 1.7 – *A: Schematic overview of the cell cycle progression. As soon as the restriction point is passed during G1 phase, the progression is independent of extracellular signals (such as mitogenic growth factors or TGF-β). B: Cell cycle phase dependent expression of cyclins. Figure adapted from Schafer (1998).*

So cell cycle progression is initiated by upstream mitogenic signalling events occurring in the ERBB network. The main downstream pathways of the ERBB receptors are the MAPK and PI3K/AKT pathways. It has been shown that ErbB2 can reversibly inhibit p27 and up-regulate cyclin D1 levels through these two pathways (Lenferink et al., 2001). Further, ERBB2 overexpression was reported to enhance cell proliferation and cell cycle progression by enhancement of the MAPK pathway (Timms et al., 2002). Once the restriction point during G1 is passed, i.e. when pRb phosphorylation is mediated by Cyclin E–Cdk2, the cell progresses through the whole cycle and is independent of the abundance of growth factors or even inhibitory signalling proteins that antagonise the proliferative signalling (like TGF-β, or transforming growth factor β, an anti-proliferative factor in epithelial cells, see e.g. Moses et al. (1990)). The tight binding of the signalling events in the ERBB network to the cell cycle regulatory circuit explains its role as cancer promoting signalling system and that major effort is put in the identification of therapeutics that control cell growth and death through the upstream signalling in the ERBB pathway.

1.4 Aims of this work

With the *DDEPN* approach, a novel network inference algorithm is developed that enables the user to reconstruct signalling networks from perturbed longitudinal data. The method was originally developed based on protein phospho-

rylation measurements of proteins of the ERBB signalling cascade that plays a crucial role in breast cancer development. A method was needed that provided the means for inferring knowledge about the treatment effects on breast cancer cell lines with different (and already clinically applied) ERBB-receptor inhibitor drugs. However, *DDEPN* is a general framework for network inference from experimental data generated in perturbed systems, and therefore also different types of data like gene expression data are suitable as input to *DDEPN*.

The aim of *DDEPN* is to model the signal flow in a biological system over time and to explicitly incorporate the effects of one or more sources of external perturbation of the system into the model. For an arbitrary number of perturbations, the effects of the perturbations onto the network nodes are inferred dynamically from the data in form of edges originating in the perturbation nodes. An extension to current MCMC structure learning approaches is developed (termed inhibMCMC) in which sampling of network structures containing two types of edges can be performed. Further, an extended prior model is developed that allows for the inclusion of external knowledge on the presence or absence of individual edges in the network. Again, the type of the edges are modelled explicitly, providing additional capabilities to other models described in the literature.

DDEPN is compared to two external DBN approaches, and its competitive performance is shown. It is also applied to two datasets: first, reconstruction of the ERBB signalling network is presented for protein phosphorylation data generated in the ERBB2 over-expressing breast cancer cell line HCC1954. It is shown that *DDEPN* is able to infer the core interactions of the ERBB signalling network and that the reconstruction can be substantially improved by the inclusion of prior knowledge. Under treatment with two ERBB receptor inhibiting drugs, the network inference suggests that combinatorial treatment with both erlotinib and trastuzumab have the strongest effect on subsequent MAPK and PI3K/AKT signalling. Second, for a set of cell cycle related genes from a transcriptional profiling experiment on microarrays, networks were reconstructed using *DDEPN* and the two external DBN approaches. In the results of this experiment, it is apparent that *DDEPN* is able to reconstruct sparse regulatory networks which yield less putative interactions than the other two DBN approaches, making the interpretation easier, and giving results that correspond better to the expected biological response.

2 Methods

The following chapter introduces methods for network inference and functional analysis of high-throughput data that are relevant for this work. First, the notions of Bayesian Networks (BN) and Dynamic Bayesian Networks (DBN), two frequently used approaches for network reconstruction, are introduced in section 2.1. Two freely available DBN implementations are used in this work to compare the results of the network reconstructions: *G1DBN* of Lébre (2009) and *ebdbNet* of Rau et al. (2010), which are discussed in the corresponding sections 2.1.1 and 2.1.2. Section 2.2 covers an overview of signalling network databases that can be used as source of prior knowledge to be incorporated in the modelling approaches. In the final section of this chapter, section 2.3, a prediction approach is presented that is used for predicting the membership of a gene or protein to signalling pathways based on their protein domain information (Fröhlich et al., 2008b) (used in the analysis workflow of section 3.7).

2.1 Bayesian Networks and Dynamic Bayesian Networks

Relationships between biological entities can be described by assembling graph structures, composed of nodes which correspond to an entity and edges that represent some kind of biological relationship. The relationship between two nodes can have different interpretations, such as transcriptional regulation, protein (de-)phosphorylation, binding of two entities or indirect causative influences between two entities. In this work only directed models are considered, i.e. a regulation is characterised by a direction from the parent to a child entity. Thus, the whole map of relationships represents a directed graph. A traditional way of describing probabilistic relationships are Bayesian Networks (BNs, Friedman et al. (2000); Pearl (1988)), where nodes correspond to random variables describing some measurable characteristic of the biological entity and edges describe the dependencies between the nodes. In BNs, graphs are not allowed to contain cycles, i.e. the graphical structures dealt with are directed acyclic graphs (DAGs), denoted as $\Phi = (V, E)$, with node set $V = \{v_i : i \in 1 \dots N\}$ and edge set $E = \{e_{ij} : i, j \in 1 \dots N\}$. Each node is associated with a conditional probability distribution (CPD) describing the marginal probability of the node given its parents. Figure 2.1 shows a toy ex-

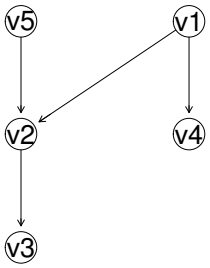

Figure 2.1 – *Toy example for a BN. The joint distribution factorises to* $P(v_1, \ldots, v_5) = P(v_1)P(v_2|v_1, v_5)P(v_3|v_2)P(v_4|v_1)P(v_5)$. *Figure adapted from Friedman et al. (2000).*

ample for a BN and its CPDs for each node. The joint probability distribution for Φ is defined as

$$P(\Phi) = \prod_{i \in 1 \ldots N} P(v_i|pa(v_i)) \tag{2.1}$$

where $pa(v_i)$ denote the parental sets of a node v_i. This means the joint distribution factorises to N conditional independent terms describing the conditional distributions of each node v_i given its parents $pa(v_i)$. The crucial point in BN inference is the definition of the CPDs. In the discrete case the random variable can take on a finite set of discrete values. Thus, the CPDs can be represented as tables containing the probabilities of a variable given each assignment of their parental values. In the continuous case, the random variables take on real values and no table can represent all joint assignments of the random variables and parental sets, so a continuous conditional density is needed (e.g. linear Gaussian densities). Given the representation of the probability distributions and a number of measurements X for the random variables, network structures can be searched for by maximising the joint probability of a model Φ given the data:

$$\hat{\Phi} = argmax_{\Phi}\{P(\Phi|X)\}.$$

BNs have been described and reviewed multiple times (Friedman et al., 2000; de Jong, 2002; Werhli, 2007), and the reader is referred to these publications for a detailed description.

The acyclicity constraint in BNs is a major obstacle for inference in biological systems, because regulatory networks often contain feedback mechanisms that are crucial for the stability and precise regulation of the system. Further, as pointed out in Werhli (2007), checking for acyclicity during network structure search represents a computational bottleneck. To model feedbacks and overcome this limitation, DBNs can be used, introduced by Friedman et al.

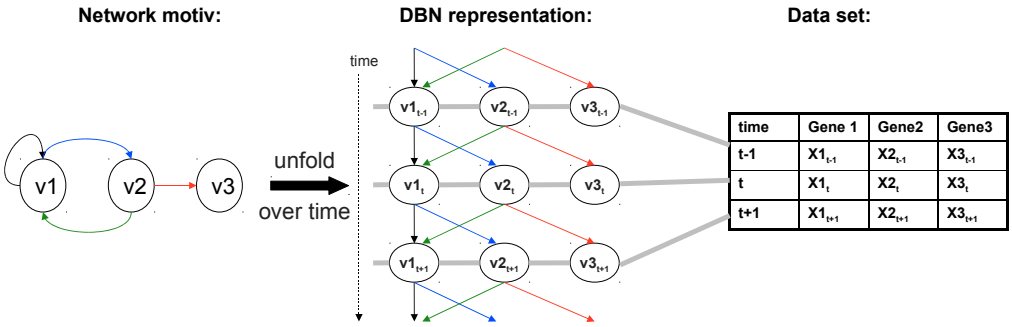

Figure 2.2 – *Left: Regulatory network for three nodes v_1, v_2, v_3. Middle: Unfolding the network over three time points results in nine nodes $v_{1,\tau}, v_{2,\tau}, v_{3,\tau}, \tau \in \{t-1, t, t+1\}$. The probability distributions are then defined as shown in equations 2.2 and 2.3. Right: toy data matrix for three proteins and three time points. Figure adapted from Lébre (2009).*

(1998) and Murphy and Mian (1999). Originally designed for time course experiments, a biological entity is now represented by one node for each time point in the experiment. That is, the network is 'unfolded' over time, making each edge at a time t only point at subsequent times $t+1$. The principle of unfolding a network over time is illustrated in figure 2.2. Unfolded network structures are again DAGs, allowing to use standard inference methods for BNs. Assuming a homogeneous stochastic process generating the random variables over time, equation 2.1 can be reformulated as

$$P(\Phi_{DBN}|t) = P(v_{1,t}, \ldots, v_{N,t}) = \prod_{i \in 1 \ldots N} P(v_{i,t}|pa(v_i)_{t-1}). \tag{2.2}$$

This means that the distribution of a variable $v_{i,t}$ of protein i at time point t is only dependent on its parents $pa(v_i)$ at time point $t-1$ (i.e. the stochastic process is Markovian). The joint distribution over all time points is then

$$P(\Phi_{DBN}) = P_0(v_{1,t_0}, \ldots, v_{N,t_0}) \prod_{t \in 1 \ldots T} P(v_{1,t}, \ldots, v_{N,t}), \tag{2.3}$$

where P_0 is the initial probability distribution for the DAG at time point 0. The terms in the product for each time point correspond to the terms from equation 2.2. See Friedman et al. (1998) for a detailed description of DBNs. Similar to BNs the joint posterior probability can be maximised during a network structure search.

2.1.1 The R-package G1DBN

The first method that is compared to *DDEPN* is the *G1DBN* approach of (Lébre, 2009). Its implementation is available as R-package 'G1DBN' from the

CRAN website (CRAN, 2011). In *G1DBN*, network structures are unfolded over time and an unfolded DAG is referred to as DAG $\tilde{\mathcal{G}}$. It describes exactly the full order conditional dependencies given the remaining past variables. First, some necessary definitions have to be provided to summarise the idea of the *G1DBN* algorithm. Let $\tilde{\mathcal{G}}$ be a DBN graph representation of probability distribution $P(\tilde{\mathcal{G}})$, with node set V and edge set E. N denotes the number of nodes, T the number of time points. The following definitions are taken from Lébre (2009):

Moral graph, (Lauritzen) The moral graph $\mathcal{G}^m$ of any DAG $\mathcal{G}$ is obtained from $\mathcal{G}$ by first 'marrying' the parents (draw an undirected edge between each pair of parents of each variable $v_{i,t}$) and then deleting directions of the original edges of $\mathcal{G}$.

Ancestral set, (Lauritzen) The subset S is ancestral if and only if, for all $\alpha \in S$, the parents of α satisfy $pa(\alpha, \mathcal{G}) \subset S$. Hence, for any subset S of vertices, there is a smallest ancestral set containing S which is denoted by $An(S)$. Then $\mathcal{G}_{An}(S)$ refers to the graph of the smallest ancestral set $An(S)$.

Conditional independence of two nodes: Two node sets U and W are said to be conditional independent given a node set S, written

$$U \perp\!\!\!\perp W|S,$$

whenever all paths from U to W intersect S in the moral graph of the smallest ancestral set containing $U \cup W \cup S$, i.e. S separates U from W.

Conditional independence between non adjacent successive variables: A node $v_{i,t}$ is conditionally independent of a preceding node $v_{j,t-1}$ if the nodes are not adjacent to each other. Then the following holds:

$$v_{i,t} \perp\!\!\!\perp v_{j,t-1}|pa(v_{i,t}, \tilde{\mathcal{G}}) \text{ and } v_{i,t} \perp\!\!\!\perp v_{j,t-1}|pa(v_{i,t}, \tilde{\mathcal{G}}) \cup S,$$

where $S \subset \{v_{k,u} : k \in 1 \dots N, u < t\}$.

qth-order conditional dependence graph $\mathcal{G}^{(q)}$: Edges in the qth-order conditional dependence graph are not drawn between two nodes $v_{i,t}$ and $v_{j,t-1}$, whenever there exists a subset $V_{q,t-1} = \{v_{q,t-1} : q < N\}$ variables, such that the $v_{i,t}$ and $v_{j,t-1}$ are conditionally independent given this subset. Define $\mathcal{G}^{(q)}, \forall q < N$:

$$\mathcal{G}^{(q)} = \left(V, \{(v_{j,t-1}, v_{i,t}) : \forall Q \in V \setminus \{j\}, |Q| = q, v_{i,t} \not\perp\!\!\!\perp v_{j,t-1} | V_{Q,t-1}\}_{i,j \in V, t \in T}\right)$$

Inferring a DAG $\tilde{\mathcal{G}}$ is done in a two step procedure in *G1DBN*. The DAG $\mathcal{G}^{(1)}$ represents the first order conditional dependency graph, describing all conditional dependencies of pairs of nodes, given exactly one remaining node. It is a superset of the true DAG $\tilde{\mathcal{G}}$, that should be inferred, and contains higher order dependencies, meaning that pairs of nodes are conditionally dependent to each other, given any set of remaining nodes. Step 1 of *G1DBN* corresponds to inferring the DAG $\mathcal{G}^{(1)}$. Consider the definition of the partial regression coefficient $a_{ij|k}$:

$$v_{i,t} = m_{ijk} + a_{ij|k} v_{j,t-1} + a_{ik|j} v_{k,t-1} + \eta_{ijk,t},$$

with some intercept m_{ijk} and centred errors $\{\eta_{ijk,t}\}_{t \geq 2}$. Conditional dependency of two variables is determined by testing the null assumption $\mathcal{H}_0^{i,j,k}$: $a_{ij|k} = 0$, using a least square estimator to find the estimate for $\hat{a}_{ij|k}$, and testing the null hypothesis using a standard t-test:

$$\frac{\hat{a}_{ij|k}}{\hat{\sigma}(\hat{a}_{ij|k})} \sim t(n-4).$$

P-values of the t-tests for each edge between nodes i and j are calculated and an ordering of the edges according to the p-values is performed. Edges are included into $\mathcal{G}^{(1)}$ if their corresponding p-value falls below a threshold α_1.

Because $\tilde{\mathcal{G}} \subset \mathcal{G}^{(1)}$, the edges of $\tilde{\mathcal{G}}$ can be selected using model selection among the edges from $\mathcal{G}^{(1)}$, by defining a second regression coefficient $a_{ij}^{(2)}$:

$$v_{i,t} = m_i + \sum_{j \in pa(v_{i,t}, \mathcal{G}^{(1)})} a_{ij}^{(2)} v_{j,t-1} + \eta_{i,t}$$

Again, p-values for each edge are calculated performing standard hypothesis testing on the null hypothesis $\mathcal{H}_0^{i,j} : a_{ij}^{(2)} = 0$ and using the statistic

$$\frac{\hat{a}_{ij}^{(2)}}{\hat{\sigma}(\hat{a}_{ij}^{(2)})} \sim t(n - 1 - |pa(v_{i,t}, \mathcal{G}^{(1)})|).$$

The DAG $\tilde{\mathcal{G}}$ contains all edges with p-values below a threshold α_2.

The detailed model is found in Lébre (2009). Here, only the idea of the inference method should be conveyed and the reader is referred to the publication for a comprehensive description of *G1DBN*.

2.1.2 The R-package ebdbNet

As a second method for comparison to *DDEPN*, *ebdbNet* is chosen. It is an Empirical Bayes DBN procedure for network inference (Rau et al., 2010). The method also was developed for microarray time series data and is available as R-package *ebdbNet* from CRAN (CRAN, 2011). Again, the inherent limitations of BNs, acyclicity and existence of equivalence classes are overcome by using a DBN formulation. The authors chose a special case of a DBN, in particular a state space model (SSM, also known as linear dynamical system) to model continuous noisy measurements and to perform inference on continuous hidden states. In general, a pair of equations, the state and dynamic equations are used to model the relationship of the measured components as well as hidden states from one time point to another, while linearity and time-invariance are assumed for the set of equations.

Assume time-course gene expression data is given for P genes, K hidden states, T time points and R replicates. The set of hidden states is denoted by $\mathbf{x}_{tr} = \{x_{tr1}, \ldots, x_{trK}\}$, the set of expression values by $\mathbf{y}_{tr} = \{y_{tr1}, \ldots, y_{trP}\}$, both for replicate r and time point t. The SSM is formulated as:

$$\mathbf{x}_{tr} = A\mathbf{x}_{t-1,r} + B\mathbf{y}_{t-1,r} + \mathbf{w}_{tr}$$
$$\mathbf{y}_{tr} = C\mathbf{x}_{t,r} + D\mathbf{y}_{t-1,r} + \mathbf{z}_{tr},$$

where $\mathbf{w}_{tr}$ and $\mathbf{z}_{tr}$ are multivariate normal, A is the $K \times K$ state dynamics matrix, B the $K \times P$ observation-to-state matrix, C is the $P \times K$ observation matrix and D the $P \times P$ observation-to-observation matrix. *ebdbNet* operates in three steps: Model selection (choice of K), estimation of hidden states and calculation of posterior distributions, which are summarised in the following three paragraphs.

Model selection: A block-Hankel matrix of auto-covariances of the time-series gene expression measurements is constructed, incorporating the assumption how long a gene or protein is able to influence the expression of other genes. This assumption has to be formulated as model parameter and is set to a value between $1 - 3$ time units forward in time (these values are common for microarray experiments). Without error, the rank of the block-Hankel matrix corresponds to the number K of hidden states. Since error in the measurements cannot be ignored, Rau et al. (2010) perform a singular value decomposition (SVD) on the block-Hankel matrix and order the Eigenvalues found in the SVD. The optimal number K of hidden states is then determined by observing the singular values and finding the point when the singular values are getting small and the variation described by the

particular singular value is not big any more. Rau et al. (2010) use a cutoff of 90% of the total variance explained to determine the number K of hidden states.

Estimation of Hidden states: Once the number of hidden states is known, a Kalman filter is used to estimate the values of the hidden states. Let $\mathbf{x}_t$ and $\mathbf{y}_t$ be the hidden state and observation values, respectively, at time t, $\hat{\mathbf{x}}_t$ be the filtered estimate of $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t^-$ be the a priori estimate of $\mathbf{x}_t$ based on the previous time step. For the filtering, the posterior means $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ are used in the following equations, and two steps are defined. In the filter step,

$$\hat{\mathbf{x}}_t^- = A\hat{\mathbf{x}}_{t-1} + B\mathbf{y}_{t-1}$$
$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^- + K(\mathbf{y}_t - C\hat{\mathbf{x}}_t^- - D\mathbf{y}_{t-1}),$$

with K being the Kalman gain matrix. In the smoothing step,

$$\hat{\mathbf{x}}_t^T = \hat{\mathbf{x}}_t + J(\hat{\mathbf{x}}_{t-1}^T - A\hat{\mathbf{x}}_t - B\mathbf{y}_t),$$

with $\hat{\mathbf{x}}_t^T$ being the smoothed estimate of the hidden state value $\mathbf{x}_t$ at time t and J being the Kalman smoothing matrix. The K and J matrices are calculated with the standard formulas. See Kalman (1960) and Bremer (2006) for the details of the Kalman filtering procedure.

Posterior distributions: The calculation of model parameters is based on the approach of Beal et al. (2005):

$$\mathbf{a}_{(j)}|\alpha \sim \mathcal{N}(\mathbf{0}, diag(\alpha)^{-1})$$
$$\mathbf{b}_{(j)}|\beta \sim \mathcal{N}(\mathbf{0}, diag(\beta)^{-1})$$
$$\mathbf{c}_{(i)}|\gamma, v_i \sim \mathcal{N}(\mathbf{0}, v_i^{-1} diag(\gamma)^{-1})$$
$$\mathbf{d}_{(i)}|\delta, v_i \sim \mathcal{N}(\mathbf{0}, v_i^{-1} diag(\delta)^{-1}).$$

Here, α, β, γ and δ are vectors that are extracted as each a column of the matrices A, B, C and D and that build a set of hyperparameters for the prior precision of the parameter matrices. Point estimates for the hyperparameters (that depend on the hidden state estimates $\hat{\mathbf{x}}$ are found with an expectation maximisation (EM) algorithm and the posterior means $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ of the parameter matrices A, B, C and D are calculated. This parameter and hidden state estimation is iterated until preset convergence criteria are fulfilled. After convergence, the network topology is derived from the posterior means of

matrix D by applying z-score cutoffs for the Gaussian posterior distributions for each edge. Scores in the lower tail of the distributions are regarded as inhibiting edges and scores in the upper tail as activations.

2.2 Signalling network databases

A lot of knowledge about signalling pathways, biochemical and metabolic as well as interaction networks has been gained in the past and a vast amount of these data are stored in public databases. This section describes a selection of databases containing manually curated signalling pathways that can be utilised as prior knowledge for various network reconstruction methods, as those used in this work. Instead of providing a holistic repository of up to date pathway databases, only three examples are presented: the Kyoto encyclopedia of genes and genomes (KEGG), Reactome and the Pathway interaction database (PID). A list of many pathway databases together with a short description and link to the respective website can be found at the Pathguide website (`http://www.pathguide.org/`, Bader et al. (2006)), providing a useful resource for pathway knowledge references.

Kyoto encyclopedia of genes and genomes

A frequently used pathway database is the Kyoto encyclopedia of genes and genomes (KEGG, Kanehisa and Goto (2000); Kanehisa et al. (2008)). It is a repository in which genomic information is linked to higher order functional information and consists of several databases organised into three layers of information, according to the type of information stored.

The building blocks of all higher order information are found in the genomic and chemical information layers. The KEGG Orthology system (KO) is used for assigning unique identifiers to each database entry in KEGG. KO-numbers are stored in the database ORTHOLOGY and information on genes or whole genomes can be stored and refined using these identifiers. This is done in the databases GENES and GENOME in the genomic information layer, for instance. Chemical information, i.e. linking of genomic content to chemical structures of endogenous molecules as well as knowledge on enzymatic reactions or structure transformations, can be found in the databases COMPOUND, DRUG, GLYCAN, ENZYME, REACTION and RPAIR. All of these six databases are referred to as KEGG LIGAND. Finally, higher order functional information is stored in the systems information layer, including the KEGG PATHWAY, MODULE, BRITE and DISEASE databases. PATHWAY contains a set of manually curated reference pathway maps between gene or molecules. For the purpose of network reconstruction, this database represents

the most useful knowledge resource in KEGG, since it comprises maps of signal transduction pathways, cellular process pathways, human disease pathways and also a large set of metabolic pathways.

All of the information in KEGG can be retrieved in KEGG markup language (KGML) format using either an FTP service or via SOAP web services. Two R-packages are available on the Bioconductor website for retrieval of KEGG pathways using the R statistical programming environment: *KEGGSOAP* for the web service API and *KEGGgraph* for ftp-based download and parsing of the KGML pathway files (Gentleman et al., 2004; Zhang and Wiemann, 2009).

Reactome

The Reactome database (Vastrik et al., 2007) is a peer-reviewed pathway database that undergoes manual curation and comprises data on both human pathways and reactions. Unlike the KEGG database, all pathways in Reactome are built up around a set of reactions that transform certain input biological entities into output biological entities. This data model is preserved for any type of pathway, where in KEGG different computational representations, e.g. for metabolic and signalling pathways, are used (i.e. metabolic pathways are represented as chemical reactions and signalling pathways as semantic graphs, where nodes have an influence onto other nodes). In Reactome, each entity can serve as input node for a reaction. The reaction transforms the input into an output entity. Further, each entity can mediate a reaction as catalyst, too. Reactome gains its flexibility to describe biological processes by its entity representation model. Each modification of a biological entity such as phosphorylation of a protein, methylation of a nucleic acid or even conformational changes of proteins are recorded as reaction that transforms an input into a separate output entity representing its modified form. This means that for example an unphosphorylated protein is a different entity as its phosphorylated counterpart. Even different sub-cellular locations can be modelled in this way by referring to a protein in the cytoplasm as one entity, and as separate entity if it is translocated into the nucleus. The transport itself is then again a simple reaction.

The pathway model of Reactome is hierarchical, i.e. each pathway can be part of a larger pathway and each pathway is cross-referenced to the Gene Ontology (GO, Ashburner et al. (2000)) biological process ontology. Likewise, larger complexes of physical entities having catalytic activity are linked to the molecular function ontology and entities itself are cross-linked to several biological databases (e.g. NCBI Entrez Gene, Ensemble), giving an easy way to obtain a functional characterisation of each entity. Pathways from Reactome can be downloaded as flat files or in various file formats, including MySQL, BioPAX (Demir et al., 2010), SBML (Hucka et al., 2003) and PSI-MITAB

(Kerrien et al., 2007), either directly or using the Reactome Web Services APIs.

Pathway interaction database

As a last pathway database example the Pathway interaction database (PID, Schaefer et al. (2009)) is given, a collaborative project between the US National Cancer Institute and Nature Publishing Group. It differs from KEGG and Reactome in its focus on signalling and regulatory pathways and does not cover metabolic pathways like the two other resources. PID includes the 'NCI-Nature Curated' pathway collection and data exported from Reactome and the BioCarta collection of pathways as external data sources. PID's pathways are organised around interaction events connecting four types of molecules: small molecules (named compounds), RNA, proteins and complexes. Five kinds of events are used in PID: a gene regulation event called transcription, that, despite its name, also includes translation, a molecule transport event called translocation, a protein-protein-interaction event with name modification and a black-box event called macroprocesses. Participants in interactions are distinguished into four roles: input, output, positive and negative regulator, where an interaction consumes its inputs, produces the outputs and uses the regulators as necessary and sufficient conditions for the interaction. Meta information on genes, proteins or small molecules is obtained by mapping to UniProt, Entrez or Chemical Abstract Service (CAS) numbers and can be augmented by additional information like chemical modification and cellular location. All interactions in the NCI-Nature Curated dataset are referenced by one or more pubmed publications together with one or more evidence codes describing the type of evidence pointing to the particular interaction. An important side note for PID is that all pathways are assumed to be valid for the non-perturbed state of the organism, i.e. the organism is healthy. Deviations from pathway structure for organisms that have a disease have to be taken care of explicitly. Export of pathways from PID for automated use is possible either in PID XML or BioPAX Level 2 formats which can be parsed and transformed into the desired format that is needed for later analyses.

Pathway compendia and further pathway resources

PID is an example for an integrative pathway database that combines information from multiple sources in a common data model and interface. Other examples of this kind of repository include the ConsensusPathDB-human (CPDB) database (Kamburov et al., 2009) that combines a set of currently 18 external resources. The strength of CPDB is that pathway databases with heterogeneous interaction types are integrated. Export of data is possible in BioPAX

format which can be utilised for parsing and further analyses. Another example is the Human pathway database (HPD, Chowbina et al. (2009)), combining information from five external databases including KEGG, Reactome and PID. These are two examples for attempts to assemble a 'bigger picture' based on a multitude of resources available so far. Since a complete review of pathway databases is outside the scope of this thesis, the reader is referred to the Pathguide website (`http://www.pathguide.org`), which provides an up to date and comprehensive overview of currently 325 pathway databases.

2.3 Gene2pathway: A method to predict signalling pathway membership for non annotated genes

A common problem in analysing data containing large numbers of biological entities like microarrays or other high-throughput technologies in genomics or proteomics is to give a functional characterisation of entities that were found significantly regulated. Often sets of entities are searched for that share functionality in order to fulfil a specific purpose, such as kinase activity, for instance. In the workflow presented for the analysis of the CAMDA dataset in section 3.7.1, after identifying differentially expressed genes among a set of around 20000 genes, the differential genes should be grouped according to the pathways in which they are active. In this section, the prediction tool *gene2pathway* (Fröhlich et al., 2008b) is described that facilitates prediction of pathway membership for genes to the set of KEGG pathways using the InterPro protein domain information (Mulder et al., 2008) of the corresponding protein. For a high number of proteins information on the protein domains is available, whereas information on pathway membership is not. Taking into account all known protein domains, for each gene/protein a domain signature can be derived that indicates which domains are present and which are not. In the *gene2pathway* approach a classification model is set up that uses the domain signatures of all proteins from one pathway as training data to learn a common pathway signature that is used later to predict whether the signature of an unknown protein belongs to the pathway or not. The method is implemented in the R package *gene2pathway*, freely available on the CRAN website.

As a reference set of pathways the KEGG database is used in *gene2pathway*. The fact that KEGG is organised hierarchically is used explicitly in this method. A misclassification of the higher-level branches like 'Metabolism' or 'Environmental Information Processing' is penalised stronger than a misclassification of a single KEGG pathway far down in the hierarchy, because membership to the general class of pathway is expected to be classified precisely, where the more specific predictions can be wrong from time to time.

Classification scheme

Given the set of all InterPro domains indexed by i, each gene p is described by a binary vector $\mathbf{x} = \{x_i : x_i \in \{0, 1\}\}$, where $x_i = 1$ if the respective InterPro domain i is present in p and $x_i = 0$ otherwise. Further, each gene is mapped to a number of positions in the KEGG hierarchy, which is described as binary position code vector $\mathbf{C} = \{C_l : l \in 1 \dots K\}$, where K is the number of individual KEGG pathways plus the number of branches at hierarchy level 2 plus the number of branches at the top level. $C_l = 1$ if the gene maps to the particular branch l or any of its child branches, where multiple mappings are possible for one protein p.

For classification of a gene/protein p with corresponding binary vector $\mathbf{x}$, a two step procedure is used. First, support vector machine classifiers (SVM) are trained to distinguish one specific branch from all others using linear kernels and soft margin parameter $C = 1$. Each SVM classifier indexed by j yields a prediction value $f_j(\mathbf{x}) \in \mathbb{R}$ that is transformed to the predicted class by taking the sign of $f_j(\mathbf{x})$. All prediction values are summarised into an input code vector $\vec{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_K(\mathbf{x}))$. Second, each input code vector $\vec{f}(\mathbf{x})$ is mapped on the best matching position code vector(s)

$$\begin{aligned} \mathbf{C}^* &= \mathbf{C}_{\hat{\mathbf{j}}} \\ \hat{j} &= \underset{j}{argmax} \langle \mathbf{C}_j, \vec{f}(\mathbf{x}) * \mathbf{w} \rangle, \end{aligned}$$

where $\{\mathbf{C}_1, \dots, \mathbf{C}_m\}$ are a dictionary of possible position code vectors (from training set of genes with KEGG and InterPro domain annotation) and $\mathbf{w}$ is a weight vector to be minimised with respect to the mismatch between predicted and true KEGG hierarchy positions in the training set. '$*$' is the component-wise multiplication.

Training procedure

In a first step for the training all K classifiers are trained and a position labelled dataset $D = \{(\vec{f}_1(\mathbf{x_1}), \mathbf{C_1}), \dots, (\vec{f}_n(\mathbf{x_n}), \mathbf{C_n}))\}$ is retrieved. Negative examples of gene mappings from the same branch are used to training each SVM. So every SVM is able to detect one specific branch in the hierarchy. In the second training step, a ranking perceptron algorithm is used to optimise the weight vector $\mathbf{w}$. The algorithm is described as follows:

Input: learning rate η, D
Output: weight vector $\mathbf{w}$
Define $F(\mathbf{x}, y) = \langle \mathbf{C}_y, \vec{f}(\mathbf{x}) * \mathbf{w} \rangle$

$\mathbf{w} = 0$
for $i = 1$ **to** n **do**
  $foes(i) = \{1, \ldots, n\} - \{p | l(\mathbf{C}_i, \mathbf{C}_p)\} = 0\}$
  $l = \underset{p \in foes(i)}{argmax} F(\mathbf{x_i}, p)$
  if $F(\mathbf{x_i}, i) - F(\mathbf{x_i}, l) < 2$ **then**
    $\mathbf{w} \leftarrow \mathbf{w} + \eta \cdot l(\mathbf{C}_i, \mathbf{C}_l) \cdot (\vec{f}(x_i) * \mathbf{C}_i - \vec{f}(x_i) * \mathbf{C}_l)$
  end if
end for

Optimising the weight vector is done by maximising the margin between position code vectors $\mathbf{C_i}$ and $\mathbf{C_j}$, $\mathbf{C_i} \neq \mathbf{C_j}$. The vector $\mathbf{w}$ is updated proportional to the loss l between a wrong position vector $\mathbf{C}_j$ and the true position vector $\mathbf{C}_i$. The loss is chosen such that a wrong prediction in the high levels of the KEGG hierarchy is penalised stronger than in the low levels in the following loss function:

$$l(\mathbf{C}, \mathbf{C}') = \sum_{i=1}^{K} c_i \mathbf{1}\{C_i \neq C_i' \text{ and } ((C_j = C_j' \forall j \in Anc(j)) \text{ or } (Anc(j) = \emptyset))\},$$

with $Anc(j)$ is the set of all ancestors of branch j and $\mathbf{1}$ the indicator function. The factor c_i is the punishment coefficient that should increase when the level in the hierarchy at which the mismatch occurs increases. Let $|T(i)|$ denote the size of the hierarchy below branch i and $|T(root)|$ the size of the complete hierarchy, then c_i is defined as:

$$c_i = \frac{|T(i)|}{|T(root)|}$$

The above described classifier can be used after training each SVM to predict the membership of each individual gene to a subset of pathways from the reference pathway set in KEGG.

3 Results

3.1 Dynamic Deterministic Effects Propagation Networks - a novel network inference approach

In the following sections a description of the framework Dynamic Deterministic Effects Propagation Networks (*DDEPN*, Bender et al. (2010, 2011)) is given. The approach was designed for network inference from longitudinal data, generated after external perturbation. A signalling or regulatory network is regarded as a directed (and possibly cyclic) graph $G = (V, E)$, where $V = \{v_i : i \in 1, \ldots, N\}$ is the set of nodes (i.e. proteins or genes) and $E = \{e_{ij} : i, j \in 1, \ldots, N, e_{ij} \in \{0, 1, -1\}\}$ the set of edges in the graph. Two types of edges are allowed (activation and inhibition), leading to the following edge encoding:

$$e_{ij} = \begin{cases} 0 \text{ if edge is not present} \\ 1 \text{ if edge is activation} \\ -1 \text{ if edge is inhibition} \end{cases} \tag{3.1}$$

Let $t \in 1 \ldots T$ denote the index for the time point and $r \in 1 \ldots R$ denote the index for the replicated measurements. The data are recorded in a matrix $X = \{x_{itr} \in \mathbb{R} : i \in 1 \ldots N, t \in 1 \ldots T, r \in 1 \ldots R\}$. A hidden boolean state is associated with each measurement x_{itr}, recorded in an unknown matrix $\Gamma^* = \{\gamma^*_{itr} \in \{0, 1\} : i \in 1 \ldots N, t \in 1 \ldots T, r \in 1 \ldots R\}$.

Figure 3.1 shows the outline of the *DDEPN* workflow. Each node has a boolean activity state that is modelled over time. Initially, all nodes are in their passive state, except for the external stimuli. These are included into the network to be reconstructed and assumed to be constantly active. Given any network hypothesis (i.e. wiring of the network, see figure 3.1, A), a synchronous update rule for the activity states is assumed. The activity at the stimuli is propagated along the edges of the network and combined at the child nodes according to a boolean logic rule. Updates are performed stepwise until a stable state is reached (figure 3.1, B, section 3.1.1), which generates a finite number of possible system states. A series of T system states is identified for each time point that maximises the likelihood for the data, given the states, using a Hidden Markov Model (HMM, figure 3.1, C) in a Viterbi Training

Expectation Maximisation (EM) algorithm. The HMM state sequence optimisation is described in section 3.1.2. During the EM, a set of model parameters Θ is estimated (figure 3.1, D), corresponding to the mean and standard deviations for two Gaussian distributions for each node (one for the active state, one for the passive state). The model parameters and system states are used to calculate the model likelihood (figure 3.1, E, section 3.1.3), which is used to perform the network structure search (figure 3.1, F), described in section 3.2.

Two different network structure search algorithms are included in *DDEPN*. A genetic algorithm (GA) for optimisation of the network structure, and a Metropolis-Hastings type of Markov Chain Monte Carlo structure sampler, that was extended by the ability to explicitly sample the space of directed (and possibly cyclic) graphs including two types of edges (activation and inhibition edges). Additionally, a prior probability model for assessing how well an inferred network reflects a given reference pathway and a prior model that penalises deviations from the scale-free property of biological networks are presented in section 3.3. Their inclusion into the *DDEPN* approach is described there, as well as the rationale for parameter adjustment to perform reasonable network inference. Finally, a short note is added on the implementation and availability at the end of this chapter.

3.1.1 Modelling the dynamics of the system by a boolean signal propagation

For a set of nodes V and an adjacency matrix Φ as defined before, the signal flow through a given network of proteins is represented as a matrix $\Gamma = \{\gamma_{ik} \in \{0,1\} : i \in 1 \dots N, k \in 1 \dots M\}$ which contains a series of possible system states $\boldsymbol{\gamma}_k = \{\gamma_i \in \{0,1\} : i \in 1 \dots N\}$. These are vectors of activation states for each node at a time step k. Define $0 < M \leq 2^N$ as number of reachable system states, determined as soon as a state is repeated during the signal propagation. Each perturbation is seen as an external influence which is included as a node into the network and whose state is constantly active (i.e. 1). Starting at the stimuli nodes, the status of all children is subsequently determined. A child is active if at least one parent connected by an activation edge is active and all parents connected via inhibition edges are inactive in the preceding step. For example, in the matrix Γ shown in figure 3.1, the state γ_{B2} of protein B at step 2 is determined by $\gamma_{B2} = \gamma_{S1} \wedge \neg\gamma_{A1} = 1 \wedge 1 = 1$ (where '$\neg$' is the logical negation which is used whenever a parent is connected via an inhibitory edge).

A formal description of the signal propagation follows: Define $S \subseteq V$ as the set of input stimuli and consider the network Φ as fixed for the propagation. In the propagation, the state matrix Γ that comprises all M reachable state

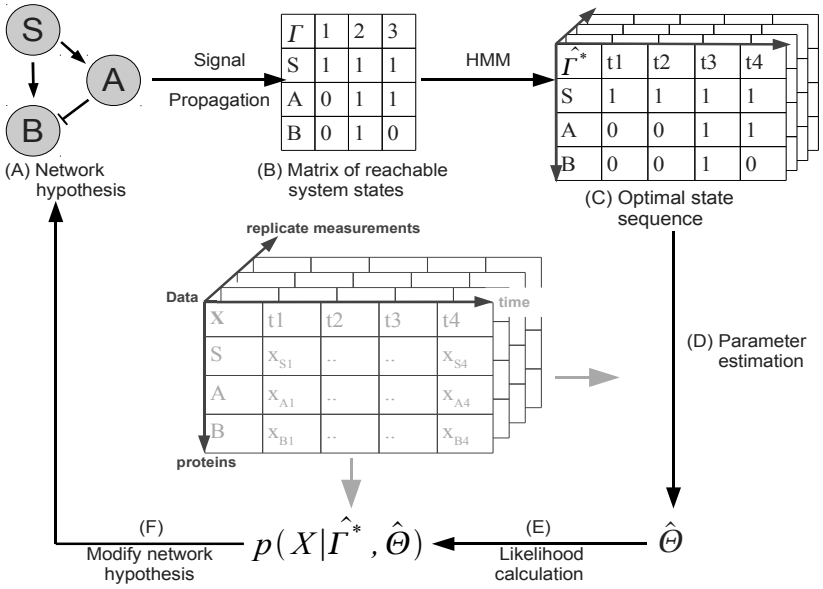

Figure 3.1 – *Overview of the approach: Given a network hypothesis (A), we generate a set of reachable system states by applying a fixed signal propagation scheme (B) which in effect reduces the number of possible system states. An optimal path through these reachable system states over time is identified by an HMM (C). Using the series of system states from the HMM, model parameters for two Gaussian distributions for each protein (one for active, one for passive) are estimated (D) and a total likelihood of our measurements given the network and model parameters is calculated (E). The likelihood score is used in the network structure search, in which the candidate networks are optimised with respect to the score (F). Two algorithms are used for the structure search: a genetic algorithm and a Metropolis-like MCMC sampling procedure.*

vectors $\boldsymbol{\gamma}_k$ is derived for the given network. The propagation is stopped at a step M, if $\exists k \leq M$, such that $\boldsymbol{\gamma}_k = \boldsymbol{\gamma}_M$, i.e. if one of the preceding states is found a second time.

All stimuli nodes $s \in S$ are active in all steps, i.e. $\gamma_{sk} = 1 \; \forall k$, and all other nodes are initialised to be 0 in the first step, i.e. $\gamma_{v_i 1} = 0 \; \forall v_i \in V \setminus S$. Let $pa(v_i)$ be the set of all parents of a node v_i and ϕ_{wv_i} an edge from a node w to v_i. For any status k and protein v_i, define

$$E^+_{k-1}(v_i) = \{\gamma_{wk-1} : \phi_{wv_i} = 1, \; \forall w \in pa(v_i)\}$$

$$E^-_{k-1}(v_i) = \{\gamma_{wk-1} : \phi_{wv_i} = -1, \; \forall w \in pa(v_i)\}$$

as the sets of states of parental nodes of v_i in step $k-1$, connected by activating edges (E^+_{k-1}) and connected by inhibiting edges (E^-_{k-1}). An entry $\gamma_{v_i k}$ in Γ is

then determined by:

$$\gamma_{v_i k} = \Big(\bigvee_{e^+ \in E^+_{k-1}(v_i)} e^+ \Big) \wedge \neg \Big(\bigvee_{e^- \in E^-_{k-1}(v_i)} e^- \Big) \tag{3.2}$$

This procedure reduces the maximal number of columns in the system state matrix Γ from 2^N to $M \leq 2^N$. However, the states in Γ do not necessarily correspond to the actual measured time points in the data. In general it is expected that a different number of reachable states as there are time points is found. For example, in the hypothetical case that the system remains in a constant state, only one state would be present in Γ. Therefore, a series of system states has to be identified which is consistent with the measured experimental data and represents the expected dynamics for the given network hypothesis. This procedure is described in the next section.

3.1.2 Searching the optimal sequence of system states using a Hidden Markov Model

Given a data matrix X and a state matrix Γ, the goal is to find an optimal state matrix Γ^*. Each entry in Γ^* represents the state of a node i at time point t and corresponds to a measurement x_{itr}, where replicate measurements indexed by r are assumed to have the same state. The index r is omitted for notational simplicity for the rest of this section, but the reader should be aware that optimisation in the HMM is done by multiplying over all replicate emission probabilities for determining the entries in the Viterbi matrix (as shown in equation 3.5).

Intuitively, Γ^* provides a classification of measurements into measurements coming from an active state and those from an inactive state. An estimate $\hat{\Gamma}^*$ for Γ^* is inferred by using an HMM $H = (W, \Gamma, A, e)$. W represents the range of possible values for observations, i.e. all positive real valued intensities generated by the array scanning software (often the pixels are encoded in 16 bit, and thus ranged within $[0, 2^{16} - 1]$). Γ is the set of possible states, as derived in section 3.1.1 and A a matrix of transition probabilities for the system states. e is referred to as the emission probability $e(\mathbf{x}_t) = p(\mathbf{x}_t | \hat{\boldsymbol{\gamma}}^*_t, \hat{\Theta})$ (equation 3.5), where $\hat{\Theta}$ is the matrix of estimated model parameters (equation 3.4). e corresponds to the likelihood of observing a data point $\mathbf{x}_t$ given its state $\hat{\boldsymbol{\gamma}}^*_t$. Note that $\mathbf{x}_t$ is a column in the measurement matrix X, i.e. a vector of intensity values.

The HMM can be used to optimise the system state sequence given a set of estimated model parameters $\hat{\Theta}$ and state transition probability matrix A. Since neither Θ and A nor the state sequence matrix $\hat{\Gamma}^*$ are known, all of them have to be estimated simultaneously. This is done using the Viterbi Training

algorithm (Durbin et al., 1998), an EM type algorithm. $\hat{\Gamma}^*$ is initialised by sampling random states from Γ, while the order of the states is preserved. A is initially set to uniform probabilities for all state transitions. The model parameters $\hat{\Theta}$ depending on $\hat{\Gamma}^*$ are estimated (see equations 3.3 and 3.4). Now $\hat{\Gamma}^*$ is updated using the HMM and the procedure iterated until convergence, as described in Durbin et al. (1998). This yields the final state matrix estimate $\hat{\Gamma}^*$ used for the likelihood calculation, described in the next section.

3.1.3 Defining the likelihood of the data for a given network hypothesis

During the HMM and the network structure search (section 3.2), a likelihood score is needed that reflects the fit of a data point to a corresponding state vector, which is regarded as emission probability. Computing the likelihoods for all data points will then reflect the fit of the measured data to the network hypothesis. Given a state matrix estimate $\hat{\Gamma}^*$, each measurement x_{itr} for protein i, time point t and replicate r comes from an 'active' normal distribution $\mathcal{N}(\mu_{i1}, \sigma_{i1})$, if its state $\hat{\gamma}^*_{itr} = 1$, and from a 'passive' normal distribution $\mathcal{N}(\mu_{i0}, \sigma_{i0})$, if $\hat{\gamma}^*_{itr} = 0$:

$$x_{itr} \sim \begin{cases} \mathcal{N}(\mu_{i0}, \sigma_{i0}),\ if\ \hat{\gamma}^*_{itr} = 0\ \ (passive) \\ \mathcal{N}(\mu_{i1}, \sigma_{i1}),\ if\ \hat{\gamma}^*_{itr} = 1\ \ (active) \end{cases} \tag{3.3}$$

The parameters of each distribution for one protein are obtained as unbiased empirical mean and standard deviation of all measurements for this protein in the given class. This yields the parameter matrix:

$$\hat{\Theta} = \{\hat{\theta}_{i0}, \hat{\theta}_{i1}\} = \{(\hat{\mu}_{i0}, \hat{\sigma}_{i0}), (\hat{\mu}_{i1}, \hat{\sigma}_{i1})\}\ \forall i \in 1 \dots N \tag{3.4}$$

Now we can write the likelihood for a data point $\mathbf{x}_t$ as :

$$\begin{aligned} p(\mathbf{x}_t|\Phi) &= p(\mathbf{x}_t|\hat{\boldsymbol{\gamma}}^*_t, \hat{\Theta}) \\ &= \prod_{i=1}^{N}\prod_{r=1}^{R} p(x_{itr}|\hat{\theta}_{\mathbf{i}\hat{\gamma}^*_{\mathbf{itr}}}) \end{aligned} \tag{3.5}$$

The total likelihood for a network hypothesis Φ can be written as:

$$\begin{aligned} p(X|\Phi) = p(X|\hat{\Gamma}^*, \hat{\Theta}) &= \prod_{t=1}^{T} p(\mathbf{x}_t|\hat{\boldsymbol{\gamma}}^*_t, \hat{\Theta}) \\ &= \prod_{t=1}^{T}\prod_{i=1}^{N}\prod_{r=1}^{R} p(x_{itr}|\hat{\theta}_{\mathbf{i}\hat{\gamma}^*_{\mathbf{itr}}}) \end{aligned} \tag{3.6}$$

3.2 Algorithms for network structure search

In the previous sections the assessment of a single network hypothesis with respect to the measured data was discussed. A likelihood model was set up making it possible to match a network to a particular score. Now, candidate networks have to be optimised to identify the best fitting network structures. The problem of learning network structures was shown to be NP-complete for BNs (Chickering, 1996) and represents a difficult problem to solve, also for non BNs, since the number of graph structures to be evaluated increases super exponentially with the number of nodes. Clearly, heuristic approaches are needed here, two of which are presented in this work. The first is an adaptation of a GA that evolves a population of candidate networks according to their likelihood score, such that an optimal network population is reached. The second is a Metropolis Hastings MCMC approach that samples the space of possible network structures such that structures with a high likelihood are preferred over structures with a low score during the sampling process.

3.2.1 Utilising a genetic algorithm for the optimisation of the network structure

For the description in this section, the fitness score is defined as the Bayesian Information Criterion (BIC) (Schwarz, 1978), derived from the likelihood and model complexity:

$$BIC = -2\,log(p(X|\Phi)) + K\,log(n)$$

where K is the number of edges in Φ and n is the number of data points in X. The BIC score was chosen because it penalises highly complex models, i.e. models with many edges. The number of edges should in general be rather small to prevent model overfitting and prefer sparse network structures. Note that for optimisation according to the BIC score, the fitness has to be minimised rather than maximised, as it would be the case for the raw likelihood or posterior probability as optimisation criterion.

A population $\mathcal{P} = \{\Phi_p : p \in 1 \dots P\}$ of P networks is initialised randomly as start population for the algorithm. Two parameters q and m with with $q, m \in [0, 1]$ have to be given. The GA includes three iteratively performed steps: selection, cross over and mutation. In the selection step, a number of $\lfloor (1-q) \cdot P \rfloor$ individuals is chosen with probability proportional to their fitness. It is required that the BICs of selected networks are smaller than a given quantile of the BICs of all individuals in the population (here, the median is used). This mimics a simple greedy search, but leaves the possibility for selecting suboptimal moves, too. The selected individuals are added to the next generation population $\mathcal{P}'$.

During the cross over $\lfloor \frac{q \cdot P}{2} \rfloor$ random pairs are chosen from $\mathcal{P}$, again proportional to each individuals' fitness. To perform crossing over of two networks, each network adjacency matrix is represented as a vector (simply attaching all columns to each other) and two-point cross over is performed for these vectors. The modified individuals are added to $\mathcal{P}'$ if their BICs are smaller than the given BIC quantile for all individuals in $\mathcal{P}'$. In case that after cross over the size of the modified population $\mathcal{P}'$ is smaller than $\mathcal{P}$, as many random individuals are added from $\mathcal{P}$ to $\mathcal{P}'$, such that the population size stays constant.

Finally, in the mutation step $\lfloor m \cdot P \rfloor$ networks are chosen from the new population $\mathcal{P}'$. For each selected network a random edge is drawn and its type is changed randomly to one of the remaining types. As an example, given an edge $\phi_{ij} = -1$, it can be either changed to $\phi'_{ij} = 1$ or $\phi'_{ij} = 0$. Mutations are allowed if the fitness of the individual improves by introducing the mutation.

These three steps are repeated until a predetermined number of iterations (usually 1000) has been run or the given quantile of all BICs in the population does not change for a preset number of times in a row (usually 10). At the end of the GA the population of candidate networks is combined into a final network by including each edge that occurs in more than a particular fraction of all networks in the population (usually 50% if not stated explicitly).

3.2.2 inhibMCMC: An extension to the Markov Chain Monte Carlo structure sampler

As an alternative to the GA an MCMC structure learning approach to sample the space of possible networks is presented. The purpose of the GA was optimisation of the network structure population. In contrast, MCMC is used to sample from the posterior distribution of network structures to describe the whole network sampling space. The sampler is based on a previous approach names Markov Chain Monte Carlo Model Composition (MC^3, Madigan et al. (1995); Werhli and Husmeier (2007)). Because two edge types are allowed (activation and inhibition), the MCMC sampler has to be adapted in the following way. Adding an edge is replaced by two moves, one for adding an activation and one for adding an inhibition. Another two moves for switching the edge type from activation to inhibition and vice versa and one for simultaneously reversing and changing the type of an edge are introduced, leaving in total six move types: *add activation*, *add inhibition*, *delete*, *revert*, *switch type* and *revswitch*. Inclusion of the novel move types is necessary to ensure that any edge can be changed to any other edge (w.r.t. to type and direction) in exactly one step. Using these move operations, for any given network structure all other structures can be constructed in a finite series of moves. Consider table 3.1 for an illustration of the edge transitions.

	$\rightarrow$	$\dashv$	$\leftarrow$	$\vdash$	$\emptyset$
$\rightarrow$	$-$	st	rev	rst	del
$\dashv$	st	$-$	rst	rev	del
$\leftarrow$	rev	rst	$-$	st	del
$\vdash$	rst	rev	st	$-$	del
$\emptyset$	$addA$	$addI$	$addA$	$addI$	$-$

Table 3.1 – *Possible edge transitions and the corresponding move to perform the transition. addA: Add activation, addI: Add inhibition, del: delete, rev: reverse, st: switch type, rst: reverse and switch type; $\rightarrow$: activation, $\dashv$: inhibition, $\emptyset$: no edge*

Now the essential relationships for the MCMC sampling procedure are repeated (see Werhli and Husmeier (2007)). The index p now denotes any iteration during the MCMC sampling. The proposal probability of any network Φ_{p+1} that differs from a network Φ_p by only one edge is:

$$Q(\Phi_{p+1}|\Phi_p) = \begin{cases} \frac{1}{|\mathcal{N}(\Phi_p)|}, & \text{if } \Phi_{p+1} \in \mathcal{N}(\Phi_p) \\ 0, & \text{if } \Phi_{p+1} \notin \mathcal{N}(\Phi_p) \end{cases}, \tag{3.7}$$

where $\mathcal{N}(\Phi_p)$ is the neighbourhood of a network Φ_p, i.e. all network structures that can be reached by a single edge operation, when starting at the structure Φ_p. An edge operation (or move) is accepted with acceptance probability

$$\mathcal{A}(\Phi_{p+1}|\Phi_p) = min\{1, R(\Phi_{p+1}|\Phi_p)\}, \tag{3.8}$$

$$R(\Phi_{p+1}|\Phi_p) = \frac{P(\Phi_{p+1}|X)}{P(\Phi_p|X)} \cdot \frac{Q(\Phi_p|\Phi_{p+1})}{Q(\Phi_{p+1}|\Phi_p)}, \tag{3.9}$$

and the posterior distribution $P(\Phi_p|X)$ is defined as:

$$\begin{aligned} P(\Phi_p|X) &= \frac{P(X|\Phi_p)P(\Phi_p)}{P(X)} \\ &\propto P(X|\Phi_p)P(\Phi_p). \end{aligned} \tag{3.10}$$

Since $P(X)$ is a constant normalising factor, describing the probability of the measured data, it can be neglected for model comparison purposes. $P(\Phi_p)$ represents the prior probability distribution for a network structure Φ_p, which is described in section 3.3.

To determine the neighbourhood $\mathcal{N}(\Phi_p)$ of a network, let the node set V and network adjacency matrix Φ be defined as above. There are three cases to be considered to determine the cardinality of the neighbourhood of a network Φ_p:

addactivation/addinhibition $|\mathcal{N}(\Phi_p)| := |\{\phi_{ij} : \phi_{ij} = 0;\ i,j \in 1 \dots N, i \neq j\}|$ (the number of node pairs that is not connected by an edge, where self-activations/inhibitions are not considered here)

deletion/switchtype $|\mathcal{N}(\Phi_p)| := |\{\phi_{ij} : \phi_{ij} \neq 0;\ i,j \in 1 \dots N\}|$ (the number of node pairs that are connected by an edge)

reverse/revswitch $|\mathcal{N}(\Phi_p)| := |\{\phi_{ij} : \phi_{ij} \neq 0 \wedge \phi_{ji} = 0;\ i,j \in 1 \dots N\}|$ (the number of node pairs that are connected by an edge, and where the reverse edge is not already present)

Depending on the type of the move, the corresponding proposal probabilities Q can be calculated.

3.3 Inclusion of prior knowledge for structure learning

Structure learning of biological networks is a difficult problem and relies either on heuristic approaches to optimise the target structure or on sampling based approaches to get a stochastic description of the target distribution of network structures and corresponding posterior probability distribution. Often the problem of unambiguous identifiability is not solvable using data driven network reconstruction only, so further ideas to approximate the optimal solution are needed. One frequently used way is to utilise external knowledge on the network structure itself. For example, knowledge about confidences of edges in the network can be used to make the inclusion of edges more likely whenever it is found frequently in external knowledge bases. Also general properties of the graph structure, such as scale-free characteristics can be used. This section deals with these two ways of how prior knowledge can be formulated as statistical models and plugged into the network inference method *DDEPN*. The first subsection describes the 'Laplace prior', which assesses the probability of an edge by comparison to a reference network. The second method is based on the scale-free property of biological networks and introduced in the second subsection 'Scale-free prior'.

3.3.1 Defining prior weights for individual edges by the laplace prior model

Based on the structure prior of Fröhlich et al. (2008a) a prior model is proposed that also incorporates different types of edges and a more fine-grained control of the prior probabilities. Networks are encoded as before. A matrix $B = N \times N \rightarrow [-1, 1]$ is needed, containing prior confidences for each edge, which can be obtained in various ways. Usually, confidences are derived from public network structure databases, like the examples described in

section 2.2. Here, one example is given how to derive B using the KEGG database (Kanehisa et al., 2008). The approach is similar to the one described in Werhli and Husmeier (2007), but preserves the information on the type of the edges. The steps described in this paragraph can be used as guidance to incorporate knowledge from other pathway databases in a similar way in order to derive prior edge confidences.

First, the signalling and disease related networks have to be downloaded from KEGG in KGML format (can be done using the R-package *KEGGgraph*, for example) and converted to adjacency matrices. The number of occurrences of each node pair v_i and v_j in all pathways is counted and recorded in a matrix $\mathcal{M} = N \times N \rightarrow \mathbb{N}$. Further, it is counted how often each node pair is connected via an activation or inhibition edge in all reference networks and the corresponding numbers are recorded in two matrices $\mathcal{M}_{act}$ and $\mathcal{M}_{inh}$, both with the same dimensions as $\mathcal{M}$. For pairs of nodes that do not occur in any reference network (i.e. $\mathcal{M}_{ij}$ is 0), the confidence score is set to 0. The prior confidence matrix B is thus defined as:

$$B = \begin{cases} \frac{\mathcal{M}_{act}}{\mathcal{M}} - \frac{\mathcal{M}_{inh}}{\mathcal{M}}, & if\ \mathcal{M} > 0 \\ 0 & else, \end{cases}$$

assuming that the type of each edge is consistent in all reference networks. This leaves positive confidences for activation edges and negative confidences for inhibiting edges. The larger the absolute value of the confidence score, the stronger is the belief in the presence of this edge.

No matter how B was derived, to calculate the prior belief for a network structure Φ all edge probabilities are assumed to be independent:

$$P(\Phi|B, \lambda, \gamma) = \prod_{i,j} P(\phi_{ij}|b_{ij}, \lambda, \gamma), i, j \in \{1 \cdots N\} \tag{3.11}$$

Now the difference between an edge in the inferred network Φ and the prior confidence in B is calculated, where a weight exponent $\gamma \in \mathbb{R}^+$ is included to obtain the weighted difference term:

$$\Delta_{ij} = |\phi_{ij} - b_{ij}|^\gamma, \tag{3.12}$$

The prior belief for an edge in the network is then modelled as

$$P(\phi_{ij}|b_{ij}, \lambda, \gamma) = \frac{1}{2\lambda} e^{\frac{-\Delta_{ij}}{\lambda}}, \tag{3.13}$$

which penalises large differences from the network structure Φ to the prior belief B.

Now upper and lower bounds for the prior function are derived, in the general case for two edge types. Let $\lambda, \gamma \in \mathbb{R}^+$. If the edge type information is used, all differences Δ_{ij} lie in the interval $\Delta_{ij} \in [0, 2^\gamma]$, because $\phi_{ij} \in \{0, 1, -1\}$ and $b_{ij} \in [-1, 1]$. Without edge type information, the signs in both Φ and B are ignored, leading to $\Delta_{ij} \in [0, 1^\gamma]$, because $\phi_{ij} \in \{0, 1\}$ and $b_{ij} \in [0, 1]$. Because the bounds for $P(\Phi)$ will not change in either case, only the case for including edge type information is shown in the following.

For the moment, let $\gamma = 1$ and consider the limits of the exponential term in equation 3.13:

$$\lambda \to \infty \quad \Rightarrow \begin{cases} e^{\frac{-\Delta_{ij}}{\lambda}} \to 1 \text{ if } \Delta_{ij} = 0 \\ e^{\frac{-\Delta_{ij}}{\lambda}} \to 1 \text{ if } \Delta_{ij} > 0 \end{cases}$$

$$\lambda \to 0 \quad \Rightarrow \begin{cases} e^{\frac{-\Delta_{ij}}{\lambda}} \to 1 \text{ if } \Delta_{ij} = 0 \\ e^{\frac{-\Delta_{ij}}{\lambda}} \to 0 \text{ if } \Delta_{ij} > 0 \end{cases}$$

This means that

$$0 \leq P(\phi_{ij}|b_{ij}, \lambda, \gamma) \leq \frac{1}{2\lambda} \forall \lambda \in \mathbb{R}^+, \gamma = 1. \tag{3.14}$$

Since $\Delta_{ij} \geq 0$, $\forall \gamma \in \mathbb{R}^+$, the bounds are valid for $\gamma \in \mathbb{R}^+$, too. Figure 3.2 shows on the left side the prior curve for equation 3.13 when $\lambda \in \{0.05, 0.1, 1\}$ and $\gamma = 1$. As it can be seen there, with increasing λ the (unnormalised) prior probability curve flattens out, giving unbiased probabilities for each value of Δ_{ij}. The maximum value is bounded by $P(\phi_{ij}) = \frac{1}{2\lambda}$. On the right side of figure 3.2, $\lambda = 0.01$ and with increasing $\gamma \in \{0.5, 1, 5, 15, 50\}$ a broader prior probability plateau at the upper bound for small differences Δ_{ij} is observed, suggesting that γ can be used to control how strong small differences from inferred edges to the prior confidence of an edge should be penalised. The parameter γ should be used such that the 'drop' in the prior occurs at a value for Δ_{ij} that reflects the variance in the confidences that are still regarded as strong evidence for the presence of an edge. This will lead to high prior weights for edges with absolute confidence values not equal to 1, and additionally leave a strong penalisation of large differences. However, this threshold has to be defined in advance depending on the reference that is used.

The prior parameter λ should be adjusted in a way that it exceeds the changes introduced by the likelihood, if strong bias towards prior knowledge is desired during inference. The adjustment differs for the GA and inhibMCMC, because for the GA the absolute difference of the posterior of a proposal network to the predefined quantile of the network posterior probabilities in the

population is influencing the decision whether to accept the proposed network or not, while in the inhibMCMC approach, this is done via the posterior ratios. So for inhibMCMC, one could inspect the likelihood and prior ratios for various settings of λ and choose λ in a way that both ratios are approximately equal. To do this, transform equation 3.13 to log scale:

$$log(P(\phi_{ij}|b_{ij}, \lambda, \gamma)) = -log(2) - log(\lambda) - \frac{\Delta_{ij}}{\lambda}$$

Now consider the prior and likelihood ratios on log scale, i.e. the differences of the log priors and log likelihoods. To make the prior capable of having substantial influence on the inference, the log prior differences should be on similar scale as the log likelihood differences. For instance, if the log likelihood differences are on the scale of 10^3, set $\lambda = 10^{-3}, \gamma = 1$, for instance, such that $\frac{\Delta_{ij}}{\lambda}$ will be in the range of the thousands. The first part of the prior $(-log(2) - log(\lambda))$ only contributes little to the total prior. This means, that the prior influence is controlled over the second part, which is zero for no difference to the prior and and can become very large for differences > 0. So mainly edge mismatches between the reference and inferred net will guide the structure search, and the strength of the influence can be controlled using different settings of λ.

For the GA, adjustment of the prior hyperparameters is different in two ways. The first difference is, that for adjusting the prior hyperparameters one has to inspect the difference terms between the scores of two consecutive iterations in stead of the ratios. An operation on a network individual in an iteration of the GA is accepted whenever the posterior of a modified network exceeds the given quantile of all posteriors of the current population. Hence, the second difference is, that acceptance or rejection of a proposal depends on a whole summary of posterior probabilities. The observed differences of such a summary will be smaller on average than the differences observed for single networks. Therefore, to obtain equal influence of the prior strength in the GA compared to inhibMCMC, λ can be set to a larger value than in inhibMCMC, again estimated using the observed differences in the likelihoods and priors during the GA.

3.3.2 Modelling the network's degree-distribution with the scale-free prior model

A different way of specifying a prior model was introduced in Kamimura and Shimodaira (2005). It is assumed that the networks have a scale-free architecture and that the degree of a node follows a power-law $P(d) \propto d^{-\gamma}$, where d is the number of edges adjacent to a node. For any graph structure Φ with fixed

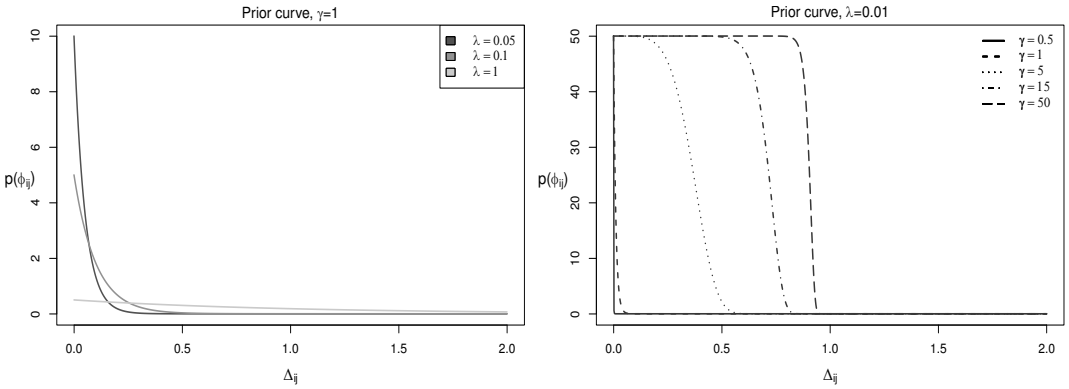

Figure 3.2 – *Unnormalised prior densities, depending on difference Δ_{ij}. Left: γ is constant, for increasing λ a 'flattening' of the prior curve can be observed. For small λ, small differences to the reference retrieve higher weight than large differences. For large λ, all differences are weighted approximately equal. Right: λ is fixed, when γ increases, a plateau at the upper bound $\frac{1}{2\lambda}$ can be seen. This means that small differences to the reference are not penalised as strong as for small γ, leaving the control that up to some deviance from the reference a high prior weight is retained.*

number of nodes N a prior probability can be calculated as follows. First, assign a probability P_i to each node $i \in 1 \dots N$:

$$P_i = \frac{i^{-\mu}}{\Sigma_{j=1}^{N} j^{-\mu}} \approx \frac{1-\mu}{N^{1-\mu}} i^{-\mu},$$

This probability decreases when i gets large, and all P_i sum up to 1, i.e. $\sum_{i \in 1 \dots N} P_i = 1$. μ is in the range $0 < \mu < 1$ and defined as $\mu = \frac{1}{\gamma - 1}$, $\gamma \in [2; \infty[$. The purpose of this probability is to describe a decreasing probability of choosing a node as interaction partner for some other node. Choosing interaction partners subsequently will be rather likely for the first few partners, but less likely when additional partners should be selected. In random graphs this probability would be uniform. Therefore, the probability of selecting additional interaction partners would be equal to the probability of selecting the first interaction partner.

Assuming independent node selection proportional to P_i, each two vertices i and j are selected in one unit time duration. To describe this time unit, an additional parameter K is needed, which can be seen as a chemical potential-like parameter, describing a rate of vertice connections. It controls the mean number of edges and leads to an increasing number of edges when K is increased. Let a node pair be selected with probability $P_i \cdot P_j$ and pair selection be performed for $N \cdot K$ 'times'. The probability of two nodes *not* being connected is then defined as

$$\overline{P_{ij}} = (1 - 2P_iP_j) \simeq e^{-2NKP_iP_j}$$

and the probability of a pair being connected as

$$P_{ij} = 1 - e^{-2NKP_iP_j}.$$

The probability of any structure $\Phi_\sigma = (V, E)$ of node set V, edge set E and a permutation $\sigma = \{\sigma_1, \ldots, \sigma_N\}$ of all nodes in Φ is then

$$P(\Phi_\sigma) = \prod_{\{v_i,v_j\}\in E} (1 - e^{-2NKP_iP_j)} \prod_{\{v_i,v_j\}\notin E} (e^{-2NKP_iP_j}),$$

because each edge is selected independently.

A number of permutations $\sigma = \{\sigma_b : b \in 1 \ldots \mathcal{B}\}$ is generated, resulting in one graph Φ_{σ_b} for each permutation. The final probability of Φ is averaged over the prior probabilities of all permutation networks:

$$P(\Phi) = \frac{1}{\mathcal{B}} \sum_{\sigma_b} P(\Phi_{\sigma_b})$$

A detailed description of the model can be found in Kamimura and Shimodaira (2005) and Lee et al. (2005). The scale-free prior can be used in cases where no information on edge confidences is available. During inference with the scale-free prior model, sparse network structures will be preferred, because high node degrees are penalised by the prior model.

3.4 Analysis of inference results of *DDEPN* using statistical testing procedures

3.4.1 Generating consensus networks from inhibMCMC and GA structure search results

The first question to be answered after inference is, of course, what is the reconstructed network. Since *DDEPN* uses sampling based approaches and results in either a population of optimised networks (for the GA) or in a series of sampled network structures (for inhibMCMC), additional steps have to be performed to summarise these networks into a single final network. The easiest way is to assign a threshold th defining the proportion of networks in the population or sample sequence, that must contain a particular edge, in order to include it. For the GA, usually a single reconstruction is run and summarisation using the inclusion threshold is trivial. However, for inhibMCMC, to assess convergence, multiple independent MCMC runs are performed and

have to be merged for a summarisation. A short description of the consensus network generation for inhibMCMC is given below.

Consider L independent sampling runs and a fixed inclusion threshold th. For each of the sampling runs, generate a network $\Phi^l_{th},\ l \in \{1 \dots L\}, th \in [0,1]$ by including all edges that occur in more than $(th * 100)\%$ of the sampled networks. To generate the consensus network, perform a simple majority vote, i.e. count all activations, inhibitions and missing edges across the L runs and select the edge type having the maximum number. In case of ties, extract all likelihoods of the 'tie networks', build the average of the likelihoods for each edge type in the tie and select the type with the highest likelihood. This gives a summarisation of the L sampling runs into a single network structure, referred to as the consensus network of an inhibMCMC sampling at a given threshold th.

The problem for both techniques to find consensus networks lies in the selection of the inclusion threshold th. A low threshold setting will lead to very dense network structures, containing many false positive hits, since support for the edges from the data is weak. Choosing a high threshold will lead to sparse networks, but also exclude potential true edges. In case of the GA, there is no clear solution to this problem, and a threshold has to be optimised by assessing the resulting networks. For inhibMCMC, in the following section an approach is presented for determining edges based on standard statistical testing techniques which aims to overcome the threshold selection problem.

3.4.2 Determining edge types

A significance testing procedure is described in this section that is used for deciding on the type of the edge between two nodes. Consider L independent sampling runs in inhibMCMC, each with a number of it iterations and burn-in phase of bi iterations. The output consists of a population of $P = it - bi$ network structures $\Phi^p, p \in \{1, \dots, P\}$ for each of the L runs. Each network consists of at most $N \times N$ edges, where self edges and edges pointing towards the stimuli are not considered here (but in principle are possible). For all edges, the number of activations and inhibitions are counted over all runs, and divided by the sum of all activations and inhibitions, to describe the proportion of activations or inhibitions in the total number of inferred edges, respectively. These edge confidences are defined as

$$\begin{aligned} \mathbf{c}^+_{ij} &:= \{c^+_{ijl}\}, l \in \{1, \dots, L\} \\ \mathbf{c}^-_{ij} &:= \{c^-_{ijl}\}, l \in \{1, \dots, L\}, \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} c^+_{ijl} &= e^+_{ijl}/(e^+_{ijl} + e^-_{ijl}), \\ c^-_{ijl} &= e^-_{ijl}/(e^+_{ijl} + e^-_{ijl}), \\ e^+_{ijl} &= |\Phi^p_{ijl} = 1|,\ \forall p \in \{1, \ldots, P\}, \\ e^-_{ijl} &= |\Phi^p_{ijl} = -1|,\ \forall p \in \{1, \ldots, P\}. \end{aligned} \tag{3.16}$$

P is the number of non burn-in network structures, l is the sampling run and $i, j \in \{1, \ldots, N\}$ are the indices of the source and destination nodes in the network. e^+_{ijl} and e^-_{ijl} count the number of activation and inhibition edges in sampling run l, respectively, and c^+_{ijl} and c^-_{ijl} are the normalised confidences of seeing an edge with the appropriate type between nodes i and j. Thus, $\mathbf{c}^+_{ij}$ and $\mathbf{c}^-_{ij}$ are two random vectors of length L collecting the edge confidences over all sampling runs. To derive which kind of edge is present between nodes i and j, it is assumed that elements from the confidence vectors for each edge are drawn independently and identically distributed. Further, if no edge is present between nodes i and j, it is assumed that no difference in the edge confidences is observed, while an activation edge between i and j would lead to a significant increase in activation confidences and inhibition would lead to a significant increase in inhibition confidences. Thus, standard statistical testing can be performed to test the null hypotheses

$$\mathcal{H}_0 : \mathbf{c}^+_{ij} = \mathbf{c}^-_{ij}.$$

A two-sample wilcoxon rank sum test is conducted for both alternative hypotheses:

$$\mathcal{H}^+_1 : \mathbf{c}^+_{ij} > \mathbf{c}^-_{ij};\ \ \mathcal{H}^-_1 : \mathbf{c}^+_{ij} < \mathbf{c}^-_{ij}.$$

Adjustment for multiple testing has to be done using appropriate approaches. In this work the approach of Benjamini and Yekutieli (2001) is chosen. Testing for edges is done under the assumption that all tests can be performed independently. However, in general it cannot be assumed that the presence of an edge is independent of the presence of all other edges. Benjamini's and Yekutieli's approach controls the false discovery rate under the assumption of dependencies of the test statistics. It is seen as an appropriate way to correct for interdependencies between occurring edges. If $\mathcal{H}_0$ is rejected and $\mathcal{H}^+_1$ is accepted on a significance level α, the edge between nodes i and j is regarded as activation. If $\mathcal{H}_0$ is rejected and $\mathcal{H}^-_1$ accepted on significance level α, the edge is an inhibition and if $\mathcal{H}_0$ is not rejected, no edge is drawn between i and j. See figure 3.3 for an example of deciding on the type of the edges.

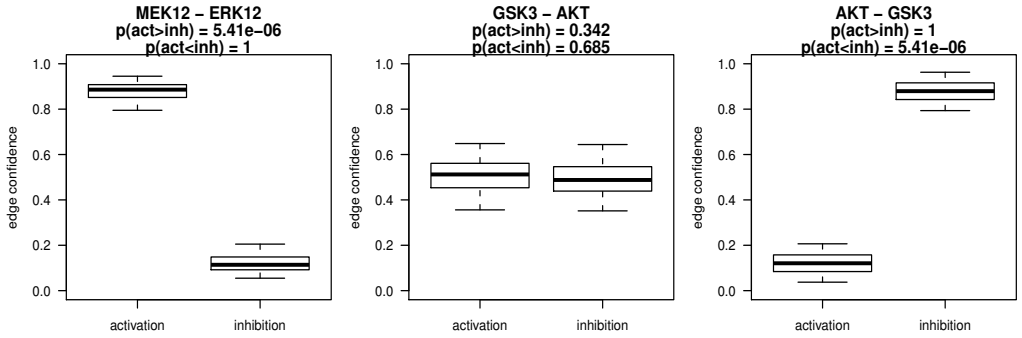

Figure 3.3 – *Example for edge type testing. Left: edge from MEK12 to ERK12. The alternative $\mathcal{H}_1^+$ is accepted, i.e. MEK12 activates ERK12. Middle: The null hypothesis $\mathcal{H}_0$ is not rejected, there is no edge from GSK3 to AKT. Right: edge from AKT to GSK3. The alternative $\mathcal{H}_1^-$ is accepted, i.e. AKT inhibits GSK3. All tests are wilcoxon rank sum tests, significance level $\alpha = 0.05$.*

3.5 Evaluation of the performance of *DDEPN* for simulated data and networks

To assess the performance of *DDEPN* from a theoretical point of view, several tests were performed using artificially constructed networks and data matrices. The purpose of using simulated data and networks was to provide the algorithm with data that depended on an entirely known network structure. A description of the network and data generation process is given in the current section. Afterwards, in section 3.5.1, the performance of the HMM and Viterbi Training for identifying the correct sequence of system states is analysed. Section 3.5.2 describes how well known network structures could be reconstructed by *DDEPN* using data that was generated under the *DDEPN* model assumptions. Section 3.5.3 outlines the *DDEPN* performance in comparison to the two external DBN approaches from section 2. Finally, in section 3.5.4 the impact of prior knowledge on the network structure is analysed.

Generation of simulation data

First, artificial networks had to be sampled, given a number of nodes N and a number of input stimuli. Starting at the inputs, activation edges between each pair of the stimuli and the remaining nodes were drawn proportional to a power law for the probability of choosing subsequent edges (where the order of the interaction partners was random). From each target node that was connected to the stimuli, again random edges were drawn to the remaining unconnected nodes, and the procedure was continued until all nodes were connected. Afterwards, a number of node pairs (20% of the activation edges) were

sampled randomly and connected by inhibition edges. By this, networks were generated that were fully connected and could contain inhibitions, feed forward and feed back loop structures, as they might occur in biological networks, too. Besides, the node degrees followed a power law distribution, as it is required for scale-free network structures.

To generate data depending on a given network, a data matrix X (as defined in chapter 3.1) was constructed. Let parameters $nstim$ be the number of distinct input stimuli and $cstim$ be the number of stimulus combinations. For instance, for stimulation by two receptor ligands, set $nstim = 2$, and for simultaneous stimulation with both ligands, set $cstim = 1$. Each stimulus gives rise to a separate experiment, so for each stimulus a separate state matrix must be constructed using the effect propagation (see section 3.1.1). A state transition matrix for each stimulus was built up by sampling T columns with replacement from each state matrix, while the order of the states was preserved. Each column in the state transition matrix was repeated R times to generate a number of replicates. Finally, all state matrices were attached to get the total state matrix Γ, and all state transition matrices were attached to generate Γ^*. Then, for each time point replicate and node, measurements x_{itr} were sampled from two Gaussian distributions, either from $x_{itr} \sim \mathcal{N}(1200, 400)$ if $\gamma^*_{itr} = 0$ or from $x_{itr} \sim \mathcal{N}(2000, 1000)$ if $\gamma^*_{itr} = 1$, if not stated explicitly. The parameters for the Gaussians (mean and variance) were chosen similar to observed measurements in a real dataset, as were the number of time points $T = 10$ and replicates $R = 9$, representing a typical sized dataset.

3.5.1 Performance of recovering the true state sequence via the HMM

A crucial step in *DDEPN* inference is the recovery of the true state sequence matrix Γ^* by the HMM. To assess the quality of the recovery, networks were constructed as described in the previous section for an increasing number of nodes. The effect propagation was performed for all simulated networks for different numbers of input stimuli. Subsequently, Γ^* matrices were sampled 100 times for each network and stimulus combination, and for each matrix Γ^*, artificial data were generated as described before. Now the HMM state sequence search was performed for all data matrices. The resulting state transition matrices $\hat{\Gamma}^*$ were compared to the corresponding reference Γ^* in terms of sensitivity $SN = (TP/(TP + FN))$ and specificity $SP = (TN/(TN + FP))$, counting the true and false occurrences of the entries in $\hat{\Gamma}^*$. Figure 3.4 depicts the recovery performance and shows constantly high values at averages of around $SN = 0.84$ and $SP = 0.95$ for networks up to 30 nodes. Increasing the number of input stimuli $nstim$ lead to similar values around $SN = 0.83$ and $SP = 0.97$. Hence, given an unknown series of system states, the HMM

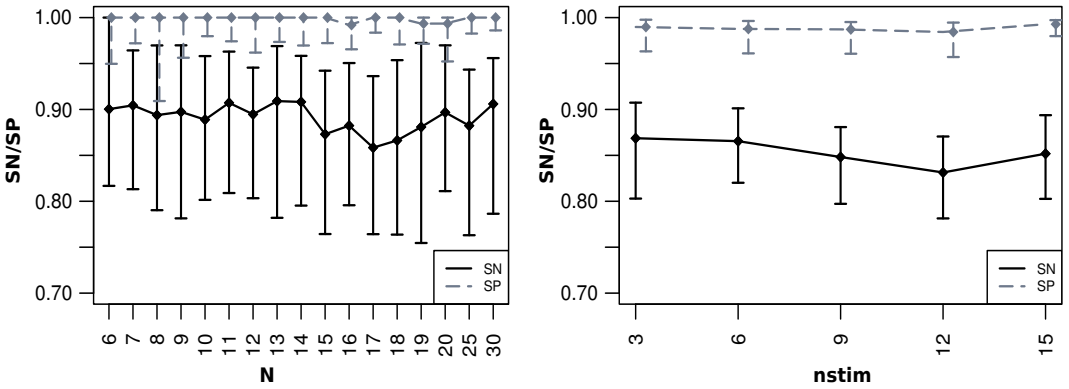

Figure 3.4 – *Performance of state recovery for increasing number of nodes N (left) and number of stimuli nstim (right).*

is able to identify the correct states, even for bigger networks with up to 30 nodes.

3.5.2 Performance of the structure search using a genetic algorithm

Artificial signalling networks and intensity measurements were generated as described in section 3.5. For network comparisons we counted the number of truly inferred edges (TP), truly not inferred edges (TN), erroneously inferred edges (FP) and erroneously not inferred edges (FN). Note that unlike the performance tests in the previous section, now edges in the network are counted, rather than entries in the state matrix.

To test the performance of the structure search using the GA without prior knowledge, 25 artificial networks were sampled for each setting of *nstim* and *cstim*, as shown below. For all networks, system state and data matrices were simulated and the network reconstruction performed. A set of network, state and data matrices is referred to as 'experiment' in this subsection.

Network reconstructions were done for artificial networks of size $N = 10$ with population sizes from $p \in \{100, 250, 500\}$, $q = 0.3$ and $m = 0.8$. No prior knowledge was included in this test. Also increasing numbers of different input stimuli were compared. The parameters were set to $nstim \in \{1, 2\}$ and $cstim \in \{0, 1\}$. The GA was run for 25 sampled networks, each time with the maximum number of generations set to 1000. The edge inclusion threshold for the final network (section 3.2) was varied in $[0, 1]$, and the respective final network for each given threshold was compared to the original net, yielding SN and SP values for the generation of Receiver Operator Characteristic (ROC) curves and Area Under Curve (AUC) values.

Figure 3.5 shows that the reconstruction performance was limited for the case of $nstim = 1, cstim = 0$ and increasing population size $p = 100$, $p = 250$ and $p = 500$ (AUCs $0.57, 0.6, 0.61$), while a slight increase could be found for the bigger population size. A true increase was reached when two distinct stimuli ($nstim = 2$) and one stimulus combination ($cstim = 1$) were included. Here, the AUCs increased to 0.75 and 0.73, respectively. As before, for higher population sizes the AUCs increased (from 0.7 to 0.75). In the lower part of figure 3.5, for a fixed threshold $th = 0.5$, SN and SP were plotted for each simulation test. In the $nstim = 1$ case, SP was high around 0.87, while SN was rather low around 0.17. For $nstim = 2$ SN increased to values around 0.4, while SP improved from 0.78 to 0.83 for growing population sizes. This showed that inclusion of multiple stimuli triggering signalling in the network at different input nodes increased the amount of information that could be used to find the signalling connectivity, and thus resulted in better identification of true edges in the network (apparent in the increasing SN values).

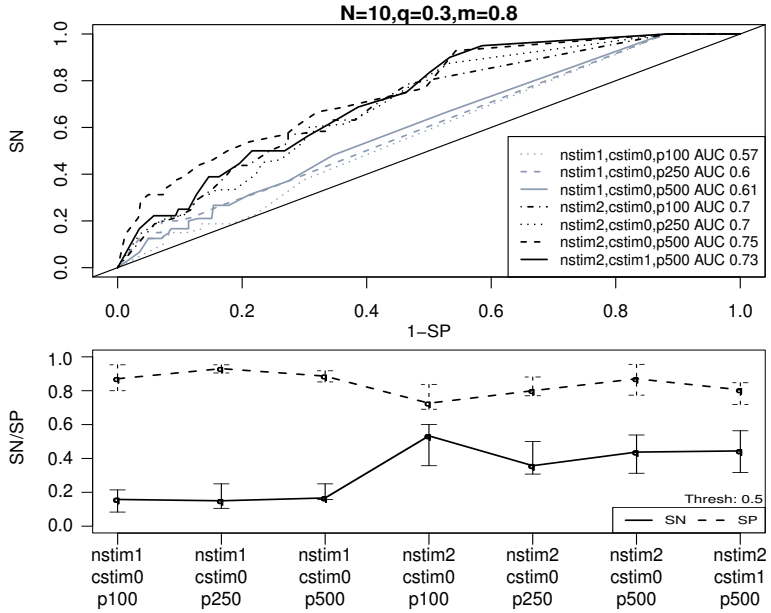

Figure 3.5 – Upper: *ROC curves and AUCs for different settings of input (nstim) and combinatorial stimuli (cstim) and population sizes (p). SN and SP were calculated as average of each 25 network reconstructions with network size of* $N = 10$. Lower: *Example SN and SP plot for* $th = 0.5$ *for all settings. For* $p = 500$, *SP was high at* ~ 0.83, *while SN increased from* ~ 0.17 *to* ~ 0.4. *This shows, that* DDEPN *found edges with strong support from the data with low FP rates. The increase in SN for bigger population sizes shows, that broader sampling of the network search space yielded better inference results.*

3.5.3 Comparison to alternative network inference approaches

To assess whether *DDEPN* performs well with respect to other methods, network inference was conducted for *DDEPN* and the DBN reconstruction approach *G1DBN* of Lébre (2009) and *ebdbNet* of Rau et al. (2010). As in the previous section, network size was chosen as $N = 10$ and inference was run for 25 simulated networks. Each network reconstruction was repeated 100 times and ROCs and AUCs were calculated as before The results are depicted in figure 3.6. For $nstim = 1, cstim = 0$, *DDEPN* performed slightly better than *G1DBN* and *ebdbNet*(AUCs 0.61 for *DDEPN*, 0.58 and 0.55 for *G1DBN* and *ebdbNet*). However, the performance was limited in this case for all methods. Using $nstim = 2$, *DDEPN* clearly outperformed *G1DBN* and *ebdbNet*, for both $cstim = 0$ (AUC=0.75) and $cstim = 1$ (AUC=0.73). This highlighted the ability of *DDEPN* to make use of the additional information gained from multiple perturbations. But the better performance has its price in terms of computation time. On average, a 10 node network with 2 input stimuli was reconstructed in around 7000 seconds using *DDEPN*, while *G1DBN* and *ebdbNet* completed this task in a few seconds. However, the network inferred in *DDEPN* was derived from a whole population of candidate networks that covers larger portions of the network search space than the other two approaches. Calculation was done on a Quad-Core AMD Opteron(tm) 2.7 Ghz machine with 64 Gb memory, on which each 14 cores were used in parallel to optimise the population of networks in the GA.

3.5.4 Assessing the prior influence

The tests from the previous two sections show the overall performance of *DDEPN* without inclusion of any prior knowledge. However, it is also possible to include knowledge on the network structure itself into the modelling approach. Using the laplace prior from section 3.3.1 the aim was to show that the inference could be influenced in a way that on the one hand the result was close to a given reference network and on the other hand allowed to confute the prior, when evidence from the data got strong enough. For all tests, one network was sampled with network size $N = 15$ and data were sampled 10 times for the same network. The reconstruction was applied using both the GA and inhibMCMC with and without prior inclusion. For the GA, the parameters were chosen as follows. Population size was set to $p = 500$, cross over rate $q = 0.3$, mutation rate $m = 0.8$ and maximal number of iterations to 1000. For inhibMCMC, each 50000 iterations were performed with a burn-in phase of 5000 iterations and 10 independent sampling starts. Thus, for each of the 10 independent datasets, a final network was generated from each 10 independent Markov Chains.

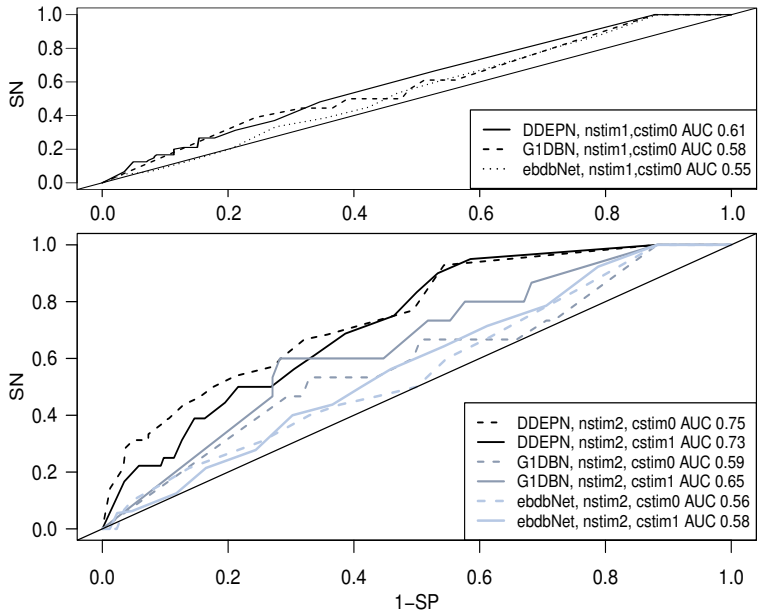

Figure 3.6 – *ROC curves and AUCs for* DDEPN *network reconstruction compared to* G1DBN *and* ebdbNet. Upper: *for* $nstim = 1, cstim = 0$ *a slight improvement of AUCs was observed, and performances were limited for all approaches.* Lower: *for* $nstim = 2, cstim = \{0, 1\}$ *a clear increase in AUC was found for* DDEPN*, showing the improved quality of the network reconstructions.*

The following rationale was applied for the laplace prior tests. First, it was assumed the prior information were true. To ensure this, the original sampled network was used as laplace prior matrix B. Sampling of the network and data were described in section 3.5. Both the prior confidences in B and the inferred edges only take on values $\in \{0, 1, -1\}$, so the absolute differences between both were either 0, 1 or 2. All differences larger than 0 should have been strongly penalised, ensured by setting $\gamma = 1$ which leads to a sharp decrease of the prior density (equation 3.13) for $\Delta_{ij} > 0$. Each mismatch in an inferred edge to the prior was thus given a weight close to 0 (see figure 3.2).

Figure 3.7 shows the results of the 10 inferences with inhibMCMC. On the left side, AUCs are depicted that show an increase for decreasing λ. For $\lambda = 0.005$ and $\lambda = 0.001$ the reference networks could be successfully reconstructed (AUCs around 1). On the right side, the likelihood and prior ratios are visualised for decreasing λ. It was suggested in section 3.3.1 to inspect the likelihood ratios and adjust λ in a way that the the quotient $\frac{-\Delta_{ij}}{\lambda}$ and thus the prior ratios are on a similar scale, which was done in this test. In the plot, for $\lambda = 0.001$ the prior ratios varied over a much broader range than the likelihood ratios, which lead to inferred networks that were nearly identical to the prior network, as it can be seen in the AUC of around 1. For increasing

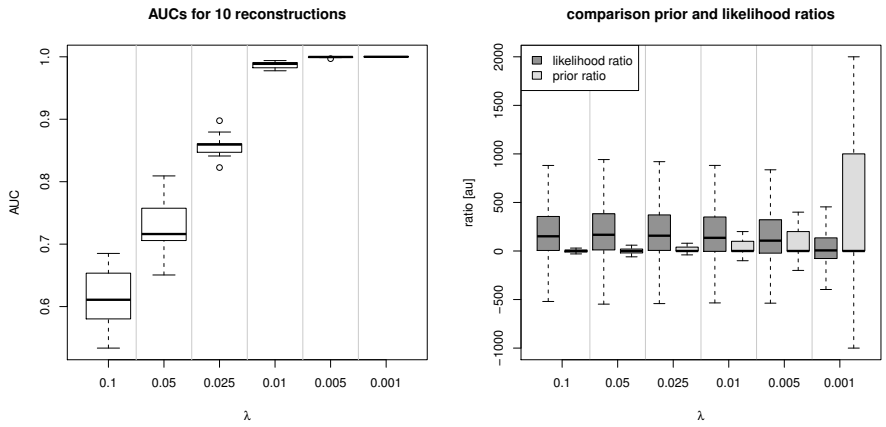

Figure 3.7 – *Diagnostics for inhibMCMC of a randomly sampled network (N=15), 50000 iterations, burn-in 5000, $\gamma = 1$, varying λ. The sampled network was used as prior confidence, i.e. the prior knowledge was 'perfect' in this test. Left: The smaller λ, the stronger the prior influence was and the closer the inferred networks were to the prior (reflected in increasing AUCs). Right: Comparison of Likelihood and Prior ratios, depending on λ. λ should be chosen such that the prior and likelihood ratios vary in a comparable range. For instance, based on the plot, set $\lambda = 0.005$.*

λ the likelihood ratios showed a larger variance than the prior ratios, which lead to decreasing AUCs and more variable inferred networks in turn. Thus, the setting of the prior parameters determines how robust the reconstruction of the networks is. The settings have to be carefully adjusted to preserve robustness, but leave enough variance to gain additional knowledge, represented in the data, too.

Turning to the GA, for one sampled network structure the inference was run 10 times for different initial network populations and the results are shown in figure 3.8. On the left hand side of the figure the AUC distributions are shown. When using the BIC score optimisation for the reconstruction, it is apparent that the performance of the GA is weak, with AUCs around 0.5, emphasising the need for the inclusion of prior knowledge to produce reliable results for larger networks. When using prior knowledge in the laplace prior, for decreasing λ, an increase in the reconstruction performance was observed as in the case for inhibMCMC. The improvement in reconstruction performance can be controlled, similar to the case for inhibMCMC, using smaller settings for λ. Using $\lambda \leq 0.01$ gave comparable reconstruction results with AUCs above 0.95, but in contrast to inhibMCMC, the reference network could not be inferred entirely for even smaller settings for λ. The GA seems to reach a local optimum, but does not find the true network, even for the test situation where the prior strength is increased subsequently.

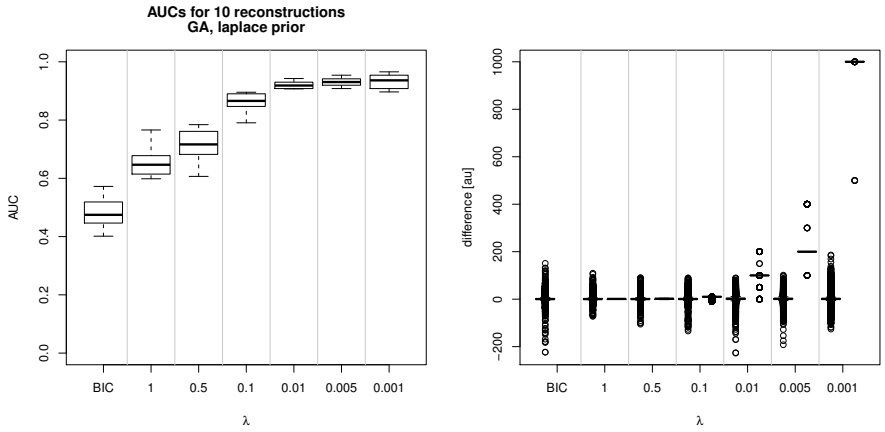

Figure 3.8 – *Results for GA reconstruction for one sampled network (N=15), population size p=500, number of iterations 1000, cross over/selection rate q=0.3, mutation rate m=0.8, $\gamma = 1$. As in figure 2, the sampled network was used as 'perfect prior knowledge'. Left: AUC values without prior (column BIC) and for various settings of λ. When λ was decreased, the AUCs increased. However, unlike the inhibMCMC example, AUCs did not approach a value of 1, giving evidence that the GA converges to a local optimum. AUCs for BIC score optimisation were low, emphasising the need for prior knowledge inclusion for larger networks. Right: Likelihood and prior differences. Since most of the observed prior differences were zero, only the non-zero values are shown. For each setting of λ, the left box corresponds to the observed distribution of likelihood differences, the right box to the prior differences. In the BIC column, only the likelihood difference distribution is shown, since no prior was used in this case.*

On the right hand side of the figure, the likelihood and prior differences are shown. As depicted in figure 3.8, setting $\lambda = 0.1$ lead to prior differences of around 10 and already had a strong influence on the reconstruction performance. So this setting for λ could be used as appropriate setting for λ. Nevertheless, it seems harder to find a proper λ setting for the GA inference, because already small changes in the prior setting lead to large variance in the performance of the GA.

3.6 ERBB signalling network reconstruction for longitudinal protein array data

The development of *DDEPN* was driven by the need to analyse a phosphoproteomic dataset that was generated with the goal to analyse the effects of different ERBB receptor inhibiting drugs onto breast cancer derived cell lines. The following section introduces the experimental setup and data generation

and shows the results of the application of *DDEPN* to infer signalling interactions in a set of ERBB signalling related proteins.

3.6.1 Inference for a phosphoproteomic dataset from ERBB signalling related proteins

To begin with, a description of the data generation in the biological experiment is given. The human breast cancer cell line HCC1954 was cultivated as recommended by ATCC (American Type Culture Collection) and cells were split three times per week. For stimulation experiments, cells were seeded in 6-well plates, cultivated for 24h and serum-starved in phenol-red free medium for additional 24h. EGF (Sigma) and HRG (Biovision) were added to the cells to a final concentration of 5 nM. After times 0, 4, 8, 12, 16, 20, 30, 40, 50 and 60 min, medium was replaced by ice-cold PBS and plates were put on ice. Afterwards, PBS was aspirated and cells were harvested by manual scraping in 40 µL lysis buffer (M-PER (Pierce), Complete Mini, PhosSTOP (Roche)). Cells were lysed for 20 min at 4°C. After centrifugation, total protein concentration was determined using the BCA method (Pierce) and all samples were adjusted to the same protein concentration. Prior to printing, samples were mixed with Tween-20 to a final concentration of 0.05%. Three biological replicates were generated on three different days. The samples were printed in triplicate onto nitrocellulose coated glass slides (Oncyte, Grace-Biolabs) with a contact spotter (2470 Arrayer; Aushon Biosystems) using 180 µm pins. Slides were blocked in 50% Odyssey Blocking Buffer (LI-COR) in PBS containing 5 mM sodium fluoride and 1 mM vanadate. Primary antibodies were diluted 1:300 in antibody diluent with background reducing components (Dako). Alexa 680 labelled secondary antibodies (Molecular Probes) were diluted 1:5000 in PBS (+0.2% NP-40, 0.02% SDS + 0.5% BSA). After drying, arrays were scanned using the Odyssey Infrared Imaging System (LI-COR) and signal intensities were determined with GenePix Pro 5.0 (Molecular Devices). Sample normalisation was done using Fast Green FCF dye (see Luo et al. (2006); Loebke et al. (2007)) to account for different protein concentrations in each spot on the array. Replicate time courses were centred around their common mean to remove systematic shifts in the intensities. 16 antibodies for specific phosphorylation sites were used to obtain signal intensities of phosphorylated protein. A list of the proteins and targeted phosphorylation sites is shown in table 3.2. The antibodies were obtained from the following companies: ERBB4 and GSK3 from Epitomics, ERBB2 from Millipore, MEK1/2 from Sigma, PKCα from Abcam all others from Cell Signalling.

Protein	Phosphosite	Protein	Phosphosite	Protein	Phosphosite
AKT	S473	EGFR	Y1068	ERBB2	Y1112
ERBB3	Y1289	ERBB4	Y1162	ERK1/2	T202,Y204
GSK3	Y279,Y216	MEK1/2	S217,S221	MTOR	S2448
p38	T180,Y182	p70S6K	T389	PDK1	S241
PKCα	S657,Y658	PLCγ	S1248	PRAS	T246
SRC	Y416				

Table 3.2 – *Proteins and phosphorylation sites used in the RPPA analysis.*

3.6.2 Using the genetic algorithm to infer basic signalling interactions in the ERBB network

DDEPN was used to reconstruct a signalling network from the experimental data. The GA was utilised as structure search algorithm first, together with the BIC score optimisation that ensures sparsity of the resulting network. Parameters were chosen as population size $p = 500$, maximum iterations 1000, cross over rate $q = 0.3$ and mutation rate $m = 0.8$. The reconstruction here was performed to show the results of *DDEPN* when no prior knowledge was included. The inferred network is shown in figure 3.9. An edge is shown if it was contained in at least 50% of the networks in the final population ($th = 0.5$), allowing only interactions with strong support from the data. Several signal cascades were seen in the network that were known from the literature. For example, the regulation HRG $\rightarrow$ ERBB1 was inferred. Olayioye et al. (1999) showed that HRG is an activator of the ERBB-Dimers 1/3 and 1/4, which supported this result. Activation of ERBB2 by EGF or HRG could be found in Jones et al. (1999) (EGF/HRG $\rightarrow$ ERBB2/3), which also supported activation of PKCα by HRG through the cascade HRG $\rightarrow$ ERBB2 $\rightarrow$ PKCα, since crosstalk between ERBB2 and PKCα in ERBB2 over-expressing breast cancer cells was reported by Magnifico et al. (2007). The result was further interesting, since the HCC1954 cells over-express ERBB2. Kim et al. (2009) reported activation of p38 by ERBB2 in ERBB2 over-expressing breast cancer cells, reflected in the activation EGF $\rightarrow$ p38. The activations of MEK1/2, ERK1/2 and p70S6K by EGF are key elements in the classical MAPK signalling cascade EGF $\rightarrow$ ERBB1/1 $\rightarrow$ GRB2 $\rightarrow$ SOS1 $\rightarrow$ RAS $\rightarrow$ RAF1 $\rightarrow$ MEK1/2 $\rightarrow$ ERK1/2 $\rightarrow$ p70S6K. EGF $\rightarrow$ ERBB1/1 $\rightarrow$ PLCγ was shown by Kim et al. (1990), which demonstrated the relevance of the activation EGF $\rightarrow$ PLCγ in the inferred network. Further EGF $\rightarrow$ AKT $\dashv$ GSK3α is found in the cascade EGF $\rightarrow$ ERBB $\rightarrow$ GRB2 $\rightarrow$ GAB1 $\rightarrow$ PI3K $\rightarrow$ AKT $\dashv$ GSK3α. More hypothetical interactions included the inferred SRC activation (ERBB3 $\rightarrow$ SRC), interpreted as activation of SRC by ERBB2 (see e.g. Luttrell et al. (1994); Mao et al. (1997); Xian et al. (1997)) through the ERBB2/3 heterodimer. Finally, PDK1 activation by receptor tyrosine kinases was shown by Cohen et al. (1997) in insulin signalling. In the reconstructed net, the activation ERBB3 $\rightarrow$ PDK1

was present, which supported the hypothesis that the cascade ERBB1/3 → PI3K → PIP3 → PDK1 (see Oda et al. (2005); Vanhaesebroeck et al. (1997)) might also play a role in cancer related signal transduction processes.

All of these inferred and literature confirmed interactions had high support by the data (occurrence in more than 75% of all networks in the final population, see edge labels in figure 3.9). These findings showed, that literature knowledge was reproduced well by *DDEPN* and in addition allowed for discussion of the newly inferred interactions.

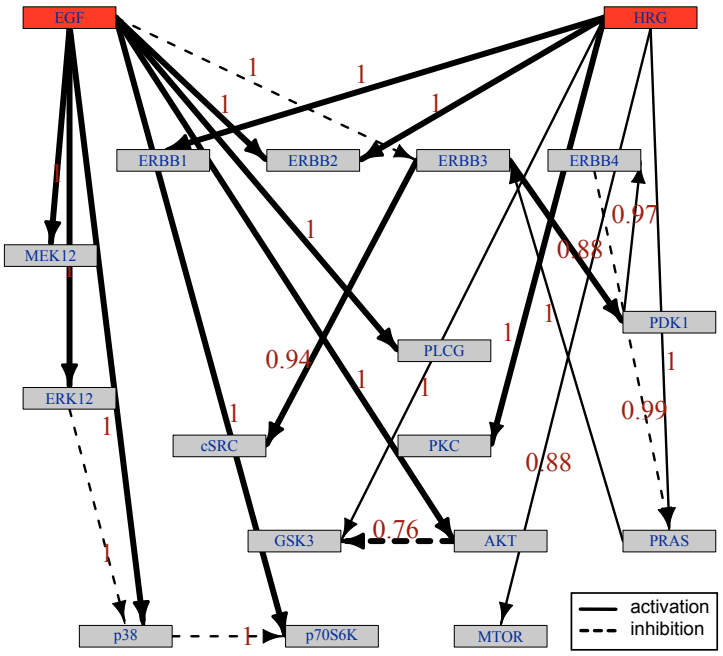

Figure 3.9 – *Network reconstructed from HCC1954 data. Interactions found in the literature are marked as thick lines. Red nodes mark the input stimuli. The numbers at the edges show the proportion of networks in the final GA population, in which the respective edge was contained.*

3.6.3 InhibMCMC inference with prior knowledge resolves correct signalling cascades

To demonstrate how the prior knowledge inclusion improves reconstruction results from real data, the inference was repeated on the HCC1954 dataset using inhibMCMC together with the laplace prior model. Prior edge confidences were generated from KEGG as described in section 3.3.1, and a final reference network was assembled as follows. Edges with prior confidence ≥ 0.1 were

included in the prior network, while the edge type information was preserved. Additionally, several edges were included manually, that were described in current literature resources. The prior ERBB network is shown in figure 3.10.

InhibMCMC inference was applied with 50000 iterations, where the first 25000 iterations were regarded as burn-in and discarded. The following parameters were chosen: $\lambda = 0.0025$, $\gamma = 1$. To assess convergence and ensure robustness of the results, 10 independent inhibMCMC chains were run in parallel, each starting at a randomly sampled initial network structure. A final network was generated by using the testing procedure presented in section 3.4.2 with a significance level $\alpha = 0.001$ and the multiple testing correction approach of Benjamini and Yekutieli (2001). The result is shown in figure 3.11, (A). Black edges correspond to edges that are present in both the prior and inferred network, blue edges are novel edges reconstructed by *DDEPN* and not present in the prior. It can be seen that the prior network was resembled well (mainly black edges) and five new edges were identified in the inferred net (blue edges, in particular EGF → ERK1/2, EGF → p70S6K, EGF → AKT, EGF → PLCG and PRAS → ERBB1).

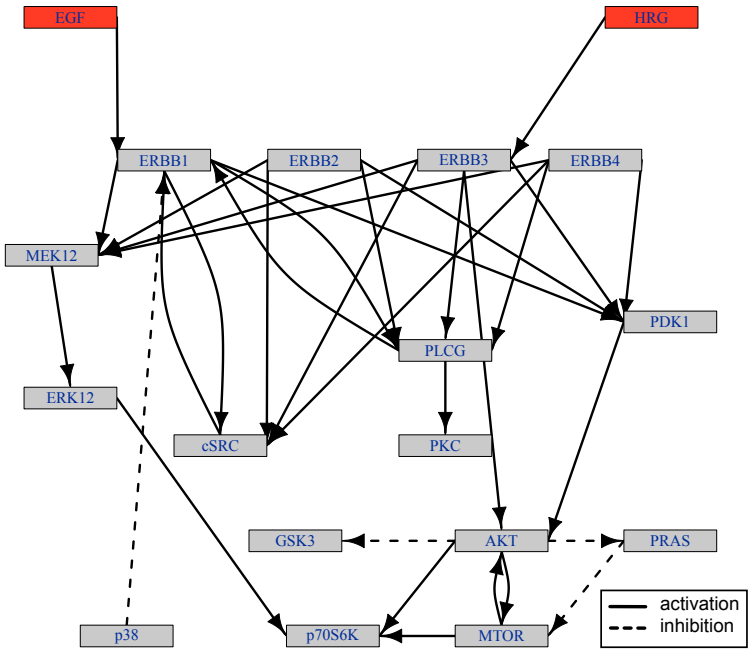

Figure 3.10 – *ERBB prior network assembled from KEGG and by manual curation that was used for the inference.*

Comparing the inferred net in figure 3.11, (A) to the inferred net from the GA run with BIC score optimisation (see figure 3.9) it is apparent that

the inference was improved using the prior knowledge. When looking at the structure of known signalling cascades, for example, the MAPK kinase cascade EGF → ERBB1 → MEK12 → ERK12 → p70S6K or the cascade HRG → ERBB3 → PDK1 → AKT were inferred, which could be expected, because these are major signalling cascades that are ubiquitously present in biological systems. The novel activations (the blue edges) mainly show transitive effects from EGF to downstream nodes of the MAPK and AKT signalling cascades, resembled also by a path from EGF down to the respective target over several intermediate nodes. The exception is the activation of PRAS → ERBB1, for which no path could be seen originating in PRAS and ending in ERBB1.

3.6.4 inhibMCMC with a prior model reveals treatment specific effects on the structure of the ERBB signalling network

Additional experiments were performed for the HCC1954 cell line in which inhibitor drugs against the ERBB receptors were included. Cells were incubated in starving medium containing trastuzumab (10ng/μL) and erlotinib (1 μM) (Roche, Penzberg, Germany). Treatment was performed with each drug alone and as combination 1 h prior to growth factor stimulation. Afterwards, EGF and HRG stimulation (each 5 nM) was performed. Lysates were prepared after 0, 4, 8, 12, 16, 20, 30, 40, 50 and 60 min as for the stimulation experiment. Again, both the single stimulations with EGF and HRG and the combined stimulation with EGF and HRG were done in three separate experiments. Each experiment was performed three times resulting in three biological replicates and spotted in three technical replicates by the spotting robot. Network reconstruction was performed as for the stimulation experiment in section 3.6.3 using inhibMCMC and the laplace prior with 50000 iterations, burn-in of 25000 iterations, $\lambda = 0.0025$ and $\gamma = 1$. The resulting networks are shown in figure 3.11. Unlike the figures from the previous experiments, edges that are contained in the prior as well as the inferred networks are marked in gray (and not in black) for a clearer visualisation of the networks. In general, the overall structure of the signalling network is retained, indicated by mainly gray edges. However, depending on the type of drug used, different effects could be observed. The upper right network (B) shows the case for treatment with erlotinib. In comparison to the prior network, six new edges were observed (blue edges, EGF → p70S6K, EGF → AKT, MEK1/2 ⊣ ERBB2, MTOR → MEK1/2, MTOR → ERK1/2 and cSRC → MEK1/2). The transitive activation of the targets of the MAPK and AKT cascades, p70S6K and AKT itself, remained active as in the case without inhibitory drug (figure 3.11, (A)). The expected inhibition of the ERBB receptors by erlotinib were inferred.

Turning to the lower left part of figure 3.11 (C), the treatment with tras-

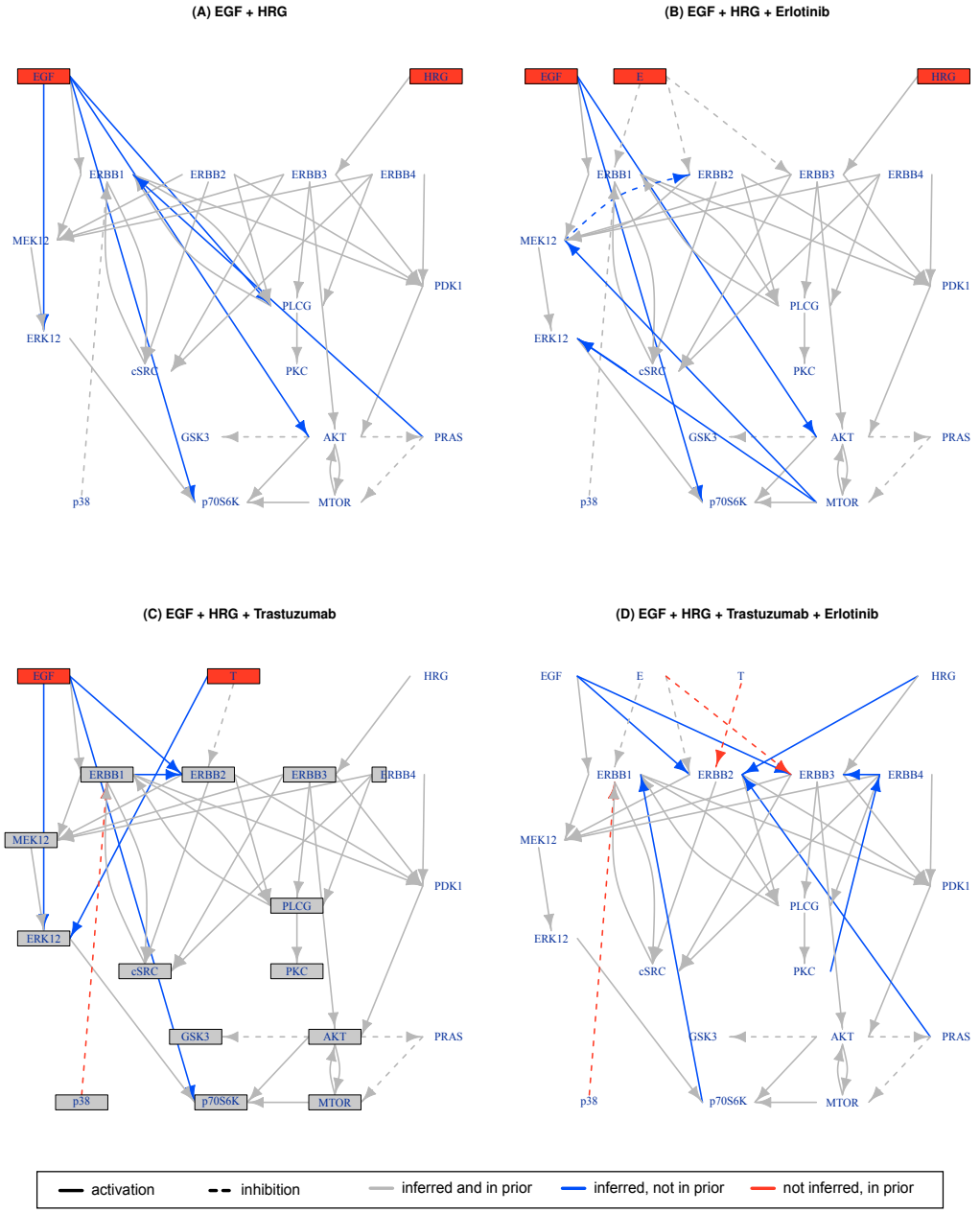

Figure 3.11 – *Inferred networks for different treatments. For all networks, three experiments were conducted and used for the reconstruction: Stimulation with EGF and HRG alone and EGF+HRG stimulation. Additional treatments: (B) Inhibition with erlotinib. (C) Inhibition with Trastuzumab. (D) Combination of Trastuzumab/erlotinib. Edges: gray - present in both inferred and prior network; blue - only in inferred; red - only in prior; dashed - inhibitions. Direct transitive activation of the MAPK downstream target p70S6K is only lost for the treatment with both drugs. Direct and transitive activation of AKT is lost when both drugs and trastuzumab alone are used, suggesting that combined treatment most effectively deactivates downstream signalling of the MAPK and AKT signalling cascades.*

tuzumab is shown. Five additional edges compared to the prior were found (EGF → ERBB2, EGF → ERK1/2, EGF → p70S6K, ERBB1 → ERK1/2 and ERBB1 → ERBB2). One edge could not be reconstructed that was suggested by the prior: p38 ⊣ ERBB1 (marked as red edge). The inhibition of ERBB2 by trastuzumab was correctly inferred. Again, a transitive activation of p70S6K was found originating in the EGF stimulus, but in contrast to the inhibition with erlotinib and the case without inhibition, the transitive AKT activation by EGF is lost.

The last example shows the treatment with both trastuzumab and erlotinib simultaneously (figure 3.11, (D)). Here, seven additional edges were found (EGF → ERBB2, EGF → ERBB3, HRG → ERBB2, ERBB4 → ERBB3, PKC → ERBB4, p70S6K → ERBB1 and PRAS → ERBB2), compared to the prior network and three edges from the prior were ruled out by the inference procedure (E → ERBB3, T → ERBB2). This shows, that the expected inhibitions from erlotinib to ERBB3 and from trastuzumab to ERBB2 were not supported by the data. Finally, the transitive activations from the stimuli to p70S6K and AKT were both not present. In general, the combinatorial treatment shows the strongest effects with regard to inhibition of the downstream components of the MAPK and AKT signalling cascades, pointing to an effective inhibition. See also the discussion for an interpretation of the results shown here (section 4.2.2).

3.7 The CAMDA challenge dataset: a microarray time course measuring the response to survival factor deprivation in endothelial cells

To give an example of the inference of regulatory networks on microarray data and to show that *DDEPN* also is capable of modelling signal transduction on the transcriptional level an analysis for the Critical Assessment of Microarray Data Analysis contest (`http://camda.bioinfo.cipf.es/camda08/` is presented and extended (Bender et al., 2008). First, a gene selection workflow is shown in section 3.7.1 that filters uninformative genes and selects functionally related groups of genes. Afterwards, network reconstruction results using *G1DBN*, *ebdbNet* and the GA and inhbibMCMC of *DDEPN* are described in section 3.7.2.

The microarray dataset was published in Affara et al. (2007). It includes time-course gene expression data generated in human umbilical vein endothelial cells (HUVEC) after survival factor deprivation (SFD). For a pool of 10 individuals of HUVEC, RNA was prepared at time points 0, 0.5, 1.5, 3, 6, 9,

12 and 24h and hybridised to UniSet Human 20K gene chips. Gene expression was measured using CodeLink expression analysis software.

3.7.1 Workflow for identifying functionally relevant gene or protein subsets as input to network inference methods

Current research in genomics or proteomics is evolving in that the number of genes or proteins that are measured in parallel is steadily increasing. On recent microarray platforms, a whole set of known genes in an organism can be measured in parallel. The number of proteins, for which abundance can be quantified in one experiment, is increasing with the advent of novel techniques like protein arrays or mass spectrometry, too. This bears a number of challenges for the analysis of such amount of data. The most important question to ask right in the beginning is, which of the genes or proteins show a response at all in any of the experimental conditions. *DDEPN* was introduced as novel network inference method in the previous sections. Because the algorithm is tailored to rather small networks (number of nodes smaller than 50), a workflow is presented that describes how subsets of genes or proteins can be selected from larger experiments that share a functional relationship. It has to be ensured that gene or protein profiles are not equal for all conditions, in which case no hypotheses on the connectivity of these components could be generated. Thus, the identification of effected components is of primary interest in such experiments. Currently, several methods exists for this purpose. For the identification of differentially expressed genes or proteins from array experiments, one could either use *limma* (Smyth, 2004) or *SAM* Tusher et al. (2001), for instance. Even for longitudinal data, identification of differentially expressed time courses is possible (e.g. Tai and Speed (2006)). These methods are all available as R-packages and thus easily accessible and applicable. The core assumption of all of these methods is that most of the genes do not change in the experiments. From this set of 'background' components, a null-distribution of expression intensities is estimated and genes differentially deviating from this null-distribution can be found using standard statistical testing procedures.

Once a set of interesting, i.e. varying components, is found, one has to ask if the genes somehow 'belong' to each other. That is, do the genes have a functional relation and thus influence each other directly, leading to e.g. correlations in their expression profiles, or are observed correlations only spurious and found by chance? For instance, the KEGG database (see section 2.2) offers gene annotation and visualises this information in pathway maps, but only annotation of about 4000 of the estimated 20000-25000

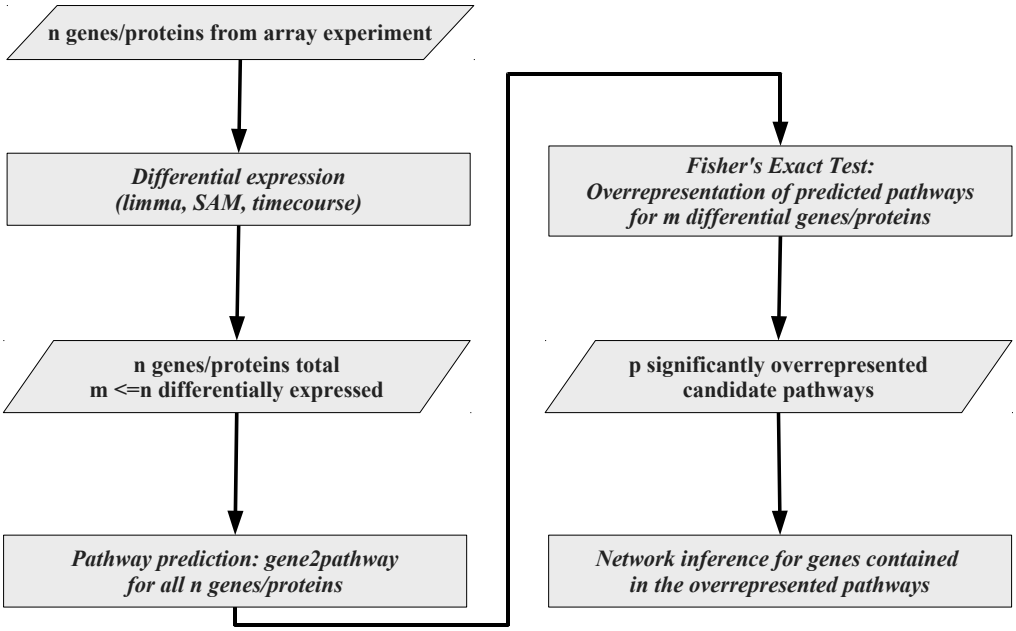

Figure 3.12 – *Workflow for identification of functional related gene/protein sets.*

protein-coding genes is available. Gene ontology (GO, The Gene Ontology Consortium (2004)) offers annotation for most genes, but not all have a known function. Geneset Enrichment Analysis can be used to determine over-represented functions or pathways in gene lists (e.g. Beissbarth and Speed (2004); Al-Shahrour et al. (2004)), but is limited by the availability of gene annotation. An approach of Hahne et al. (2008); Fröhlich et al. (2008b) is able to predict pathway membership of genes based only on the protein domain annotation. The method was described in section 2.3. The InterPro database (Mulder et al., 2008) offers protein-domain annotation for about 19000 genes, so the use of such a tool gives a closer characterisation of the interesting components regarding their function. Figure 3.12 shows the workflow for the identification of functionally related gene/protein subsets from high-throughput data, as it was proposed in Bender et al. (2008). First, the array measurements were analysed using one of the methods for identification of differentially expressed genes/proteins (*limma* was used in the work described here). Second, a prediction of pathway membership of each gene/protein was performed, using the R-package *gene2pathway*, available on CRAN. Note that each gene/protein can be mapped to multiple pathways, making over-representation analysis of pathways possible. Next,

Fisher's exact test was used to identify over-represented pathways in the list of differentially expressed genes/proteins, compared to the list of all genes or proteins. Entities belonging to a significantly over-represented pathway could now be selected for further analysis, including the reverse engineering methods described above.

Candidate genes in the time-course expression data were selected by first normalising the raw expression values using variance stabilisation normalisation (VSN, Huber et al. (2002)) and successively analysing differential gene expression with *limma* (Smyth, 2004). Genes with a normalised intensity in the lower quartile of the observed intensity range in all time-points were excluded from the analysis as they have non-informative expression profiles. Each pair of time points was analysed, and genes showing an FDR (Benjamini and Hochberg, 1995) smaller than 0.001 in at least one of the comparisons were taken as differentially expressed. From the 20265 genes on the array 18310 genes were kept as informative genes after filtering for constant expression profiles and bad quality flags on the arrays. 1002 genes were found differentially expressed after *limma* analysis. The mapping of the microarray's ProbeID to the Entrez-GeneID resulted in 14015 unique genes that could be analysed by *gene2pathway*. These were fed into the KEGG-pathway membership prediction (see section 2.3), in which InterPro domains for 10630 genes were found. 3385 had pathway memberships defined by KEGG, with 268 being differential. Predictions for 4206 genes were made using the domain signatures and 353 of them were differentially expressed. For each of the predicted pathways Fisher's Exact Test was performed to find out whether a particular pathway was significantly over-represented in the sets of differentially expressed genes. This was done once for the genes that were directly annotated in KEGG and additionally for those that were predicted to be a member of the pathway by their domain signature using *gene2pathway*. The results are shown in table 3.3.

pathway	p_1	p_2	K	DS
Cell cycle	0,0031	0,0004	22	30
Metabolism	1	0,0316	96	364
Cell Growth and Death	0,3877	0,0447	26	43
Nucleotide Metabolism	0,3159	0,0562	22	22
Insulin signalling pathway	0,3159	0,2787	2	2
...	...	...	...	...

Table 3.3 – *P-values for pathway over-representation: p_1 for pathway membership defined only by KEGG; p_2 for pathway membership by KEGG and domain signature prediction; K: number of genes found as member of the pathway in KEGG annotation, DS: number of genes assigned to a pathway by KEGG and the domain signature prediction with* gene2pathway.

A significant overrepresentation was found for the pathways *Cell Cycle*, *Metabolism*, *Cell Growth and Death* and *Nucleotide Metabolism* after pathway assignment with *gene2pathway*. As seen in table 3.3, the significance for the pathways is increased when the domain signature prediction is incorporated. It also makes sense to find the pathway *Cell Cycle* and its parent map *Cell Growth and Death* over-represented, since the microarray data originated in an apoptosis study, which is part of *Cell Death and Growth* and closely related to *Cell Cycle*. This suggests, that genes from the *Cell Growth and Death* tier show the highest activity in the time-course. For further investigation and network reconstruction exactly those differentially expressed genes, that were part of the *Cell Cycle* pathway were taken. Since *Metabolism* is a branch that can hardly be distinguished by the use of domain signatures (Hahne et al., 2008), no further examination of these pathways was performed. In total a selection of 30 genes was done. The genes are shown in table 3.4.

KEGGid	HGNC symbol	KEGGid	HGNC symbol	KEGGid	HGNC symbol
hsa:699	BUB1	hsa:701	BUB1B	hsa:890	CCNA2
hsa:891	CCNB1	hsa:991	CDC20	hsa:1026	CDKN1A
hsa:1028	CDKN1C	hsa:1111	CHEK1	hsa:1647	GADD45A
hsa:2288	FKBP4	hsa:3434	IFIT1	hsa:3437	IFIT3
hsa:4085	MAD2L1	hsa:4171	MCM2	hsa:4172	MCM3
hsa:4173	MCM4	hsa:4174	MCM5	hsa:4176	MCM7
hsa:4678	NASP	hsa:5347	PLK1	hsa:5933	RBL1
hsa:9133	CCNB2	hsa:9134	CCNE2	hsa:9700	ESPL1
hsa:10051	SMC4	hsa:10926	DBF4	hsa:23594	ORC6L
hsa:55075	UACA	hsa:55761	TTC17	hsa:81570	CLPB

Table 3.4 – *Genes selected by* limma *and* gene2pathway

3.7.2 Comparison of network inference using *DDEPN*, *G1DBN* and *ebdbNet*

After selection of the functionally relevant subset of genes, further analyses using network inference methods were performed to identify regulatory interactions between the genes. The inferred interactions were compared to a reference network derived from KEGG and by manual curation in order to highlight potential crosstalk of the known cell cycle nodes to the additional nodes found by the gene selection process. Three approaches for network reconstruction were used. The DBN method *G1DBN* (Lébre, 2009), used in the original analysis proposal, is presented first. Afterwards, *ebdbNet* (Rau et al., 2010) as well as the novel method *DDEPN*, developed and described in this work, are compared to the original results.

Assembling a reference network

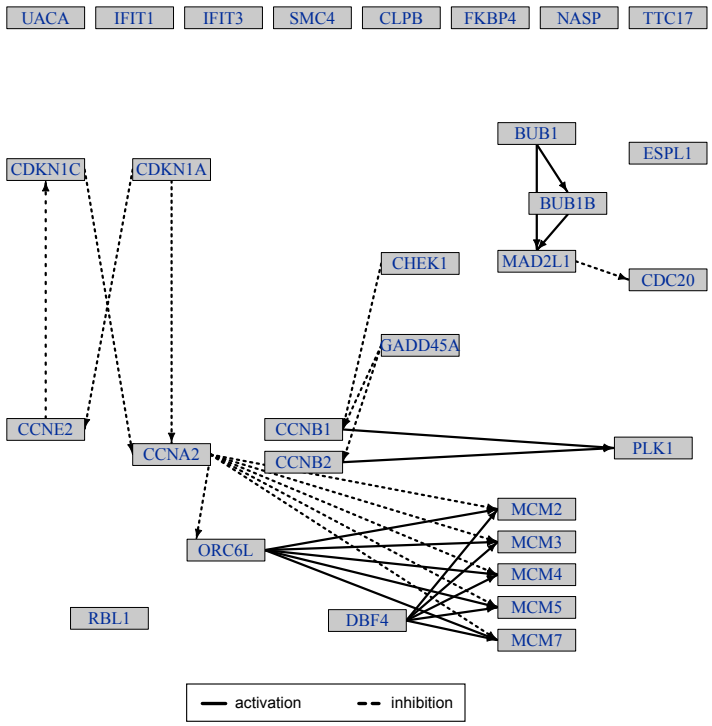

Figure 3.13 – *Manually assembled reference network for the subset of cell cycle genes.*

To summarise the expectations on the regulatory interactions, a reference network was assembled and used as prior network for the *DDEPN* inference. To derive the prior network, edge confidences were generated as described in section 3.3.1 using the signalling and disease pathways from the KEGG database. These edge confidences were discretised to the levels 1, -1 or 0 for an activation edge, inhibition edge or missing edge, respectively. The purpose of doing so was to derive a prior network that clearly defined where edges were to be expected and where no edge was expected. Additionally, some edges were added manually, in order to define a regulatory network that fit best the expectations of the author on the interactions in the selected subset of cell cycle related genes. The reference network is shown in figure 3.13.

G1DBN

G1DBN was chosen as network inference method in the original workflow because it was designed for the analysis of microarray time course data and directly available as R-package. The following parameters were used for inference. In the first step of the algorithm, the DAG $(G)^{(1)}$ was inferred using

the least squares M-estimator with an edge selection threshold $\alpha_1 = 0.5$. In the second step of the algorithm, the full-dependency graph $\tilde{\mathcal{G}}$ was estimated from $\mathcal{G}^{(1)}$, using a significance level $\alpha_2 = 0.1$. Figure 3.14 shows a comparison of the inferred network of *G1DBN* to the interactions found in the prior network. Solid lines correspond to activations, dashed lines to inhibitions. Black edges occur in both the prior and inferred network. Blue edges represent edges that were not present in the prior and additionally inferred, red edges were present in the prior and not inferred in *G1DBN*. Note that the edge type is ignored for the comparisons, because *G1DBN* is not able to infer different types of edges. Only three edges were reconstructed and present in the prior network, CCNA2 → MCM5, CCNE2 → CDKN1C and ORC6L → MCM7. However, it is apparent that the overlap between the inferred network and the prior cell cycle interactions is sparse. Looking at some of the edges that were not inferred, but are present in the prior pathway, it can be seen that inhibitory effects from cell cycle regulator genes are lost. This holds for the cyclin-dependent kinase inhibitor 1A (CDKN1A, also known as p21) and cyclin-dependent kinase inhibitor 1B (CDKN1B, also known as p27) genes, which normally bind to the Cyclin E-CDK2 and Cyclin D-CDK4 complexes and inhibit cell cycle progression. Also the inhibition of CCNB2 by CHEK1 is not reconstructed by *G1DBN*, which usually occurs in check point mediated cell cycle arrest controlled by CHEK1.

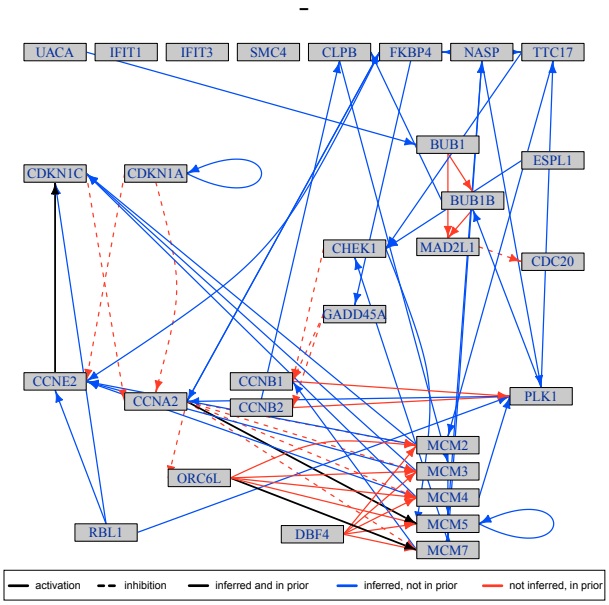

Figure 3.14 – G1DBN *network for the CAMDA dataset. Two edges were found in concordance with the reference net (black edges). Several edges were not found by* G1DBN, *but were present in the reference, e.g. CHEK1 ⊣ CCNB2.*

ebdbNet

The second DBN approach *ebdbNet* was used to infer a signalling network from the CAMDA data. Parameters were set up as described in Rau et al. (2010). A network of type 'feedback' was inferred, using parameters $K = 0$ for no hidden states, and convergence criteria $\Delta_1 = 0.15$ and $\Delta_2 = \Delta_3 = 0.05$ as suggested in the reference. The result for selecting the 0.25% and 0.975% quantiles of the resulting z-scores as inhibition and activation edges is shown in figure 3.15. Line style and colour encoding are chosen as in figure 3.14. The network is not as sparse as the *G1DBN* network, but only one edge was found that is also present in the reference network, CDKN1A ⊣ CCNE2, a well known inhibitory relationship of the cell cycle inhibitor CDKN1A (also known as p21). Also the inverse edge CCNE2 ⊣ CDKN1A is found, which points to a down-regulating influence of the cyclin CCNE2 onto p21. As in *G1DBN*, check point inhibition by CHEK1 in the interaction CHEK1 ⊣ CCNB2 is not inferred from data, pointing to an active cell cycle progression.

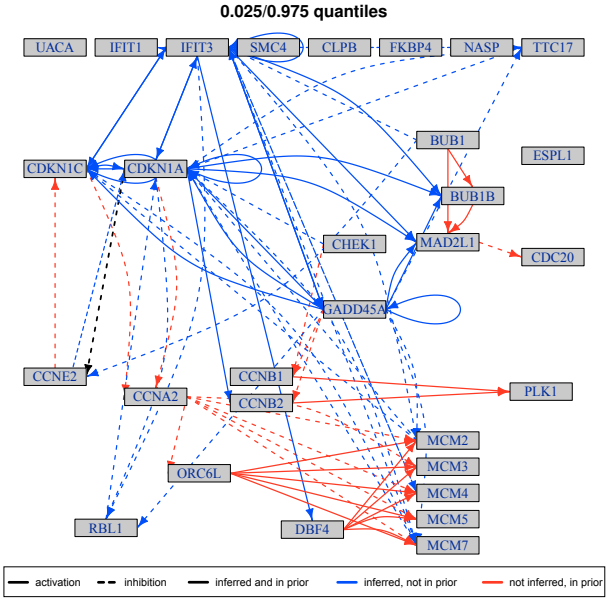

Figure 3.15 – ebdbNet *network. The approach tends to infer more edges than* G1DBN, *which makes interpretation of the edges more difficult. Only one edge was found in agreement with the prior network, CDKN1A ⊣ CCNE2. However, for example the inhibition CHEK1 ⊣ CCNB2 is not inferred (marked red), despite its existence in the reference, which is in concordance to the* G1DBN *method.*

DDEPN

Finally, the network reconstruction was performed for the *DDEPN* method using both the GA and inhibMCMC structure samplers. Figure 3.16, (A) shows the results of the inhibMCMC run. Prior knowledge was incorporated using the laplace prior (see section 3.3.1) with the network from figure 3.13 as prior network. Parameters were set to 50000 iterations, burn-in of 25000 iterations, the hyperparameters for the laplace prior to $\lambda = 0.01$ and $\gamma = 1$. Ten independent sampling runs were performed in parallel and the final network assembled as described in section 3.4.2. The resulting network is much sparser than the networks from *G1DBN* and *ebdbNet*. 14 edges were inferred, that also were in the prior network. One novel interaction was found (BUB1 $\rightarrow$ RBL1) and 15 edges from the prior network were confuted by the data. The edge CCNA2 $\dashv$ MCM5 was found consistently with the *G1DBN* result, again pointing to an active CCNA2 protein that occurs in cell cycle progression. In *G1DBN*, the type of the effect could not be inferred, but is found in DDEPN as inhibitory effect. The cascade of proteins BUB1 $\rightarrow$ BUB1B $\rightarrow$ MAD2L1 $\dashv$ CDC20 is found by *DDEPN*, too, in accordance to the prior network.

The reconstruction result using the GA of *DDEPN* is shown in figure 3.16, (B). The GA was run using the laplace prior model with the same prior network as for inhibMCMC. Parameters were set to $p = 500$, $q = 0.3$ and $m = 0.8$. The maximum number of iterations was set to 1000 and the prior hyperparameters $\lambda = 0.5$ and $\gamma = 1$. After inference, the final network was obtained as shown in section 3.4.1 by inspecting the final GA population of networks and including edges that were present in more than 85% of all networks in the population. Like inhibMCMC, the GA yields a much sparser result than *G1DBN* and *ebdbNet* do. 19 edges were found in agreement with the literature (black edges) and 10 edges were not supported by the data. An overlap to inhibMCMC of 8 interactions was observed. In concordance with all other methods, the CHEK1 $\dashv$ CCNB2 inhibition was not inferred, providing evidence that this interaction is indeed shut down, although the prior network would suggest the edge to be present.

3.8 Implementation as R-package 'ddepn'

The *DDEPN* method is implemented as R-package 'ddepn' (R Development Core Team, 2010) available on the CRAN homepage (`http://cran.r-project.org`). It is supported by Linux and Windows architectures, including support for parallelisation using multiple CPU cores on Linux systems. The set of KEGG signalling networks used for the prior confidence determination described in section 3.3.1 is included in the package as dataset *kegggraphs*. Further, the HCC1954 dataset described in section 3.6.1 is included as dataset

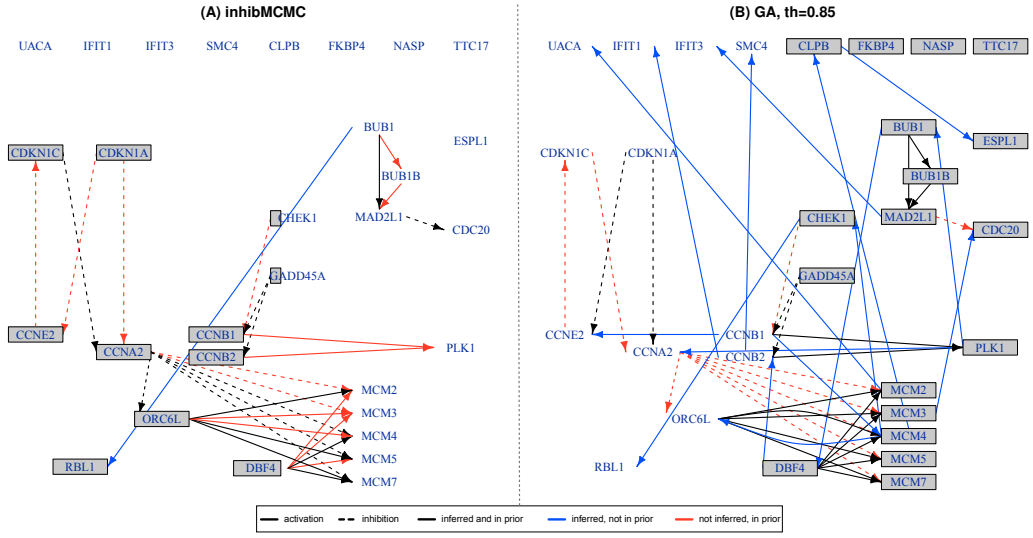

Figure 3.16 – *Result for* DDEPN *inference. (A) inhibMCMC run with laplace prior,* $\lambda = 0.01$, $\gamma = 1$. *The shown network was obtained as described in section 3.4.2. (B) GA run with laplace prior,* $\lambda = 0.01$, $\gamma = 1$. *The final net was obtained as described in section 3.4.1 with a threshold of* $th = 0.85$.

hcc1954, providing the data for 16 phosphoproteins measured at 10 time points within 1 hour after stimulation with two ligands. Together with its documentation the package offers an easy obtainable resource for network inference including exemplary data and resources for the usage of prior knowledge.

4 Discussion

The novel network inference algorithm *DDEPN* was presented which can be used to reconstruct either signalling networks or regulatory networks from high-throughput data generated after external perturbation. A perturbation can be seen as abstract external influence that is imposed on the biological system by some kind of treatment. The effect is neither restricted to one particular type (i.e. activation or inhibition) nor to a specific node in the set of measured proteins or genes. The algorithm infers edges from the external perturbation to nodes if the measurements support it and thus allows for a data driven determination of the treatment effects. Starting at the perturbation node, an activity state is propagated along the edges of a hypothesised network in order to derive deterministically the boolean ON/OFF states of all proteins or genes over the measured time frame. This propagation is done by following a predefined boolean rule that approximates the biological signal transduction process. The activity states relate directly to the network hypothesis, because they depend entirely on the network connectivity. They are used to assess the fit of the data to the proposed network through a score derived from a novel likelihood model. *DDEPN* is able to include prior knowledge in two ways, as prior on the network structures which provides the possibility to bias the inference procedure towards an external reference network, or as prior biasing towards the more general graph-theoretic scale-free property. In this chapter, the approach is discussed from three perspectives. First, in section 4.1, advantages compared to other available methods are elucidated and also limitations of the approach are discussed. Second, an interpretation of the results of *DDEPN* for the HCC1954 and CAMDA datasets is given in sections 4.2 and 4.3. Third, current available external knowledge sources and the process of setting up a suitable reference network for the inference are discussed in section 4.4, with an emphasis on the way how prior networks were generated in this work.

4.1 *DDEPN* as flexible means to perform network inference from high-throughput data generated after external perturbation

Network reconstruction methods have been frequently used to analyse the interplay of genes or proteins as a system. Important examples of such inference methods are Boolean networks (de Jong, 2002; Huang, 1999), BNs and DBNs (introduced in section 2.1, see also Pearl (1988), Friedman et al. (2000)). Usually, it is difficult to generate hypotheses from simple observational data without perturbation of the system. Geier et al. (2007) studied reverse engineering methods on simulated data for time courses and external perturbations and came to the conclusion that additional perturbation of the system is beneficial. Already early studies showed that a system is best studied under external influences that directly or indirectly interfere with normal cellular processes (Pe'er et al., 2001; Sachs et al., 2005). A review on cellular network analysis under perturbation can be found in Markowetz (2010). *DDEPN* fits in the methodological approaches that deal with data generated after directed perturbation of several nodes. Examples for these methods include Nested Effects Models (NEM, Markowetz et al. (2005); Fröhlich et al. (2008a)), where each gene in the network to be reconstructed is perturbed individually by siRNA mediated knockdown, or studies from Tegner et al. (2003) or Nelander et al. (2008), who modelled perturbation effects as linear combinations of inputs or as non-linear effects, respectively. However, *DDEPN* extends current approaches by its ability to determine the targets for the perturbation effects dynamically during the inference (in form of edges) and to model the system's dynamics as boolean signal transduction process that describes the activity of the nodes over time. Further, several experiments (i.e. treatments) can be integrated into one inference run to use the additional information in the network reconstruction process. In the next subsections these properties are discussed.

4.1.1 Perturbation effects are estimated explicitly in *DDEPN*

In *DDEPN* perturbations are explicitly included into the network structure and influences originating in the perturbation node are estimated from the data directly. Often individual perturbation of many or even all nodes in the network is not feasible due to time and cost restrictions, and for many experiments, a perturbation in form of a specific treatment often affects the whole cell or tissue under study. Different environmental conditions like temperature variation or nutrition deprivation, or treatment with drugs that are supposed to target one or more proteins specifically, but have side effects,

are examples for these system-wide perturbations. In this case, the targets of the influences cannot be determined explicitly in advance. The advantage of *DDEPN* is that effects are derived from the measured data directly, and even if expected targets could be included as prior guidance for the inference (expressed as higher prior probability for the respective interaction), learning additional and possibly unexpected perturbation targets is possible. Comparison of *DDEPN* to the DBN methods presented earlier (sections 2.1.1 and 2.1.2), or to another DBN application to perturbed time course data (Dojer et al., 2006), shows that all models infer possibly cyclic directed graph structures from longitudinal data and are able to cope with external perturbations. However, the DBN approaches only represent the perturbation effects implicitly in the network structure that is inferred. As a side note, the type of external perturbation is not restricted to inhibiting perturbations in *DDEPN*. In principle, both activating and inhibiting effects can be introduced simultaneously, for example by stimulation of cell via a receptor ligand and simultaneous inhibition by a receptor inhibiting drug. Applying multiple interventions in this manner might lead to competing influences on particular nodes that are the focus of the analysis. *DDEPN* determines the state for each experiment (i.e. perturbation state) separately, but performs parameter estimation for the active and passive states of each node together for all experiments. If a node, for instance, is passive under several conditions, the parameter estimate will be more robust, since measurements from all of these conditions are used for parameter estimation. The results from section 3.5.2 show, that the inclusion of multiple experiments in form of external stimuli increases the performance of the inference. Thus, *DDEPN* is able to integrate several measurement runs and determines influences from each of the perturbation nodes separately. Finally, edges in the inferred networks are modelled as either activation or inhibition rather than an abstract and general influence with no type (as is the case for many methods, e.g. *G1DBN*), leading to a more detailed representation of the dynamics of the system and thus the underlying biological processes.

4.1.2 Integration of prior knowledge is done flexibly

As shown in section 3.3, inclusion of prior knowledge is possible in *DDEPN* and from the results in section 3.6.3 it can be seen that doing so helps to overcome limitations like insufficient temporal resolution of the time course to resolve signalling cascades. However, no guarantee can be given, that the prior knowledge fed into the inference will be correct (see also section 4.4). Even when using different databases, different reference networks can be obtained and the results of the inference will vary. Furthermore, users have their own preferences for the selection of a suitable knowledge base on which the prior

model is built up, so a flexible way of prior knowledge inclusion is desired that makes it possible to easily exchange an outdated or bad prior model with a newer and better one. *DDEPN* offers this possibility in that the prior matrix B for the laplace prior (section 3.3.1) can be obtained from virtually all network structure databases and can thus easily be exchanged. Even the combination of several databases is possible, making integrative prior models possible.

The laplace prior is one example for prior knowledge inclusion which gives prior weights for individual edges. However, there are more possibilities to construct priors on the network structure itself. A good overview on types of priors and their usage can be found in Mukherjee and Speed (2008). In this study, besides the prior on individual edges, four alternative ways of prior models are discussed. One of these is a prior on the degree distributions of nodes, which also is implemented in *DDEPN* as the scale-free prior model. A sparsity prior is suggested as well, meaning that the highest indegree of the nodes in the graph is bounded by some value. Also this feature is implemented in *DDEPN* in form of a parameter during inference (called 'fanin' in the respective function call), fulfilling exactly this purpose. The third model is a prior on edges connecting classes of vertices. This means that e.g. only ligand to receptor bindings are allowed, because ligands would not bind to a cytoplasmic protein. At last, prior knowledge on the existence of edges between particular higher-level features of networks can be used. For example, it might be required that there exists a path from the receptor level of a cell down to the cytoplasmic or nuclear level, since this is main function of a signalling pathway. Even if the last two models are not explicitly included in *DDEPN*, increased probability for edges between classes of nodes or higher-level network features can be achieved by appropriate adjustment of the laplace prior matrix B and corresponding hyperparameters, giving higher edge confidences to the respective edge sets.

An implicit inclusion of prior knowledge already happened in *DDEPN* at the level of the signal propagation procedure (see section 3.1.1). Here, in fact a very crude assumption is made. In particular, it is assumed that all combinations of parental activity states at a child node follow the same logical rule in order to define the corresponding output. A child node becomes active if at least one of its parents connected by activation edges is active at the preceding time point and none of its parents connected by an inhibition edge is active at the preceding time point (see section 3.1.1). This is clearly simplifying the true competitive way of determining the child's activity state in real biological systems. Further, all node states are updated in a synchronous fashion (see e.g. de Jong (2002)), which is not necessarily the case in real biological systems. The propagation scheme, however, is a modular part of the *DDEPN* workflow, and can, in principle, be exchanged by more sophisticated rule sets. A challenge in doing so is the fact, that the connectivity of the network is naturally not known, since this is exactly the subject of the search algorithms. Hence,

it is not possible (and not feasible for larger networks) to predefine boolean functions for all possible combinations of parental configurations. There are interesting methodologies that deal with this kind of problem by inferring a set of boolean functions from data directly (Hunter and Klein, 1998). It could be an interesting subject of further research to include this kind of methodology into the *DDEPN* approach to overcome the limitation of the very strict and heavily simplified model of biological signal propagation.

4.1.3 *DDEPN* offers more than meets the eye - additional features obtained by the inference

The main purpose of *DDEPN*, of course, is the reconstruction of signalling networks from data. There are some useful features, however, that are computed on the fly and reported after an inference run. In the ERBB signalling network example from section 3.6 a network between phosphorylated proteins in cancer cells was inferred. Looking at the activation of a protein depending on its phosphorylation state, high phosphorylation not always means activation, and low phosphorylation does not always mean that the protein is inactive. It can be the other way round, too, as it is for example the case for the GSK3-α protein (Plyte et al., 1992). Since *DDEPN* does not make any assumptions on the order of active and passive intensity levels, the active and passive levels are detected automatically. Of course, if only one experiment is present as basis for the inference, it can be hard to determine the correct assignment of the intensity levels to the activity state. However, inclusion of multiple experiments under perturbations will help to determine the states correctly. As noted above, the determined pattern of active and passive states for all proteins and time points are reported by *DDEPN* at the end of an inference run. The boolean dynamics of the system can hence be analysed in more detail after the network is reconstructed and the activity states can be associated with corresponding intensity levels, helping to investigate the qualitative effects of a specific treatment (compare figure 4.1). In section 3.1.1 it is noted that the state space of 2^N for N nodes is reduced to a number $M \leq 2^N$ by performing the signal propagation as long as no state vector is repeated. From the theory on boolean networks it is known that both point attractor states (i.e. steady states) as well as dynamic attractor states (i.e. state cycles) can be reached (de Jong, 2002). Even if the signal propagation is stopped whenever a state is produced a second time, the state mapping to the real time points using the HMM procedure (section 3.1.2) can result in a periodic assignment of system states. So even if no steady state is reached in the dynamics of the system, it is possible to observe a dynamic attractor state during the time course and *DDEPN* is able to find this periodic state switching behaviour.

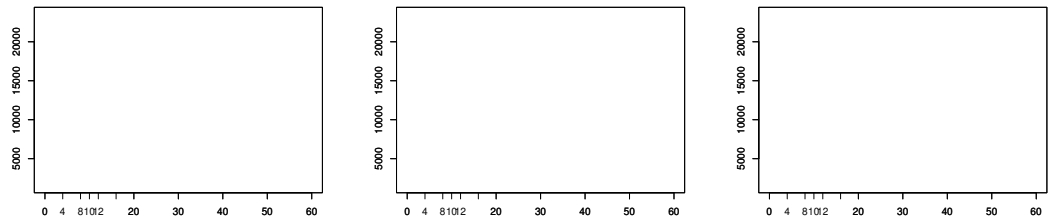

Figure 4.1 – *Active and passive profiles of pERK12, obtained during inference in* DDEPN. *Different response curves are observed for pERK12 depending on the stimulation ligands used. The red lines correspond to the inferred mean value (solid) and standard deviation (dotted) for the active state, the green lines to the model parameters for the passive state. The boxplots in gray show the distributions of the measurements, the black curve a spline fit to the intensity profile. Left: Simultaneous stimulation with EGF and HRG. Middle: Stimulation with EGF. Right: Stimulation with HRG. In the right figure, all measurements are classified as passive, even if an activation peak is visible.*

4.2 Interpretation of the inferred networks for the HCC1954 dataset

Once the reconstruction of a network is done, a final network is reported representing the optimal network with respect to the measured data. However, this is not the end of the analysis, but interpretation of the network interactions is necessary afterwards. Consider the results from section 3.6.2. In this example, no prior model was used in the GA approach of *DDEPN* to infer a signalling network for the ERBB signalling cascade in the HCC1954 breast cancer cell line. Only the implicit assumption that networks are sparse was put into the inference by applying the BIC score optimisation, and thus large numbers of edges were avoided in the reconstructed networks. All of the interactions depend purely on the measured data. As it was touched briefly in the results section, an interaction is just an abstract notion of an influence of one node onto the other. Consider for example the edge ERK1/2 $\dashv$ p38 in figure 3.9. ERK1/2 is a MAP-kinase activated by extracellular signals, as they are induced by growth factors like EGF. It is part of the MAPK cascade and able to phosphorylate a multitude of downstream proteins triggering diverse cellular responses, including cell proliferation and differentiation (Weinberg, 2007b; Seger and Krebs, 1995). p38 is also a MAP-kinase, but is involved in inflammatory responses and plays a role as tumour suppressor and in cancer development (Bradham and McClay, 2006). The roles of ERK1/2 and p38 seem to be of competitive nature and p38 controls cell growth which is induced by ERK1/2 activation (Aguirre-Ghiso et al., 2004). A direct inhibition of p38 by ERK1/2 is not yet described and seems unlikely, but seeing this

inhibition in the inferred network points to a down-regulated p38 and strongly active ERK1/2, which would endorse strong cell growth in the HCC1954 cancer cells.

However, there were also cases, where interactions would have been expected, but were not found in the network. For example in the classical MAPK cascade, MEK1/2 phosphorylates ERK1/2 directly. In the inferred network in figure 3.9, the interaction between MEK1/2 and ERK1/2 was not found, but only direct activations of the two proteins by EGF. The reason was that phosphorylation was only measured at time points 8 and 12 minutes. Activation of MEK1/2 and ERK1/2 is expected around 10 minutes after stimulation, but in the data the peaks for both proteins occur at the 12 minute time point. Thus, this cascade could not be resolved at a higher resolution. Another problem arises when proteins of a signalling cascade were not measured on the array, as seen for example for several of the components in the MAPK cascade (e.g. RAS, RAF, and others). So even if a direct edge between two proteins is found, it has to be carefully assessed whether this edge is a direct influence or an indirect interaction over multiple intermediate steps. The data only represents the abundance of phosphorylated protein in the cells, which might increase or decrease in response to some treatment. All interactions from such data are abstract influences between two proteins that have to be validated in further experiments.

4.2.1 Prior knowledge inclusion helps to retrieve a robust scaffold of the ERBB network

Inclusion of prior knowledge was described in section 3.6.3 and shown in figure 3.11, (A) for the HCC1954 data example. In the figure, gray edges encode overlaps between a manually curated reference network (figure 3.10) and the inferred network, blue edges encode novel, and red edges lost interactions in the inferred graph, compared to the reference. The network closely resembles the manually constructed reference network, as seen by the predominantly black edges. This scaffold is retained in the networks that were reconstructed under different treatments, too (see parts (B)-(D) in figure 3.11). Only few additional edges occur or disappear under the various treatments, suggesting that *DDEPN* inference is biased towards the literature knowledge, but allows for changes supported strongly by the data. However, five additional interactions were found by *DDEPN* that are worth a discussion. Four of these originate in the stimulus node EGF and point to major proteins from the MAPK and AKT signalling cascades (ERK1/2, p70S6K, AKT, PLCG). All of these interactions point to indirect influences from the stimulating ligand over several intermediate steps. ERK1/2 and p70S6K, are activated through the cascade EGF $\rightarrow$ ERBB1 $\rightarrow$ MEK1/2 $\rightarrow$ ERK1/2 $\rightarrow$ p70S6K, the classical MAPK cascade

(see e.g. Weinberg (2007b); Seger and Krebs (1995)), AKT via the PI3K/AKT pathway (Weinberg, 2007b; Manning and Cantley, 2007) and PLCG through EGF → ERBB1 → PLCG (Kim et al., 1990). The interactions are inferred, although the reference network does not contain them, which shows that the data give strong support for this interaction.

One could see these kind of indirect influences as strong support for a downstream effect, like MAPK or AKT activation, even if the direct interaction between the proteins does not exist. However, this phenomenon shows a limitation of *DDEPN*, because the modelling approach is not able to substitute the direct influence by a cascade over several intermediate steps, but more or less infers the transitive closure of the network. The fifth interaction, PRAS → ERBB1, could be interpreted as follows. PRAS is reported as inhibitor of the mTOR kinase (Sancak et al., 2007), stimulated by insulin. It binds to the raptor protein of the mTORC1 complex and inhibits the kinase activity of mTOR, in turn. Since p70S6K is the major downstream target of the mTOR kinase, and p70S6K activation promotes translational activity at the ribosome (Berven and Crouch, 2000), it could have a positive effect on the receptor tyrosine kinase protein synthesis, explaining the direct effect of PRAS onto ERBB1. However, this is a highly hypothetical interpretation, and external validation would be necessary to provide further evidence for this.

The network reconstructed after inclusion of the prior knowledge, however, shows that inference can be guided by additional knowledge. Nevertheless, the sources of prior knowledge have to be chosen carefully (see also section 4.4) and after the reconstruction, each interaction has to be assessed and validated to explain the results of the inference.

4.2.2 Combinatorial treatment using trastuzumab and erlotinib has the strongest impact on signalling processes in HCC1954 cells

Two inhibitors were used in this work to treat the HCC1954 breast cancer cell line: the monoclonal antibody trastuzumab (Carter et al., 1992) and the small molecule drug erlotinib (Hidalgo, 2003). Both were used as single treatments and in combination. The data measured after treatment was fed into the *DDEPN* reconstruction to generate the networks shown in figure 3.11, (B)-(D). On a first glance on the networks, it is apparent that the network (D), generated under treatment with both drugs simultaneously, contains the most effected edges when compared to the prior reference (edges coloured blue and red), as it would be expected just by the fact that two treatments were applied at the same time. An interesting question to answer is, whether the results from the simultaneous treatment can be related to the single treatments, in order to figure out effects that are specific to the combinatorial treatment

and beneficial with respect to the overall cellular response. In particular, is an effective inhibition of the mainly deregulated and predominantly active MAPK and AKT signalling cascades observed? The general scaffold of the network, imposed by the reference network is similar under all treatments (gray edges) and could be robustly inferred. To assess the overall effect of a drug, a closer look to the novel interactions (blue edges) and the missing edges (red) has to be taken. Exemplary, the focus put on the activation of ERK1/2, AKT and p70S6K, three major regulators of the cellular response. They induce a multitude of effects, such as proliferation and differentiation (ERK1/2, Seger and Krebs (1995)), translation, cell motility and proliferation (p70S6K, Berven and Crouch (2000)) and cell survival, growth and proliferation (AKT, Manning and Cantley (2007)).

Under erlotinib treatment (figure 3.11, (B)), activation of both AKT and p70S6K is retained, as is the case for the stimulation without any inhibitors (figure 3.11, (A)). Only ERK1/2 activation is lost, compared to the non-inhibited experiment, suggesting that activation of p70S6K is triggered by the AKT pathway in this case. For trastuzumab treatment (figure 3.11, (C)), on the other hand, the activating influence of EGF onto AKT was not reconstructed, while EGF $\rightarrow$ p70S6K and EGF $\rightarrow$ ERK1/2 still were found. Finally, with both drugs (figure 3.11, (D)), activations of AKT, ERK1/2 and p70S6K were not inferred any more, pointing at a successful inhibition of the respective influences by the chosen treatment. Further, the interaction p38 $\dashv$ ERBB1 is omitted in contrast to the reference network. Frey et al. (2006) show, that p38 is required for EGF induced down-regulation of EGFR in four different cell lines and usually controls the balance between cell proliferation and migration via EGF receptor degradation. Interestingly, this mechanism is present in the uninhibited experiment and the erlotinib experiment (figure 3.11, (A) and (B)), but lost in the trastuzumab and combinatorial treatment (figure 3.11, (C) and (D)), suggesting that trastuzumab treatment is affecting the migrative potential of the cells. Decreased migration is reported to come with an increased proliferation through sustained activation of ERK1/2 (shown by Frey et al. (2004) in intestinal epithelial cells), controlled by the Y1045 phosphorylation site of EGFR (Frey et al., 2006). The remaining ERK1/2 activation under trastuzumab treatment gives some evidence for this, but since Y1045 was not measured in this experiment, this hypothesis would have to be verified through additional experiments. In summary, the combinatorial treatment seems to have a more potent effect to down-regulate both MAPK and AKT signalling cascades than single treatment with each drug and is a promising option for treatment of this aggressive type of breast cancer cell line.

4.3 Interpretation of interactions in the CAMDA data example

In section 3.7.2 network inference on the CAMDA dataset (Affara et al., 2007) was shown for three different algorithms, *G1DBN* and *ebdbNet* as well as the novel *DDEPN* algorithm. In this section an assessment and interpretation of the reconstruction results is given, with respect to the ability of the *DDEPN* method to infer regulatory networks from gene expression data. The resulting networks are compared to a literature derived reference network (figure 3.13) to assess the biological relevance of each inferred interaction. Here, edges are linked to literature resources, in order to give further evidence for their existence, and a comparison between the inferred interactions in the three approaches is conducted.

The first resulting network, obtained by *G1DBN*, is shown in figure 3.14. Black edges encode overlaps to the reference, blue edges encode novel, and red edges missing edges in the inferred networks when compared to the reference network. It can be seen that three edges were reconstructed by *G1DBN* that could be linked to literature resources. These include the edges CCNA2 $\rightarrow$ MCM5, CCNE2 $\rightarrow$ CDKN1C and ORC6L $\rightarrow$ MCM7. The first points to an active cyclin A2 protein, which plays a role in G1/S and G2/M transition and exhibits an effect on the protein MCM5 that is needed for the pre-replication complex, formed during G1 phase before replication start (Li and Jin, 2010). The second interaction CCNE2 $\rightarrow$ CDKN1C is described in Lahoz et al. (1999) and reflects the antagonistic expression pattern described in the referenced publication (note that the type of the interaction cannot be determined using *G1DBN*). It points to an active cyclin-E/CDK2 complex and down-regulated CDK inhibitor CDKN1C (also referred to as Kip2 or p57, see e.g. Guo et al. (2010) for a review on p57 functions). The last interaction, ORC6L $\rightarrow$ MCM7 is part of the replication initiation machinery (Li and Jin, 2010; Nishitani and Lygerou, 2002). Both the hexameric ORC complex, containing ORC6L, and the MCM complex, containing MCM 2-7, are required for the replication licensing process that ensures that DNA replication is performed only once per cell cycle. The interaction reflects this binding of the two complexes. Additionally, interactions between CHEK1 and CCNB1 as well as between GADD45A and CCNB1/2 were not inferred, meaning that the regulation of the B-type cyclins might not be present (see Jin et al. (2002); Sanchez et al. (1997) for a description of GADD45A and CHEK1 functions, respectively). These are all examples that point to an active progression through the cell cycle, which is in fact counterintuitive, keeping in mind that the cells were exposed to survival factor deprivation (SFD), where one would expect down-regulated cell cycle progression.

In the second DBN reconstruction approach *ebdbNet*, a much more densely

connected network was inferred, shown in figure 3.15. Only one edge from the reference network was found, CDKN1A ⊣ CCNE2. CDKN1A (also known as p21 or CIP1) is a cell cycle inhibitor of the CIP/KIP family (Harper et al., 1993; Park and Lee, 2003) that regulates G1 CDK proteins. None of the interactions in the pre-replication complex, indicated by interactions between ORC6L and the MCM proteins were found. Inhibition of the cyclin CCNE2 and missing interactions between ORC and MCM complex members points to an inactive replication machinery, which could be induced by the exposure to SFD during the experiment. However, a wealth of novel interactions is found, making it hard to assess which of these edges could be meaningful.

Finally, the two structure search algorithms inhibMCMC and the GA from *DDEPN* were applied and the resulting networks are shown in figures 3.16. The networks are, in general, much sparser connected than their counterparts from *G1DBN* and *ebdbNet*, supporting the results from section 3.5.2. In these tests on simulated data, it was apparent that sensitivity levels were rather low. Hence, the reconstruction missed quite a number of edges, that should have been found. On the other hand, the high values for specificity indicated that the inferred edges were those with strong support from the experimental measurements and that few false positive edges were found. Similar, in the real data, it can be expected that the interactions that are reconstructed are supported well by the data and represent meaningful hypotheses on the network structure. There are two regulatory patterns that were present in the reference network and in common between inhibMCMC and the GA. For example, in both algorithms the cyclin CCNA2 is inhibited, either by CDKN1C (inhibMCMC) or by CDKN1A (GA). This points to an inactive cell cycle progression through G1 and thus cell cycle arrest without progression into S-phase. Further, GADD45 inhibits CCNB1 and CCNB2 in both inhibMCMC and the GA. This is also supported by e.g. Jin et al. (2002), who report decreased nuclear CCNB1 levels depending on increased Gadd45 protein activity leading to G2/M arrest depending on nuclear CCNB1 levels. Thus, the cells seem to stop cell cycle progression after SFD.

In summary, it could be shown that *DDEPN* is also applicable to infer regulatory networks from gene expression measurements. Interactions were discussed for all inference approaches and suggested that *DDEPN* coincides best with the expectations one could have for the biological experiment. Further, *DDEPN* alleviates interpretation of the networks, because it tends to yield sparse network structures.

4.4 Determining reference networks using external knowledge bases

The final section in this discussion of the *DDEPN* method deals with the generation of prior networks that can be used to guide the inference procedure. Several studies have shown, that inclusion of prior knowledge is both helpful and sometimes necessary to produce reliable and meaningful results (Gat-Viks et al., 2006; Werhli and Husmeier, 2007; Mukherjee and Speed, 2008; Steele et al., 2009; Sheridan et al., 2010). *DDEPN* offers the possibility to include different sources of prior knowledge into the inference. Besides modelling the general scale-free characteristic as prior, the laplace prior offers the opportunity to include knowledge on network structures directly, in order to increase or decrease the weight for specific interactions. But how does one create a suitable prior network that covers the important interactions? Clearly, the choice of the source of prior knowledge heavily determines the created network. There is a multitude of online databases available for this purpose (see section 2.2 for some examples, or also Adriaens et al. (2008); Ooi et al. (2010)). Pathway information is generally available for gene regulation, metabolic processes, signal transduction or protein protein interactions. So, as first criterion, how a suitable reference pathway is assembled, a selection of the particular database has to be made, such that the pathway type to be inferred matches the type present in the database. Naturally, using a metabolic pathway map to augment knowledge on signalling processes does not make much sense. Further, it is difficult to define a 'true' reference network, valid for all possible biological conditions. Cellular systems undergo changes, especially during the development of cancer and a curated pathway under normal condition might include interactions that are not valid under disease conditions. In this work, the prior network for the HCC1954 dataset (section 3.6) was built up from the KEGG database, including only the signalling and disease related networks. Additionally, manual inspection was performed to fine-tune the resulting prior network. In the CAMDA data example (section 3.7), the KEGG cell cycle pathway was inspected and a prior network was constructed manually, including only the proteins that were selected in the functional gene filtering procedure (section 3.7.1).

This way of creating prior networks shows, that it is hard to device a fully automated workflow for this purpose. If an automatic retrieval of reference pathways were conducted for the same type of pathways, it would still be not clear, under which biological context the respective pathways were created. For example, in the PID database (Schaefer et al., 2009), the pathways are created under the assumption, that the pathway is present in a 'normal' biological state of the specimen of interest. Abnormalities caused for example by pathological responses are included only in a small number of pathways. In a

more general sense, each pathway interaction is derived from an experiment that was conducted for a particular biological condition. These include issues like the cell or tissue type, the disease state, the treatment of the specimen and many more. In cancer cells, for instance, an increased number of mutations is found which cause tumourigenesis. Rather than concentrating on the mutation and the affected gene or protein itself, it is important to consider the effect of the alteration onto its respective pathway, in which the entity is functional (Vogelstein and Kinzler, 2004). Interactions that are present under normal biological conditions can be shut down or activated by the mutations. For example, mutated oncogenes usually exhibit constant activity or activity where the un-mutated oncogene would be inactive. Mutated tumour suppressors, on the other hand, lack activity where it would be expected under normal conditions. These types of interference with the normal conditions yield changes in the normal structure of a biological pathway that have to be considered, when using the normal pathway as prior knowledge for the chosen network reconstruction. An additional challenge is that in general pathway databases are prone to be incomplete or even erroneous (Adriaens et al., 2008). Even if a huge amount of information has been collected over the years (consider Galperin (2008) for an overview on biological databases), the particular information on pathway structures, interaction partners and biological processes is far from being complete.

Since the discovery of alterations in the network structure as well as identification of novel or falsification of present interactions is the very purpose of network reconstruction, care has to be taken, that the prior knowledge does not force interactions to be present where the measured data provide evidence against it and vice versa. As described in section 3.5.4 for the *DDEPN* method it is suggested to adjust the prior strength, such that the average changes in the data likelihood are similar to the average changes of the prior probabilities over all iterations in the algorithm. This ensures, that prior evidence for an edge can be 'overwritten' by sufficiently strong evidence from the data. The reference network is still assumed to be the 'correct' underlying network, and disease or treatment specific alterations are expressed in additional or missing edges in the reconstructed network. This rationale for setting up the modelling approach and interpretation of the inference results was followed exemplary in the previous sections of this discussion (4.2, 4.3).

When considering the type of pathways as well the biological context, and making sure that an appropriate selection of pathways was performed, still the integration of multiple pathways into a single reference network is a challenge (Adriaens et al., 2008). An integration of a number of pathways from one source database is comparably straightforward, since the data format is the same and nodes and interactions can simply be merged. However, when different databases are seeked to be integrated, additional effort has to be taken to translate various data storage formats into a common representation of all

information that preserves at best all of the semantics of the different repositories. Due to the fact that data storage formats are not standardised (while some de facto standards exist, see e.g. Demir et al. (2010); Hucka et al. (2003)), the conversion to a common format is still a rather tedious task. Additionally, often conversions between file formats cause loss of information, because various formats are not able to store all of the contained information. And finally, pathway databases undergo a constant cycle of revision and curation, which leads to constantly changing knowledge repositories. Thus, automated conversion is highly desirable, but at the moment it is highly recommended to create integrated pathways by manual curation and comparison to recent literature, in order to ensure high quality of the generated pathways.

4.5 Conclusions

In this dissertation, the novel approach *DDEPN* was presented for the reconstruction of signalling networks from high-throughput omics data. The method is able to include multiple external perturbations into the inference and to estimate the effects of these from the measured data directly. It uses time-resolved quantifications of the abundance of measured nodes and models the system's dynamics as boolean variable over time, leaving a boolean state vector for each time point containing the states of all nodes. A likelihood score relates the measurements to this series of system states, and, in turn, depends on the network structure that is assumed. By using this score, identification of high-scoring network structures with respect to the measured data is performed. Two algorithms are included into *DDEPN*. First, a genetic algorithm performs network structure optimisation, and second an extended MCMC structure sampler performs sampling of the network search space for networks with different types of edges. To bias the reconstruction towards known signalling networks from the literature, inclusion of external knowledge is possible by two prior models on the network structure. The algorithm was tested from a theoretical point of view via simulation studies, showing good performance results for the reconstruction process. Further, a comparison was conducted to two related DBN inference approaches, yielding improved reconstruction performance. To assess the usefulness of *DDEPN* for real data examples, two model systems were chosen, for which experimental data were available. First, a signalling network from real protein phosphorylation measurements, generated by RPPAs for cell lysates from the human breast cancer cell line HCC1954, was done. *DDEPN* successfully identified parts of the canonical MAPK and AKT signalling cascades. After inclusion of prior knowledge from the KEGG database, the resolution of the signalling cascades could be improved substantially. For data generated after external inhibition by the two ERBB-receptor inhibiting drugs trastuzumab and erlotinib, treatment specific effects of the drugs

could be observed. In particular, for the simultaneous treatment with both drugs, a potent inhibition of the MAPK and AKT downstream cascades was observed. For single treatment, the MAPK activation remained, while AKT activation was inhibited by trastuzumab only. As second application example, a dataset from human umbilical vein endothelial cells exposed to survival factor deprivation was chosen, measuring cell cycle related gene expression across several time points. *DDEPN* found interactions that gave evidence for a cell cycle arrest of the cells, which was in concordance with the expectations of the experiment. The alternative DBN approaches on the other hand showed evidence for a still active cell cycle progression. Further, *DDEPN* inference led to sparser network structures, making interpretation easier and reducing the number of false positive interactions.

List of Figures

List of Tables

List of Abbreviations

aCGH	Array Comparative Genomic Hybridisation
API	Application Programming Interface
APC	Anaphase Promoting Complex
AR	Amphiregulin
AUC	Area Under (ROC) Curve
ATCC	American Type Culture Collection
BioPAX	Biological Pathway Exchange
BN	Bayesian Network
BTC	Betacellulin
CAMDA	Critical Assessment of Microarray Data Analysis
CAS	Chemical Abstracts Service
CDK	Cyclin Dependent Kinase
CGH	Comparative Genomic Hybridisation
CKI	Cyclin-Kinase Inhibitor
CPD	Conditional Probability Distribution
CPDB	Consensus Path DB
CRAN	Comprehensive R Archive Network
DAG	Directed Acyclic Graph
DBN	Dynamical Bayesian Network
DDEPN	Dynamic Deterministic Effects Propagation Networks
DNA	Deoxyribonucleic Acid
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EM	Expectation Maximisation
EPR	Epiregulin
ER	Oestrogen Receptor
ERK	Extracellular-Signal-Regulated Kinases
GA	Genetic Algorithm
GO	Gene Ontology
HB-EGF	Heparin-binding EGF-like Growth Factor
HMM	Hidden Markov Model
HPD	Human Pathway Database
HRG	Heregulin
IPA	Ingenuity Pathways Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
KGML	KEGG Markup Language

LIST OF ABBREVIATIONS

KO	KEGG Orthology
HUVEC	Human Umbilical Vein Endothelial Cells
MAPK	Mitogen-activated Protein Kinases
MCM	Minichromosome Maintenance proteins
MCMC	Markov Chain Monte Carlo
MC^3	Markov Chain Monte Carlo Model Composition
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NEM	Nested Effects Models
NRG	Neuregulin
ODE	Ordinary Differential Equation
ORC	Origin Recongnition Complex
PI3K	Phosphatidylinositol 3-kinases
PID	Pathway Interaction Database
PSI-MITAB	Proteomics Standards Initiative Molecular Interactions tab delimited data exchange format
Rb	Retinoblastoma protein
RNA	Ribonucleic Acid
RNAi	RNA interference
ROC	Receiver Operator Characteristic
RPPA	Reverse Phase Protein Array
RTK	Receptor Tyrosine Kinase
SBML	Systems Biology Markup Language
SFD	Survival Factor Deprivation
SNP	Single Nucleotide Polymorphism
SSM	State Space Model
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TGF-α	Transforming Growth Factor alpha
VSN	Variance Stabilisation Normalisation

Bibliography

Adriaens, M. E., M. Jaillard, A. Waagmeester, S. L. M. Coort, A. R. Pico, and C. T. A. Evelo (2008, Oct). The public road to high-quality curated biological pathways. *Drug Discov Today 13*(19-20), 856–862.

Affara, M., B. Dunmore, C. Savoie, S. Imoto, Y. Tamada, H. Araki, D. S. Charnock-Jones, S. Miyano, and C. Print (2007, Aug). Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? *Philos Trans R Soc Lond B Biol Sci 362*(1484), 1469–1487.

Aguirre-Ghiso, J. A., L. Ossowski, and S. K. Rosenbaum (2004, Oct). Green fluorescent protein tagging of extracellular signal-regulated kinase and p38 pathways reveals novel dynamics of pathway activation during primary and metastatic growth. *Cancer Res 64*(20), 7336–7345.

Akavia, U. D., O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er (2010, Dec). An integrated approach to uncover drivers of cancer. *Cell 143*(6), 1005–1017.

Akutsu, T., S. Miyano, and S. Kuhara (1999). Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac Symp Biocomput*, 17–28.

Al-Shahrour, F., R. Dìaz-Uriarte, and J. Dopazo (2004, Mar). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics 20*(4), 578–580.

Ashburner, M., 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 (2000, May). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet 25*(1), 25–29.

Avraham, R. and Y. Yarden (2011). Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat Rev Mol Cell Biol 12*, 104–117.

Bader, G. D., M. P. Cary, and C. Sander (2006, Jan). Pathguide: a pathway resource list. *Nucleic Acids Res 34*(Database issue), D504–D506.

Bansal, M., V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo (2007). How to infer gene networks from expression profiles. *Mol Syst Biol 3*(19), 78.

Bansal, M., G. D. Gatta, and D. di Bernardo (2006, Apr). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics 22*(7), 815–822.

Bibliography

Beal, M. J., F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild (2005, Feb). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics 21*(3), 349–356.

Beissbarth, T. and T. P. Speed (2004, June). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics 20*(9), 1464–1465.

Bender, C., H. Fröhlich, M. Johannes, and T. Beißbarth (2008). Extending pathways with inferred regulatory interactions from microarray data and protein domain signatures. In *CAMDA: Critical Assessment of Microarray Data Analysis*, pp. 48–52.

Bender, C., F. Henjes, H. Fröhlich, S. Wiemann, U. Korf, and T. Beißbarth (2010, Sep). Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics 26*(18), i596–i602.

Bender, C., S. v.d. Heyde, F. Henjes, S. Wiemann, U. Korf, and T. Beißbarth (2011, Mar). Inferring signalling networks from longitudinal data using sampling based approaches in the r package 'ddepn'. under review.

Benjamini, J. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics 29*, 1165—1188.

Beroukhim, R., C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. M. Henry, R. M. Pinchback, A. H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson (2010, Feb). The landscape of somatic copy-number alteration across human cancers. *Nature 463*(7283), 899–905.

Berven, L. A. and M. F. Crouch (2000, Aug). Cellular function of p70S6K: a role in regulating cell motility. *Immunol Cell Biol 78*(4), 447–451.

Bevilacqua, V., G. Mastronardi, F. Menolascina, P. Pannarale, and G. Romanazzi (2009, June). Bayesian Gene Regulatory Network Inference Optimization by means of Genetic Algorithms. *Journal of Universal Computer Science 15*(4), 826–839.

Bradham, C. and D. R. McClay (2006, Apr). p38 MAPK in development and cancer. *Cell Cycle 5*(8), 824–828.

Bremer, M. M. (2006, December). *Identifying regulated genes through the correlation structure of time dependent microarray data*. Ph. D. thesis, Purdue University.

Brown, P. O. and D. Botstein (1999, Jan). Exploring the new world of the genome with DNA microarrays. *Nat Genet 21*(1 Suppl), 33–37.

Busch, H., D. Camacho-Trullio, Z. Rogon, K. Breuhahn, P. Angel, R. Eils, and A. Szabowski (2008). Gene network dynamics controlling keratinocyte migration. *Mol Syst Biol 4*, 199.

Callinan, P. A. and A. P. Feinberg (2006, Apr). The emerging science of epigenomics. *Hum Mol Genet 15 Spec No 1*, R95–101.

Carter, P., L. Presta, C. M. Gorman, J. B. Ridgway, D. Henner, W. L. Wong, A. M. Rowland, C. Kotts, M. E. Carver, and H. M. Shepard (1992, May). Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proc Natl Acad Sci U S A 89*(10), 4285–4289.

Castelo, R. and T. Kocka (2003, December). On inclusion-driven learning of bayesian networks. *J. Mach. Learn. Res. 4*, 527–574.

Cheung, S. W., C. A. Shaw, W. Yu, J. Li, Z. Ou, A. Patel, S. A. Yatsenko, M. L. Cooper, P. Furman, P. Stankiewicz, P. Stankiewicz, J. R. Lupski, A. C. Chinault, and A. L. Beaudet (2005). Development and validation of a CGH microarray for clinical cytogenetic diagnosis. *Genet Med 7*(6), 422–432.

Chickering, D. M. (1996). Learning Bayesian Networks is NP-Complete. In *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130. Springer-Verlag.

Chickering, D. M. (2003, March). Optimal structure identification with greedy search. *J. Mach. Learn. Res. 3*, 507–554.

Chowbina, S. R., X. Wu, F. Zhang, P. M. Li, R. Pandey, H. N. Kasamsetty, and J. Y. Chen (2009). HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics 10 Suppl 11*, S5.

Christianson, T. A., J. K. Doherty, Y. J. Lin, E. E. Ramsey, R. Holmes, E. J. Keenan, and G. M. Clinton (1998, Nov). NH2-terminally truncated HER-2/neu protein: relationship with shedding of the extracellular domain and with prognostic factors in breast cancer. *Cancer Res 58*(22), 5123–5129.

Citri, A. and Y. Yarden (2006, Jul). EGF-ERBB signalling: towards the systems level. *Nat Rev Mol Cell Biol 7*(7), 505–516.

Clynes, R. A., T. L. Towers, L. G. Presta, and J. V. Ravetch (2000, Apr). Inhibitory Fc receptors modulate in vivo cytoxicity against tumor targets. *Nat Med 6*(4), 443–446.

Cohen, P., D. R. Alessi, and D. A. Cross (1997, Jun). PDK1, one of the missing links in insulin signal transduction? *FEBS Lett 410*(1), 3–10.

Cooley, S., L. J. Burns, T. Repka, and J. S. Miller (1999, Oct). Natural killer cell cytotoxicity of breast cancer targets is enhanced by two distinct mechanisms of antibody-dependent cellular cytotoxicity against LFA-3 and HER2/neu. *Exp Hematol 27*(10), 1533–1541.

Cooper, G. F. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning 9*, 309–347. 10.1007/BF00994110.

CRAN (2011). The comprehensive R archive network. http://cran.r-project.org.

Daniel, D. C. (2002, Oct). Highlight: BRCA1 and BRCA2 proteins in breast cancer. *Microsc Res Tech 59*(1), 68–83.

Bibliography

Davies, H., G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B. A. Gusterson, C. Cooper, J. Shipley, D. Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G. J. Riggins, D. D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J. W. C. Ho, S. Y. Leung, S. T. Yuen, B. L. Weber, H. F. Seigler, T. L. Darrow, H. Paterson, R. Marais, C. J. Marshall, R. Wooster, M. R. Stratton, and P. A. Futreal (2002, Jun). Mutations of the BRAF gene in human cancer. *Nature 417*(6892), 949–954.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol 9*(1), 67–103.

Demir, E., M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, F. Schacherer, I. Martinez-Flores, Z. Hu, V. Jimenez-Jacinto, G. Joshi-Tope, K. Kandasamy, A. C. Lopez-Fuentes, H. Mi, E. Pichler, I. Rodchenkov, A. Splendiani, S. Tkachev, J. Zucker, G. Gopinath, H. Rajasimha, R. Ramakrishnan, I. Shah, M. Syed, N. Anwar, O. Babur, M. Blinov, E. Brauner, D. Corwin, S. Donaldson, F. Gibbons, R. Goldberg, P. Hornbeck, A. Luna, P. Murray-Rust, E. Neumann, O. Reubenacker, M. Samwald, M. van Iersel, S. Wimalaratne, K. Allen, B. Braun, M. Whirl-Carrillo, K.-H. Cheung, K. Dahlquist, A. Finney, M. Gillespie, E. Glass, L. Gong, R. Haw, M. Honig, O. Hubaut, D. Kane, S. Krupa, M. Kutmon, J. Leonard, D. Marks, D. Merberg, V. Petri, A. Pico, D. Ravenscroft, L. Ren, N. Shah, M. Sunshine, R. Tang, R. Whaley, S. Letovksy, K. H. Buetow, A. Rzhetsky, V. Schachter, B. S. Sobral, U. Dogrusoz, S. McWeeney, M. Aladjem, E. Birney, J. Collado-Vides, S. Goto, M. Hucka, N. L. Novère, N. Maltsev, A. Pandey, P. Thomas, E. Wingender, P. D. Karp, C. Sander, and G. D. Bader (2010, Sep). The BioPAX community standard for pathway data sharing. *Nat Biotechnol 28*(9), 935–942.

di Bernardo, D., M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins (2005, Mar). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol 23*(3), 377–383.

Dickler, M., M. Cobleigh, K. Miller, P. Klein, and E. Winer (2009). Efficacy and safety of erlotinib in patients with locally advanced or metastatic breast cancer. *Breast Cancer Research and Treatment 115*, 115–121. 10.1007/s10549-008-0055-9.

Dojer, N., A. Gambin, A. Mizera, B. Wilczyński, and J. Tiuryn (2006). Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics 7*(52), 249.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison (1998). *Biological sequence analysis* (first ed.). Cambridge University Press.

Ferlay, J., H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin (2010, Dec). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer 127*(12), 2893–2917. http://globocan.iarc.fr.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello (1998, Feb). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature 391*(6669), 806–811.

Frey, M. R., R. S. Dise, K. L. Edelblum, and D. B. Polk (2006, Dec). p38 kinase regulates epidermal growth factor receptor downregulation and cellular migration. *EMBO J 25*(24), 5683–5692.

Frey, M. R., A. Golovin, and D. B. Polk (2004, Oct). Epidermal growth factor-stimulated intestinal epithelial cell migration requires Src family kinase-dependent p38 MAPK signaling. *J Biol Chem 279*(43), 44513–44521.

Friedman, N. and D. Koller (2003). Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks. *Machine Learning 50*(3), 95–126.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze expression data. *J Comput Biol 7*(3-4), 601–620.

Friedman, N., K. Murphy, and S. Russell (1998, July). Learning the Structure of Dynamic Probabilistic Networks. In *Proceedings of the Proceedings of the Fourteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, Fourteenth Conf. on Uncertainty in Artificial Intelligence, San Francisco, CA, pp. 139–14. Morgan Kaufmann. Jul 24-26 1998, Madison, WI.

Fröhlich, H., M. Fellmann, H. Sültmann, A. Poustka, and T. Beißbarth (2008a, Jan). Estimating Large Scale Signaling Networks through Nested Effect Models with Intervention Effects from Microarray Data. *Bioinformatics 24*(22), 2650–2656.

Fröhlich, H., M. Fellmann, H. Sültmann, A. Poustka, and T. Beißbarth (2008b, Oct). Predicting pathway membership via domain signatures. *Bioinformatics 24*(19), 2137–2142.

Fröhlich, H., Özgür Sahin, D. Arlt, C. Bender, and T. Beißbarth (2009, Oct). Deterministic Effects Propagation Networks for Reconstructing Protein Signaling Networks from Multiple Interventions. *BMC Bioinformatics 322*(10).

Galperin, M. Y. (2008, Jan). The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res 36*(Database issue), D2–D4.

Gardner, T. S., D. di Bernardo, D. Lorenz, and J. J. Collins (2003, Jul). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science 301*(5629), 102–105.

Gat-Viks, I., A. Tanay, D. Raijman, and R. Shamir (2006, Mar). A probabilistic methodology for integrating knowledge and experiments on biological networks. *J Comput Biol 13*(37), 165–181.

Geier, F., J. Timmer, and C. Fleck (2007). Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst Biol 1*(51), 11.

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

Gillespie, D. and S. Spiegelman (1965, Jul). A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. *J Mol Biol 12*(3), 829–842.

Bibliography

Grzegorczyk, M. and D. Husmeier (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move, journal = Mach. Learn. *71*(2-3), 265–305.

Guo, H., T. Tian, K. Nan, and W. Wang (2010, Jun). p57: A multifunctional protein in cancer (Review). *Int J Oncol 36*(6), 1321–1329.

Hahne, F., A. Mehrle, D. Arlt, A. Poustka, S. Wiemann, and T. Beissbarth (2008). Extending pathways based on gene lists using InterPro domain signatures. *BMC Bioinformatics 9*, 3.

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

Hanahan, D. and R. A. Weinberg (2011, Mar). Hallmarks of cancer: the next generation. *Cell 144*(5), 646–674.

Harper, D. (2001, Nov). Online Etymology Dictionary. `http://www.etymonline.com/index.php?term=cell`. accessed Feb 14th, 2011.

Harper, J. W., G. R. Adami, N. Wei, K. Keyomarsi, and S. J. Elledge (1993, Nov). The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell 75*(4), 805–816.

Hatakeyama, M. (2007, Feb). System properties of ErbB receptor signaling for the understanding of cancer progression. *Mol Biosyst 3*(2), 111–116.

Heckerman, D. (1996). A tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research.

Henjes, F. (2010, Oct). *Analysis of ERBB signalling and the impact of targeted therapeutics using protein microarrays*. Ph. D. thesis, Ruprecht-Karls-Universität Heidelberg.

Hidalgo, M. (2003, Nov). Erlotinib: preclinical investigations. *Oncology (Williston Park) 17*(11 Suppl 12), 11–16.

Hoheisel, J. D. (2006, Mar). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet 7*(3), 200–210.

Huang, S. (1999, Jun). Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J Mol Med 77*(6), 469–480.

Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18 Suppl 1*(4), S96–104.

Hucka, M., A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. L. Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and S. B. M. L. Forum (2003, Mar). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics 19*(4), 524–531.

Hunter, L. and T. Klein (Eds.) (1998). *Reveal, a general reverse engineering algorithm for inference of genetic network architectures.* World Scientific Publishing.

Ideker, T., T. Galitski, and L. Hood (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet 2*, 343–372.

Imoto, S., T. Goto, and S. Miyano (2002). Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac Symp Biocomput*, 175–186.

Imoto, S., T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano (2004, Mar). Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J Bioinform Comput Biol 2*(58), 77–98.

International HapMap Consortium (2005, Oct). A haplotype map of the human genome. *Nature 437*(7063), 1299–1320.

Izumi, Y., L. Xu, E. di Tomaso, D. Fukumura, and R. K. Jain (2002, Mar). Tumour biology: herceptin acts as an anti-angiogenic cocktail. *Nature 416*(6878), 279–280.

Jin, S., T. Tong, W. Fan, F. Fan, M. J. Antinore, X. Zhu, L. Mazzacurati, X. Li, K. L. Petrik, B. Rajasekaran, M. Wu, and Q. Zhan (2002, Dec). GADD45-induced cell cycle G2-M arrest associates with altered subcellular distribution of cyclin B1 and is independent of p38 kinase activity. *Oncogene 21*(57), 8696–8704.

Jobs, M., W. M. Howell, L. Stromqvist, T. Mayr, and A. J. Brookes (2003, May). DASH-2: flexible, low-cost, and high-throughput SNP genotyping by dynamic allele-specific hybridization on membrane arrays. *Genome Res 13*(5), 916–924.

Johnson, J. M., J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker (2003, Dec). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science 302*(5653), 2141–2144.

Jones, J. T., R. W. Akita, and M. X. Sliwkowski (1999, Mar). Binding specificities and affinities of EGF domains for ErbB receptors. *FEBS Lett 447*(2-3), 227–231.

Jung, S.-H., S.-H. Shin, S.-H. Yim, H.-S. Choi, S.-H. Lee, and Y.-J. Chung (2009, Jul). Integrated analysis of copy number alteration and RNA expression profiles of cancer using a high-resolution whole-genome oligonucleotide array. *Exp Mol Med 41*(7), 462–470.

Kaderali, L., E. Dazert, U. Zeuge, M. Frese, and R. Bartenschlager (2009, Jun). Reconstructing Signaling Pathways from RNAi Data using Probabilistic Boolean Threshold Networks. *Bioinformatics*.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering 82*(Series D), 35–45.

Kamburov, A., C. Wierling, H. Lehrach, and R. Herwig (2009, Jan). ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res 37*(Database issue), D623–D628.

Kamimura, T. and H. Shimodaira (2005). A Scale-free Prior over Graph Structures for Bayesian Inference of Gene Networks. Online.

Bibliography

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res. 36*, D480 – D484.

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

Karaman, M. W., S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, and P. P. Zarrinkar (2008, Jan). A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol 26*(1), 127–132.

Kerrien, S., S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. J. Salama, S. Moore, A. Ceol, A. Chatr-Aryamontri, M. Oesterheld, V. Stümpflen, L. Salwinski, J. Nerothin, E. Cerami, M. E. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler, and H. Hermjakob (2007). Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol 5*, 44.

Kim, I.-Y., H.-Y. Yong, K. W. Kang, and A. Moon (2009, Mar). Overexpression of ErbB2 induces invasion of MCF10A human breast epithelial cells via MMP-9. *Cancer Lett 275*(2), 227–233.

Kim, J. W., S. S. Sim, U. H. Kim, S. Nishibe, M. I. Wahl, G. Carpenter, and S. G. Rhee (1990, Mar). Tyrosine residues in bovine phospholipase C-gamma phosphorylated by the epidermal growth factor receptor in vitro. *J Biol Chem 265*(7), 3940–3943.

Kitano, H. (2002a, Nov). Computational systems biology. *Nature 420*(6912), 206–210.

Kitano, H. (2002b, Mar). Systems biology: a brief overview. *Science 295*(5560), 1662–1664.

Klos, K. S., X. Zhou, S. Lee, L. Zhang, W. Yang, Y. Nagata, and D. Yu (2003, Oct). Combined trastuzumab and paclitaxel treatment better inhibits ErbB-2-mediated angiogenesis in breast carcinoma through a more effective inhibition of Akt than either treatment alone. *Cancer 98*(7), 1377–1385.

Korf, U., S. Derdak, A. Tresch, F. Henjes, S. Schumacher, C. Schmidt, B. Hahn, W. D. Lehmann, A. Poustka, T. Beissbarth, and U. Klingmüller (2008, Nov). Quantitative protein microarrays for time-resolved measurements of protein phosphorylation. *Proteomics 8*(21), 4603–4612.

Kraus, M. H., W. Issing, T. Miki, N. C. Popescu, and S. A. Aaronson (1989, Dec). Isolation and characterization of ERBB3, a third member of the ERBB/epidermal growth factor receptor family: evidence for overexpression in a subset of human mammary tumors. *Proc Natl Acad Sci U S A 86*(23), 9193–9197.

Kuhn, K., S. C. Baker, E. Chudin, M.-H. Lieu, S. Oeser, H. Bennett, P. Rigault, D. Barker, T. K. McDaniel, and M. S. Chee (2004, Nov). A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res 14*(11), 2347–2356.

Lahoz, E. G., N. J. Liegeois, P. Zhang, J. A. Engelman, J. Horner, A. Silverman, R. Burde, M. F. Roussel, C. J. Sherr, S. J. Elledge, and R. A. DePinho (1999, Jan). Cyclin D- and E-dependent kinases and the p57(KIP2) inhibitor: cooperative interactions in vivo. *Mol Cell Biol 19*(1), 353–363.

Lébre, S. (2009, Jan). Inferring dynamic genetic networks with low order independencies. *Stat Appl Genet Mol Biol 8*(1).

Lee, D.-S., K.-I. Goh, B. Kahng, and D. Kim (2005, June). Scale-free random graphs and Potts model. *Pramana - journal of physics 64*(6), 1149–1159.

Lenferink, A. E., D. Busse, W. M. Flanagan, F. M. Yakes, and C. L. Arteaga (2001, Sep). ErbB2/neu kinase modulates cellular p27(Kip1) and cyclin D1 through multiple signaling pathways. *Cancer Res 61*(17), 6583–6591.

Li, C. and J. Jin (2010). DNA replication licensing control and rereplication prevention. *Protein & Cell 1*, 227–236. 10.1007/s13238-010-0032-z.

Li, M., C. Balch, J. S. Montgomery, M. Jeong, J. H. Chung, P. Yan, T. H. M. Huang, S. Kim, and K. P. Nephew (2009). Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Med Genomics 2*, 34.

Lockhart, D. J. and E. A. Winzeler (2000, Jun). Genomics, gene expression and DNA arrays. *Nature 405*(6788), 827–836.

Loebke, C., H. Sueltmann, C. Schmidt, F. Henjes, S. Wiemann, A. Poustka, and U. Korf (2007, Feb). Infrared-based protein detection arrays for quantitative proteomics. *Proteomics 7*, 558–564.

Luo, S., N. B. Wehr, and R. L. Levine (2006, Mar). Quantitation of protein on gels and blots by infrared fluorescence of Coomassie blue and Fast Green. *Anal Biochem 350*(2), 233–238.

Luttrell, D. K., A. Lee, T. J. Lansing, R. M. Crosby, K. D. Jung, D. Willard, M. Luther, M. Rodriguez, J. Berman, and T. M. Gilmer (1994, Jan). Involvement of pp60c-src with two major signaling pathways in human breast cancer. *Proc Natl Acad Sci U S A 91*(1), 83–87.

Madigan, D., J. York, and D. Allard (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique 63*(2), 215–232.

Magnifico, A., L. Albano, S. Campaner, M. Campiglio, S. Pilotti, S. Ménard, and E. Tagliabue (2007, Jun). Protein kinase Calpha determines HER2 fate in breast carcinoma cells with HER2 protein overexpression without gene amplification. *Cancer Res 67*(11), 5308–5317.

Manning, B. D. and L. C. Cantley (2007, Jun). AKT/PKB signaling: navigating downstream. *Cell 129*(7), 1261–1274.

Mao, W., R. Irby, D. Coppola, L. Fu, M. Wloch, J. Turner, H. Yu, R. Garcia, R. Jove, and T. J. Yeatman (1997, Dec). Activation of c-Src by receptor tyrosine kinases in human colon cancer cells with high metastatic potential. *Oncogene 15*(25), 3083–3090.

Markowetz, F. (2010, 02). How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLoS Comput Biol 6*(2), e1000655.

Markowetz, F., J. Bloch, and R. Spang (2005, Nov). Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics 21*(21), 4026–4032.

Bibliography

Molina, M. A., J. Codony-Servat, J. Albanell, F. Rojo, J. Arribas, and J. Baselga (2001, Jun). Trastuzumab (herceptin), a humanized anti-Her2 receptor monoclonal antibody, inhibits basal and activated Her2 ectodomain cleavage in breast cancer cells. *Cancer Res 61*(12), 4744–4749.

Moses, H. L., E. Y. Yang, and J. A. Pietenpol (1990, Oct). TGF-beta stimulation and inhibition of cell proliferation: new mechanistic insights. *Cell 63*(2), 245–247.

Mukherjee, S. and T. P. Speed (2008, Sep). Network inference using informative priors. *Proc Natl Acad Sci U S A 105*(38), 14313–14318.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. O. nd Robert Petryszak, J. D. Selengut, C. J. A. Sigrist, P. D. Thomas, F. V. nd Derek Wilson, C. H. Wu, and C. Yeats (2008). New developments in the InterPro database. *Nucleic Acids Res. 35*, D224 – D228.

Murphy, K. and S. Mian (1999). Modelling gene expression data using dynamic bayesian networks. Technical report, University of California, Berkeley.

Nahta, R. and F. J. Esteva (2007, May). Trastuzumab: triumphs and tribulations. *Oncogene 26*(25), 3637–3643.

Nelander, S., W. Wang, B. Nilsson, Q.-B. She, C. Pratilas, N. Rosen, P. Gennemark, and C. Sander (2008). Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol 4*, 216.

Nicholson, R. I., J. M. Gee, and M. E. Harper (2001, Sep). EGFR and cancer prognosis. *Eur J Cancer 37 Suppl 4*, S9–15.

Nielsen, U. B., M. H. Cardone, A. J. Sinskey, G. MacBeath, and P. K. Sorger (2003, Aug). Profiling receptor tyrosine kinase activation by using Ab microarrays. *Proc Natl Acad Sci U S A 100*(16), 9330–9335.

Nishitani, H. and Z. Lygerou (2002, Jun). Control of DNA replication licensing in a cell cycle. *Genes Cells 7*(6), 523–534.

Oda, K., Y. Matsuoka, A. Funahashi, and H. Kitano (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol 1*, 2005.0010.

Olayioye, M. A., I. Beuvink, K. Horsch, J. M. Daly, and N. E. Hynes (1999, Jun). ErbB receptor-induced activation of stat transcription factors is mediated by Src tyrosine kinases. *J Biol Chem 274*(24), 17209–17218.

Ooi, H. S., G. Schneider, T.-T. Lim, Y.-L. Chan, B. Eisenhaber, and F. Eisenhaber (2010). Biomolecular pathway databases. *Methods Mol Biol 609*, 129–144.

Oren, M. (2003, Apr). Decision making by p53: life, death and cancer. *Cell Death Differ 10*(4), 431–442.

Park, M. and S. Lee (2003, Jan). Cell cycle and cancer. *J Biochem Mol Biol. 36*(1), 60–65.

Paweletz, C. P., L. Charboneau, V. E. Bichsel, N. L. Simone, T. Chen, J. W. Gillespie, M. R. Emmert-Buck, M. J. Roth, E. F. P. III, and L. A. Liotta (2001, Apr). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene 20*(16), 1981–1989.

Pearl, J. (1988, September). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann.

Pe'er, D., A. Regev, G. Elidan, and N. Friedman (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics 17 Suppl 1*, S215–S224.

Perou, C. M., T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein (2000, Aug). Molecular portraits of human breast tumours. *Nature 406*(6797), 747–752.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson (1998, Oct). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet 20*(2), 207–211.

Pleasance, E. D., R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton (2010, Jan). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature 463*(7278), 191–196.

Plowman, G. D., J. M. Culouscou, G. S. Whitney, J. M. Green, G. W. Carlton, L. Foy, M. G. Neubauer, and M. Shoyab (1993, Mar). Ligand-specific activation of HER4/p180erbB4, a fourth member of the epidermal growth factor receptor family. *Proc Natl Acad Sci U S A 90*(5), 1746–1750.

Plowman, G. D., G. S. Whitney, M. G. Neubauer, J. M. Green, V. L. McDonald, G. J. Todaro, and M. Shoyab (1990, Jul). Molecular cloning and expression of an additional epidermal growth factor receptor-related gene. *Proc Natl Acad Sci U S A 87*(13), 4905–4909.

Plyte, S. E., K. Hughes, E. Nikolakaki, B. J. Pulverer, and J. R. Woodgett (1992, Dec). Glycogen synthase kinase-3: functions in oncogenesis and development. *Biochim Biophys Acta 1114*(2-3), 147–162.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rau, A., F. Jaffrèzic, J.-L. Foulley, and R. W. Doerge (2010). An Empirical Bayesian Method for Estimating Biological Networks from Temporal Microarray Data. *Statistical Applications in Genetics and Molecular Biology 9*.

Sachs, K., O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan (2005, Apr). Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*(5721), 523–529.

Bibliography

Sancak, Y., C. C. Thoreen, T. R. Peterson, R. A. Lindquist, S. A. Kang, E. Spooner, S. A. Carr, and D. M. Sabatini (2007). PRAS40 Is an Insulin-Regulated Inhibitor of the mTORC1 Protein Kinase. *Molecular Cell 25*(6), 903 – 915.

Sanchez, Y., C. Wong, R. S. Thoma, R. Richman, Z. Wu, H. Piwnica-Worms, and S. J. Elledge (1997, Sep). Conservation of the Chk1 checkpoint pathway in mammals: linkage of DNA damage to Cdk regulation through Cdc25. *Science 277*(5331), 1497–1501.

Schaefer, C. F., K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow (2009, Jan). PID: the Pathway Interaction Database. *Nucleic Acids Res 37*(Database issue), D674–D679.

Schafer, K. A. (1998, Nov). The cell cycle: a review. *Vet Pathol 35*(6), 461–478.

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

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Segal, E., D. Pe'er, A. Regev, D. Koller, and N. Friedman (2005). Learning module networks. *Journal of Machine Learning Research 6*(25), 557–588.

Seger, R. and E. G. Krebs (1995). The MAPK signaling cascade. *FASEB journal 9*(9), 726–735.

Sheridan, P., T. Kamimura, and H. Shimodaira (2010). A scale-free structure prior for graphical models with applications in functional genomics. *PLoS One 5*(11), e13580.

Sherr, C. J. and J. M. Roberts (1999, Jun). CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes Dev 13*(12), 1501–1512.

Shigematsu, H., T. Takahashi, M. Nomura, K. Majmudar, M. Suzuki, H. Lee, I. I. Wistuba, K. M. Fong, S. Toyooka, N. Shimizu, T. Fujisawa, J. D. Minna, and A. F. Gazdar (2005, Mar). Somatic mutations of the HER2 kinase domain in lung adenocarcinomas. *Cancer Res 65*(5), 1642–1646.

Slamon, D. J., G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire (1987, Jan). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science 235*(4785), 177–182.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol 3*(40), Article3.

Southern, E. M., S. C. Case-Green, J. K. Elder, M. Johnson, K. U. Mir, L. Wang, and J. C. Williams (1994, Apr). Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Res 22*(8), 1368–1373.

Spieth, C., R. Worzischek, and F. Streichert (2006). Comparing evolutionary algorithms on the problem of network inference. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, GECCO '06, New York, NY, USA, pp. 305–306. ACM.

Steele, E., A. Tucker, P. 't Hoen, and M. Schuemie (2009, April). Literature-based priors for gene regulatory networks. *Bioinformatics 25*(14), 1768–1774.

Sørlie, T. (2004, Dec). Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur J Cancer 40*(18), 2667–2675.

Sørlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale (2001, Sep). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A 98*(19), 10869–10874.

Tai, Y. C. and T. P. Speed (2006). A multivariate empirical bayes statistic for replicated microarray time course data. *The Annals of Statistics 34*(5), 2387–2412.

Tedford, N. C., A. B. Hall, J. R. Graham, C. E. Murphy, N. F. Gordon, and J. A. Radding (2009, Mar). Quantitative analysis of cell signaling and drug action via mass spectrometry-based systems level phosphoproteomics. *Proteomics 9*(6), 1469–1487.

Tegner, J., M. K. S. Yeung, J. Hasty, and J. J. Collins (2003, May). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci U S A 100*(10), 5944–5949.

The Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research 32*, D258–D261.

Timms, J. F., S. L. White, M. J. O'Hare, and M. D. Waterfield (2002, Sep). Effects of ErbB-2 overexpression on mitogenic signalling and cell cycle progression in human breast luminal epithelial cells. *Oncogene 21*(43), 6573–6586.

Tusher, V. G., R. Tibshirani, and G. Chu (2001, Apr). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A 98*(9), 5116–5121.

Ullrich, A., L. Coussens, J. S. Hayflick, T. J. Dull, A. Gray, A. W. Tam, J. Lee, Y. Yarden, T. A. Libermann, and J. Schlessinger (1984). Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. *Nature 309*(5967), 418–425.

Valabrega, G., F. Montemurro, and M. Aglietta (2007, Jun). Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. *Ann Oncol 18*(6), 977–984.

Vanhaesebroeck, B., S. J. Leevers, G. Panayotou, and M. D. Waterfield (1997, Jul). Phosphoinositide 3-kinases: a conserved family of signal transducers. *Trends Biochem Sci 22*(7), 267–272.

Vastrik, I., P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol 8*(3), R39.

Vogelstein, B. and K. W. Kinzler (2004, Aug). Cancer genes and the pathways they control. *Nat Med 10*(8), 789–799.

Weinberg, R. A. (2007a). *The biology of cancer*, Chapter 8: pRb and Control of the Cell Cycle Clock. Garland Science, Taylor & Francis Group, LLC.

Weinberg, R. A. (2007b). *The biology of cancer*, Chapter 6: Cytoplasmic Signaling Circuitry Programs Many of the Traits of Cancer. Garland Science, Taylor & Francis Group, LLC.

Bibliography

Werhli, A. V. (2007). *Reconstruction of gene regulatory networks from postgenomic data*. Ph. D. thesis, University of Edinburgh, Institute for Adaptive and Neural Computation, School of Informatics.

Werhli, A. V. and D. Husmeier (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol 6*, Article15.

World Health Organization Databank (2010). WHO Statistical Information System. http://www.who.int/whosis.

Xian, W., M. P. Rosenberg, and J. DiGiovanni (1997, Mar). Activation of erbB2 and c-src in phorbol ester-treated mouse epidermis: possible role in mouse skin tumor promotion. *Oncogene 14*(12), 1435–1444.

Yakes, F. M., W. Chinratanalab, C. A. Ritter, W. King, S. Seelig, and C. L. Arteaga (2002, Jul). Herceptin-induced inhibition of phosphatidylinositol-3 kinase and Akt Is required for antibody-mediated effects on p27, cyclin D1, and antitumor action. *Cancer Res 62*(14), 4132–4141.

Yamamoto, T., S. Ikawa, T. Akiyama, K. Semba, N. Nomura, N. Miyajima, T. Saito, and K. Toyoshima (1986). Similarity of protein encoded by the human c-erb-B-2 gene to epidermal growth factor receptor. *Nature 319*(6050), 230–234.

Yu, J., V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis (2004, Dec). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics 20*(18), 3594–3603.

Zhang, J. D. and S. Wiemann (2009, Jun). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics 25*(11), 1470–1471.

List of publications

Henjes, F., Bender, C., Mannsperger, HA., von der Heyde, S., Schmidt, C., Szabó, V., Tschulena, U., Wiemann, S., Hasmann, M., Beißbarth, T. and Korf, U. (2011, Jan). Impact of EGFR and ERBB2 targeting drugs on signaling networks in ERBB2 positive breast cancer cell lines. *under revision.*

Bender, C., Heyde, S. v.d., Henjes, F., Wiemann, S., Korf, U. and Beißbarth, T. (2011, Mar). Inferring signalling networks from longitudinal data using sampling based approaches in the R package 'ddepn'. *under review.*

Bender, Christian; Henjes, Frauke; Fröhlich, Holger; Wiemann, Stefan; Korf, Ulrike and Beißbarth, Tim (2010, Sep). Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics 26 (18)*, i596–i602.

Fröhlich, H.; Sahin, Ö.; Arlt, D.; Bender, C.; Beißbarth, T. (2009, Oct). Deterministic Effects Propagation Networks for Reconstructing Protein Signaling Networks from Multiple Interventions. *BMC Bioinformatics, 322 (10).*

Bender, Christian; Fröhlich, Holger; Johannes, Marc & Beißbarth, Tim (2008, Dec). Extending pathways with inferred regulatory interactions from microarray data and protein domain signatures. In *CAMDA: Critical Assessment of Microarray Data Analysis*, pp 48-52.

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder ähnlicher Form bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

. . . . . .
Ort, Datum

. . . . . .
Unterschrift