AUS DEM LEHRSTUHL FÜR STATISTISCHE BIOINFORMATIK
Prof. DR. RAINER SPANG
DER FAKULTÄT FÜR MEDIZIN
DER UNIVERSITÄT REGENSBURG

Considering Unknown Unknown: Reconstruction of Non-confoundable Causal Relations In Biological Networks

Inaugural – Dissertation
zur Erlangung des Doktorgrades
der Biomedizinischen Wissenschaften

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von

Mohammad Javad Sadeh

im Jahr
2012

AUS DEM LEHRSTUHL FÜR STATISTISCHE BIOINFORMATIK
Prof. DR. RAINER SPANG
DER FAKULTÄT FÜR MEDIZIN
DER UNIVERSITÄT REGENSBURG

Considering Unknown Unknown: Reconstruction of Non-confoundable Causal Relations In Biological Networks

Inaugural – Dissertation
zur Erlangung des Doktorgrades
der Biomedizinischen Wissenschaften

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von

Mohammad Javad Sadeh

im Jahr
2012

Dekan:	Prof. Dr. Dr. Torsten E. Reichert
Betreuer:	Prof. Dr. Rainer Spang
Tag der mündlichen Prüfung:	22.04.2013

Abstract

Our current understanding of cellular networks is rather incomplete. We miss important but so far unknown genes and mechanisms in the pathways. Moreover, we often only have a partial account of the molecular interactions and modifications of the known players. When analyzing the cell, we look through narrow windows leaving potentially important events in blind spots. This might severely bias both the computational and manual reconstruction of underlying biological networks. Network reconstruction is naturally confined to what we have observed. Little is known about how the incompleteness of our observations confounds our interpretation of the available data. In this dissertation I ask the question: Which features of a network can be confounded by incomplete observations and which cannot? In order to answer this question, I first summarize the methodology of Nested Effects Models (NEMs) proposed by Markowetz *et al.* to reconstruct non-transcriptional networks using subset relationships from perturbation data and bring out its limitation to model biological processes in the presence of hidden mechanisms. In the context of Nested Effects Models, I show that in the presence of missing observations or hidden factors a reliable reconstruction of the full network is not feasible. Nevertheless, I show that certain characteristics of signaling networks like the existence of cross talk between certain branches of the network can be inferred in a non-confoundable way. I introduce and describe new statistical methodologies called Non-confoundable Neteorks Analysis (No-CONAN) and Partial Nested Effects Models (pNEM) for analyzing cell signaling pathways and gene expression. No-CONAN is based on a simple polynomial test for inferring non-confoundable characteristics of signaling networks. I then introduce a new data structure to represent partially reconstructed signaling networks. Finally, I evaluate the methods presented in this dissertation on simulated data and two biological studies, a first application to embryonic stem cell differentiation in mice and a recent study on the Wnt signaling pathway in colorectal cancer cells. I demonstrate that taking unknown hidden mechanisms into account changes our account of real biological networks.

Zusammenfassung

Unser derzeitiges Verständnis zellulärer Netzwerke ist unvollständig. Wir übersehen derzeit wichtige aber bisher unbekannte Gene und Mechanismen der Signalwege. Darüber hinaus ist uns oft nur ein Ausschnitt der molekularen Interaktionen und Modifikationen bekannt. Bildlich betrachten wir Zellen nur durch ein kleines Fenster und übersehen dadurch wichtige Vorgänge. Sowohl die Computer gestützte also auch die manuelle Rekonstruktion des zugrunde liegende biologischen Netzwerks wir dadurch möglicherweise stark verfälscht. Die Rekonstruktion von Netzwerken ist naturgemäßauf unsere Beobachtungen limitiert. Inwieweit die Unvollständigkeit unserer Beobachtungen die mögliche Interpretation der vorhandenen Daten beeinflusst ist weitestgehend unbekannt. In dieser Arbeit möchte ich die Frage beantworten, welche Merkmale eines Netzwerks durch unvollstndige Daten beeinflusst werden können und welche nicht. Dazu fasse ich zunächst Nested Effects Models (NEMs) von Markowetz *et al.* zur Rekonstruktion nicht-transkriptionaler Netzwerke zusammen, die auf Teilmengen Beziehungen in Daten aus Perturbatiosexperimenten basiert. Dabei arbeite ich die Grenzen von NEMs im Bezug auf unbeobachtete Mechanismen heraus. In diesem Kontext zeige ich, dass in Gegenwart von fehlenden Beobachtungen oder Faktoren die zuverlässige Rekonstruktion des vollständigen Netzwerks nicht möglich ist. Nichts desto trotz zeige ich, dass bestimmte Charakteristika von Signal-Netzwerken wie z.B. die Wechselwirkungen bestimmter Zweige des Netzwerks auch unter Berücksichtigung von confounders eindeutig abgeleitet werden können. Ich führe 'Non-confoundable Network Analysis (No-CONAN)' und 'Partial Nested Effects Models (pNEM)' zur Analyse von Zell Signalwegen und beschreibe diese. No-CONAN und pNEM basieren auf einem einfachen polynomial Test zur Störfaktors unabhängigen Ableitung der Charateristik von Signal-Netzwerken. Anschließend führe ich eine neue Datenstruktur zur Darstellung partiell rekonstruierter Signal-Netzwerke ein. Schlussendlich evaluiere ich die in dieser Arbeit eingeführten Methodologien auf simulierten Daten und wende sie auf Daten zweier biologischer Studien an: einer Analyse der Differenzierung embryonischer Stammzellen in Mäusen, sowie eine kürzlich erfolgten Arbeit zur Analyse des Wnt Signalweges in Zellen des Kolorektalen Karzinoms. Dabei zeige ich, wie das Einbeziehen unbekannter Mechanismen unsere Sichteweise auf echte biologische Netzwerke verändert.

To my Parents

Acknowledgements

This work was carried out in the Institute of Functional Genomics and Bioinformatics in the department of Statistical Bioinformatics, all part of the University of Regensburg. It was supported by BMBF grants (EraSys:0315714B) and the Bavarian Genomenetwork BayGene. I would like to thank all past and present colleagues for their moral, friendly and technical support. Especially, I am grateful to my supervisor *Rainer Spang* for giving me the opportunity to work in his group, suggesting the topic, his scientific support, and the opportunity to write this thesis under his guidance. Further thanks goes to *Achim Tresch* and *Peter Oefner* for their supervision and counsel during my PhD work. I thank *Florian Markowetz* for introducing me to the world of NEMs when I visited him at IPM institute in Iran. During the time I worked on this thesis, I enjoyed fruitful discussions with many people. In particular, I gratefully acknowledge *Benedict Anchang*, *Christian Hundsrucker*, *Claudio Lottaz*, *Matthias Maneck*, *Julia C. Engelmann*, *Katharina Meyer*, *Giusi Moffa* and *Daniela Herold*.

Above all, I am exceptionally grateful to my girlfriend Sepideh, my parents and my sisters for their relentless support and love.

Publications Parts of this thesis have been published or accepted to peer-reviewed journals. The main content of Chapter 4 and 5 have accepted to *RECOMB*.

Mohammad Javad Sadeh, Giusi Moffa and Rainer Spang. *Considering Unknown Unknowns - Reconstruction of Non-confoundable Causal Relations in Biological Networks.* Accepted for RECOMB2013.

My contribution to chapter 2 have been published in the Proceedings of National Academy of Science (PNAS).

Benedict Anchang, Mohammad J. Sadeh, Juby Jacob, Achim Tresch, Marcel O. Vlad, Peter J. Oefner, and Rainer Spang. *Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models*. Proc Natl Acad Sci USA, 106(16):6447-6452, Apr 2009.

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

GLOSSARY

Glossary

α	Probability to observe a false positive effect
β	Probability to observe a false negative effect
κ	Calibration parameter
Θ	Parameter for E-gene positions
$\theta_k = i$	Position parameter for E-gene E_k linked to S_i
D	Perturbation data matrix
D_{ikls}	Observed E-gene expression level E_k after perturbation S_i for replicate experiment l in time point t_s. Sometimes also represented as e_{ikls}
E	Effect genes known as E-genes. Also known as observables O
K	Set of rate parameters between S-S-genes and S-E-genes
S	Signaling genes known as S-genes
APC	adenomatosis polyposis
Axin1	axin inhibition protein-1
BIC	Bayesian information criterion; a criterion for model selection among a class of parametric models with different numbers of parameters
CK1α	casein kinase 1
CSAN	Current State of Art Network
DAG	Directed acyclic graph
DNA	deoxyribonucleic acid
DNEMs	Dynamic Nested Effects Models
Dsh	dishevelled

GLOSSARY

EM Expectation-Maximization; an algorithm for solving incomplete data problems

ESC Embryonic stem cells

FDNEMs Fast Dynamic Nested Effects Models

FN False negative

FP False positive

Fz frizzled

GSK3β glycogen synthase kinase β

GTN Ground True Network

LIF Leukemia Inhibitory Factor; a cytokine that affects cell growth and development

MAP Maximum a posteriori probability. The MAP estimate is a mode of the posterior distribution

MC EMiNEM Monte Carlo Expectation-Maximization Nested Effects Models

MCMC Markov chain Monte Carlo; a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution

mRNA messenger RNA

NEMs Nested Effects Models; a class of models for the analysis of non-transcriptional signalling networks. NEMs infer the graph of upstream/downstream relations for a set of signalling genes from perturbation effects.

No-CONAN Non-confoundable Network Analysis

pNEM Partial Nested Effects Models

RA Retinoic Acid; is a metabolite of vitamin A (retinol) that mediates the functions of vitamin A required for growth and development

RNA Ribonucleic acid

RNA ribonucleic acid

RNAi RNA interference

SEM Structural EM

TCF T-cell specific transcription factors

TFs Transcription factors

TN True negative

TP True positive

1

Introduction

This thesis is concerned with inferring signaling networks from interventional data in the presence of unknown hidden mechanisms. I develop methodology to address the following problem: The current understanding of biological networks can be confounded by hidden mechanisms that we are not aware of. The first chapter is divided into two parts. First, I discuss the statistical motivation of reconstructing a network in the presence of hidden variables. I describe statistical models to reconstruct networks from incomplete data (section 1.1). The second part gives a concise background on gene regulation and cell signaling and explains the experimental technique of RNA interference (RNAi). I then discuss the biological motivation of analyzing networks with possibly many hidden mechanisms involved (section 1.2). In recent years there has been growing interest in reconstructing biological networks from data, but no work has been done so far to adapt statistical methodology for very incomplete data.

1.1 Complex phenomena are attracting scientists

Complex phenomena are attracting scientists around the world. There are many reasons why complexity is a popular topic of research. The main one seems to be very simple and says "we live in a complex world". Examples of complex systems include the stock market, social insect and ant colonies, manufacturing businesses and social system such as political parties or communities. However, complex phenomena not only exist in our environment but also inside us. Because of our brains with millions of neurons and many more neuronal connections and cell structures with many components, we are probably the most complex systems in the universe. Therefore by understanding complex phenomena we can better understand ourselves.

Studying complexity can be surprising at least in two ways. The first way is connected with the moment of discovery. When scientists find something new they find it

surprising. However the phenomenon of surprise is not only restricted to scientists who study something new. In the second way, complexity can be surprising due to its nature. For instance, when changing minor details in a system has major impacts on its global behavior. Real-life problems are further complicated by the fact that we are usually given only a partial view of the world. For instance, a poker player will have to bet and win the game without necessarily knowing the cards in the other player's hands. He will never have access to his opponent's strategy, but will constantly try to infer it from the opponents moves and history of games. As a matter of fact, in making decisions for a domain of interest, an influencing factor may be hidden and never be observed. These hidden factors are all over the place in real-life domains. They often play the central role in hidden mechanisms and influence many of the observations. This rises the following questions: how can the hidden factors influence our decision making? What of our current understanding of complex systems can be confounded by hidden factors? It is the goal of this dissertation, to answer these questions when our current knowledge about the complex system is not complete.

1.1.1 Probabilistic Graphical Models

In order to deal with the challenges outlined above, we rely heavily on our life experience. For example, if the doctors suspect that a patient has cancer based on his symptoms or the results of a screening test or clinical examination, they will order certain diagnostic tests to find out whether the patient has abnormal cells and, if so, whether they are cancerous or non-cancerous. If the tests show that the cells are cancerous, they may need more tests to find out more about the cancer cells. All these real-life examples motivated researchers to define new field of research called machine learning.

Machine learning Machine Learning is a scientific discipline that addresses the following question: How can we define frameworks to automatically learn and to improve with experience? Learning in this context is not learning by heart but recognizing complex patterns and making decisions based on data. The difficulty lies in the fact that the set of all possible decisions given all possible inputs is too complex to describe. To tackle this problem the field of Machine Learning develops algorithms that discover knowledge from data and experience, based on sound statistical and computational principles (1). Therefore, the aim is to learn from example, just as the baby applies its inherited skills to understand its environment. A central paradigm in Machine Learning is that of probabilistic graphical models (2, 3) that have become popular in recent years, and are being used in numerous applications (4).

Probabilistic graphical models Probabilistic graphical models (PGMs) provide a natural tool for dealing with two problems specific to real-world applications, uncertainty and complexity. Uncertainty is unavoidable in real-world applications and

we can almost never predict with certainty what will happen in the future. Many important aspects of the world are not observed with certainty. Probability theory gives us the basic foundation to model our beliefs about the different possible states of the world, and to update these beliefs as new evidence is obtained. These beliefs can be combined with individual preferences to help guide our actions, and even in selecting which observations to make. The PGM framework uses ideas from discrete data structures to efficiently encode and manipulate probability distributions. These methods have been used in an enormous range of application domains, which include: medical diagnosis, image understanding, reconstruction of biological networks and many more. This framework provides an essential tool for learning how to reason coherently from limited and noisy observations (5).

Probabilistic graphical models compactly represent a joint distribution over a set of variables in a domain of interest, and facilitate the efficient computation of flexible probabilistic queries. A graph is comprised of nodes connected by links (also known as edges or arcs). In a probabilistic graphical model, each node represents a random variable, and the links express probabilistic relationships between these variables. The graph captures the way in which the joint distribution over all of the random variables can be decomposed into a product of factors each depending only on a subset of the variables. Generally, probabilistic graphical models use a graph-based representation as the foundation for encoding a complete distribution over a multi-dimensional space and a graph that is a compact or factorized representation of a set of independences that hold in the specific distribution. Three main branches of graphical representations of distributions commonly used, are, Bayesian networks, Markov random fields and Factor graph (1).

- **Bayesian networks** Also known as directed graphical models, in which the links of the graphs have a particular directionality indicated by arrows. Bayesian networks are directed acyclic graphs whose nodes represent random variables and edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Directed graphs are useful for expressing causal relationships between random variables.

- **Markov random fields** The other major class of graphical models are Markov random fields, also known as undirected graphical models, in which the links do not carry arrows and have no directional significance. A Markov random field is similar to a Bayesian network in its representation of dependencies; the differences being that Bayesian networks are directed and acyclic, whereas Markov networks are undirected and may be cyclic. Thus, a Markov network can represent certain dependencies that a Bayesian network cannot (such as cyclic dependencies).

- **Factor graph** For the purposes of solving inference problems, it is often convenient to convert both directed and undirected graphs into a different representation called a factor graph. Both directed and undirected graphs allow a global function of several variables to be expressed as a product of factors over subsets of those variables. Factor graphs make this product explicit by introducing additional nodes for the factors themselves in addition to the nodes representing the variables.

Probabilistic graphical models in general consist of two components. The first is a graph in which each vertex corresponds to a random variable. This graph represents a set of conditional independence properties of the represented distribution. The graph captures the structure of the probability distribution, and is exploited for efficient inference and decision making. The second component is a collection of local interaction models that describe the conditional probability of each variable given its parents in the graph. Together, these two components represent a unique probability distribution (2).

If the network structure of the model is a directed acyclic graph, the model represents a factorization of the joint probability of all random variables. More precisely, if the events are $X_1, ..., X_n$ then the joint probability satisfies

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa_i) \tag{1.1}$$

where pa_i is the set of parents of node X_i in the graph. In other words, the joint distribution factors into a product of conditional distributions over the variables in the domain. For example, Figure 1.1A shows the structure of a simple directed graphical model for a cancer domain (5, 6). It is easy to conclude from the graph that the relation between smoking and the appearance of lumps in an x-ray is mediated by the cancer node. It is also easy to see that cancer is simultaneously effected by several possible direct causes. In contrast, cancer is the only direct influencing factor on indigestion. The distribution represented by our simple structure of the cancer domain consists of the random variables smoking (S), exposure (E), alcohol (A), cancer (C), lumps (L), indigestion (I) and bleeding (B) can be written as

$$P(C, S, E, A, L, B, I) = P(S)\cdot P(E)\cdot P(A)\cdot P(C|S, E, A)\cdot P(L|C)\cdot P(B|C)\cdot P(I|C) \tag{1.2}$$

The unique advantage of the factorization of the distribution gives probabilistic graphical models a compact representation of the joint distribution. The compact representation facilitates efficient probabilistic computations (2, 7). Given a joint distributions,

a central task of interest is that of inference, or answering probabilistic queries (6). For example we might want to examine the influence of one factor on another to quantify the value of future decisions. All these tasks are typically intractable even for small domains if the joint distribution is naively represented. While inference in general graphical models is NP-hard (8), the compact representation of the distribution allows us to compute varied probabilistic queries for relatively large and complex domains.

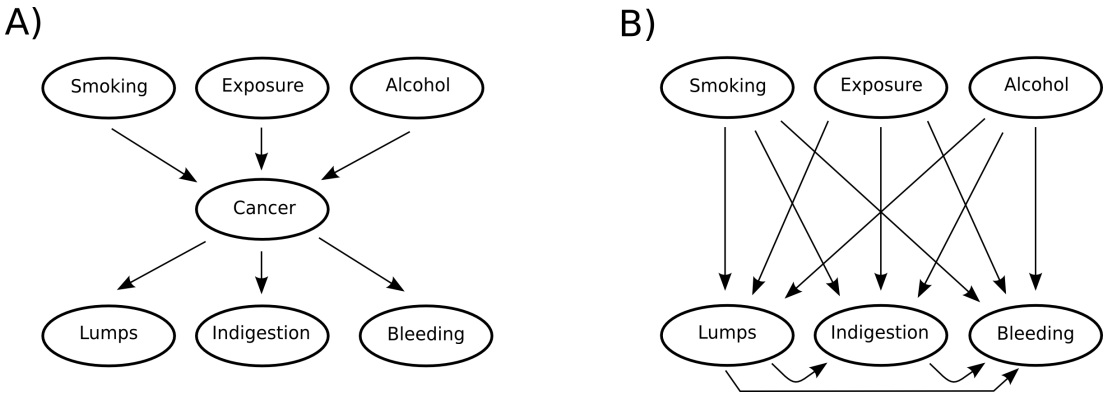

Figure 1.1: Plausible network for cancer domain - **A** shows a simple structure for cancer and its cause and effects. In this model cancer separates its causes from symptoms. **B** shows the resulting structure when cancer is removed from the model and is no longer able to mediate between its causes and symptoms.

.

Maximum likelihood estimation In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters. In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects values of the model parameters that produce a distribution that gives the observed data the greatest probability.

In recent years there has been a growing interest in learning baysian networks from data (9, 10, 11). Current methods are successful at learning both the structure and parameters from complete data. The complete data describes the values of all variables in the network. The main challenge in this case is to learn the structure of the network. The number of possible structures depends on the number of variables exponentially. In order to search the space of all possible structures, heuristic greedy procedures are typically used. However, these heuristic procedures can get trapped in local maxima. The situation is different when the data is incomplete. When

some observations are missing or some variables are altogether unobserved, learning is significantly harder: local maxima often trap the learning procedure and lead to infer models incorrectly.

1.1.2 Hidden Variables

When we consider real-life domains, we also have to cope with the fact that the size of the problem may limit our ability to learn an effective model in practice. Consequently, much of the research in recent years has been directed at learning probabilistic graphical models in complex scenarios where some of the data may be missing (12, 13). In case where data are incomplete or partially observed, they contain hidden variables whose value is never observed. In contrast to observed variables, the hidden variables are not known to be part of the domain.

Why should we bother with hidden variables that are never observed? These variables may play an important role in the model, and therefore they may be critical for our understanding of the domain. Consider again the model of the cancer domain shown in Figure 1.1A. This simple model encodes the fact that an observation of the cancer node separates possible causes (smoking, exposure to sunlight, excessive consumption of alcohol) from a few plausible symptoms (lumps, bleeding, indigestion). Now imagine cancer is a hidden variable in this simple model. If we remove cancer influences on other nodes, we might be able to recognize a correlation between smoking and the appearance of lumps. We might also be able to deduced a relation between repeated bleeding and indigestion. Considering these correlation we may end up with a model similar to the one shown in Figure 1.1B.

There are many reasons why the true model in Figure 1.1A is more appealing than the one where cancer is hidden. First, it can tell us more about the structure domain, particularly the way that different variables influences each other. Second, the representation of the domain is more compact in the true model. Since most of the nodes are connected to most of the others in Figure 1.1B, this structure is significantly more complex than the true model. As it is clear in the above example, the inclusion of a hidden variable in the network can simplify the structure, reducing the complexity of the network that needs to be learned (5). This motivates the learning of the hidden variables in a case of incomplete data.

1.1.3 Learning Bayesian Networks in the presence hidden variables

Learning hidden variables in the context of Bayesian networks has at least two advantages. First, learning these variables effectively can result in a succinct model for representing the distribution over the known entities, which in turn facilitates efficient inference and robust estimation. By introducing hidden variables that do not appear explicitly in the model we can often learn simpler models. Second, by learning new hidden variables, we can improve our understanding of the domain, potentially revealing important hidden entities. The importance of incorporating hidden variables in the model was recognized early on in the probabilistic graphical models community (e.g., (14, 15)).

When a hidden variable is known to exist, we can introduce it into the network and apply known Bayesian networks learning algorithms. If the network structure is known, algorithms such as *EM* (16, 17) or *gradient ascent* (18) can learn parameters. If the structure is not known, the *Structural EM (SEM)* algorithm of (19) or *Structure-based approach* (20) can be used to perform structure learning with incomplete data.

The EM algorithm Dempster *et al.* (16) present a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. Since each iteration of the algorithm consists of an expectation step followed by a maximization step they call it EM algorithm. The EM algorithm tries to find maximum likelihood of parameters in statistical models, where the model depends on unobserved hidden variables. Typically these models involve hidden variables in addition to unknown parameters. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points.

The EM process is remarkable because of its simplicity and wide range of applications. Other advantages of the EM approach are that it is easy to program and often allows simple adaptation of complete data methods and it provides fitted values for missing data. The two main disadvantages, slow convergence and local maxima, have often been criticized (21). The EM algorithm can easily get trapped in sub optimal local maxima. It also requires an initial estimate of the model parameters. Since multiple local maxima of the likelihood function are frequent in practice and the algorithm converges only to one local maximum, the quality of the initial estimate can greatly influence the final results.

Adaptive probabilistic networks with a hidden variables Binder *et al.* (18) investigate the problem of learning probabilistic networks with known structure and hidden variables. They present a new learning algorithm for probabilistic networks.

They focus on the problem of learning networks where some of the variables are hidden. They also restrict attention to the problem of learning the probabilistic parameters, assuming that the network structure is known. They have demonstrated a gradient-descent learning algorithm for probabilistic networks with hidden variables that uses localized gradient computations piggybacked on the standard network inference calculations.

Structural EM algorithm Friedman *et al.* (19) introduced a framework for searching over structures with incomplete data. The key idea of this method is to use the best estimate of the distribution to complete the data, and then use procedures that work efficiently for complete data on this completed data. This follows the basic intuition of the EM algorithm for learning parameters in a fixed parametric model (16). At each step, it can either find better parameters for the current structure, or select a new structure. The former case is a standard parametric EM step, while the later is a structural EM step. A charming feature of the method is it attempts to directly optimize the true Bayesian score within EM iterations.

In particular, although Friedman *et al.* provided convergence proofs for the abstract version of the algorithm, it is still not clear whether these proofs apply given the approximations need to perform this algorithm in practice. Empirical experience shows that the procedure does consistently converge. An additional aspect glossed over in this framework is the computation of the expected statistics. This requires large number of computations during learning. This is the main bottleneck in applying this technique to large scale domains.

Structure-based approach Structure-based approach aims to detect hidden variables that interact with the observed variables. A very natural approach to detect hidden variables is to search for "structural signatures" of hidden variables (substructures) in the learned network that tend to suggest the presence of a hidden variable. Elidan *et al.* (20) make this basic idea concrete, and show how to integrate it with structure-search algorithms. They use standard Bayesian model selection algorithms to learn a structure over the observable variables. They then search the structure for substructures that seem as if they might be induced by a hidden variable. They temporarily introduce the hidden variable in a way that breaks up the substructures, and then continue learning based on that new structure. If the resulting structure has a better score, they keep the hidden variable. They finally integrate this idea with existing learning algorithms such as structural EM. This approach can be considered as a preprocessing step for structural EM, substantially reducing the structural EM search space. It suffers from the detection of hidden variables in certain situations. It can not detect situations where the hidden variable provides more succinct model of a distribution that can be described by a network without a hidden variable (as in the simple example in Figure 1.1).

Learning hidden variable approaches do not work with many hidden variables A major question in learning hidden variables is how to decide on the number of hidden variables. The above approaches are able to learn models with 1 hidden variable, 2 hidden variables, etc., and then to select the network with the highest score. Latent variable approaches are only practical if the number of hidden variables is small, an assumption that is questionable, and it is often violated in the domain real life particularly in molecular biology (5, 19).

What is surprising, is that despite the influx of research for learning probabilistic graphical models in recent years, few works address the challenge of considering many hidden variables in these models. Imagine a tool that not only reveals the structure of the network but also takes into account the existence of many hidden variables. It is the goal of this dissertation to present methods that will form the framework towards this goal, in the content of molecular biology.

1.2 Molecular biology and hidden mechanisms

The previous section motivates the study of probabilistic graphical models in the presence of hidden variables. This section gives the motivations of reconstructing the biological networks with the present of unknown mechanisms. As background information, we first describe important biological terms that are used for this dissertation. We then discuss about the biological motivation of reconstructing network with a present of hidden mechanisms.

The cell The basic unit of life is the cell. All living creatures are made of cells which are small membrane-bounded units filled with a concentrated aqueous solution of chemicals called *cytoplasm*. The membrane functions as a selective barrier to substances that enter the cell and exit from it. Each cell is an independent entity, capable of creating copies of itself by growing and dividing into two identical daughter cells. The complete characteristics of an organism is carried by each of its cells. This hereditary information is stored within the *deoxyribonucleic acid (DNA)* molecule. DNA molecules are informational molecules encoding the genetic instructions used in the development and functioning of all cells. In higher multicellular organisms, each cell carries the same DNA content, storing the complete biological information essential for life [1]. All cells transform DNA to *proteins*, which determine cells structure and function. Proteins are crucial elements to the existence of each organism as they build the cell and drive most of its functions.
Organisms can be divided into two classes:

Central dogma in molecular biology The central dogma of molecular biology is a framework for understanding the transition from the instructions coded in the DNA

to the production of proteins. DNA produces *ribonucleic acid (RNA)* which in turn produces proteins. The functional units in the DNA that code for RNA or proteins are called *genes*. This process consists of two main steps, *transcription* and *translation*, shown in Figure 1.2. Transcription is the process by which the information contained in a section of DNA is transferred to a newly assembled piece of messenger RNA (mRNA). It is facilitated by *transcription factors*, which are molecules that can bind to specific DNA sequences. The transcription factors can either activate or repress the process of transcription. After splicing the mRNA in transcription the newly synthesized RNA is transferred by other proteins to the cytoplasm. Translation is a process that synthesize a protein from mRNA template in the cytoplasm. An enzyme called ribosome attaches to the RNA and uses the information in it as a template for the synthesis of the protein.

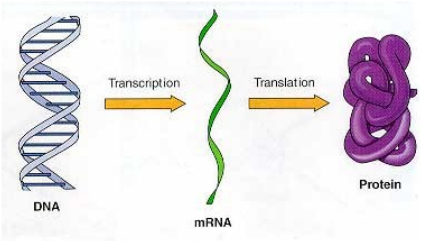

Figure 1.2: The classic view of the central dogma of molecular biology - The coded genetic information hard-wired into DNA is transcribed into individual transportable unit, composed of messenger RNA (mRNA); each mRNA contains the program for synthesis of a particular protein (or small number of proteins).

.

Gene regulation Regulation of gene expression (or gene regulation) comprises all processes that cells use to calibrate the amount of RNA species in them. Although a functional gene product can be an RNA, the majority of known mechanisms regulate protein coding genes. Gene regulation is essential for prokaryotes and eukaryotes as it increases the versatility and adaptability of an organism by allowing the cell to express proteins when needed. Since only a fraction of genes is expressed within a cell in a certain condition, it is important for all cells to regulate their gene expression in response to variable environmental conditions.

Regulation takes place at all levels, e.g., in signal propagation, in transcription, in translation, and in protein degradation (22). At each single step many regulatory

processes can concur. A transcription factor, for example, can be regulated transcriptionally and non-transcriptionally. Transcriptional regulation means control of the transcription factor's mRNA level. Non-transcriptional regulation means controlling the activity level of the transcription factor protein by binding to a ligand, by dissociation of an inhibitor protein, by a protein modification like phosphorylation, or by cleavage of a larger precursor (22, 23). Of particular interest for this thesis are signal transduction pathways.

Signal transduction pathways Signal transduction occurs when an extracellular signaling molecule activates a cell surface receptor (22). In turn, this receptor alters intracellular molecules creating a response. If the receptor is found in the membrane of the cell, a series of signals must be passed through cytosol into the nucleus which in turn activate transcription. This series of signals is called a signal transduction pathway.

Cellular signaling pathways regulate essential processes in living cells. In many cases, alterations of these molecular mechanisms cause serious diseases including cancer. Understanding the organization of signaling pathways is hence a principal problem in modern biology. With methods making use of *RNA interference (RNAi)*, one can identify new pathway component and order pathways in regulatory hierarchies. RNAi (24) is a cellular mechanism of post-transcriptional gene silencing that moderates the activity of their genes. Discovering RNAi plays an important role in functional genomic research for many reasons. For instance, a physiological role it plays in gene regulation is one of the most important discovery of last decades. Moreover, screening RNAi of target genes can be applied on a genomic scale and allows rapid identification of genes contributing to cellular processes and pathways (25).

Gene knockdown The RNA interference pathway is often exploited in experimental biology to study the function of genes in cell. Double-stranded RNA is synthesized with a sequence complementary to a gene of interest and introduced into a cell or organism, where it is recognized as exogenous genetic material and activates the RNAi pathway. Using this mechanism, researchers can cause a drastic decrease in the expression of a targeted gene. Studying the effects of this decrease can shed light on the physiological role of the gene product. This means, RNAi perturbation experiments push a genes expression level towards zero. Only in knockouts, however, the intervention leads to a completely non-functional gene. In RNAi experiments the gene is still active, but silenced. It is less active than normal due to human intervention. Hence, we do not fix the state of the gene, but push it towards lower activities. Since RNAi may not totally abolish expression of the gene, this technique is sometimes referred to as a "knockdown", to distinguish it from "knockout" procedures in which expression of a gene is entirely eliminated.

1.2.1 Molecular biology with many hidden variables

Biological motivation A complex system is not understood solely by passive observation; it needs active manipulation by the researchers. In biology this fact has long been known. Functional genomics has a long tradition of inferring the inner workings of a cell by breaking it down using an external stimulus (23). A cells response to an external stimulus is carried out by a complex network. The stimulus is propagated via signal transduction to activate transcription factors which bind to promoters, thus activating or repressing the transcription and translation of genes, which in turn can activate secondary signaling pathways, and so on. We distinguish between the transcriptional level of signal transduction known as gene regulation and the non-transcriptional level, which is mostly mediated by post-translational modifications (26, 27). While gene regulation leaves direct traces on expression profiles, non-transcriptional signaling does not and there might be many players involved which cannot be traced in the non-transcriptional level. When we assume that all the network players are known and no hidden players are involved, networks can be modeled by formal statistical framework for network reconstruction, while networks with unknown pathways players cannot.

The appearance of methods making use of RNA interference (RNAi) enables researchers to selectively silence known genes of interest on a large scale in order to find out the complex mechanism in the cell (28). Gene expression monitoring techniques such as DNA microarrays allow us to measure the effects of a perturbation on a genome-wide scale. This enables the reversal of the engineer interdependencies between gene products on a non-transcriptional level. The known genes of interest are silenced individually, and the respective downstream effects on gene expression are measured using genome-wide gene expression data. By observing the nested structure, there is a significant up- or down-regulation of affected genes, allowing for the reconstruction of the upstream signaling pathway. Here, the main challenge is to investigate the unknown biological mechanisms that govern the relation between the known genes (29).

Cell is a complex system Due to the complexity of the cell, a complete insight into the signaling pathway networks, with detailed knowledge of individual players in the networks, is still out of reach. While the networks normally contain many players, silencing and observing the downstream effects of all genes is unfeasible. Typically, we do not have all the players to analyze biological mechanisms like signaling pathways. We are lacking important but so far unknown players in the pathways. Therefore, our current understanding of virtually all cellular signaling pathways is almost certainly incomplete.

We look at the cell through narrow windows When analyzing the cell, we look through narrow windows leaving potentially important events in blind spots. However, unknown players are not completely independent of what we observe inside the windows. Over the last decade, with the help of genomics studies, these windows have become larger, but a complete insight into the complex biological networks has still not yet been attained.

Motivation of non-confoundable analysis in the context of Nested Effects Models Unknown genes, whose involvement in cellular processes has not yet been determined, can make the interplay of the known genes appear different from what they really are. The effect of unknown genes might be mixed up with the effect of known genes. Moreover, separating these effects can be difficult and can confound our perspective of the networks. Statistically speaking, these unknown genes are hidden variables, and for network reconstruction they are hidden nodes. Network reconstruction in the presence of hidden variables may result in confounding, which is a major source of bias (15). This rises a new question in network reconstruction: What of our current understanding of biological networks can be confounded by hidden mechanisms and what can not. We believe that these questions can only be addressed meaningfully in the context of formal statistical network reconstruction framework. In order to address these questions we introduce a new methodology based on Nested Effects Models (NEMs). NEMs differ from other statistical approaches like Bayesian networks by encoding subset relations instead of partial correlations. This charming feature of the method enables us to infer the graph of upstream/downstream relations for a set of signaling genes from perturbation effects. It is the goal of this dissertation to present a non-confoundable network analysis for very incomplete data in the context of NEMs.

1.3 Thesis Organization

In summary, there are two problems to be addressed when reconstructing biological networks for very incomplete data. First, how the hidden players could be misleading our current understanding of biological networks. Second, how to infer the network structures in a way that can not be confounded by hidden players. This thesis introduces a novel methodology to address both questions. It is organized as follows:

1.3.1 Nested Effects Models

Chapter 2 gives an overview of Nested Effects Models and their implementations. The basic framework of NEMs was substantially extended in several studies. In this chapter, I present a review of the methodology with a discussion of similarities, differences and limitations of the proposed algorithms.

1.3.2 Complications arising from hidden variables to Nested Effects Models

In chapter 3, I demonstrate the importance of considering hidden variables for network reconstructing when they are known to exist. I discuss complications that arise by considering these variables in the context of Nested Effects Models. I then investigate the possible influences of hidden variables on the data patterns generated from perturbation experiments.

1.3.3 Partial Nested Effects Models

In chapter 4, I investigate what is arguably the most straightforward approach for considering the existence of hidden variables in the context of NEMs. I introduce a simple edge-by-edge partial network reconstruction algorithm called *Non Confoundable Network Analysis (No-CONAN)* to derive non confoundable network properties. I then define a data structure that encodes the partially resolved networks called *partial Nested Effects Models (pNEM)*. This chapter then goes further to demonstrate how No-CONAN and pNEM perform on different simulation studies.

1.3.4 Cell differentiation in embryonic stem cells

In this chapter, I demonstrate the practical use of pNEM in biological application. I first give some background on cellular decision making and cell differentiation. I then demonstrate the performance of pNEM in a first application to embryonic stem cell differentiation in mice. This chapter then goes further to show that taking unknown players into account changes our account of biological networks.

1.3.5 Distorted canonical WNT-signaling in colorectal cancer cells

Chapter 6 shows the performance of No-CONAN in another biological application on the Wnt-signaling pathway. The chapter starts with a short introduction to the Wnt-signaling in colorectal cancer cells. It then demonstrates the performance of non-confoundable network analysis in the context of a recent study on the Wnt signaling pathway in colorectal cancer cells.

2

Nested Effects Models

This chapter gives an overview of Nested Effects Models and their implementations. The theory of NEMs has been applied and extended in several studies. Section 2.1 gives an overview of all NEMs in different formulations from literature. The section goes further and discusses the similarities, differences and limitations of all the methods. Section 2.2 and 2.3 review different searching algorithms in the context on NEMs and section 2.4 describes the dynamic nested effect models that enables the analysis of perturbation time series data.

2.1 Nested Effects Models (NEMs)

A cells response to an external stimulus is complex. The stimulus is propagated via signal transduction to activate transcription factors, which bind to promoters thus activating or repressing the transcription and translation of genes, which in turn can activate secondary signaling pathways, and so on. We distinguish between the transcriptional level of signal transduction known as gene regulation and the nontranscriptional level, which is mostly mediated by post-translational modifications. While gene regulation leaves direct traces on expression profiles, non-transcriptional signaling does not (23).

A hypothetical pathway Figure 2.1A shows a hypothetical biochemical pathway adapted from Wagner (23, 30, 31). It consists of two transcription factors, a protein kinase and a protein phosphatase and the genes encoding these proteins. The figure shows the three biological levels of interest: genome, transcriptome and proteome. The thick arrows show information flow through the pathway. The transcription factor expressed by gene 1 binds to the promoter region of gene 2 and activates it. Gene 2 encodes a protein kinase, which phosphorylates a protein phosphatase (expressed by gene 3). This event activates the protein phosphatase, which now dephosphorylates

the transcription factor produced by gene 4. It binds to gene 5 and induces expression. The three biological levels of DNA, mRNA and protein are condensed into a graph model on five nodes. Gene expression data only shows the mRNA level. A model inferred from expression data will only have two edges, connecting gene 1 to gene 2 and then gene 2 to gene 5. Since genes 3 and 4 only contribute on the protein level, a model based on correlations on the mRNA level will ignore them.

Interventions at genes in the pathway shed light on the pathway topology. This is exemplified by an RNAi intervention at gene 3 in Figure 2.1B Silencing gene 3 will cut information flow in the pathway and result in an expression change at gene 5. This is reflected in the model by extending it to include an edge from gene 3 to gene 5. Note that we have no observation of direct effects of the intervention at gene 4 in mRNA data. The only information we have are secondary effects at the transcriptional end of the pathway. *Nested Effects Models (NEMs)* were first introduced by Markowetz *et al*. as a framework to order genes in regulatory hierarchies from secondary effects.

NEMs use a probabilistic framework to compare a given network hypothesis with the observed nested structure of downstream effects. Perturbing one gene may have an influence on a number of downstream genes, whereas perturbing others affects a subset of those. NEMs framework distinguishes two kinds of genes: The first are the candidate pathway genes for perturbation and the second are genes which show effects of such interventions in expression profiles. We call the perturbed genes *S-genes* for signaling genes and denote them by $\mathbf{S} = S_1, \ldots, S_n$. The genes that change expression after perturbation are called *E-genes* and we denote them by $\mathbf{E} = E_1, \ldots, E_N$. We further denote the set of E-genes displaying expression changes in response to the perturbation of S_i by $\mathcal{D}_i$. In a nutshell: NEMs infer that S_1 acts upstream of S_2:

$$S_1 \longrightarrow S_2 \quad \text{if} \quad \mathcal{D}_2 \subset \mathcal{D}_1$$

All downstream effects of a perturbation in S_2 can also be triggered by perturbing S_1 (Figure 2.2). This suggests that the perturbation of S_1 causes a perturbation of S_2 and acts upstream of S_2. S-genes can take values 1 and 0 according to whether signaling is interrupted or not. State 0 corresponds to a node, which is reached by the information flow through the pathway. We call the subset of S-genes, which are in state 1 when S-gene S is silenced, the influence region of S. The set of all influence regions is called a silencing scheme Φ. The silencing scheme summarizes the effects of interventions predicted from the pathway hypothesis. Mathematically speaking, Φ is a transitively closed graph that defines a partial order on S-genes from the expected nested structure of downstream effects. Following Markowetz *et al*. (32), the positions of the E-genes are included as model parameters and it is assumed that each E-gene is attached to a single S-gene only. Knocking down a specific S-gene S_k interrupts signal flow in the downstream pathway, and hence an effect on the E-genes attached to S_k and all S-genes depending on S_k is expected. Let us assume

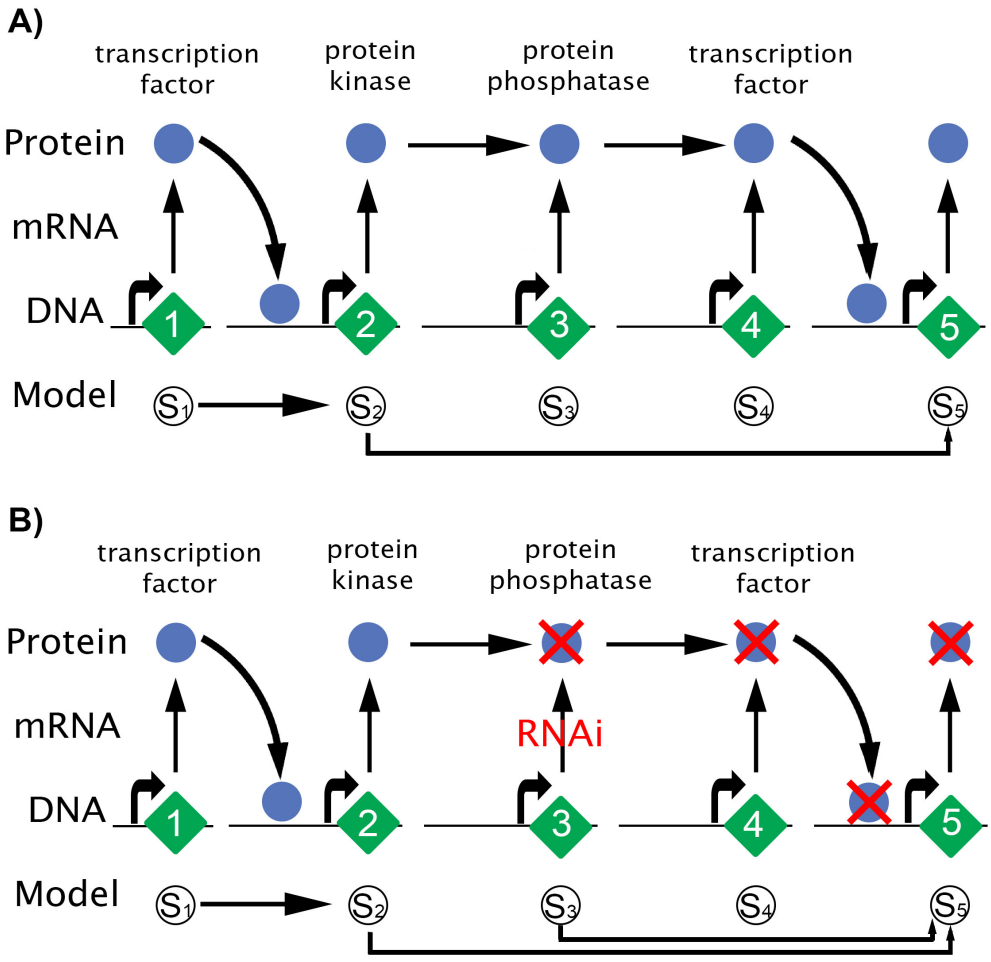

Figure 2.1: A hypothetical biochemical pathway - **A** shows a hypothetical biochemical pathway adapted from Wagner (30). It shows four levels of interest: three biological and one of modeling. Inference from gene expression data alone only gives a very limited model of the pathway. The contributions of genes 3 and 4 are overlooked. The situation changes if we can use interventional data for model building. **B** shows that silencing gene 3 by RNAi will cut information flow in the pathway and result in an expression change at gene 5. This is visible on the mRNA level and can be integrated in the model. Thus, the expanded model shows an edge from gene 3 to gene 5. The figure is adapted from (23).

n knock-downs are performed and there exist m E-genes in total. The outcomes of these experiments are summarized in an $m \times n$ data matrix D. According to Bayes formula a specific network hypothesis can be scored as:

$$P(\Phi|D) = \frac{P(D|\Phi)P(\Phi)}{P(D)}$$

where $P(D)$ is a constant that does not depend on Φ. Consequently, the (marginal) likelihood $P(D|\Phi)$ together with the network prior $P(\Phi)$ play the central role in the inference.

The position of the E-genes is introduced as a model parameter $\Theta = \{\Theta_i | \Theta_i \in 1, ..., n, i = 1, ..., m\}$, for example $\Theta_i = j$, if E-gene i is attached to S-gene j. Assuming independence of E-genes one can write down the conditional likelihood $P(D|\Phi, \Theta)$ given a fixed network hypothesis Φ and model parameters Θ as:

$$P(D|\Phi, \Theta) = \prod_{i=1}^{m} P(D_i|\Phi, \Theta_i)$$

However, in practice it is unknown which E-genes are being controlled by which S-genes. In a perturbation experiment we predict effects at all E-genes, which are attached to an S-gene in the influence region. Expected effects can be compared with observed effects in the data to choose the topology, which fits the data best. Owing to measurement noise we cannot expect to find an expected topology to be in complete agreement with all observations. We allow deviation from predicted effects by introducing error probabilities α and β for false positive and negative situations, respectively. We model the expression levels of E-genes on the various perturbation experiments k as binary random variables E_{ik} . The distribution of E_{ik} is determined by the silencing scheme Φ and the error probabilities α and β. For all E-genes and targets of intervention, the conditional probability of E-gene state e_{ik} given silencing scheme Φ can then be written in tabular form as:

Table 2.1: The distribution of binary effect data - The distribution of E_{ik} is determined by the silencing scheme Φ and the error probabilities α and β.

$$P(e_{ik}|\Phi, \theta_i = j) = \left\{ \begin{array}{cc|l} e_{ik} = 1 & e_i = 0 & \\ \hline \alpha & 1-\alpha & \text{if } \Phi \text{ predicts } \textbf{no effect} \\ 1-\beta & \beta & \text{if } \Phi \text{ predicts } \textbf{effect} \end{array} \right.$$

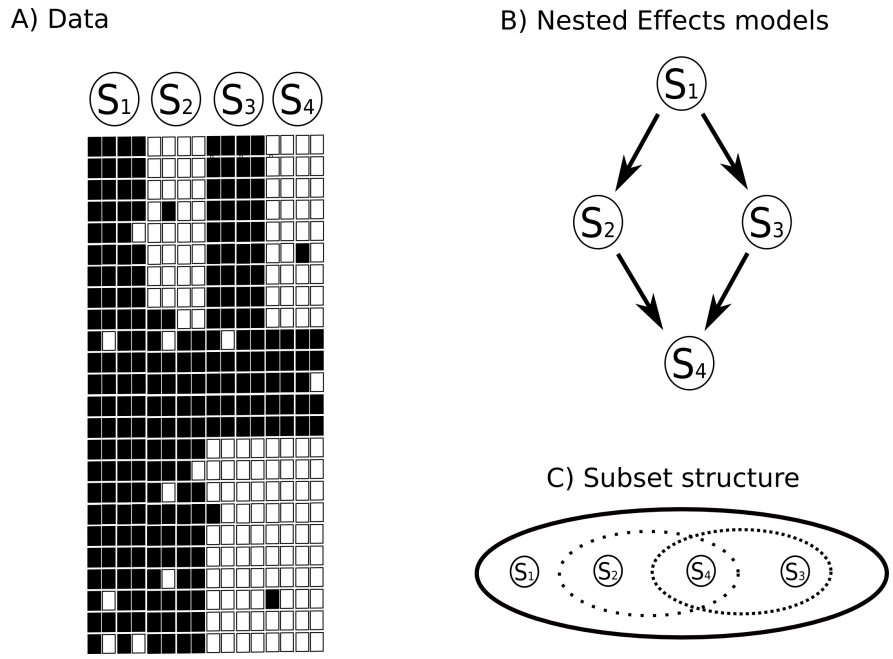

Figure 2.2: An introduction to Nested Effects Models - Plot **A** shows a toy dataset consisting four replicate of phenotypic profiles for four perturbed genes (S1,S2, S3 and S4). Each profile is binary with black coding for an observed effect and white for an effect not observed. An important feature of the data is the subset structure visible between the profiles in the data set: the effects observed when perturbing genes S1 are a superset to the effects observed for all other genes. The effects of perturbing S4 are a subset to all other genes effects. The pair S2 and S3 have different but overlapping effect sets. The directed acyclic graph (DAG) shown in plot **B** represents these subset relations, which are shown in plot **C**.

This means that if E_i is not in the influence region of the S-gene silenced in experiment k, the probability of observing E_{ik}=1 is α(probability of false alarm); the probability to miss an effect and observe $E_{ik} = 0$ even though E_i lies in the influence region is β (probability of missed signal). In the following, we summarize NEMs based on their statistical approach for dealing with Θ in scoring a given network.

2.1.1 The marginal likelihood scoring

NEMs aim at reconstructing the silencing scheme Φ of S-genes. To deal with these issues, we interpret the position of edges between S- and E-genes as nuisance parameters, and average over them to obtain a marginal likelihood. In the Bayesian framework of Markowetz *et al.* (2005) (32), networks are scored by marginal posterior probabilities. The marginal likelihood involves marginalization over the whole parameter space Θ.

$$P(D|\Phi) = \int_{\Theta} P(D|\Phi, \Theta)P(\Theta|\Phi)d\Theta. \tag{2.1}$$

The Equation 2.1 is based on the following assumptions given in (32):

1. Given the silencing scheme Φ, and fixed positions of E-genes Θ, the observations in D are sampled independently and distributed identically:

$$P(D|\Phi, \Theta) = \prod_{i=1}^{m} P(D_i|\Phi, \theta_i) = \prod_{i=1}^{m}\prod_{k=1}^{l} p(e_{ik}|\Phi, \theta_i), \tag{2.2}$$

   where D_i is the ith row in data matrix D.

2. Parameter independence. The position of one E-gene is independent of the positions of all the other E-genes at any given time:

$$P(\Theta|\Phi) = \prod_{i=1}^{m} P(\theta_i|\Phi) \tag{2.3}$$

3. Uniform prior. The prior probability to attach an E-gene is uniform over all S-genes:

$$P(\theta_i = j|\Phi) = \frac{1}{n} \qquad \text{for all} \quad i \quad \text{and} \quad j \tag{2.4}$$

   The last assumption can be dropped to include existing biological knowledge about regulatory modules (33, 34).

With the above assumptions the marginal likelihood can be calculated as follows. The numbers above the equality sign indicate which assumption was used in each step.

$$\begin{aligned} P(D|\Phi) &= \int_{\Theta} P(D|\Phi,\Theta)P(\Theta|\Phi)d\Theta \\ &\overset{[1,2]}{=} \prod_{i=1}^{m}\int_{\theta_i} P(D_i|\Phi,\theta_i)P(\theta_i|\Phi)d\theta_i \\ &\overset{[3]}{=} \frac{1}{n^m}\prod_{i=1}^{m}\sum_{j=1}^{n} P(D_i|\Phi,\theta_i=j) \\ &\overset{[1]}{=} \frac{1}{n^m}\prod_{i=1}^{m}\sum_{j=1}^{n}\prod_{k=1}^{l} p_{\alpha,\beta}(e_{ik}|\Phi,\theta_i=j) \end{aligned} \tag{2.5}$$

Since we have a finite number of S-genes, here we can sum over all E-gene positions.

2.1.2 Maximum aposteriori (MAP) inference scheme

Alternative to averaging over E-genes positions, Tresch *et al.* (35) proposed to estimate the edges between S- and E-genes, Θ, in a maximum aposteriori MAP sense, which is then used to calculate $P(D|\Phi,\hat{\Theta})$ and $P(\Phi|D,\hat{\Theta})$:

$\hat{\Theta} = argmax P(D|\Phi,\Theta)P(\Theta),$

$P(\Phi|D,\hat{\Theta}) = \frac{P(D|\Phi,\hat{\Theta})P(\Phi)}{P(D)}$

The maximum likelihood scoring approach allows to distinguish between network hypotheses that differ only in transitive edges which is not the case in the previous section. In order to infer NEMs using MAP inference, Tresch *et al.* (35) defined Nested Effects Model (NEM) F as a product of Φ and Θ:

$$F = \Phi\Theta \tag{2.6}$$

Assuming data independence and using Bayes rule, the likelihood of the model F is represented as $P(D|F)$ and factors out as follows:

$$\begin{aligned} P(D|F) &= P(D|\Phi\Theta) \\ &= \prod_{(j,i)\in\Phi\times\Theta} P(D_{j,i}|j = F_{ji}) \\ or\log(P(D|F)) &= \sum_{(j,i)\in\Phi\times\Theta} logP(D_{j,i}|j = F_{ji}) + const, \end{aligned} \tag{2.7}$$

if we define $P(j = x|i)$ for all $x \in \{0, 1\}$ and $(j, i) \in \Phi \times \Theta$ with $j = F_{ji}$ interpreted as S-gene j is linked to E-gene i. The quantity $\log(P(D|F))$ can also be expressed as a likelihood ratio for convenience using matrix algebra as follows:

$$\log(P(D|F)) - \log(P(D|N)) = tr(FR), \tag{2.8}$$

where $R = \log \frac{P(D_{ji}|e_{ij}=1)}{P(D_{ji}|e_{ij}=0)}$, "$tr$" denoting the trace function of a quadratic matrix and N the NULL matrix corresponding to the model predicts no effects at all. This allows us to present the likelihood as :

$$\log(P(D|\Phi, \Theta)) = tr(\Phi\Theta R) + const \tag{2.9}$$

The likelihood function in this form depends on the data only via the log ratios. This provides a flexible way of handling different input data such as binary data, p-values or any other data type as long as it can be converted to a likelihood ratio. The posterior of the model (Φ, Θ) becomes

$$\log(P(\Phi, \Theta|D)) = \log(P(D|\Phi, \Theta)) + \log(P(\Phi)) + \log(P(\Theta)) + const, \tag{2.10}$$

and the task is to find the MAP estimate for $\log(P(\Phi, \Theta|D))$,

$$(\hat{\Phi}, \hat{\Theta}) = \text{argmax}_{\phi,\Theta}(\log(P(D|\Phi, \Theta)) + \log(P(\Phi)) + \log(P(\Theta))), \tag{2.11}$$

In order to find the optimal graph we need to maximize the E-gene positions and vice versa.

2.1.3 NEMs as a Bayesian network

A flexible formulation of NEMs in the language of Bayesian networks can constitute a natural generalization of the original NEM model. The original formulation of the NEM suffers from some restrictions which were imposed for the sake of computability. Zeller *et al.* (36) proposed a new formulation for NEM in the context of Bayesian networks which provides a motivation for these restrictions by explicitly stating prior assumptions that are inherent to the original formulation.

Recalling chapter 1, a Bayesian network describes the joint probability distribution of a finite family of random variables (the nodes) by a directed acyclic graph Φ and by a family of local probability distributions, which we assume to be parameterized by a set of parameters Θ (2, 37). In the context of the MAP inference (35), we have to model a deterministic signaling hierarchy, in which some components (E) can be probed by measurements, and some components (S) are perturbed in order to measure the reaction of the system as a whole. All these components (S) and (E) will be hidden nodes if there is no observation for an acyclic graph $H = S \cup E$. In order to account for the data, we introduce an additional layer of observation variables $(observationvariables, O)$ in the following way: each effect node $e \in E$ has an edge pointing to a unique observable node $e^{'} \in O$ (Figure 2.3). Hence, $O = \{e^{'}|e \in E\}$, and we call $e^{'}$ the observation of e. Following the general Bayesian network framework, let $pa(x)$ be the set of parents of a node x and for notational convenience add a zero node z, $p(z = 0) = 1$, which has no parents, and which is a parent of all hidden nodes (but not of the observable measurements). For the hidden nodes, define local probabilities corresponding to deterministic relationships as follows

$$\begin{aligned} p(x = 1|pa(x)) &= \begin{cases} 1 & \text{if any parent is active} \\ 0 & \text{otherwise,} \end{cases} \\ &= \max(pa(x)) \text{ for } x \in H^{'}, \end{aligned} \tag{2.12}$$

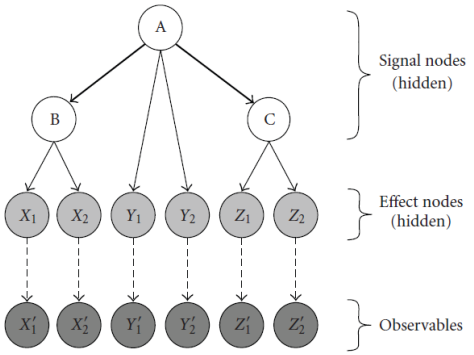

Figure 2.3: Bayesian Nested effects models - Example of a Nested effects model in its Bayesian network formulation. A,B,C represent the S-genes and $X_1, X_2, Y_1, Y_2, Z_1, Z_2$ represent the effect nodes. The bold arrows determine the graph Φ, the solid thin arrows encode Θ. Dashed arrows connect the effects to their reporters. This figure is reproduced from (35).

Obviously, all hidden nodes are set to 0 or 1 deterministically, given their parents. The local probabilities $p(e^{'}|e \in E)$, $e \in E$ can come from both discrete or continuous

distributions (35). The Bayesian network NEM is parameterized by its topology Φ and its local probability distributions, which we assume to be given by a set of local parameters Θ. The final goal is to maximize $P(\Phi|D)$. In the presence of prior knowledge and if we assume independent priors for the topology and the local parameters, we can write

$$\begin{aligned} P(\Phi,\Theta|D) &= \frac{P(D|\Phi,\Theta)P(\Phi)P(\Theta)}{P(D)} \\ &\propto P(D|\Phi,\Theta)P(\Phi)P(\Theta) \end{aligned} \tag{2.13}$$

from which it follows that

$$\begin{aligned} P(\Phi|D) &= \int P(\Phi,\Theta|D)d\Phi \\ &\propto P(\Phi)\int P(D|\Phi,\Theta)P(\Theta)d\Theta \end{aligned} \tag{2.14}$$

In a case that the integral in Equation 2.13 can be solved analytically, it can be used by standard optimizations for the approximation of the $argmaxP(\Phi,\Theta|D)$. If the expression is difficult to solve, resort to a simultaneous maximum a posteriori estimation of Φ and Θ (35), that is,

$$\begin{aligned} (\hat{\Phi},\hat{\Theta}) &= \mathrm{argmax}_{\Phi,\Theta}P(\Phi,\Theta|D) \\ &= \mathrm{argmax}_{\Phi}(\mathrm{argmax}_{\Theta}P(D|\Phi,\theta)P(\Theta))P(\Phi). \end{aligned} \tag{2.15}$$

2.1.4 Factor graph NEMs

A signed version of the Nested Effects Model and an associated efficient structure inference method, named Factor Graph-Nested Effects Model(FG-NEM) (38) was developed to distinguish between activating and inhibiting regulation in a pathway. Recall that the original NEM by Markowetz *et al.* (32) include two sets of parameters. The parameter set Φ records all pair-wise interactions among the S-genes and the parameter set Θ describes how each E-gene is attached to the network of S-genes. Φ is a binary matrix with entry ϕ_{AB} set to one if S-gene A acts above S-gene B and zero otherwise. Φ must also be transitively closed. The model by Markowetz *et al.* (32) does not distinguish between stimulatory and inhibitory interactions. To tackle this drawback, Vaske *et al.* (38) suggest a model, in which ϕ_{AB} takes six possible values for each unique unordered S-gene pair A,B also known as interaction modes. The possible values are: 1) A activates B, $A \rightarrow B$; 2) A inhibits B,$A \dashv B$; 3) A is equivalent to B, A=B; 4) A does not interact with B, $A \neq B$; 5) B activates A,$B \rightarrow A$; and 6) B inhibits A, $B \dashv A$. The Factor graph NEMs allow for the reconstruction of a broader set of S-gene interactions from the secondary effects of E-gene expression corresponding to the observed data denoted as D. Similarly like

the other NEM approaches discussed so far, a maximum aposteriori is used to identify the Φ that maximizes the posterior $P(\Phi|D)$ represented as :

$$\begin{aligned}\hat{\Phi} &= \text{argmax}_{\Phi} P(\Phi|D) \\ &= \text{argmax}_{\Phi} \sum_{\Theta} \sum_{H} P(\Phi, \Theta, H|D). \end{aligned} \tag{2.16}$$

where Θ refers to the attachment point of each E-gene into the network and H refers to the hidden E-gene states corresponding to up, down regulations or no change. Applying the same assumptions as in Markowetz *et al.* (32) we have:

$$\begin{aligned}\hat{\Phi} &= \text{argmax}_{\Phi} P(\Phi) \sum_{\Theta} P(\Theta|\Phi) \sum_{H} P(H|\Phi, \Theta) P(D|H) \\ &= \text{argmax}_{\Phi} P(\Phi) \sum_{\Theta} \sum_{H} P(H|\Phi, \Theta) P(D|H) \\ &= \text{argmax}_{\Phi} P(\Phi) \prod_{e \in E} \sum_{\Theta} \sum_{H} P(H_e|\Phi, \theta_e) P(D_e|H_e)\end{aligned}$$

given independence of E-genes, E.

$$= \text{argmax}_{\Phi} P(\Phi) \prod_{e \in E} L_e(\Phi) \tag{2.17}$$

where D_e and H_e are the row vectors of data matrix and hidden states for E-genes respectively and θ_e records the attachment of an E-gene to an S-gene and L_e summarizes the marginal likelihood of the data restricted only to E-gene e under a given model Φ and θ_e. Note that L_e can be reformulated as a product of pair-wise S-gene terms (38).

2.1.4.1 Structure of factor graph NEMs and Network inference

Scoring a given S-gene graph can be achieved based on max-sum message passing in a factor graph (39) which provides an efficient means for estimating highly probable S-gene configurations. The parameters that determine the S-gene interactions, Φ, are explicitly represented as variables in the factor graph. Identifying a high-scoring S-gene network is therefore converted to the task of identifying likely assignments of the Φ variables in the factor graph. A factor graph is a probabilistic graphical model whose likelihood function can be factorized into smaller terms (factors) representing local constraints on a set of random variables. A factor graph can be represented as an undirected, bi-partite graph with two types of nodes: variables and factors. A variable is adjacent to a factor if the variable appears as an argument of the factor. Figure 2.4 shows the factor graph representation of a Bayesian network. Factor graphs represent both the variables as nodes and the factors as nodes, with edges from each factor to the variables in that factor's domain, resulting in a bipartite graph. Factor

graphs generalize probability mass functions as the joint likelihood function requires no normalization and the factors need not be conditional probabilities. Each factor encodes a local constraint pertaining to a few variables. In the factor graph NEM

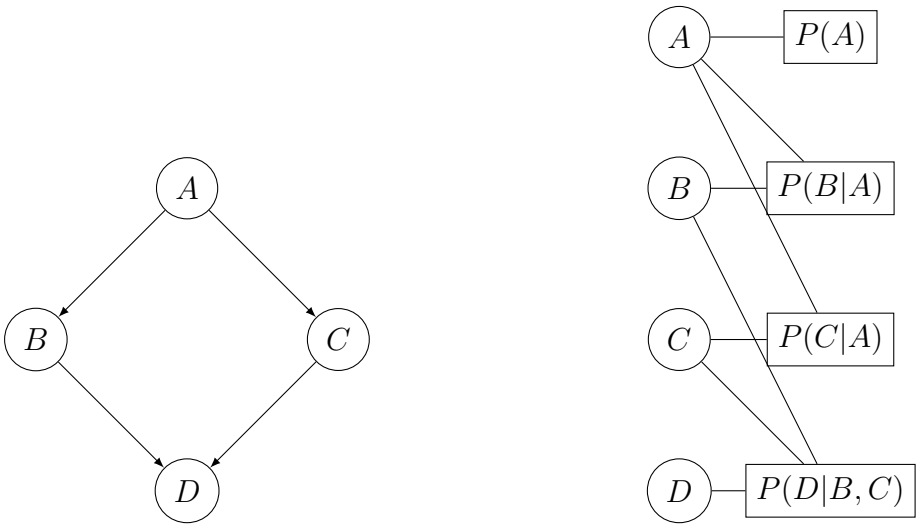

Figure 2.4: Bayesian network next to corresponding factor graph- A Bayesian network (left) and the corresponding factor graph (right). The decomposition of the joint probability, P(A, B, C, D) = P$(D|B, C)$ P$(B|A)$ P$(C|A)$ P(A) is made explicit in the bipartite factor graph.

a Φ that maximizes the posterior is found using max-sum message passing using all terms from Equation 2.17 in log space. The complete model of a factor graph NEM by Vaske *et al.*(2009) contains three types of variables and three classes of factors. The variables include: the continuous random observation of E-gene expression under a given intervention and replicate experiment, the unknown hidden state of E-gene under a particular intervention which is a discrete variable with domain $\{1, 0, -1\}$ and the interaction modes between two S-genes. The factors consists of: the Expression factors which model expression as a mixture of Gaussian distributions, the Interaction Factors which constrain E-gene states to five possible types of interaction modes between two S-genes and the Transitivity factors which constrain pair-wise interactions to form consistent triplets of S-genes. During message passing, messages which are simply local belief potentials associated with variable interactions are passed between all nodes(variables) in the graph using two inference steps. In the first step, messages from observation nodes are passed through the expression factors and hidden E-gene state variables, to calculate all messages in a single upward pass . In the second step, messages are passed between only the interaction variables and transitivity factors until convergence using Equation 2.17. The final S-gene network is derived by transitive reduction of all redundant edges from Φ. Figure 2.5 reproduced from (38) gives an

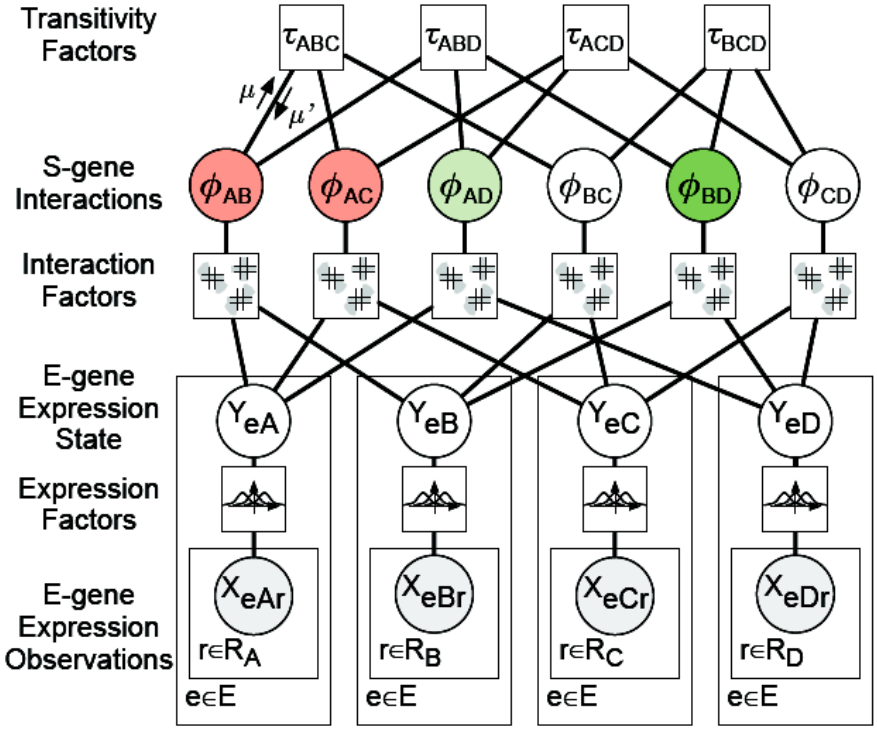

Figure 2.5: Structure of factor graph for network inference in Factor graph NEMs - The factor graph consists of three classes of variables (circles) and three classes of factors (squares). X_{eAr} is a continuous observation of E-gene e's expression under intervention A and replicate r. Y_{eA} is the hidden state of E-gene e under intervention A, and is a discrete variable with domain $\{1, 0, -1\}$. ϕ_{AB} is the interaction between two S-genes A and B. In this figure red, green and white shading denotes activation, inhibition and no interaction respectively. Expression Factors model expression as a mixture of Gaussian distributions. Interaction Factors constrain E-gene states to interaction modes between two S-genes. Transitivity Factors constrain pair-wise interactions to form consistent triangles. The arrows labeled μ and μ' are messages encoding local belief potentials on ϕ_{AB} and are propagated during factor graph inference. This figure is reproduced from (38).

overview of the structure of factor graph NEM with expression factors, interaction factors, and transitivity factors. For acyclic factor graphs, the marginal, max-marginal and conditional probabilities of single or multiple variables can be calculated exactly with the max-sum algorithms (39). Message-passing algorithms have been shown to demonstrate excellent empirical results in various practical problems even on graphs containing cycles such as feed-forward and feed-back loops (40, 41, 42).

2.2 Network learning algorithms in NEMs

NEMs score networks by their posterior probabilities given data. So far it was assumed that in principle all possible network topologies could be enumerated completely and then scored individually (43). However, the exhaustive search limits the method to small networks of up to 5 S-genes. For five S-genes, there are already more than 1,000,000 possible networks topologies exist. For ten S-genes, more than 10^{27} For larger network, search heuristics are used to explore model space.

The Bayesian scoring scheme presented in (34) does not distinguish between two network hypotheses, if they only differ in transitive edges. This issue is known as likelihood equivalence and reflects the fact that subset relationships, which are represented by a NEMs, are transitive in principle. However, a restriction to the limited class of transitively closed graph networks does not solve the problem of many possible networks for larger number of S-gene (43). As an alternative, one could make use of MAP inference presented in (35), which allows to distinguish between graphs that differ only in transitive edges. However, this line of investigation has not been reported to date. Instead, several approaches have been proposed by Frölich *et al.* (33, 34, 35, 38), which are all restricted to estimating just one high-scoring network.

2.2.1 Pairwise and triple search

The idea of pairwise and triple search is to concentrate on small sub-models involving only pairs, triples of nodes (34). In the pairwise approach, We infer pairwise relations by choosing between four models for each gene pair (S_1, S_2): either $S_1 \rightarrow S_2$ (effects of S_1 are a superset of the effects of S_2), or $S_1 \leftarrow S_2$ (effects of S_1 are a subset of the effects of S_2), or $S_1 \leftrightarrow S_2$ (the effects of S_1 and S_2 are undistinguishable) or $S_1 \cdot\cdot S_2$ (S_1 and S_2 are unrelated). For every pair (S_1, S_2), we compute the Bayesian score detailed is section 2.1 and select the maximum aposteriori (MAP) model $M_{S_1,S2} \in \{S_1 \rightarrow S_2, S_1 \longleftarrow S_2, S_1 \leftrightarrow S_2, S_1 \cdot\cdot S_2\}$. The advantage of this approach is the increase in speed and the possibility to infer networks involving a very large number of nodes. However, pairwise learning treats all edges independently of each other. This causes low accuracy of the reconstruction. To improve on this limitation the triple search approach was introduced (34).

The division into subgraphs can also be into all triples of nodes. Triple search decomposes the complete network in all three combinations of S-genes. This is a natural way to extend an inference method beyond the independence assumption between edges. For each triple, the highest scoring models can then be found among all 29 possible transitive edge interactions between three S-genes. At the end, triple search combines these models into one final graph with the help of model averaging and thresholding. However, edgewise model averaging and thresholding is not guaranteed to yield a transitively closed graph. An approximate transitive network among the S-genes can be computed by using transitive approximations of directed graphs(44)

2.2.2 Greedy hillclimbing search

Greedy hillclimbing search is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution (45). In the context of NEMs, greedy hillclimbing search (33, 35) starts with an initial network (which in most cases will be the graph with no edges) and then successively adds those edges, which increases the likelihood of the data the most. If no improving edge exists, the algorithm terminates. Using greedy hillclimbing search does not guarantee finding a global maximum. This way a local maximum of the likelihood function in network space can be reached.

2.2.3 Module networks

A divide-and-conquer approach is module networks which were first described in Frölich *et al.* (2007) (33) and slightly modified in Frölich *et al.* (2008) (46). It begins with a hierarchical clustering of the preprocessed expression profiles of all S-genes. The idea behind this approach is that S-genes with a similar E-gene response profile should be close in the signaling pathway. After clustering, we then move down the cluster tree hierarchy until we find a cluster with only four signals at most. Figure 2.6 illustrates the idea with an assumed network of 10 signals. The exhaustive search approach is then applied independently on these submodules and the optimal subnetworks are reconnected using pairwise node testing as well as transitive closure until the topology for the total network is completed.

2.3 Monte Carlo sampling combine with an EM algorithm for NEMs (MC EMiNEM)

In the context of NEM following the presentation in (35), the main objective is the reconstruction of the signal graph Φ. Several approaches try to maximize the (marginal)

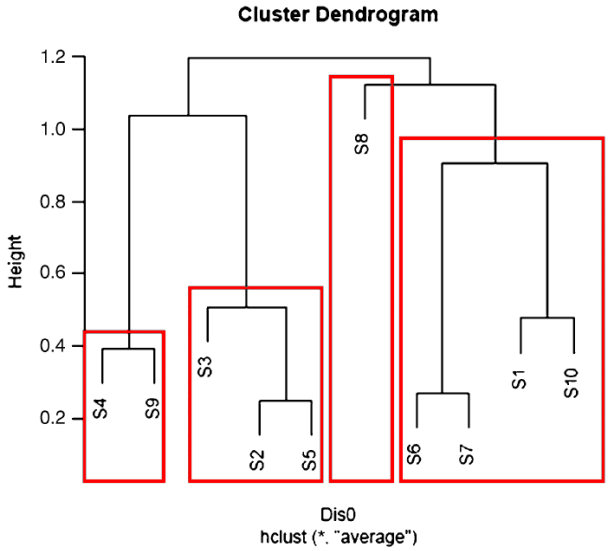

Figure 2.6: Basic idea of module networks - by successively moving down the cluster hierarchy, we identify the clusters (modules) of signals, which are marked in red. They contain four signals at most and can be estimated by exhaustively searching for the highest scoring model. This figure is reproduced from (33).

structure posterior $P(\Phi|D)$ by integrating out the effect graph Θ describing the assignment of the effect nodes to the signal nodes (43). This marginalization however is a time consuming step that increases the complexity of the respective algorithms by at least a factor of the number of effect nodes (E-genes), making the analysis of larger effects sets slow or even impossible. In order to avoid this drawback, Niederberger *et al.* (47) introduce *MC EMiNEM* for learning NEM. First, they develop an efficient Expectation-Maximization (EM) algorithm for the optimization of the NEM structure posterior (EMiNEM) This enables very fast detection of local maxima of the posterior probability function, even for large expression data sets. Second, they combine EMiNEM with mode-hopping Markov Chain Monte Carlo (MC EMiNEM) for an efficient optimization of the structure posterior.

An EM algorithm for NEMs Similar to the presentation in (35), here we want to find the maximum a posterior estimate $\hat{\Phi}$ for the signals graph,

$$\hat{\Phi} = argmax P(\Phi|D) = argmax \sum_{\Theta} P(\Phi, \Theta|D) \tag{2.18}$$

This is the classical situation in which Expectation-Maximization is applicable (16). Given some guess Φ^t for $\hat{\Phi}$, the EM algorithm describes how to find an improved

guess Φ^{t+1} so that the sequence $(P(\Phi^t|D))_{t=1,2,...}$ is monotonically increasing, and converges to a local maximum of $P(\Phi|D)$.

The expectation (E-)step of the EM algorithm involves calculating the expected log-posterior with respect to the distribution of Θ, given the current guess Φ^t:

$$Q(\Phi, \hat{\Phi}) = E_{P(\Theta|D,\Phi^t)}[logP(\Phi, \Theta|D)] \tag{2.19}$$

The maximization (M-)step of the EM algorithm then consists of finding the maximizer $\Phi^{t+1} = argmax_{\Phi}Q(\Phi, \hat{\Phi})$. This is usually a much easier task than solving Equation 2.18 directly (47).

Monte Carlo sampling of the signal posterior's local maxima As previously mentioned in chapter 1, the EM algorithm is guaranteed to find a local maximum. However, the outcome of the EM algorithm may therefore strongly depend on its initialization, and it may be far from the global optimum. In order to deal with this problem, the classical Metropolis-Hastings MCMC sampling step added to EMiNEM to introduce MC EMiNEM. In this step, consecutive parameter samples $..., \Phi_n, \Phi_{n+1}, ...$ are drawn from the distribution $P(\Phi|D)$. Given Φ_n, a random process generates a new proposal $\hat{\Phi}$. The Hastings ratio, a quantity that involves Φ_n and $\hat{\Phi}$, then determines the probability of acceptance $\Phi_{n+1} = \hat{\Phi}$ or rejection $\Phi_{n+1} = \Phi_n$ of the new proposal.

The MC EMiNEM algorithm instead applies an EM step to each new proposal $\hat{\Phi}$, which maps it to the nearest local maximum $\hat{\Phi}'$. The acceptance/rejection step is then modified by plugging $\hat{\Phi}_n$ and $\hat{\Phi}'$ into the Hastings ratio, instead of Φ_n and $\hat{\Phi}$. Niederberger *et al.* (47) show that the series of local maxima $..., \hat{\Phi}_n, \hat{\Phi}_{n+1}, ...$ associated to the underlying Markov chain $..., \Phi_n, \Phi_{n+1}, ...$ is approximately drawn from $P(\hat{\Phi}|D)$, where $\hat{\Phi}$ ranges exclusively over the space of local maxima.

2.4 Dynamic Nested Effects Models

NEMs infer the graph of upstream/downstream relations for a set of signaling genes from perturbation effects. Since non-transcriptional signaling is too fast to be analyzed by delays of downstream effects, time series are not used. NEMs monitor static perturbation effects. Specifically, a perturbation signal is supposed to propagate deterministically through the whole S-gene graph. Cycles in the graph imply that perturbation effects are indistinguishable. As a matter of principle, it is impossible to detect feedback loops in the graph. Thus, it is highly desirable to have time series measurements of perturbation effects, which help resolve biological feedback

loops and distinguish direct from indirect effects. This motivates the need of dynamic NEMs when analyzing slow-going biological processes like cell differentiation.

2.4.1 Dynamic Nested Effects Models (DNEM)

In Anchang *et al.* (48) we develop a new Bayesian method known as the Dynamic Nested Effects models (DNEMs). This approach is an extension to NEMs to infer the dynamics of a given network which is an important limitation in NEMs.

Figure 2.7 illustrates the idea of DNEM in an elementary example. The graph on the left of the tables is a transitively closed graph on 3 Signaling genes (S_1, S_2, S_3). The tables give the time series binary data of effects for all target genes (E_1, E_2, E_3) after intervention on all signaling genes. In each table, 1 indicate that a signal has reach the E-gene by time t_j, while 0 indicates that the expression of this gene has not yet changed. Looking at the last time point t_5 one sees the accumulation of effects for all target genes forming a nested structure of effects which is in conformity with the hierarchy of the graph topology. Signals starting in S_1 reach E_2 one time unit after they have arrived at E_1 suggesting that signal propagation from S_1 to S_2 takes one unit of time. The same argument using the data from perturbation of S_2 suggests that it takes two time units to propagate from S_2 to S_3. Consequently, going from S_1 to S_3 *via* S_2 takes 3 time units. However, the time delay from perturbation of S_1 to observing effects in E_3 is only 1 time unit (marked in blue). This suggests the existence of a direct signal flow from S_1 to S_3. Evidence comes from the two blue ones. In case they were zeros, the time delay between S_1 and S_3 would have been the sum of times spent when going via S_2. In this case, there would be no evidence for a shortcut pathway and we would decide on the more parsimonious graph. Furthermore, the existence of a direct path combining with that of the indirect path gives evidence of the presence of a Feed-Forward Loop. Thus we can use estimated time delays to demonstrate the existence of FFLs. A real world analysis is more difficult than the toy example. Signal propagation is a stochastic process, measurements are prone to noise, and we do not know which E-genes are controlled by which S-genes. These sources of uncertainty are addressed by DNEMs.

DNEM assume exponentially distributed time delays for individual signal propagation steps. The rate constants of the exponential distributions differ from case to case and are the main parameters of the model. All edges of a transitively closed network are associated with an individual rate constant, whose posterior distribution is inferred using Gibbs sampling. The input of a DNEM consists of (a) a set of microarray time series that measure the response of cells to molecular perturbations, and (b) a transitively closed directed acyclic graph on vertex set $\mathbf{S}$ representing a hypothetical hierarchical structure of upstream/downstream relations. This graph can be derived

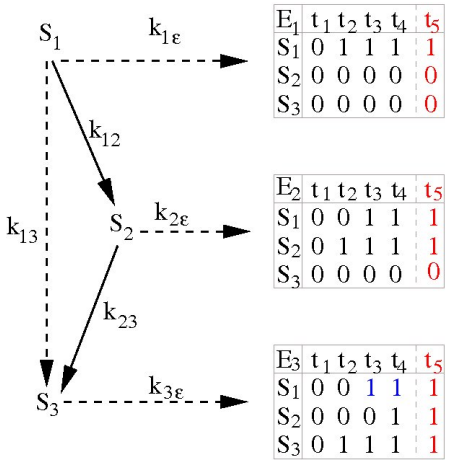

Figure 2.7: Idea of DNEMs in an elementary example - Shown is the hierarchical structure of a network and discrete time series data for three E-genes. A one indicates that a signal has reached the E-gene, while a zero indicates that the expression of this gene has not yet changed. Note, that the graph topology is consistent with the nested structure of ones in the final time point t_5, shown in red.

from any of the methods outlined in chapter 2 or from literature. The output consists of (a) the joint posterior distribution of rate constants describing the dynamics of signal propagation, and (b) a not necessarily transitive subgraph of the input graph that describes signal flow rather than upstream/downstream relation.

Model parameters for DNEM Let $D(i, k, l, s)$ denote the expression measurement of E_k in time point t_s of the l'th replication of a time series recorded after perturbation of S_i. We assume that the time spent for propagating a signal from node S_i to node S_j is exponentially distributed with a rate constant k_{ij}.

Recall that we do not observe the time spent for signal propagation between S-genes directly. Instead, we observe the time delay between a perturbation of an S-gene and the occurrence of downstream effects in E-genes. Following Markowetz *et al.* (49) we introduce parameters $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_N)$ to link E- to S-genes. If $\theta_k = i$, then E_k is linked to S_i. Moreover, we assume that every E-gene is linked to a single S-gene. The set of E-genes attached to the same S-gene is a regulatory module under the common regulatory control of the S-gene. The module of E-genes attached to S_i is denoted by $\mathcal{E}_i$. Finally, we introduce additional rate constants $k_{i\varepsilon}$ that represent the time delay between activation of S_i and regulation of its target module $\mathcal{E}_i$. A single common rate is used for all E-genes in the module.

We denote the complete set of rate constants including rates between S-genes and

rates between S- and E-genes by $\mathbf{K}$. A priori, we do not know which E-genes fall into which modules. The joint posterior distribution of $\mathbf{\Theta}$ and $\mathbf{K}$ will be inferred from the data. While the θ_k are discrete parameters by nature, rate constants are usually modeled as continuous parameters. However, for the sake of computational efficiency, we confine the rates to a discrete set of values denoted by $(\kappa_0, \dots \kappa_T)$. If the data includes time points $(t_1, \dots, t_T)$, we choose $(\kappa_0, 1/t_1, \dots, 1/t_T)$, where κ_0 is set to a high value (i.e. 1,000) that represents the very fast signal transduction through post translational protein modification like phosphorylation. Overall, we have a set of discrete parameters only $(\mathbf{K}, \mathbf{\Theta})$.

Prior distributions for model parameters Assuming independent prior distributions for $\mathbf{K}$ and $\mathbf{\Theta}$, Bayes's theorem yields

$$P(\Theta, K|D) = \frac{P(D|K, \Theta)P(K)P(\Theta)}{P(D).}$$

The prior distribution $P(\mathbf{\Theta})$ can be chosen to incorporate prior knowledge on the interactions of S- with E-genes. The prior provides an interface, through which the model can be linked to different biological data types in integrative modeling approaches.

The prior distribution $P(K)$ yields an interface for incorporating biological knowledge. If one knows that S_1 and S_2 fall into the same molecular signaling pathway, one can set $P(k_{12} = \kappa_0)$ to one, because signaling will operate on a high rate.

Marginal likelihood If we consider a fixed linear path g in Φ, which connects the S-gene S_i with the E-gene E_k:

$$S_i \xrightarrow{k_1} S_{j_1} \cdots \xrightarrow{k_{q-1}} S_{j_{q-1}} \xrightarrow{k_q} E_k,$$

where for simplicity of notation we reduce the double indices to single indices and write $k_1, k_2, \dots, k_q$ to denote the rate constants. We are interested in the time needed for propagating a signal from S_i down the path to E_k. More precisely, we want to calculate the probability, that the signal has reached E_k before some fixed time point t^*. If Z_g is the sum of q independent, and exponentially distributed random variables with rate constants $k_1, \dots, k_q$, then this probability equals $P(Z_g < t^*)$. The density function of Z_g is given by the convolution of independent exponential distributions

$$\Psi(t)_g = \int_0^\infty \cdots \int_0^\infty \delta\left(t - \sum_{u=1}^{q} \tau_u\right) \prod_{u=1}^{q} \psi_u(\tau_u)\, d\tau_1 \dots d\tau_q,$$

where $\psi_u(\tau) = k_u \exp(-k_u\tau)$ is the density of an exponential with rate k_u. Laplace-transformation yields a closed form for the cumulative distribution function of Z_g

$$F_g(t) = \sum_{b=1}^{q} \prod_{a \neq b} \left\{ \frac{k_a}{k_a - k_b} \right\} \left[1 - \exp(-tk_b)\right]. \tag{2.20}$$

See (48) for a proof. Note that the right hand side is not defined if two or more of the k_u are identical. However, as right and left limits exist and are identical, we can evaluate the probability by adding tiny distinct jitter values to the k_u.

In the general case a signal can be propagated from S_i to E_k via multiple alternative paths. In this case the fastest path determines the time delay for downstream effects to be seen. We enumerate all linear paths connecting S_i to E_k. For each path we construct a random variable Z_u as described above. The approximation of the probability that the signal has arrived at E_k before time t^* via at least one of the paths is given by

$$P_{S_i \to E_k}(t^*) = 1 - \prod_u (1 - F_u(t^*)) \tag{2.21}$$

In the general case, paths share edges, which lead to dependencies of signal propagation times. Nevertheless, simulations show that Equations (2.21) is a good approximation of the distribution of time delays, except maybe in some very unfortunate topological constellations. It is an approximation based on the assumption that the interactions among merging pathways can be neglected similar to the mean-field approximation from many body theories in statistical physics.

Equations (2.20) and (2.21) describe the stochastic nature of signal propagation in the cell. Before calculating the likelihood, we need to consider a second source of stochasticity, namely measurement error. Following Markowetz *et al.* (49), we denote the probabilities for false positive and false negative signals by α and β respectively. Assuming conditional independence, the likelihood factorizes into

$$\begin{aligned} P(D|K,\Theta) &= \prod_{D=1} P_{S_i \to E_k}(t_s)(1-\beta) + (1 - P_{S_i \to E_k}(t_s))\alpha \\ &\times \prod_{D=0} P_{S_i \to E_k}(t_s)\beta + (1 - P_{S_i \to E_k}(t_s))(1-\alpha), \end{aligned}$$

where the first product is over all data points, for which we observe a downstream effect, and the second product over those for which we do not.

Gibbs sampling With N E-genes, n S-genes and L edges in the input graph, the model comprises $N+n+L$ discrete parameters. For simplicity of notation, we reduce the double indices of rate constants to single indices such that the joint posterior is written

$$P(k_1, \ldots, k_{L+n}, \theta_1, \ldots \theta_N | D).$$

We initialize the parameters with random values from their domains. Then we iteratively cycle through all rate constants updating them by sampling from the conditional posterior distributions

$$p(k_i|\mathbf{K} - \{k_i\}, \mathbf{\Theta}, D).$$

With only discrete parameters, updating is straight forward: We calculate all values

$$p(k_i = \kappa_j)\, p(D|\mathbf{K} - k_i, \mathbf{\Theta}, k_i = \kappa_j),$$

normalize them to sum up to one, and draw a new value for k_i from this distribution. The iteration is completed by similarly updating all θ_k. We sample 10,000 times from the joint posterior distribution of parameters, discard the first 1,000 draws as burn in time, and summarize the remaining ones for inference of signal propagation. Choosing suitable values for the tuning parameters α and β protects the conditional posterior distributions from singularity, and ensures good mixing properties of the Gibbs sampler.

Inference of signal flow Under the natural assumption that perturbation effects propagate down the signaling network to all descendants of a perturbed gene, the nested structure of downstream effects resolves the network only up to its transitivity class. Network topologies with identical transitive closures produce the same nesting of downstream effects and, hence, can not be distinguished. Temporal data hold the potential of further resolving these transitivity classes. DNEM starts from a transitively closed network. Posterior distributions are calculated across a discrete set of rate constants including a very small rate constant κ_{T+1}. As explained above, $k_{ij=}\kappa_{T+1}$ reflects a network constellation, in which no signal is flowing through the edge from S_i to S_j. Note that if a rate constant is set to κ_{T+1}, the corresponding edge is not contributing to the likelihood according to Equation 2.21. The edge is effectively excluded from the model. Hence, in addition to estimating average time delays the Gibbs sampling procedure facilitates network refinement. If the posterior probability of the edge from S_i to S_j is $P[k_{ij=}\kappa_{T+1}|D] > p^*$, $p^* > 0.5$, we exclude the edge from the network. Of course the choice of p^* is subjective.

2.4.2 Fast Dynamic Nested Effects Models

A practical drawback of the DNEM is the long running times that Gibbs sampling needs to infer dynamics of very large networks. Frölich *et al.* (50) introduce a Fast Dynamic Nested Effects Models (FDNEMs) which circumvents the time consuming Gibbs sampling step for inference of signal propagation rates on the edges of a network(50). Moreover, this computationally attractive approach infers signaling cascades from high-dimensional perturbation time series measurements, hence allowing to discriminate direct from indirect perturbation effects and to resolve feedback loops. This approach directly extends the NEMs framework introduced by Markowetz *et al.*

(51) from the static to the dynamic case by unrolling the network structure over time. This approach does not aim to infer the rates of signaling. It only estimates the time lag between a perturbation and an observed downstream effect, there by providing the possibility to unroll the signal flow in the upstream signaling cascade over time. It uses a greedy hill climbing strategy in combination with a non-parametric bootstrap to assess confidences of inferred edges. The formulation of this dynamic model is just an extension of NEMs to handle cycles.

Model parameters for FDNEMs Similar to the DNEMs, let $D(i,k,l,s)$ denote the expression measurement of E_k in time point t_s of the l'th replication of a time series recorded after perturbation of S_i. t_s is replaced with t corresponding to the *index* of time point in a discrete time series, not the time point itself. These measurements could be p-values, counts or any other kind of statistics quantifying the effect of a knock-down for E-gene E_k under perturbation of S-gene S_i at time t. Suppose the true underlying pathway is given by Figure 2.8A. The signal flow is unrolled in this network over time (Figure 2.8B) in the following way: The node set $\mathcal{E}(t) = \{ E(t), E \in \mathcal{E}\}$, $\mathcal{S}(t) = \{ S(t), S \in \mathcal{S}\}$ of the dynamic network consists of a copy of the static network nodes, one for each time point $t = 1, ..., T$. An E-gene $E(t)$ is linked to $S(t)$ whenever E is linked to S in the static situation, i.e, it is determined by the same matrix $\Theta = |\mathcal{S}| \times |\mathcal{E}|$ as in the static case following (35). The actual unrolling takes place in the wiring of the S-genes. Informally, the static adjacency matrix Φ is converted to a $|\mathcal{S}| \times |\mathcal{S}|$ weighted adjacency matrix $\Psi = (\psi_{ij})$, where 0 means no edge and a value $\psi_{ij} > 0$ implies an influence of node i on E-genes downstream of node j delayed by ψ_{ij} time steps. Specifically, $T \geq \psi_{ij} \geq \Phi_{ij}$ for $i, j \in \mathcal{S}$. A non-zero entry ψ_{ij} implies that there are edges $S_i(t) \rightarrow S_j(t + \psi_{ij})$, $t = 1, ..., T - \psi_{ij}$. Furthermore, the convention $\psi_{ii} = 1$ is made. A positive time lag between nodes i and j in the model describes the number of time steps, after which a knock-down of node i results in an observed effect downstream of node j. This implies there are no assumptions made about the physical time it takes a signal at node j to produce a downstream effect at an E-gene. In contrast to classical Dynamic Bayesian Networks (52), an edge in the model may not connect consecutive time layers, but it may skip a certain amount of time steps (as it is the case for the entry $\psi_{S_2S_3} = 2$ in Figure 2.8B, which implies the edge $S_2(1) \rightarrow S_3(3)$. In other words, the model does not rely on a first order Markov assumption. In this way the unknown and variable time delays in perturbation responses are modelled due to the upstream signaling. In the following I refer to the model as FDNEM.

Marginal likelihood Considering the same parameterization like in static NEMs given in section 2.1, and assuming independence of time point measurements, the

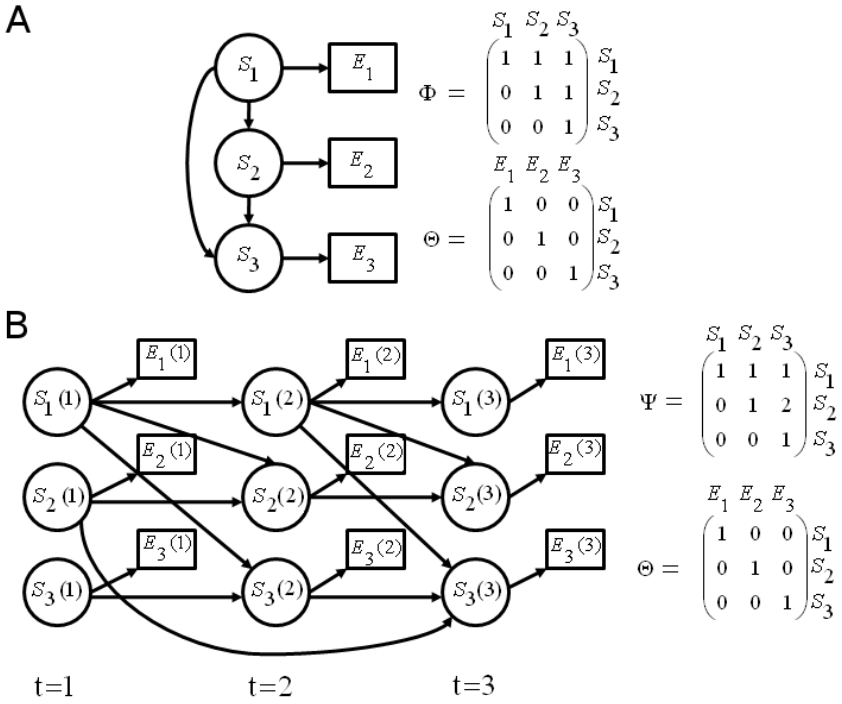

Figure 2.8: Standard NEM with 3 nodes - **A** static NEM is parameterized by a directed graph between S-genes encoded by Φ, together with a directed graph attaching each E-gene to an S-gene given by Θ. **B** Unrolling of the signal flow in the network from **A** along time. This corresponds to the network topology and parameterization of FDNEM.

marginal likelihood Equation 2.5 is extended to include time as :

$$p(D|\Psi,\Theta) = \prod_{i\in\mathcal{E}} \sum_{s\in\mathcal{S}} \prod_{l\in\mathcal{L}} \prod_{t=1}^{T} p(D_{il}(t)|\Psi,\Theta_{is}=1)P(\Theta_{is}=1) \tag{2.22}$$

To compute $p(D_{il}(t)|\Psi,\Theta_{is}=1)$ according to the proposed unrolling of the signal flow, a time dependent Boolean perturbation state for each S-gene s is introduced, which encodes an active state when perturbed as 0 and 1 when unperturbed. A knock-down of s corresponds to a switch $1 \rightarrow 0$. Since the perturbation state of s at a particular time step t is not observable, we identify it with the value $[s(t)]$ of a random variable $s(t)$. Let $pa(s)(t)$ denote the set of parents nodes of s at time t (i.e. the set $\{p|0 < \psi_{ps} < t\}$; which can be empty, if appropriate). Then, according to the unrolling of the signal flow over time, we write:

$$\begin{aligned} p(D_{il}(t)|\Psi,\Theta_{is}=1) &= \sum_{[s(t)]\in 0,1} p(D_{ikl}(t)|s(t)=[s(t)],\Theta_{is}=1) \\ &\times \quad P(s(t)=[s(t)]|pa(s)(t)) \end{aligned} \tag{2.23}$$

In the absence of more precise information we define:

$$\begin{aligned} P(s(t)=0|pa(s)(t)=[r]) &= \begin{cases} 1 & \exists p \in pa(s)(t) : [p]=1 \\ 0 & \text{otherwise} \end{cases} \\ P(s(t)=1|pa(s)(t)=[r]) &= 1 - P(s(t)=0|pa(s)(t)=[r]) \end{aligned} \tag{2.24}$$

The above definition can be interpreted as s is perturbed at time t, if any of its parents (including s itself) are perturbed. Assuming independence of observations the marginal likelihood $p(D_{ikl}(t)|s(t)=[s(t)],\Theta_{is}=1)$ can be calculated using the methods of static NEMs.

Using Priors for network structures and time delays In the last chapter a weighted adjacency matrix Ψ is introduced as a summary representation of a given network structure and time delays between S-genes and E-genes. Learning the structure of Φ is equivalent to learning the matrix Ψ based on the likelihood given in Equation 2.23. While scoring a given network, we assume observing an effect after longer time delays is less likely than smaller time delays. Moreover redundant edges are left out of the model since they do not change the likelihood of the model. These considerations are taken into account during the specification of $P(\Psi)$. Following Floerich *et al.*(2007) (33), prior probabilities for each edge are specified as follows :

$$p(\Psi|\nu) = \prod_{i,j} \frac{1}{2\nu} \exp \frac{-|\psi_{ij} - \hat{\psi}_{ij}|}{\nu}$$

where $\nu > 0$ is an adjustable scaling parameter. The parameter ν can be chosen according to the **BIC** criterion(53):

$$BIC = -2 \log p(D|\Phi) + \log(|\mathcal{E}|) \sum_{i,j} \mathbf{1}|\psi_{ij} - \hat{\psi}_{ij}| > 0$$

where $\sum_{i,j} \mathbf{1}|\psi_{ij} - \hat{\psi_{ij}}| > 0$ is an estimate of the number of parameters in the model. Usually we favor sparse network structures.

Network Learning for FDNEMs Learning the network structure Φ that fits the data best is equivalent to finding an optimal weighted adjacency matrix Ψ where the entries of Ψ_{ij} can take discrete values 0, ..., T . The greedy hill climbing strategy(section 2.3.3) is used . By this approach three search operators are used: edge weight increase ($\Psi_{ij} \mapsto \Psi_{ij} + 1$, if $\Psi_{ij} < T$), edge weight decrease ($\Psi_{ij} \mapsto \Psi_{ij} - 1$, if $\Psi_{ij} > 0$), edge reversal (exchange of Ψ_{ij} and Ψ_{ji}). At each step we apply all possible operators and accept the solution that increases the posterior likelihood most. This requires $O(|S|^2)$ likelihood evaluations per search step, where each likelihood computation according to Equation 2.23 has a time complexity of $O(T|\mathcal{E}||S|^2)$ on its own. Hence each search step requires $O(T|\mathcal{E}||S|^4)$ time. This is much faster than using the Gibbs sampling approach.

To further assess the confidence of the inferred network hypothesis on real experimental data, non-parametric bootstrapping (1000 times) is used. Thus, from the whole set $\mathcal{E}$ of available downstream effects bootstrap samples $\mathcal{E}' \subset \mathcal{E}$ of size $|\mathcal{E}|$ are randomly drawn with replacement. On each bootstrap sample a network hypothesis using greedy hill climbing is estimated. This allows the estimation of confidence intervals for each Ψ_{ij}.

2.5 A road map to network reconstruction using Nested effects models

Figure 2.9 organizes all the NEMs methods with respect to the following basic questions: Does the data include gene knockin or knockdown experiments or both? Does each experiment type involve single or multiple knockdown observations? If the former is true, does the data consist of the complete knowledge from the entire real network under study or it has only a partial view of the real network? Does the model allow for changes over time? Furthermore, does the dynamic and static NEMs include a discrete or continuous model? Some branches in the tree are missing corresponding to areas where methodology has not yet been established. The reason is not that it would be impossible to follow, but simply that we found no approach doing it. The

main contribution of this dissertation are network reconstruction for very incomplete data. They can be found in left-most branch of the tree in Figure 2.9 (red box). They are static probabilistic models for interventional data from incomplete data. The proposed approach in the next chapters partially reconstructs the upstream/downstream relations of non-transcriptional signaling networks from interventional data.

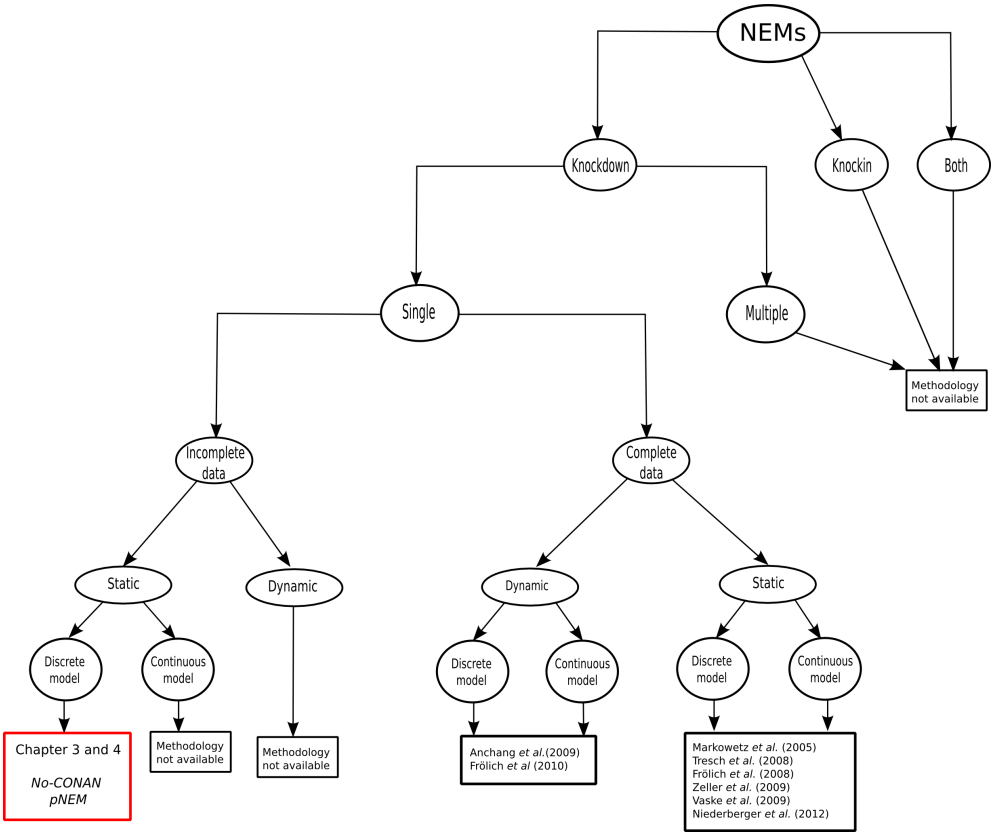

Figure 2.9: A guide to the literature on NEMs - The methods discussed in this chapter all fall into the right branch of node denoted as "Complete data" corresponding to methodology for single knockdown data. The next two chapters will deal with learning the static of a network from incomplete data, improve on accuracy of network reconstruction and making non-confoundable inferences on networks. The main contribution of this dissertation is modeling partially reconstruction of upstream/downstream relations of non-transcriptional signaling networks from interventional data.

3

Complications arising from hidden variables in Nested Effects Models

The previous chapter reviewed Nested Effects Models and their implementations. This chapter starts by demonstrating the importance of considering hidden variables for network reconstruction. The existence of such variables cause several complications for network reconstruction, especially in the case of interventional studies (section 3.1). I discuss complications that arise by considering these variables in the context of Nested Effects Models. I then investigate the possible influences of hidden variables on the data patterns generated from perturbation experiments (section 3.2). This gives a motivation to introduce a non-confoundable network analysis (section 3.3).

3.1 Network reconstruction and hidden confounding variables

One of the most basic statistical topics is the study of relationships between variables. Two variables may appear to be related in a certain way, but when a third, variable is taken into account, the apparent relation between these two variables might disappear or even reverse.This third variable is known as a confounding variable. In statistics, a hidden confounding variable is an extraneous variable in a statistical model that correlates positively or negatively with unknown variables (54). Such a relation between two observed variables is termed a spurious relationship. The hidden confounding variables are not directly observed but are rather inferred from other variables that are observed.

The hidden confounding variables are a set of hidden nodes in the context of a formal

statistical network reconstruction, like Bayesian networks (55), gaussian graphical networks (56), boolean networks (57), or nested effects models (58).

Ground Truth Network (GTN) The hidden nodes together with the observed nodes form a directed large network. The edges of the network encode causal relations. This means that if there is a directed edge from A to B, then perturbing A leads to changes in B. We call this large network the ground truth network (GTN).

Current State of the Art Network (CSAN) In practice the GTN is almost always unknown. Observed and modeled is only a subset of the GTN nodes resulting in a "Current State of the Art Network". This network only connects observed nodes. Importantly, in the GTN the hidden nodes can affect the observed nodes. A CSAN is reconstructed correctly if it is identical to the sub-network formed by the observed nodes in the GTN.

Hidden variables simplify structure One might naively think that if a variable is never observed, we can simply ignore its existence. At a certain level, this intuition is correct. We can construct a network over the observable variables. It captures the statistical dependencies among the observed variables (20). However, this approach is weak from a variety of perspectives: consider, for example, the network in Figure 3.1A. Assume that the data are generated from such a GTN, but that the node H is hidden. Artificial silencing data were generated for the extended networks as described previously (58). We generate data for GTN that include both observed and unobserved nodes. We then use NEM to reconstruct the sub-network (CSAN) of observed nodes only from data of the observed nodes. Figure 3.1B shows the reconstructed graph. From a pure representation perspective, this network is clearly less useful and contains incorrect edges. Hence, as a representation of the underlying process it is incorrect.

3.1.1 When hidden variables are known to exist

The problem of hidden nodes in network analysis has long been recognized, e.g. in causal inference theory (15). When a hidden variable is known to exist, we can introduce it into the network and apply known Bayesian network learning algorithms. If the network structure is known, algorithms such as EM (16, 17) or gradient ascent (18) can learn parameters. If the structure is not known, the structural EM algorithm can be used (59) to account for some missing observations. Moreover, the concept of structural signatures facilitates the detection and approximate location of a hidden variable in a network (60).

Latent variable approaches can not be applied for many hidden variables Latent variable approaches are only practical if the number of hidden nodes is small,

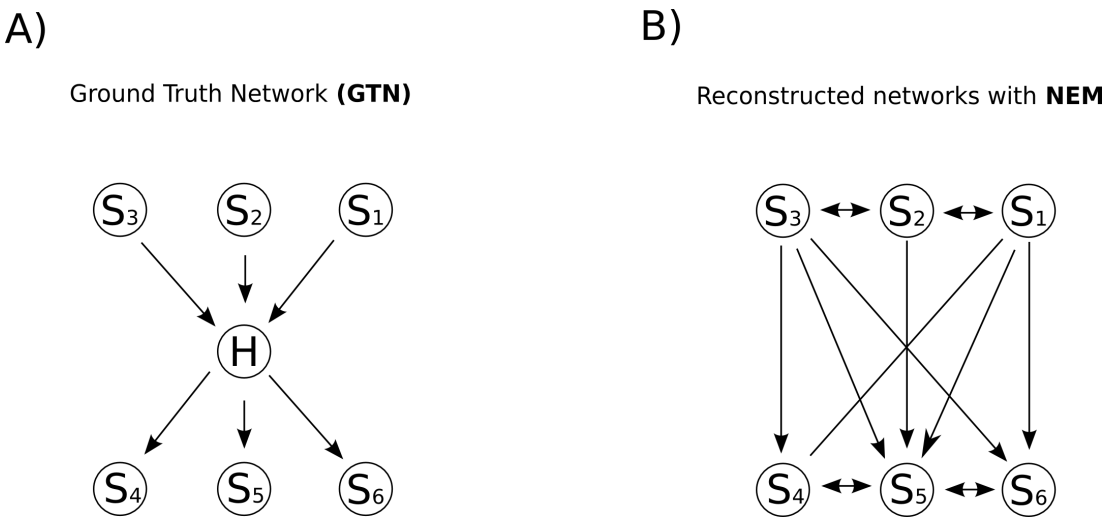

Figure 3.1: Hidden variables simplify structure - **A** shows a Ground Truth Network (GTN) that the node H is hidden. The hidden node mediates between its parents (S_1, S_2 and S_3) and its child (S_4, S_5 and S_6). Artificial silencing data were generated for the extended networks as described previously (58). **B** shows the reconstructed network using NEMs from artificial silencing data. Compared to the GTN, the reconstructed graph contains many incorrect edges.

an assumption that is critical, since in many domains including molecular biology we do not know how much we do not know.

The existence of many hidden variables might cause several complications for network reconstruction. It is essential to understand and survey the influences of hidden variables on network reconstruction before we make any further inferences. Here, we investigate complications and difficulties that arise from hidden variables for network reconstruction in the context of nested effect models (NEMs) (58). NEMs differ from the more general network reconstruction methods in the way that they are learned from interventional data. The next section summarizes the complications arising from hidden nodes in the NEMs framework and shows that network reconstruction can be awed in the presence of hidden variables.

3.2 Complications arising from hidden confounding variables in the context of Nested Effect Models

Recalling chapter 2, NEMs learn upstream/downstream relations in non-transcriptional signaling pathways from the nesting of transcriptional downstream effects when perturbing the signaling genes. In a nutshell: NEMs infer that a gene A operates upstream of a gene B in a pathway if the downstream effects resulting from silencing gene B are a noisy subset of those resulting from silencing gene A. Following (58) we call the perturbed genes in the signaling pathway S-genes and the genes that show expression changes in response to perturbation E-genes.

3.2.1 Data patterns in the language of Nested Effects Models

A NEM is a directed and possibly cyclic network that connects the S-genes representing the flow of information in the underlying signaling pathway. E-genes can be linked to single S-genes forming leaf nodes of the network.

Silencing data patterns The underlying data consists of gene expression profiles of gene silencing assays and corresponding controls. Typically, a pathway is stimulated both in cells where it is intact (controls) and in cells where it is partially disrupted by the silencing of one of its S-genes. If the silencing of an S-gene blocks the flow of information from the pathway-initiating receptor to the E-gene, the E-gene no longer changes expression in response to stimulation. In the language of nested effect models the E-gene shows a silencing effect with respect to the S-gene and

the crucial assumption is that E-genes must attach to, at most, one S-gene. In each experiment, one S-gene is silenced by RNAi and silencing effects on thousands of E-genes are measured. The expression data D_ij are assumed to be discretized to 0 and 1, with a 1 indicating that a silencing effect of S_j was observed on E_i. Signal propagation within the pathways is assumed to be deterministic, hence the silencing of S_j is expected to produce silencing effects in all E-genes downstream of S_j. Consequently, every network topology is associated with an expected data pattern across all silencing assays: the silencing scheme (58). If the network is acyclic, the silencing scheme defines a partial order relation on the S-genes reflecting the expected nesting of downstream effects. Noise comes into play at the level of observations. NEMs allow for both false positive and false negative observations accounting for them by fixed rates α and β in the likelihood. Hence, NEMs aim to detect a noisy subset relation in the observations D_{ij} and represent it as a directed network, where the directed edges can be interpreted as upstream/downstream relations of genes in the pathway.

3.2.2 Hidden nodes compromise NEM based network reconstruction

For general Bayesian networks, it is well known that hidden nodes can confound the reconstruction of networks (5, 61). Here we show that this problem still exists for the more specialized NEM. We generated data for networks that include both observed and unobserved nodes and reconstructed the sub-network of observed nodes: We generated 100 random networks of 4 nodes and extended them by $n = 0, 4, 8, 16$ additional hidden nodes. Artificial silencing data was generated for the extended networks as described previously (58). Only the data for the 4 observable nodes were used to reconstruct 4 node networks. This was compared to the corresponding sub-networks of the larger networks. The extended networks represent the ground truth signaling pathway while the 4 node sub-networks represent the small window through which we observe it.

For each silencing data set, we infer the NEM model M_{NEM} by the triple approach (58, 62). We then compute the positive predictive value of M_{NEM} with respect to M_{true} as the fraction of true positive edges out of all edges in M_{NEM}:

$$\text{positive predictive value} \quad (M_{NEM}) = \frac{TP}{TP + FP} \tag{3.1}$$

where TP are the true positive edges, and FP are the false positive edges. The positive predictive value is 1 whenever all edges of M_{NEM} are also part of M_{true}.

Figure 3.2 shows positive predicted values of network reconstruction (y-axis) for different noise levels (x-axis). The red line corresponds to network reconstruction without hidden nodes, while the green, blue and purple lines refer to 4, 8 and 12 additional hidden nodes respectively. We observe a marked decrease in network reconstruction performance when hidden nodes confound the flow of information of the observed nodes.

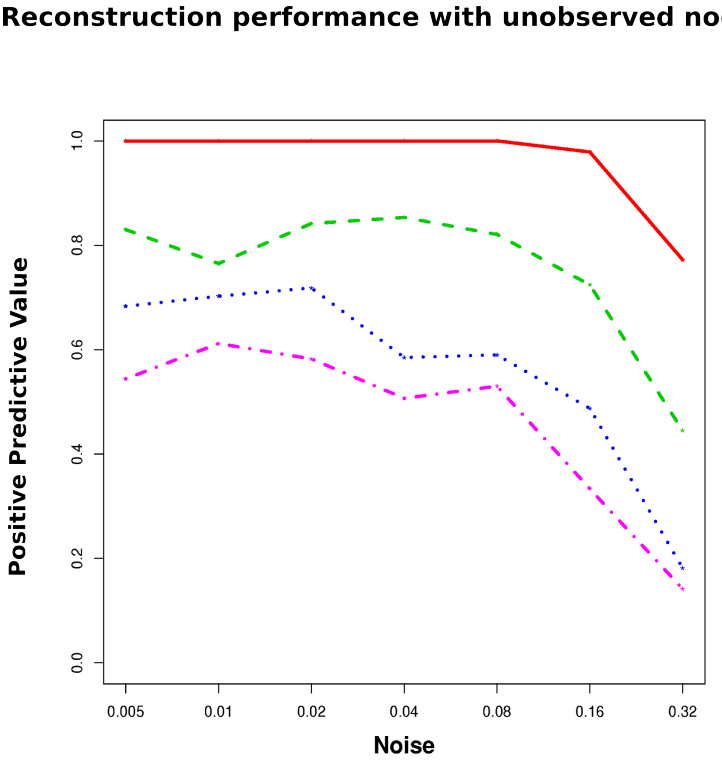

Figure 3.2: In simulations hidden nodes compromise network reconstruction - Shown is the accuracy of standard NEM based network reconstructions if hidden nodes are present. The x-axis shows the degree of noise used in the stimulations. The y-axis shows the positive predictive value of reconstructed edges of the subnetwork of observed nodes. The different lines correspond to different numbers of hidden confounders (red 0, green 4, blue 8, purple 12)

.

3.2.3 The smallest possible network consists of a pair of S-genes

The smallest possible network consists of a pair of S-genes. If we assume there is no hidden node involved, the pair of S-genes (S_1, S_2) can stand in one of the four relations:

$$S_1 \rightarrow S_2, S_1 \leftarrow S_2, S_1 \leftrightarrow S_2, S_1 \cdot \cdot S_2$$

Figure3.3 shows the four possible relations of a pair of S-genes with a specific E-gene for each. Depending on the E-gene attachment, a single E-gene can display one of the following data patterns: (1,1), (1,0), (0,1) and (0,0) in response to the perturbations of S_1 (first position) and S_2 (second position).

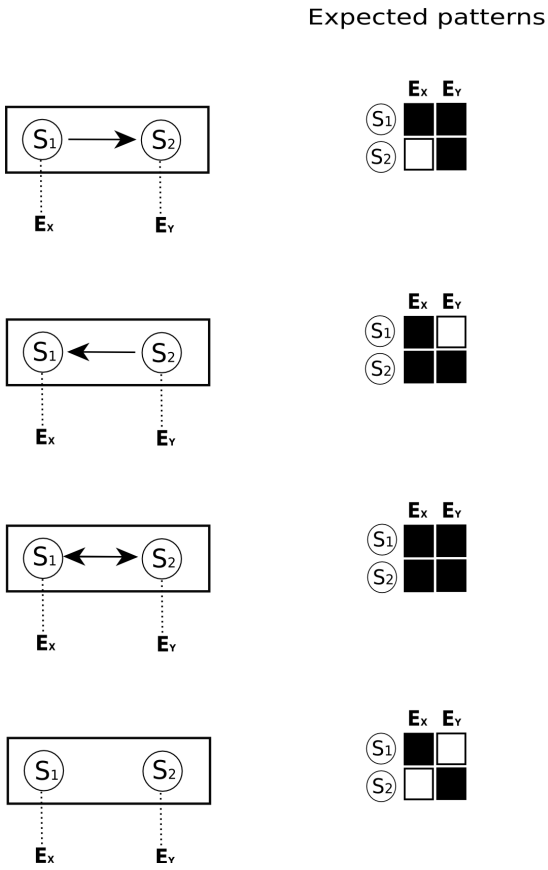

Figure 3.3: Silencing data patterns - Considering there are no hidden nodes, a pair of S-genes can stand in four relations. This figure shows the four relations and the corresponding silencing data patters for E-genes.

3. COMPLICATIONS ARISING FROM HIDDEN VARIABLES IN NESTED EFFECTS MODELS

Large networks In the context of Nested Effect Models following the presentation of Markowetz *et al*. (51), we look at large networks through narrow windows consist of pair or triple of S-genes for computational reason. This may cause a confounding problem for network reconstruction, although more S-genes are observed. When analyzing a pair of nodes the other nodes are not taken into account. Like hidden nodes they can confound our analysis and lead to incorrectly reconstructed upstream/downstream relationships in the network being modeled.

Figure 3.4 shows a GTN consist of 12 S-genes. Assume the E-genes position are known, one for each S-gene. Given the GTN, one can calculate the expected silencing data pattern after perturbation on S-genes. Consider, for example, the pair of (S_8, S_{11}) in Figure 3.4A. We expect an effect on E_8, E_{11} and E_{10} when perturbing S_8, while only E_{11} shows effect when perturbing of S_{11}. In practice, our observations will be noisy: there can be false positive (FP) and false negative (FN) observations. Given the noisy silencing data patterns, one can look at the GTN through a small window consisting of the pair (S_8, S_{11}) and infer their relation using NEM. However, the S-genes outside the window are hidden in this inference. Their effect can confound inference. They can mislead NEM to reconstruct the relations $S_8 \cdot\cdot S_{11}$ instead of $S_8 \rightarrow S_{11}$ for this pair (Figure 3.4B). Similar consideration for a pair (S_1, S_2) in Figure 3.4C shows how the E-genes downstream in the network can mislead NEM to distinguish between directed relation $S_1 \rightarrow S_2$ and the feedback loop $S_1 \leftrightarrow S_2$. These examples show that our observations are misleading when analyzing a small aspect independently of the entire network.

The uninformative E-genes Depending on the position of the pair of S-genes in the large graphs, there are two kinds of hidden nodes that can create a problem. The first type are located downstream of the pair under study. Since these hidden nodes are common children of the perturbed S-genes, their target effects respond to perturbation on upstream S-genes (Figure 3.4C). The corresponding E-genes attached to this type of hidden nodes produces the data pattern (1,1). The second type of hidden nodes are located upstream of the pair under study and their target effects do not respond to perturbation on the downstream S-genes. The E-genes which are attached to this type can only produce the uninformative pattern (0, 0); we called them *uninformative E-genes*. Figure 3.4B shows that the reconstruction can be confounded by large number of uninformative E-genes.

The (0,0) data pattern matters One might think, if the uninformative E-genes have (0,0) patterns and bring no information, we can simply ignore their existence. With this intuition, which is the case for binary NEMs (58), we remove all uninformative data patterns (0, 0) in order to remove confounding bias. We include only

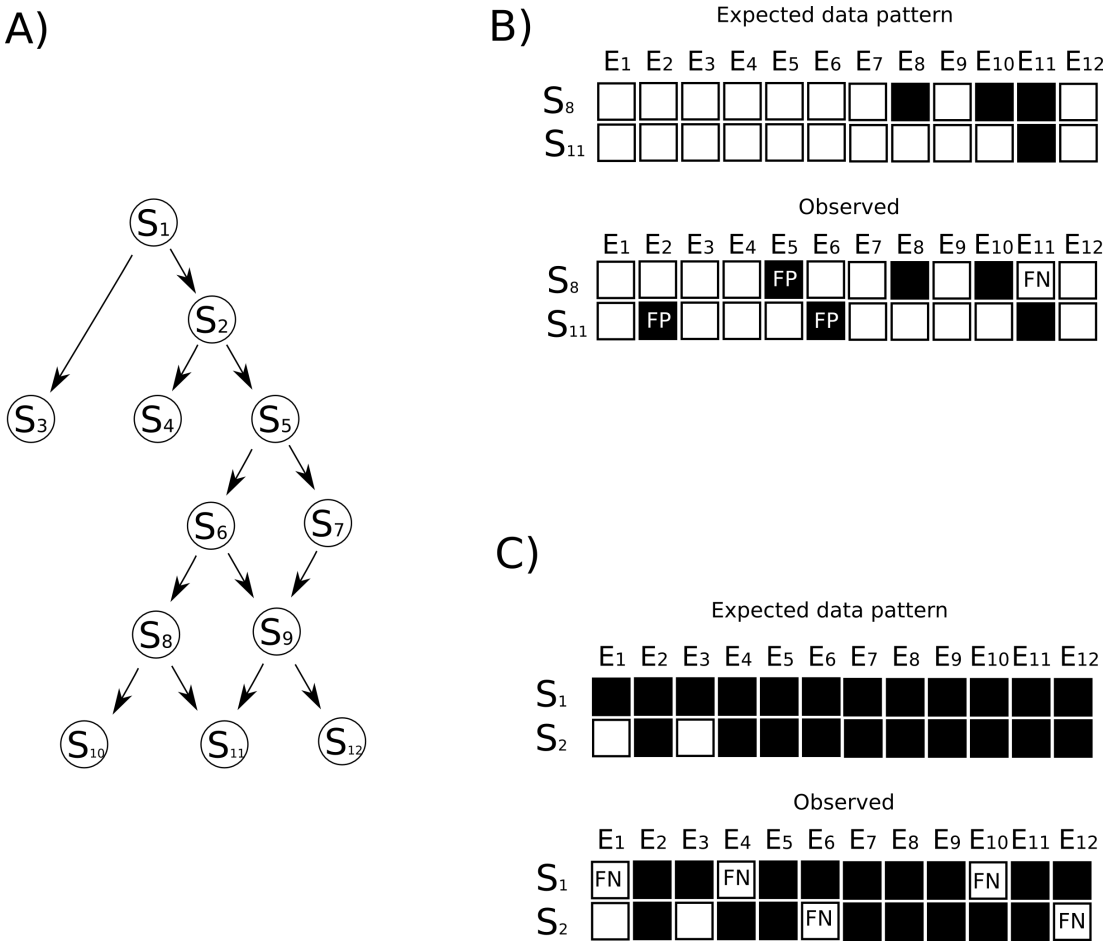

Figure 3.4: Small windows from entire network can be confounded by hidden nodes - **A** shows a network consist of 12 S-genes and 12 E-genes. The E-genes position are known, one for each S-gene. **B** shows the expected and observed silencing data patterns while analyzing the relation between S_8 and S_{11}. We are expecting to observe an effect on E_8, E_{11} and E_{10} by perturbing on S_8, while only E_{11} react to the perturbation on S_{11}. The effects for other S-genes do not show any response while analyzing S_8 and S_{11}. Given the noisy silencing data patterns, one can look at the GTN through a small window consist of the pair (S_8, S_{11}) and infer their relation using NEM. The S-genes outside the window are hidden for this analysis and their effects do not show any response to perturbation on S_8 and S_{11}. This inference can be confounded by hidden nodes, in order to reconstruct the relations $S_8 \cdot\cdot S_{11}$ instead of $S_8 \rightarrow S_{11}$ for this pair. **C** shows similar consideration for a pair S_1 and S_2. Here, the E-genes in downstream of the network can mislead NEM to distinguish between $S_1 \rightarrow S_2$ and $S_1 \leftrightarrow S_2$.

informative E-genes which have at least one 1 in their patterns. However, this approach is weak: consider, for example, the GTN in Figure 3.5 consisting of 4 S-genes A, B, C, D and 12 E-genes. We can formulate a prediction of what effects to expect after perturbation on a pair (B, C): E_7-E_{12} react to the the perturbation of B, while perturbing C only cause reaction of E_{10}-E_{12} (bottom left Figure 3.5). The noisy observations include false positive (FP) and false negative (FN) yield four uninformative E-genes E_3,E_4,E_6 and E_8 (bottom left Figure 3.5). Comparison between the expected data pattern and observations reveals that there are two types of (0,0) E-genes: the ones produced by noise from informative E-genes (E_8), and the ones which are assign to the hidden S-genes (E_3,E_4 and E_6). The latter have nothing to do with the pair under study, but the former are informative and can not be excluded. By excluding all (0,0) E-genes, we miss some of the informative E-genes.

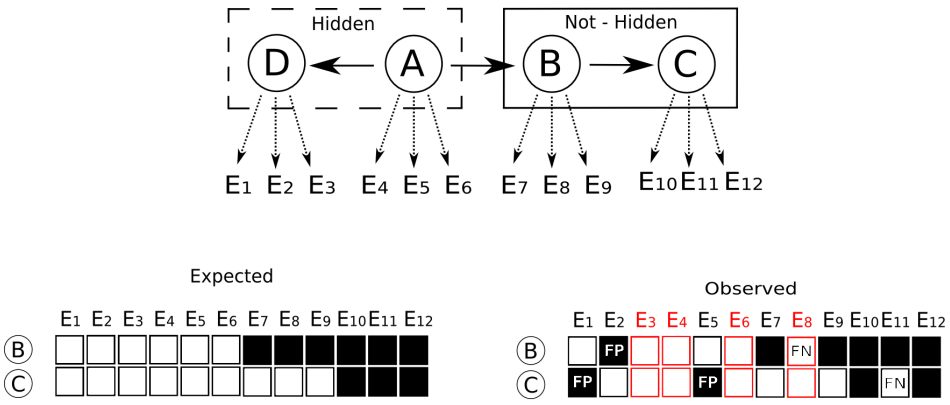

Figure 3.5: Archetypical uninformative E-genes - The top part of the figure shows a GTN consisting of 4 S-genes A, B, C, D and assignments of S-genes to specific E-genes (the dashed arrows). When analyzing a narrow window including B and C, other S-genes are hidden. Given GTN, we can formulate a prediction of what effects to expect after perturbing inside the window: perturbing B should cause E_7-E_{12} effects, while perturbing C should only cause E_{10}-E_{12} (bottom left plot). Comparison between the expected data pattern and observations shows two different types of uninformative E-genes: the ones produced by noise from one of the informative E-genes (E_8), and the ones which are assign to the hidden S-genes (E_3,E_4 and E_6).

3.3 Motivation for non-confoundable network analysis

So far we have summarized the difficulties that hidden nodes introduce in network reconstruction. NEMs can easily become trapped and reconstruct incorrect networks when there are many hidden nodes involved in the networks. This raises the questions: How can one remove the effect of hidden nodes for the network reconstruction? In other words, what kind of inference can be made that is non-confoundable by hidden nodes? These questions can be address by introducing a new class of graphs that can explain observed silencing data patterns accounting for the hidden nodes. In the next chapter, we investigate different silencing data patterns in order to find confoundable and non-confoundable network features. We go further and introduce a new set of upstream/downstream relations for a pair of nodes that can take account for hidden nodes. We then introduce the concept of non-confoundable networks analysis.

3. COMPLICATIONS ARISING FROM HIDDEN VARIABLES IN NESTED EFFECTS MODELS

4

Partial Nested Effects Models

The previous chapter dealt with the complications arising from hidden nodes for network reconstruction in the context of Nested Effects Models. In this chapter, I investigate what is arguably the most straightforward approach for considering the existence of hidden variables in this context. The main contributions are outlined in section 4.2 by extending the models for a pair of nodes that can explain all possible silencing patterns of intervention effects when the hidden players are known to exist. This approach is an extension of NEMs introduced in chapter 2 to infer non-confoundable upstream/downstream relations of non-transcriptional signaling networks from interventional data; an important limitation in NEMs. I introduce a simple edge-by-edge partial network reconstruction algorithm called Non Confoundable Network Analysis (No-CONAN) to derive non confoundable network properties. I then define a data structure that encodes the partially resolved networks called partial Nested Effects Models (pNEM) (section 4.3). Finally, I demonstrate its power in the controlled setting of simulation studies (section 4.4).

4.1 The Unknown-Unknowns of molecular biology

In February 2002, Donald Rumsfeld, the then US Secretary of Defense, stated at a Defense Department briefing: "There are known knowns. There are things we know that we know. There are known unknowns. That is to say, there are things that we now know we dont know. But there are also unknown unknowns. There are things we do not know we dont know" (63). The concept of unknown unknowns is eminent in many fields of research. Much scientific research is based on investigating known unknowns. In other words, scientists develop a hypothesis to be tested, and then, in an ideal situation, experiments are best designed to test the null hypothesis. At the outset the researcher does not know whether or not the results will support the null hypothesis. However, it is common for the researcher to believe that the result that

will be obtained will be within a range of known possibilities. Occasionally, however, the completely unexpected result is an unknown unknown.

The concept of unknown unknowns in molecular biology In the context of biological networks, known knowns make up our literature knowledge on physical and functional interactions of signaling molecules. Known unknowns might be what our current research projects are about, but unknown unknowns are those cellular mechanisms that we do not even anticipate today. They can be best appreciated from a historical perspective: Today, the role of many micro RNAs and epigenetic modifications of chromatin structure are known known mechanisms in many pathways. In other instances they are still in the realm of known unknowns. But going back 15 years in history they were unknown unknowns. Models of signaling pathways did not include them and the vast majority of molecular biologists did not anticipate the important role they would play.

When unknown unknowns become known Once unknown unknowns become known, two different scenarios can occur: (i) The new observations can add to our understanding of a network; or (ii) they can fundamentally change our perspective on the networks. In scenario (i) the network acquires more nodes and edges but the already existing parts of the network do not change. In scenario (ii), we learn that our old working hypothesis of the network was confounded by the mechanisms we were not aware of. The hidden effects of unknown unknown players made the interplay of the known players appear different from what they really are. This might severely bias both the computational and manual reconstruction of underlying networks. Their effects might be mixed up with the effect of known players. Also, separating these effects is difficult and may result in confounding, which is a major source of bias (15). This raises a new question what of our current understanding of biological networks can be condfounded by hidden mechanisms and what cannot.

4.1.1 What of our current understanding can be confounded by Unknown-Unknowns?

We believe this question can only be addressed meaningfully in the context of a formal statistical network reconstruction framework, such as Bayesian networks (55), Gaussian graphical networks (56), Boolean networks (57), or NEMs (58). In these frameworks, unknown unknowns are a set of hidden nodes. Recalling last chapter, the hidden nodes together with the observed nodes form the Ground Truth Network (GTN). However, in practice, observed and modeled nodes are only a subset of the GTN nodes, resulting in a Current State of the Art Network (CSAN).

Formal statistical biological network reconstruction frameworks attempt to identify networks over the known genes without assuming any unknown genes rather than considering the existence of unknown unknown genes. Therefore, they require CSAN to be as big as possible over the GTN. In the context of a nested effect model-based network reconstruction, we showed in the last chapter that in the presence of unknown unknown players, a reliable reconstruction of the full network is not feasible. In practice, unknown unknown genes interact with several of the known genes; therefore their effects might be mixed up with the effect of the known ones and cause incorrect networks to be identified from the given data. Since the reconstruction is not possible and can be confounded by unknown unknown players, one might wonder whether there is alternative way of analyzing the networks in a non-confoundable way. Here we ask the question: Which features of a network can be confounded by incomplete observations and which cannot?

Partial Ancestral Graphs (PAG) The reconstruction of a correct sub-network from very incomplete observations may be too ambitious. Alternatively, one can strive to derive features of a network that are correct, no matter what is going on outside the observation window. Colombo *et al*. (61) introduced the concept of partial ancestral graphs (PAG) extending work in (64). A PAG describes the common causal features of all directed acyclic graphs (DAGs) in a Markov equivalence class. This equivalence class comprises all DAGs that cannot be reliably distinguished if one accounts for possible effects of hidden nodes. The PAG does not fully reconstruct a network. Its information content lies in the network features it excludes, since this exclusion is guaranteed not to be an artifact caused by hidden nodes. The inference is not confoundable. (61) describes a computationally efficient algorithm that allows for the asymptotically consistent estimation of sparse, high-dimensional PAGs. A charming feature of the method is that it works exclusively using observational data.

PAGs cannot be applied to perturbation experiments Practical drawbacks of PAGs are the limited biological interpretability of general Bayesian networks learned from gene expression data and the inability to exploit functional information revealed in cell perturbation experiments. In fact, no applications to molecular biology have been reported to date.

Non-confoundable inference in the context of NEMs Here we follow the concept of a partial but non-confoundable network reconstruction in the context of nested effect models (NEM) (58). NEM differ from the more general networks of Colombo *et al*. in two ways: (i) They are learned from interventional data; (ii) all edges except for those involving leaf nodes encode deterministic information flow, e.g. local transition probabilities are zero or one. NEMs assume that the cellular information flow is deterministic, and stochasticity only comes in via noisy observations

(36). These features make non-confoundable network inference simpler and allow straightforward applications in systems biology.

In the next section we introduce a simple edge-by-edge partial network reconstruction algorithm called *Non-Confoundable Network Analysis (No-CONAN)* to derive non-confoundable network properties. Analogously to PAGs, we define a data structure that encodes the partially resolved networks in section 4.3.

4.2 Non-Confoundable Network Analysis (No-CONAN)

It is our aim to model upstream/downstream relations in signaling pathways in their totality. However, our approach will be to do this edge-by-edge: We analyze all pairs of S-genes S_1 and S_2 separately using only the data from silencing S_1 and S_2. Since our analysis will not be confoundable by genes outside of the observation window, it will also not be affected by the remaining S-genes that we voluntarily do not take into account.

For a pair of genes S_1 and S_2 we distinguish five possible upstream/downstream relations summarized in Figure 4.1A. (R1) S_1 is upstream of S_2, (R2) S_1 is downstream of S_2, (R3) S_1 and S_2 lie in a feedback loop in which case they are both upstream and downstream of each other indicated by the double arrow, (R4) S_1 and S_2 lie in independent modules of the network and do not interact with each other at all, and (R5) S_1 and S_2 are in different branches of a signaling network but jointly regulate at least one possibly hidden S-gene H. The five relations are encoded by the different edge types:

$$\mathcal{R} := \{R_1, \cdots, R_5\} = \{S_1 \to S_2, S_1 \leftarrow S_2, S_1 \leftrightarrow S_2, S_1 \cdot\cdot S_2, S_1 \to H \leftarrow S_2\} \tag{4.1}$$

With only two S-genes, an E-gene can show 4 different silencing patterns: It responds to both perturbations (1,1), only to one of them (1,0) and (0,1) or to neither (0,0).

4.2.1 Confoundable and non-confoundable network features

In practice it is not known where the missing player is located and, consequently, no perturbation data is available for the missing node. Let us assume the smallest window includes a pair of S-genes. Given each relation for a pair of S-genes, part of the data pattern might become infrequent and part of the data might not, regardless of the number of missing players.

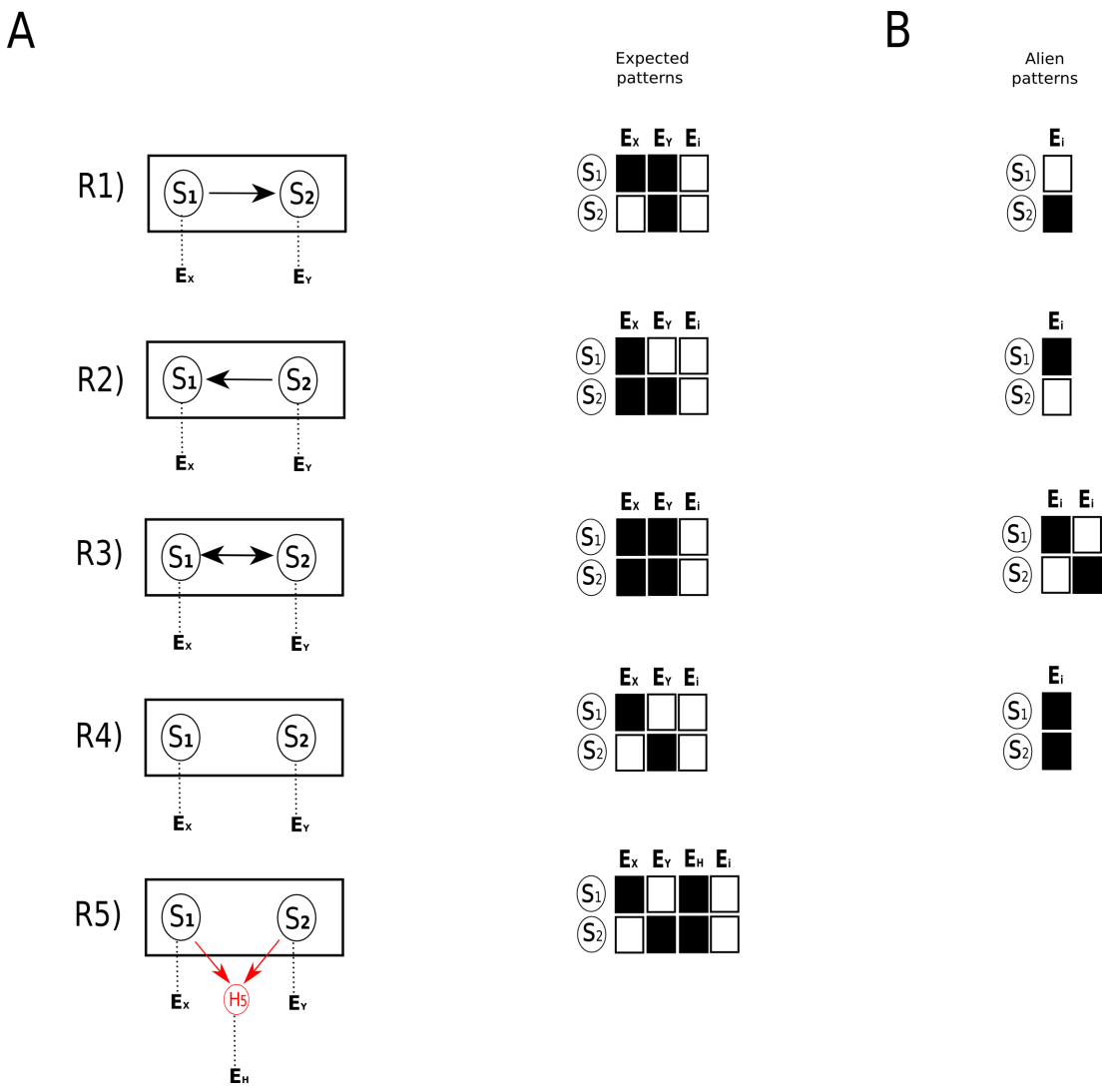

Figure 4.1: Pairwise upstream/downstream relations and their alien patterns - **A** shows the five possible possible relations R1,..., R5 together with their expected silencing patterns. (R1) S_1 is upstream of S_2, (R2) S_1 is downstream of S_2, (R3) S_1 and S_2 lie in a feedback loop indicated by the double arrow, (R4) S_1 and S_2 are disconnected, and (R5) S_1 and S_2 are in different branches of a signaling network but jointly regulate at least one possibly hidden S-gene H. **B** shows the corresponding alien patterns for each relation. Note that only relation R5 can produce all 4 possible silencing patterns. For the remaining relations at least one pattern is alien and not expected.

Alien patterns Each upstream/downstream relation induces an expected subset of four patterns (1,1), (1,0), (0,1) and (0,0). For example, in relation to (R1) an E-gene can be unconnected to both S_1 and S_2 in which case it does not show a silencing effect neither when silencing S_1 nor when silencing S_2, yielding the expected pattern (0,0). It can be attached to S_1 in which case it is expected to show an effect when silencing S_1 but not when silencing the downstream gene S_2, yielding the expected pattern (1,0). And lastly, it can be linked to S_2 and show silencing effects both when silencing S_1 and S_2, yielding the pattern (1,1). Figure 4.1A gives the set of expected silencing patterns for all five upstream/downstream relations. Note that only relation R5 can produce all 4 possible silencing patterns. For the remaining relations at least one pattern is not expected. We call these unexpected patterns *alien patterns* (Figure 4.1B).

4.2.1.1 There are nine possible locations for a hidden player

We next investigated the possible influence of hidden factors on the sets of expected and alien patterns (Figure 4.2). There are nine possible positions of a hidden confounder. The silencing patterns associated with these positions are shown in Figure4.2. The most important observation is that any position of hidden confounders in the network does not change the sets of expected silencing patterns (Figure4.2).

Note that the hidden note marked in red produces the alien pattern of R4; however, this position of a hidden confounder transfers R4 into R5. In other words, we have accounted for this problem by distinguishing the two relations R4 and R5 from the beginning. The conclusion that no alien patterns can occur through confounding facilitates our non-confoundable analysis: If the observation of an alien pattern cannot occur through confounding effects it must be due to noise in the observation. Note that the assumption of deterministic signal propagation is crucial here. In relation R1 we assume that a perturbation of S_1 is deterministically propagated to S_2, which rules out the silencing pattern (0,1).

4.2.2 Alien silencing patterns are the clue to a non-confoundable network analysis

The key idea of non-confoundable network analysis is the definition of alien silencing patterns that cannot be confounded by unobserved nodes. The existence of unknown unknown players in the network does not change the fact that alien patterns can only occur due to noise. Moreover it does not affect the probability with which they occur. If we observe an alien pattern exceedingly often, we can hold this as evidence against the network hypothesis. By observing the number of alien patterns for each

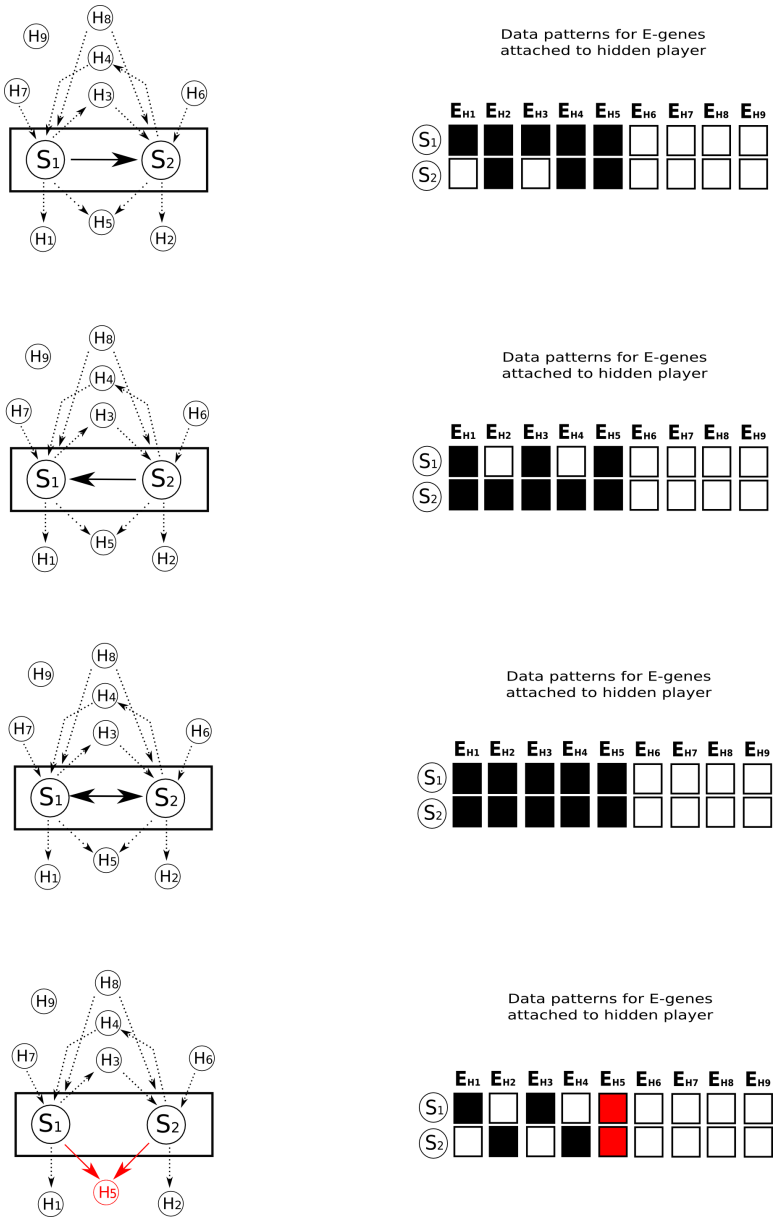

Figure 4.2: The possible influence of hidden factors on the sets of expected patterns - Hidden nodes are introduced in all possible configurations and the expected patterns of E-genes attached to the hidden nodes are shown. In R4 the Hidden node marked in red produces the alien pattern of R4. Note that this constellation leads to the constellation in R5.

relation, we have enough information to exclude incorrect relations on the basis of the incomplete data. When the reconstruction might be confounding we can exclude incorrect network features which are not compatible with the data. This means that if, for the constitutive relation, we observe a significant amount of alien patterns we can exclude it. We can ask whether or not the data supports the exclusion of the relations, based on the number of alien patterns. However, this inference cannot be confounded with the presence of unknown unknown players. In the next section we derive a simple polynomial test for inferring such non-confoundable characteristics of networks.

4.2.3 The accumulation of alien patterns is evidence against respective upstream/downstream relations

For a pair of S-genes we can systematically consider all five upstream/downstream relations and see whether they conform with the observed data. Each of the relations R1-R4 has at least one alien pattern. Every observation of an E-gene that displays this alien pattern is evidence against the respective relation. A few alien patterns can occur due to observation noise but a large number of alien patterns is unlikely. We will set up a test to detect significantly high occurrences of alien patterns.

Binary NEMs (58) model observe noise by a false positive rate α, the probability that an observed effect is a noise artifact and a false negative rate β, the probability that we will miss a true silencing effect. Further, the occurrence of observation errors is assumed to be independent across E-genes. We can derive limits for the probability that a certain number k of alien patterns occur given a relations $R \in \mathcal{R} - R_5$ in (4.1).

Non-symmetric relation: $S_1 \longrightarrow S_2$

If $R = R1$, for a certain E-gene this relation can produce (1,0), (1,1) and (0,0). When R1 holds true, the alien pattern (0,1) needs to be produced by noise from one of the three expected patterns (1,0), (1,1) and (0,0). Starting from (1,0) requires both a false positive and a false negative observation which happens with probability $\gamma_1 = \alpha \cdot \beta$, starting from (1,1) we need one true positive and one false negative observation which occurs with probability $\gamma_2 = \beta \cdot (1 - \beta)$. Finally, generating the alien pattern (0,1) from (0,0) requires one true negative and one false positive observation and occurs with probability $\gamma_3 = (1 - \alpha) \cdot \alpha$.

If we have n E-genes in total, we can think of observing an alien pattern as a repetition of Bernouli trials with a success rate γ_1, γ_2 or γ_3. So the probability that k of them show (0,1) and they produce by noise from (1,0) is:

$$P(n_{01} = k) = \binom{n}{k}\gamma_1^k(1-\gamma_1)^{n-k}$$

Setting $\gamma_{R1} = \max(\gamma_1, \gamma_2, \gamma_3)$ yields the following boundary that we have

$$P(K \geq k | S_1 \longrightarrow S_2) \leq \sum_{i=k}^{n} \binom{n}{i} \gamma_{R1}^i (1-\gamma_{R1})^{n-i}, \tag{4.2}$$

where k is the observed number of alien patterns, n the total number of E-genes and γ_{R1} an upper bound for the probability of observing the alien pattern.

Non-symmetric relation: $S_1 \longleftarrow S_2$

Similar consideration can be applied for this relation. If $R = R2$, for a certain E-genes this relation can produce (0,1), (1,1) and (0,0). When R2 holds true, the alien pattern (1,0) needs to be produced by noise from one of the three expected patterns (0,1), (1,1) and (0,0). Starting from (0,1) requires both a false positive and a false negative observation which happens with probability $\gamma_1 = \alpha \cdot \beta$, starting from (1,1) we need one true positive and one false negative observation which occurs with probability $\gamma_2 = (1-\beta) \cdot \beta$. Finally, generating the alien pattern (0,1) from (0,0) requires one true negative and one false positive observation and occurs with probability $\gamma_3 = \alpha \cdot (1-\alpha)$. Setting $\gamma_{R2} = \max(\gamma_1, \gamma_2, \gamma_3)$ similarly we obtain:

$$P(K \geq k | S_1 \longleftarrow S_2) \leq \sum_{i=k}^{n} \binom{n}{i} \gamma_{R2}^i (1-\gamma_{R2})^{n-i}, \tag{4.3}$$

where k is the observed number of alien patterns (1,0), n the total number of E-genes and γ_{R2} an upper bound for the probability of observing the alien pattern.

Symmetric relation: feedback loop $S_1 \longleftrightarrow S_2$

Similar consideration can be applied for symmetric relations. If $R = R3$, for a certain E-genes this relation can produce (1,1) and (0,0). When R3 holds true, the alien patterns (0,1) and (1,0) need to be produced by noise from one of the two expected patterns (1,1) and (0,0). Starting from (1,1) the both alien patterns require a false negative and a true positive observation which happens with probability $\gamma_1 = \beta \cdot (1-\beta)$, starting from (0,0) we need one true negative and one false positive observation which occurs with probability $\gamma_2 = (1-\alpha) \cdot \alpha$. Setting $\gamma_{R3} = \max(\gamma_1, \gamma_2)$ yields the following boundary that we have

$$P(K \geq k | S_1 \longleftrightarrow S_2) \leq \sum_{i=k}^{n} \binom{n}{i} \gamma_{R3}^i (1-\gamma_{R3})^{n-i}, \tag{4.4}$$

where k is the observed number of alien patterns, n the total number of E-genes and γ_{R3} an upper bound for the probability of observing the alien pattern. Note that the difference between this relation and others lies in the in the probabilities of observing the alien patterns when the false positive and false negative are equal. In this case we can compute the exact probabilities of observing the alien patterns instead of boundary.

Symmetric relation: Disconnected $S_1 \cdot \cdot S_2$

Finally, if $R = R4$, for certain E-genes this relation can produce (0,1), (1,0) and (0,0). When R4 holds true, the only alien pattern (1,1) needs to be produced by noise from one of the three expected patterns. Starting from (1,0) requires one true positive and one false positive which happens with probability $\gamma_1 = (1-\beta)\cdot\alpha$, starting from (0,1) we need one false positive and one true positive which occurs again with probability $\gamma_2 = \alpha \cdot (1-\beta)$. Finally, generating the alien pattern (1,1) from (0,0) requires two false positive which occurs with probability $\gamma_3 = \alpha \cdot \alpha$. Setting $\gamma_{R4} = \max(\gamma_1, \gamma_2, \gamma_3)$ yields the following boundary that we have

$$P(K \geq k | S_1 \cdot \cdot S_2) \leq \sum_{i=k}^{n} \binom{n}{i} \gamma_{R4}^{i} (1-\gamma_{R4})^{n-i}, \tag{4.5}$$

where k is the observed number of alien patterns, n the total number of E-genes and γ_{R4} an upper bound for the probability of observing the alien pattern.

Calibration parameter is needed to exclude relations For all $R \in \mathcal{R} - \{R_5\}$, γ_R is a bound for the probability of observing the alien pattern of R. If some of the above probabilities become sufficiently small, we gather evidence against the respective relations. We exclude a relation R, if and only if

$$P(K \geq k | R) < \kappa, \tag{4.6}$$

where κ is a calibration parameter that is set to 0.05 in all applications in the next chapters. Note that $R5$ cannot be rejected since it does not have an alien pattern.

We never exclude R_5 Note that $R5$ cannot be rejected since it does not have an alien pattern. This relation can produce all silencing data patterns (1,0), (0,1),(1,1) and (0,0).

Non-confoundable Network Analysis (No-CONAN) Given each relation and observation, we can calculate the boundaries for probabilities of observing alien patterns for each pair. We can also use this information against each relation and test whether or not the number of alien patterns occurs due to random fluctuation. With

these probabilities we quantify the possibilities of excluding relations given observation. *No-CONAN* is a testing approach that aims to exclude incorrect models rather than reconstructing models that can be confounded by unknown players. Exclusion of an network hypothesis cannot be confounded by unknown players. This makes our approach much more practical and effective than reconstructing an incorrect model with incomplete observation.

4.3 Partial Nested Effects Models (pNEM)

Partial network reconstruction If $R5$ is correct and we can reject relations $R1$-$R4$, leaving only relation $R5$ as compatible with the data, we have fully resolved the relation of S_1 and S_2. In cases where $R5$ is incorrect, the best we can achieve is a situation where all but one relation from $R1$-$R4$ is rejected leaving us with one edge type and the ever present possibility that $R5$ is true. However, this does not need to be the case. It is possible that we cannot reject several relations, leaving us with higher uncertainties about the true structure of the signaling network. We do not further resolve the network but confine ourselves to describing what we know and what we don't know. To do this we introduce the new data structure of a partial Nested Effects Model (pNEM). A pNEM is a graph connecting all S-genes but using a variety of different edge types. Each edge type is describing a set of relations that could not be rejected. This language of edge types is summarized in Figure 4.3. For example, if we exclude all relations except $R5$, there is no edge between S_1 and S_2. If we reject all relations except $R3, R4$ and $R5$ we draw a red double-sided edge, and so on. Sixteen different edge types are needed to encode our partial network knowledge. In the next section we show an example of a pNEM.

Power of the test Equation (4.6) has the form of a statistical test. When choosing a κ of a sufficiently small value we bound the probability of excluding a correct relation. The null hypothesis is that the tested relation is correct and that all observed alien patterns are due to noise alone. However, a small κ also leads to poorly resolved networks with only a few excluded relations. This raises the issue of the power of the test. An edge between two S-genes is well resolved if the true relation generates many E-genes with silencing patterns that are alien to many alternative relations. For example, if the relation $S_1 \rightarrow S_2$ holds true, every E-gene that is attached to S_1 and produces the expected pattern (1,0) produces evidence against the competing relations $S_1 \leftarrow S_2$ and $S_1 \leftrightarrow S_2$ since (1,0) is alien to both these relations, but not against the relation $S_1 \cdot\cdot S_2$, since (1,0) is not alien to it. However, E-genes attached to S_2 with the expected pattern (1,1) produce evidence against $S_1 \cdot\cdot S_2$. If we have enough E-genes of both types we will be able to reject all relations except

the correct one and the non-rejectable relation $R5$. Inspecting Figure 4.1 points to a problem with edges that are of the type $S_1 \leftrightarrow S_2$, since in this constellation only the patterns (1,1) and (0,0) are produced and none of the alien patterns of the two directed relations $R1$ and $R2$. Since NEMs operate on transitively closed networks the relation $S_1 \leftrightarrow S_2$ is indicative of genes involved in a feedback loop. In other words our method is not capable of reliably detecting feedback loops; a non-circular constellation can often not be ruled out. Nevertheless, our method is valid also for biological networks with feedback loops. It does not produce spurious results in this case, but reports that it can-not resolve the loop reliably. If, in contrast, the true network is not cyclic, our method has the potential to exclude a loop reliably.

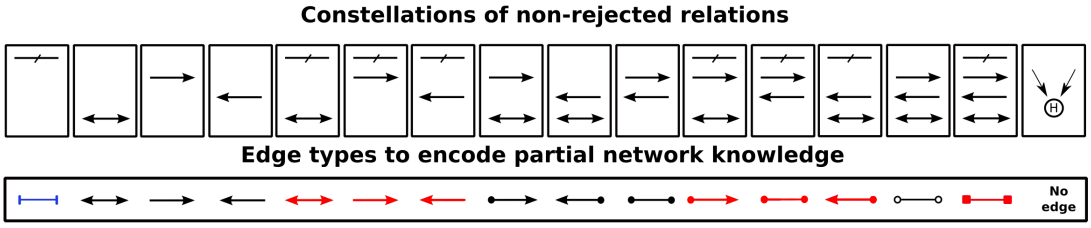

Figure 4.3: The pNEM code - The top row of boxes shows combinations of relations and the bottom boxes show the corresponding edge types we use to encode that all edges in the set could not be excluded by NoCNA.

.

4.3.1 Advantages of non-confoundable network analysis for incomplete data

Non-confoundable inference The information in a pNEM lies in the upstream/downstream relations between signaling genes that it excludes. A pNEM encodes what we know but also what we cannot know unless we can be sure that we have observed all nodes of the network. Importantly, a pNEM cannot be confounded by hidden nodes. The uncertainties left with certain edges are the price we have to pay to ensure that our results are not confoundable by mechanisms outside the window of observations.

Note that non-confoundability makes the edge-by-edge reconstruction strategy attractive. When reconstructing a network edge by edge we observe the pathway through a series of very small windows: Ones that open our view on only two S-genes, but many E-genes. In other words, we make the window of observation even smaller than necessary. However, our inference strategy is not confounded by events outside these

windows, neither by those that we did not observe nor by those that we voluntarily did not take into account when deciding on the existence and orientation of an individual edge.

Computationally fast Another great advantage of pNEM is the increase in speed. The number of models we have to test for n S-genes is $\binom{n}{2} \cdot 4$, which grows quadratically in the number of perturbed genes and remains feasible even for hundreds of genes. Additionally, building up the partially reconstructed final graph is easy, since it is defined by the set of all pairwise models.

4.3.2 Limits of non-confoundable network analysis for incomplete data

The method we described can only partially reconstruct features of the pathway, not the full topology. This stems from inherent limits of reconstruction from indirect observations. We discuss here *partially reconstruction* and *loop reliably*.

Partial reconstruction There is only one situation that pNEM can fully reconstruct the relation where all the relations $R1$-$R4$ rejected. In this case relation $R5$ has the only possibility to present data. Regardless of this situation, the best we can achieve is a situation where all but one relation from $R1$-$R4$ is rejected leaving us with one edge type and the ever present possibility that $R5$ is true. However, this does not need to be the case. It is possible that pNEM cannot reject several relations, leaving us with higher uncertainties about the true structure of the signaling network. pNEM does not further resolve the network but confine inferences to describing what is known and what is unknown.

Loop reliably Since NEMs operate on transitively closed networks, the relation $S_1 \leftrightarrow S_2$ is indicative of genes involved in a feedback loop. In other words our method is not capable of reliably detecting feedback loops; a non-circular constellation can often not be ruled out (for more detail see section 4.4). Nevertheless, our method is valid also for biological networks with feedback loops. It does not produce spurious results in this case, but reports that it cannot resolve the loop reliably. If in contrast the true network is not cyclic, our method has the potential to exclude a loop reliably.

4.4 Simulation Experiments

The last sections introduced a polynomial test approach for inferring non-confoundable characteristics of signaling networks. We will demonstrate its potential in two steps. First, we investigate accuracy and sample size requirements in different controlled simulation settings. We test the performance of No-CONAN in the context of simulation experiments using artificial data. In such simulations the true state of the network or ground truth network is known, unlike in biological scenarios. Moreover, the artificial data fully conforms to all assumptions of NEMs, which is certainly not the case for real biological data. In a second step, we compare the performance of No-CONAN with NEMs when the hidden nodes are known to exist.

4.4.1 Accuracy and sample size requirements

A first test of validity of a complex data model is to test its performance in simulation scenarios where data is artificially generated according to the model assumption. This section evaluates how our algorithm responds to different levels of noise in the data, different numbers of E-genes, different numbers of S-genes and how accurate it is. In order to answer these questions, we introduce the general set-up for data generation and choosing parameters. We then investigate the performance of our approach in different simulation studies.

4.4.1.1 Set-up for data generation

Data generation consists of four steps:

1. **S-genes:** Randomly generate a directed acyclic graph T with n_S nodes and n_{ed} edges. This is the core topology of S-genes.

2. **E-genes:** Connect n_E E-genes uniformly to core T. This forms an extended topology T'.

3. **Unrelated E-genes:** Connect another n_{UE} E-genes which are unrelated to the networks T'. These have an expected silencing pattern of (0,0) but display occasional silencing effects due to noise.

4. **Data:** Generate one random dataset D from the extended topology T'. We use only one repetition per knock-out experiment. For each knock-out experiment the response of all E-genes is simulated from T' using error probabilities α and β. The false negative rate and false positive rate are equal and varied between very low and very high noise.

4.4.1.2 Dependency on the noise levels

In a first simulation we examine the performance of No-CONAN on 100 random networks of size $n_S = 10$ and generate data for these networks using noise levels varying between 0.005 (very low) and 0.32 (very high). We attach a total of n_E=100 E-genes uniformly to the S-genes of the network and add another n_{UE}=900 E-genes which are unrelated to the networks. The number of unrelated E-genes might effect the power of the testing (see section 4.4.2.5).

We then run No-CONAN on every pair of nodes in each of the 100 networks and reject all relations possible using $\kappa = 0.05$. The results are organized according to the true underlying relations in Figure 4.4. Each of the five plots corresponds to one true relation. The x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors. For example the left-most plot corresponds to all situations where the true relation between nodes is $S_1 \longrightarrow S_2$. Rejection rates for this relation are marked in blue and we can see that the relation is not falsely rejected even for very high noise levels. In contrast the 3 competing relations marked in red, purple and green are virtually always rejected except for very high noise levels and even for maximal noise we reject them in about half of the cases. We do similarly well for the next two relations. If the true relation is the feedback loop $S_1 \leftrightarrow S_2$, we still have hardly any false positive rejections but lose almost all power in rejecting the two directed relations. As described in the previous section, this is expected since a feedback loop does not produce the alien patterns of these relations.

4.4.1.3 Dependency on the number of E-genes

In the next simulation we looked at the influence of the different number of E-genes in the previous study. Here we follow the four steps for data generation in section 4.4.1.2 with only one change. In the step two we connect $n_E \in \{100, 500, 1000, 5000\}$ E-genes uniformly to the core topology. We then run No-CONAN on every pair of nodes in each of 400 networks and reject all relations possible using $\kappa = 0.05$.

The results are organized according to the true underlying relations in Figure 4.5. Each of the four rows corresponds to different number of E-genes in the simulated data sets (100, 500, 1000, 5000), while columns represent true relations. In each plot, the x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors. For example the left-most column corresponds to all situations where the true relation between nodes is $S_1 \longrightarrow S_2$. Rejection rates for this relation are marked in blue and we can see that the relation is not falsely rejected with different number of E-genes even for very high noise levels. In contrast the 3 competing relations marked in red, purple and

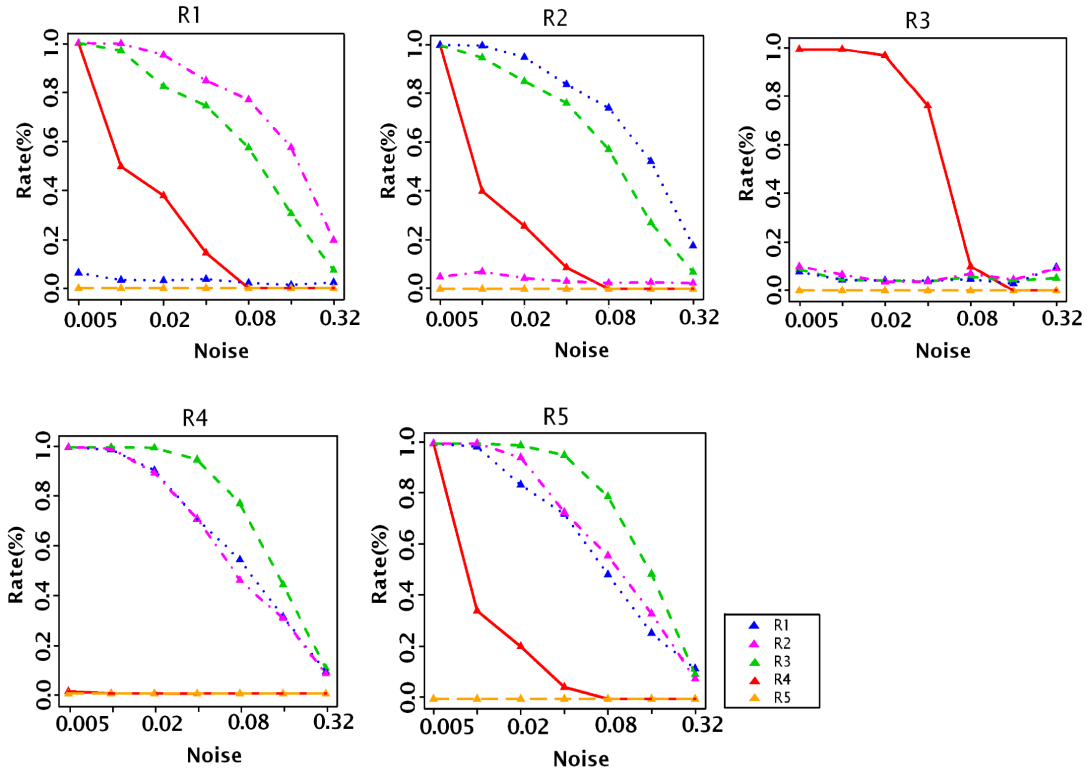

Figure 4.4: Dependency on the noise levels - Small network simulations. Each of the five plots corresponds to one true relation. The x-axis shows the different noise levels. The y-axis shows the relative frequency of rejecting the relations (R1: blue, R2: purple R3: green, R4: red and R5: orange).

green are virtually always rejected by increasing the number of E-genes. Overall these results show that we still have hardly any false positive rejections but gain power in rejecting the incorrect relations by increasing the number of E-genes.

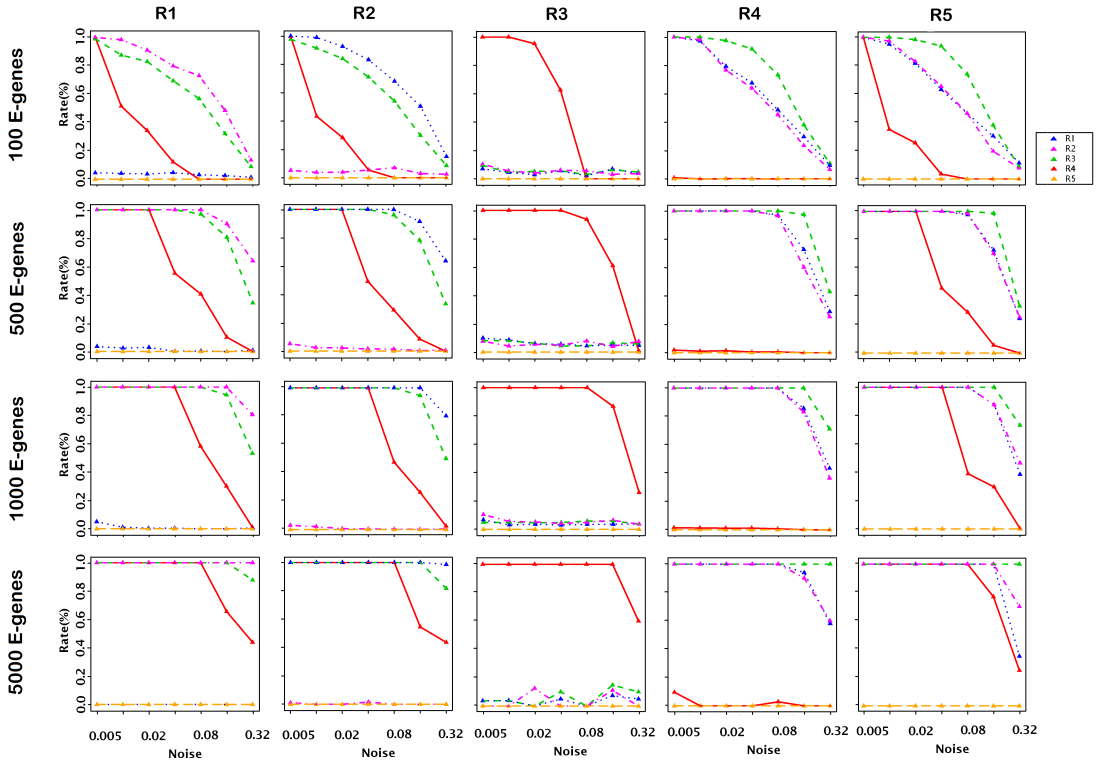

Figure 4.5: Dependency on the number of E-genes - Rows correspond to the number of E-genes in the simulated data sets (100, 500, 1000, 5000), while columns represent true relations. In each plot, the x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors (R1: blue, R2: purple R3: green, R4: red and R5: orange). Overall these results show that we still have hardly any false positive rejections but gain power in rejecting the incorrect relations by increasing the number of E-genes.

4.4.1.4 Dependency on the number of S-genes

In a third simulation we investigated the performance of No-CONAN for varying number of S-genes. In contrast to the previous simulation study, here the number of E-genes is fixed to n_E=200 and we vary the number of S-genes with n_S=4,8 and 20. The other data generation steps are similar to section 4.4.1.2.

The results are organized according to the true underlying relations in Figure 4.6. Each of the three rows corresponds to different number of S-genes in the simulated data sets (4, 8, 20), while columns represent true relations. In each plot, the x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors. For example the second-left column corresponds to all situations where the true relation between nodes is $S_1 \longleftarrow S_2$. Rejection rates for this relation are marked in purple and we can see that the relation is not falsely rejected for larger graph with 20 S-genes even for very high noise levels. In contrast, except for very low noise levels rejection rates of incorrect relations decline by increasing the number of S-genes. Nevertheless, except for very high noise levels we reject substantial fractions of relations for large graph.

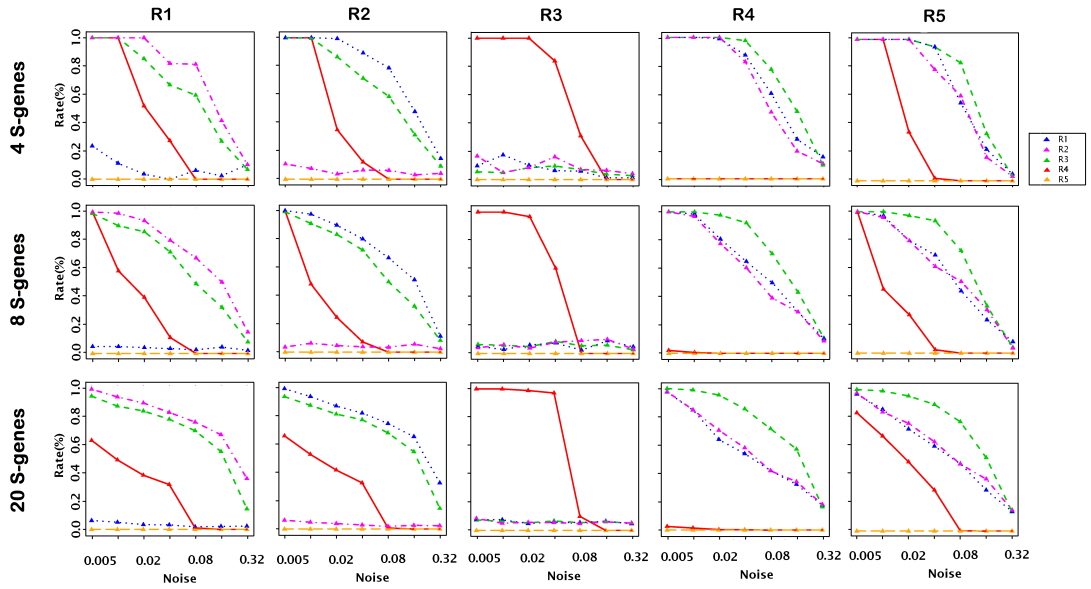

Figure 4.6: Dependency on the number of S-genes - Rows correspond to the number of S-genes in the simulated data sets (4, 8, 20), while columns represent true relations. In each plot, the x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors (R1: blue, R2: purple R3: green, R4: red and R5: orange). The results show that the true relations are not falsely rejected for larger graph with 20 S-genes even for very high noise levels. In contrast, except for very low noise levels rejection rates of incorrect relations decline by increasing the number of S-genes.

4.4.1.5 The number of unrelated E-genes might effect the power of the testing

The number of unrelated E-genes might effect the power of the testing in oder to generate uninformative E-gene with expected silencing pattern (0,0). This pattern can be expected to observe by all relations in $R \in \mathcal{R}$ (Figure 4.1). The number of such pattern does not change the number of alien patterns for a given relation but might effect the probability of observing alien patterns. In order to investigate the influence of unrelated E-genes to our algorithm, we generate different data sets with arbitrarily numbers of unrelated E-genes.

Here we follow the four steps for data generation in section 4.4.1.2 with only one change. In the step three we add $n_{UE} \in \{0, 100, 500, 1000\}$ unrelated E-genes to the data set. We then run No-CONAN on every pair of nodes in each of 400 networks and reject all relations possible using $\kappa = 0.05$. The results are organized according to the true underlying relations in Figure 4.7. Each of the four rows corresponds to different number of unrelated E-genes in the simulated data sets (0, 100, 500, 1000), while columns represent true relations. In each plot, the x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors. Overall these results show that the number of unrelated E-genes effect the power of the test. We still have hardly any false positive rejections but gain power in rejecting the incorrect relations by decreasing the number of E-genes with (0,0) pattern.

4.4.1.6 Evaluations on a large network

Finally, we examine the performance of No-CONAN in the context of the larger 25 S-genes (nodes) network shown in Figure 4.8. We generate 100 data sets from this graph using noise levels varying between 0.005 (very low) and 0.32 (very high). We then attach a total of n_E=500 E-genes uniformly to the S-genes of the network and add another n_{UE}=900 E-genes which are unrelated to the networks. Note that the network contains two feedback loops, one towards the root and another close to a leaf. The crucial difference from the smaller networks in section 4.4.1.2 is the ratio of S-genes inside the window (2 in both cases) and those outside of it (8 vs. 23). In fact, this unfavorable ratio of observed versus unobserved nodes compromises the resolution of the pNEMs generated by No-CONAN. Importantly, we still hardly ever falsely reject a correct relation. However, except for very low noise levels rejection rates of incorrect relations decline. Nevertheless, except for very high noise levels we reject substantial fractions of relations thus partially learning the structure of the network.

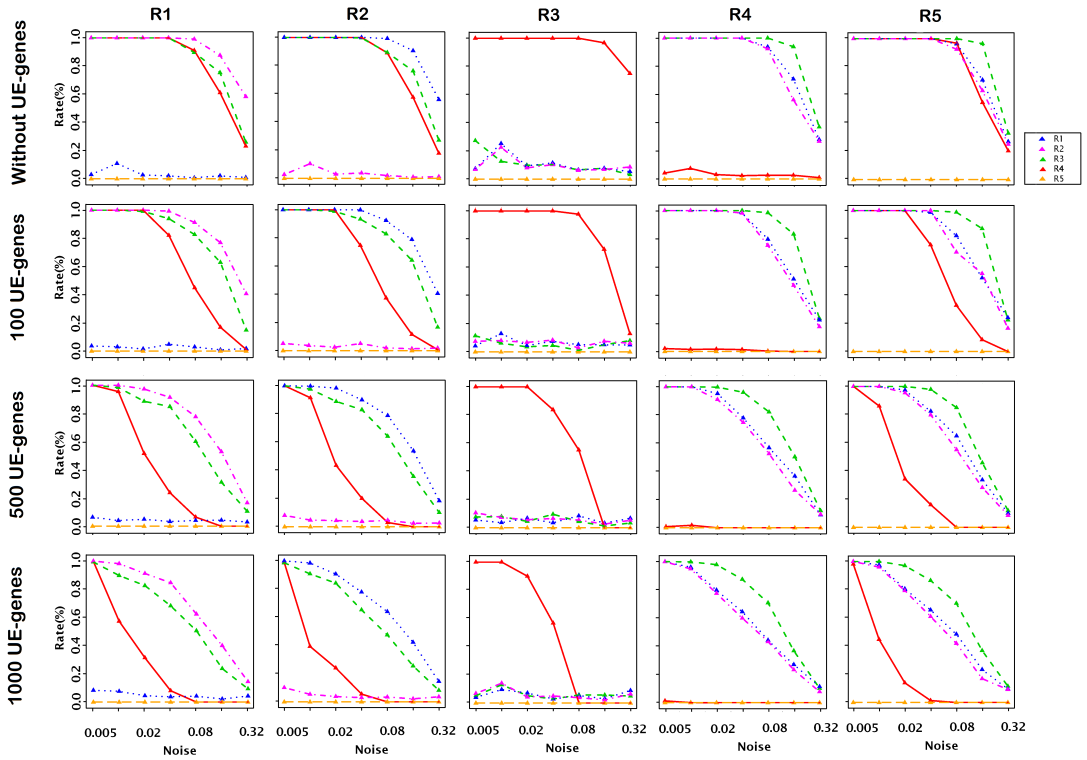

Figure 4.7: The number of unrelated E-genes might effect the power of the testing - Rows correspond to the number of unrelated E-genes in the simulated data sets (0, 100, 500, 1000), while columns represent true relations. In each plot, the x-axis shows the different noise levels while the y-axis shows the relative frequency of rejecting the different relations, which are marked by different colors (R1: blue, R2: purple R3: green, R4: red and R5: orange). Results show that the number of unrelated E-genes effect the power of the test. We still have hardly any false positive rejections but gain power in rejecting the incorrect relations by decreasing the number of E-genes with (0,0) pattern.

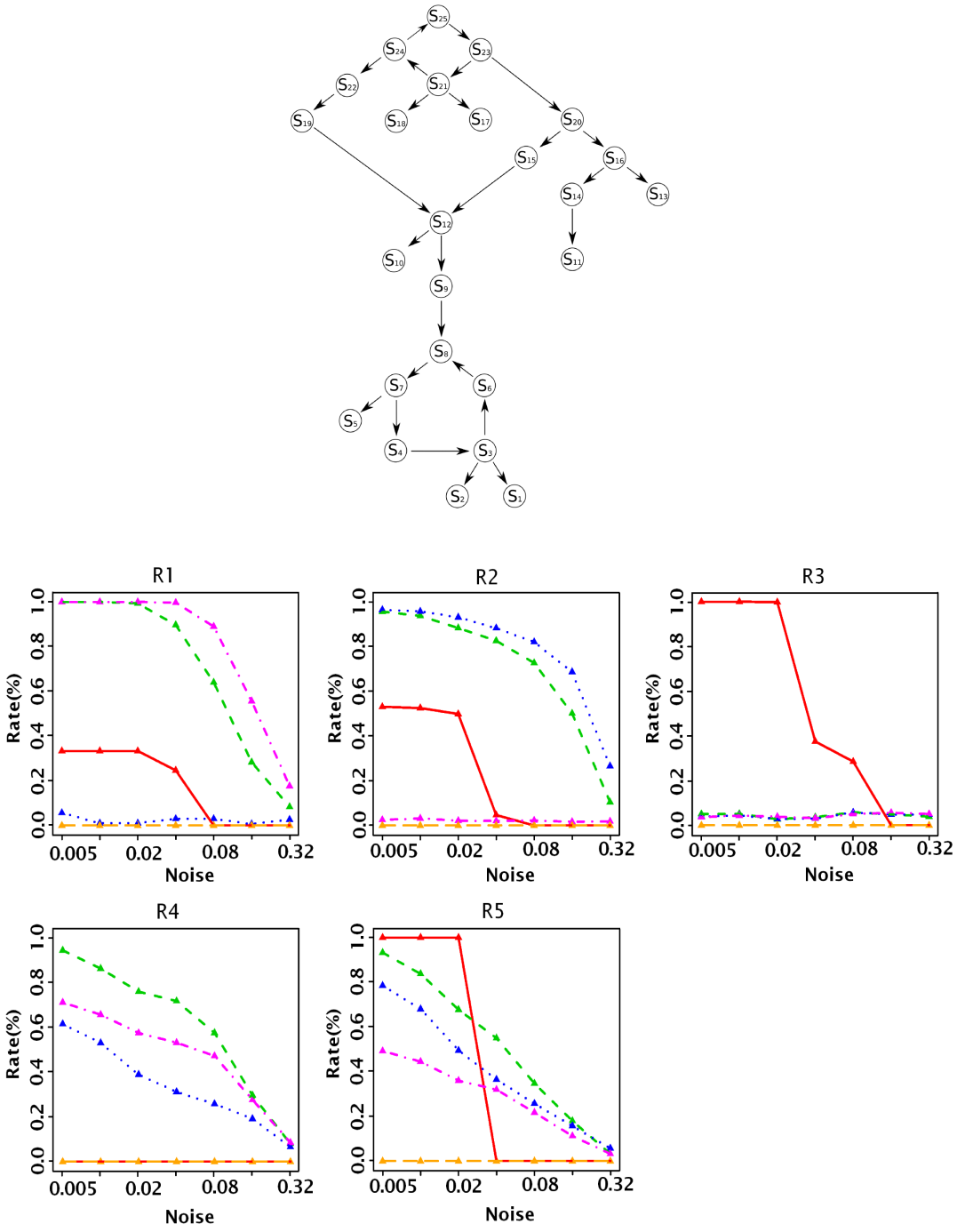

Figure 4.8: Evaluations on a relatively large network - The graph on top of the figure shows the ground truth network consist of 25 S-genes. Each of the five plots on bottom of the figure corresponds to one true relation. The x-axis shows the different noise levels. The y-axis shows the relative frequency of rejecting the relations (R1: blue, R2: purple R3: green, R4: red and R5: orange.)

4.4.2 Comparison between NEM and pNEM

We start with an example to illustrate how No-CONAN works. Consider the GTN in Figure 4.9. Note that it has only one hidden node, but this node is in a central position of the network. We attach a total of 350 E-genes uniformly to the S-genes and generate artificial data using moderate noise levels of 0.15 for both false negative and false positive observations as described previously (58). Then, using only data for the observable nodes, we reconstruct the network using a triplet search in a standard NEM approach (65) and using No-CONAN. Figure 4.9 compares the NEM to the pNEM. The NEM incorrectly predicts a feedback loop-like structure. The pNEM, in contrast, did not do this. It actually resolved the relation between S_5 and S_6 as $R5$, thus predicting the existence of the hidden node at that position. Also all other predicted relations are correct, with the exception of S_6 and S_3, where the pNEM is undecided on whether a directed relation exists (incorrect) or not (correct). Compared to GTN, the results show that the inference from NEMs suffers from hidden players and can be confounded. As is clear in the 4.9, the fully reconstructed model with NEM has three incorrect edges, while the partially reconstructed model with pNEM has no incorrect edges.

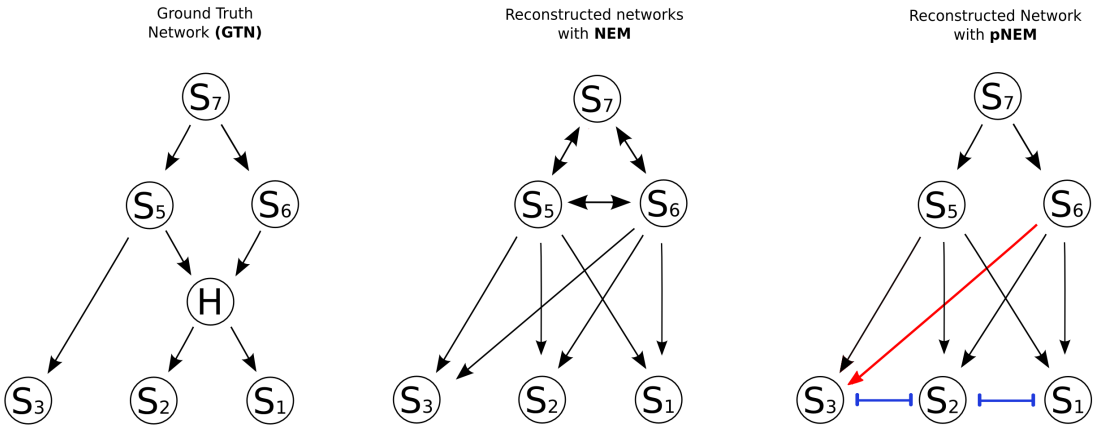

Figure 4.9: Comparison between NEM and pNEM - The left part of the figure shows a GTN that has only one hidden node. We attach a total of 350 E-genes uniformly to the S-genes and generate artificial data using a moderate noise levels of 0.15 for both false negative and false positive observations. The plot on the middle and right side of the figure shows the reconstructed network using triple search in a standard NEM approach and pNEM respectively, only for observable nodes. The NEM incorrectly predicts a feedback loop-like structure. In contrast the pNEM actually resolved the relation between S_5 and S_6 as $R5$ thus predicting the existence of the hidden node at that position. Also all other predicted relations are correct, with the exception of S_6 and S_3, where the pNEM is undecided whether a directed relation exists (incorrect) or not (correct).

5

Cell differentiation in embryonic stem cells

Here we apply the methodology of the last chapter in the context of cell differentition in embrynic stem cells. This chapter gives some background to cellular decision-making and cell differentiation (section 5.1). The cellular decision-making process of embryonic stem cells in mice may have many unknown unknown players. This process of differentiation has been previously modeled, but no work has yet been done to adapt a statistical methodology that considers the unknown unknown factors (section 5.2). I demonstrate the practical use of No-CONAN in the context of a first application to embryonic stem cell differentiation in mice. I demonstrate that taking unknown unknowns into account changes our account of real biological networks (section 5.3).

5.1 Introduction

Life at the cellular level is stochastic. Diffusion, gene expression, signal transduction, the cell cycle, and the extracellular environment are stochastic processes that change in time in ways that can be difficult to predict (66, 67). While a cell's environment determines its response, information from the environment comes from different, fluctuating, and perhaps contradictory, signals. This information is processed using biochemical networks whose components themselves fluctuate in concentration and intracellular location. For example, cellular decision-making is involved in several biological processes such as cell division, cell proliferation, apoptosis or cell differentiation. Each of these processes is regulated and controlled both by intracellular networks and extracellular signaling molecules whose mechanisms are still not clear.

5. CELL DIFFERENTIATION IN EMBRYONIC STEM CELLS

Cellular differentiation Cell differentiation is a process in which a stem cell develops into a specific type of cell in response to specific triggers from the body or the cell itself (22). This is the process which allows a single-celled zygote to develop into a multicellular adult organism which can contain hundreds of different types of cells. In addition to being critical to embryonic development, cell differentiation also plays a role in the function of many organisms, especially complex mammals, throughout their lives.

When a single cell has the capability of developing into any kind of cell, it is known as totipotent. For example, in mammals the zygote and the embryo during early stages of development are totipotent. Cells which can differentiate into several different cell types, but not all, are considered to be pluripotent. In both cases, the nucleus is the same, containing all of the genetic information needed to encode the entire organism, but only certain genes are activated (22).

When an embryo develops, cell differentiation is critical, because it allows the developing organism to create numerous, different, cell types, from neurons which will make up the brain to epidermal cells which will create the upper layers of skin. Once mature, the organism will have germ cells, somatic cells, and adult stem cells. Germ cells are haploid cells which are used in reproduction, while somatic cells make up most of the cells in the body, with over 250 known kinds of cell in the human body alone.

Stem cells Stem cells are special cells in multicellular organisms that are capable of differentiating into a wide range of other cells as needed. In other words, the cells themselves are not specialized like blood cells and nerve cells, but they can make specialized cells to form an embryo or repair damage to an adult organism. This property has suggested that they could be useful in medical treatment, and many nations have established stem cell funding to explore the possibility of research and development (68).

All multicellular organisms actually start out as a cluster of stem cells. As they divide and multiply, they differentiate themselves to make organs, muscle and bone until a complete embryo is formed. Adults also have stem cells, although their precise origin is not fully understood. These adult cells are triggered in response to serious injury to replace damaged tissues. There are three types of stem cells. Embryonic stem cells are taken from an embryo. Cord blood stem cells come from the umbilical cord, which is rich in these cells because it is of fetal origin. Adult stem cells are also known as somatic stem cells, and they are found in a range of locations around the adult body. The exact science and distribution of the adult cells is still a topic of intense research.

The mechanism underlying such a coordination is still not fully understood. During the process of cell differentiation, stem cells need to decide when and how to move from the state of self-renewal into differentiation. Such a complex process is governed by transcription networks known as developmental transcription networks (69), which need to make irreversible decisions on a slow timescale of one or more cell generations.

5.1.1 Molecular mechanism in early stem cell differentiation in mice

A zygote is the initial cell formed when two gamete cells are joined by means of sexual reproduction. In multicellular organisms, it is the earliest developmental stage of the embryo (22). The zygote can give rise to a complex organism through cell division, proliferation and cell differentiation. Since the zygote is totipotent, it can developed into the placenta. The totipotency is maintained in cells known as blastomeres of the two-cell-stage embryo. After mechanical separation of the blastomeres of the two-cell-stage embryo, each blastomere is able to give rise to an adult organism, for example a mouse (70). These cells are known as embryonic stem cells (ESC). They have the ability to self-renew as well as differentiate into different cell types of the vertebrate embryo leading to the formation of an entire organism. Embryonic stem cells are pluripotent stem cells derived from the inner cell mass of the blastocyst, an early-stage embryo (71). The cells of the embryonic inner cell mass from which mouse ESC are derived are called pluripotent because of their ability to give rise to all of the cells of an embryo and adult (72).

In fact, ESC can self-renew continuously for years if they are cultured under conditions that prevent their differentiation. For instance, mouse embryonic stem cells were grown in the presence of leukemia inhibitory factor (LIF), thus retaining their undifferentiated self-renewing state (positive controls). Differentiation-associated changes in gene expression were measured by replacing LIF with retinoic acid (RA), thus inducing differentiation of stem cells (negative controls) (73). It is reported that the transcription factor networks play a role in the maintenance of ESC pluripotency (74, 75, 76, 77, 78, 79, 80, 81). There are transcription factors (TFs) involved in the process that are pivotal for maintaining ESC in their self-renewal state when overexpressed. These include: *Nanog*, a homeobox transcription factor expressed throughout the pluripotent cells of the inner cell mass with the particular goal of preventing endoderm differentiation (76, 82); *Oct3/4*, also called *Pou5f1*, an important regulator of pluripotency that acts as a gatekeeper to prevent ESC differentiation(79); and *Sox2*, a member of the Sox (SRY-related HMG box) family of proteins that bind to DNA through the 79-amino acid HMG (high mobility group) domain. *Sox2* is co-expressed with *Oct4* in the inner cell mass (83). These TFs form a core transcriptional network

associated with pluripotency in ESC (78, 81, 84, 85). Alternatively, the differentiation of mouse ESC can be induced by the expression of certain transcription factors.

So far, two types of transcription factors have been recognized as main players in ESC: First, TFs with target genes that are expressed in undifferentiated ESC; second, TFs with target genes that are not expressed in undifferentiated cells but induced in differentiated ESC. The overexpression of the first type TFs will maintain ESC in their self-renewal state while overexpression of the second type will likely trigger the differentiation of ESC. These transcription factors function in combination with other processes and on the accessibility of their target genes, which are made accessible by the modification of their DNA, histones, or chromatin structure. The challenge is to understand how these TFs interact with one another to regulate the processes between self-renewal and differentiation. However, due to the complexity of these processes there might be many unknown unknown players involved, which would make these processes difficult to understand. Moreover, understanding the mechanisms underlying the processes of pluripotency, self-renewal and subsequent differentiation in embryonic stem cells is central to utilizing them therapeutically.

5.2 Previous works to model murine stem cell development in mice

Ivanova *et al*. (80) down-regulated six factors (Nanog, Oct4, Sox2, Esrrb, Tbx3, and Tcl1) that need to be jointly expressed in murine ESCs to keep the cells in a self-renewal state. They combined perturbation of these gene products with a time series of micro-array gene expression measurements. Mouse embryonic stem cells (ESC) were grown in the presence of the leukemia inhibitory factor (LIF), thus retaining their undifferentiated self-renewing state (positive controls). Cell differentiation-associated changes in gene expression were detected by inducing differentiation of stem cells by removing LIF and adding retinoic acid (RA) (negative controls). Finally, RNAi-based silencing of the six regulatory genes was used in (LIF+, RA-) cell cultures to investigate whether silencing of these genes partially activates cell differentiation mechanisms.

In response to the interventions, the cells go into differentiation and the resulting shifts in the transcriptome were monitored in time series of expression profiles. Differentiation includes the successive destruction of the self-renewal network. Micro-array expression measurements at 67 time points at 1-day intervals were taken for the two controls (positive and negative) and the six RNAi assays.

This process of differentiation has been previously modeled twice using nested effect models (48) and (50). Both models have in common the fact that they are dynamic nested effects models exploiting the temporal information of the time series, although they differ in the likelihood functions used. Neither of them considered the possibility of unobserved factors.

The model of Anchang et al. In Anchang *et al.* (48), we extended static NEMs to the modeling of perturbation time series measurements. Dynamic nested effects models allow for the resolution of feedback loops in the signaling cascade, as well as for the discrimination of direct and indirect signaling. As already mentioned in chapter 2, rate constants of signaling propagation are model parameters in DNEM.

We applied the DNEM approach to the Ivanova dataset on molecular mechanisms of self-renewal in murine embryonic stem cells (80). Since long computation times for Gibbs sampling prohibit the reconstruction of the networks topology from scratch by using DNEMs, we first used the triple search approach for the standard nested effect approach (49) applied to the final time point to determine a transitive closed topology for the network. The Figure 5.1 shows the binary data from the last time point and the reconstructed network. DNEM is based on binary data, which requires gene expression profiles to be discretized. In the Figure 5.1A, the reconstructed network in Figure 5.1B shows a staircase-like pattern of nested sets consistent with the linear cascade *Nanog* $\rightarrow$ *Sox2* $\rightarrow$ *Oct4* $\rightarrow$ *Tcl1*.

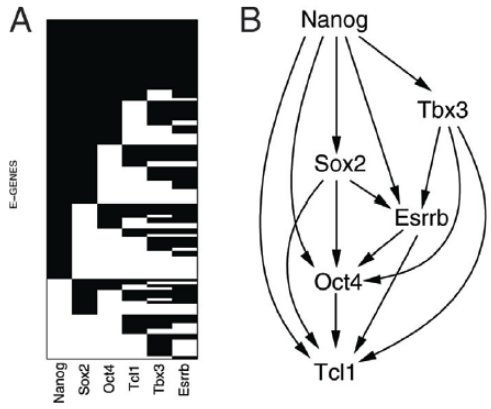

Figure 5.1: Stem cell data analysis - **A** Discretized data of the last time point across *E-genes* (rows) and *S-gene* perturbations (columns), with black representing downstream effects and white no effects. **B** The transitively closed nested effects model estimated from the data shown in **A** using static NEM.

Next, we exploited the DNEM Gibbs sampler trajectories associated with the network topology to infer average time delays and regulatory control of E-genes. Based on the marginal posterior probability, we excluded an edge if the posterior is above a certain threshold. The resulting network is shown in Figure 5.2.The time delay data has overruled the static NEM. For instance, they have removed the edge between *Nanog* and *Tbx3*. An application of DNEMs to embryonic stem cell development in mice reveals a feed-forward loop-dominated network, which stabilizes the differentiated state of cells and points to *Nanog* as the key sensitizer of stem cells for differentiation stimuli.

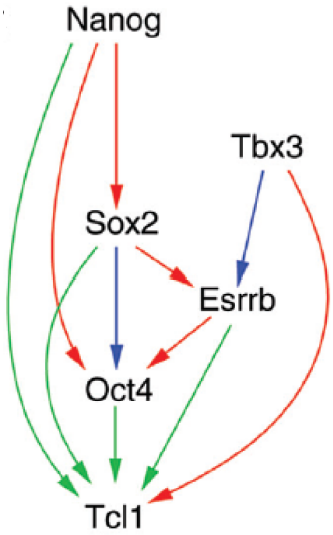

Figure 5.2: DNEM inference on signal propagation - The final network structure estimated by time delay analysis using DNEM. Edge colors correspond to estimated average time delays: fast signal propagation (green), intermediate signal propagation (blue) and slow signal propagation (red).

.

The model of Fröhlich et al. Fröhlich *et al.* (50) developed a novel approach to infer signaling cascades from high-dimensional perturbation time series measurements via Fast Dynamic Nested Effects Models (FDNEMs), hence allowing them to discriminate direct from indirect perturbation effects and to resolve feedback loops. This approach directly extends the NEMs framework introduced by Markowetz *et al.* (58) from the static to the dynamic case by unrolling the network structure over time. It also allows for a fast and efficient computation of the likelihood function without any time-consuming Gibbs sampling.

Figure 5.3 shows the reconstructed network between six transcription factors playing a key role in murine stem cell development using FDNEMs. Fröhlich *et al.* found good agreement with results published by Achang *et al.* (2009) and with the biological literature. There are several similarities to the inferred network shown in Anchang *et al.* (2009), which was obtained via the DNEM method, namely the cascades Tbx3 →Esrrb →Oct4 →Tcl1, Nanog →Oct4 →Tcl1 and Sox2 →Oct4 →Tcl1. A further striking similarity is that the transcription factor Oct4 regulating Tcl1 is itself jointly regulated by the three transcription factors Nanong, Sox2 and Esrrb. In contrast to Anchang *et al.*, in this network Nanog is not placed upstream of Sox2 and does not have any indirect outgoing edges. Indeed, the only shortcut in this network is Sox2 →Tcl1. This network is thus more sparse than the one shown by Anchang *et al.* The reason for the differences between our network and Fröhlich *et al.* is a mixture of a different likelihood model combined with a sparsity prior.

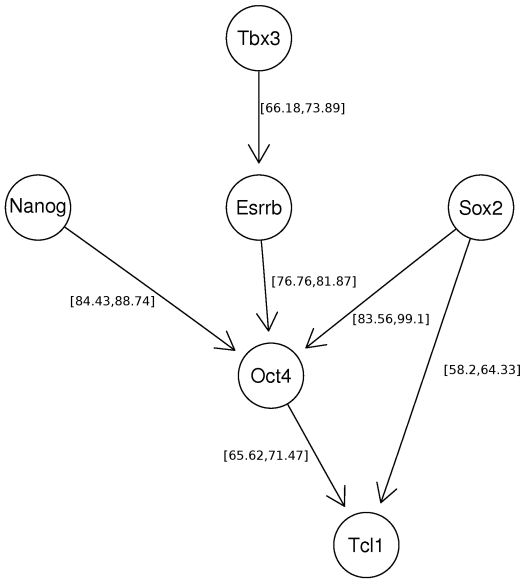

Figure 5.3: Inferred network for murine stem cell development using FDNEMs - Inferred network for murine stem cell development with 95% confidence intervals for the presence of the edges.

5.3 Application of No-CONAN to cell differentiation in embryonic stem cells

We test No-CONAN in a study on molecular mechanisms of self-renewal in murine embryonic stem cells (ESCs) (80). In the NEM framework the six regulatory gene products, *Nanog*, *Oct4*, *Sox2*, *Esrrb*, *Tbx3* and *Tcl1* are *S-genes*, while all genes showing significant expression changes in response to LIF depletion are used as *E-genes*. Downstream effects of interest are those where the expression of an *E-gene* is pushed from its level in self-renewing cells to its level in differentiated cells. The goal is to model the temporal occurrence of these effects across all time series simultaneously.

Data preprocessing We use log2 transformed values of MAS5.0 normalized data obtained from `www.nature.com/nature/journal/v442/n7102/suppinfo/nature04915.html`. In a comparison of the (LIF+, RA-) to the (LIF-, RA+) cell cultures 137 genes showed a greater than twofold up or down regulation across all time points. These were used as *E-genes* in our analysis. The two times series without RNAi were used to discretize the time series of perturbation experiments following a simple discretization method detailed in the next section, thereby setting an *E-gene* state to 1 in an RNAi experiment, if its expression value is far from the positive controls, and 0 otherwise. Genes that did not show any 1 after discretization across all experiments were removed, leaving 122 *E-genes* for further analysis.

Binary data We transform the continuous expression data to binary values. We set an *E-gene* in a certain silencing experiment and time point to 1, if its expression value is sufficiently close to the negative controls, i.e. the intervention interrupted the information flow, otherwise we set it to 0. Let $C(i,k,s)$ denote the continuous expression measurement of E_k at time point t_s of a time series recorded after perturbation of S_i. Moreover, let $C^+(k,s)$ and $C^-(k,s)$ denote the corresponding measurements in positive and negative controls respectively. We set

$$D(i,k,s) = \begin{cases} 1 & \text{if } \; C(i,k,s) < \kappa \cdot C^+(k,s) + (1-\kappa) \cdot C^-(k,s) \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

κ can be optimized by varying its value from 0 to 1 and choosing the smallest value where all negative controls are correctly recognized.

Non-confoundable network analysis We run NoCONAN on the data of the last time point of all time series. Note that the final time point of an admissible pattern accumulates information along the time series, because it reports a one whenever a downstream signal has reached the E-gene at any time. Figure 5.4A shows the pNEM produced from No-CONAN while C and D are the transitive closures of the networks derived in (48) and (50).

Notably, many edges of the pNEM are optimally resolved and often agree with those in the two previous models. E.g. the linear backbone of the network Nanog $\rightarrow$ Sox2 $\rightarrow$ Oct4 observed in the Anchang *et al.* model could be resolved unambiguously even when taking hidden confounders into account.

In contrast, the role of the remaining genes Tcl1, Tbx3 and Esrrb could not be determined unambiguously with the available observations. For example, our pNEM proclaims that there is an interaction between Esrrb and Tbx3 but can not determin its nature. It could be a feedback loop as well as any directed edge depending on how a potential unknown gene is influencing the process. Moreover, the pNEM differs from the two NEMs in that it predicts the existence of certain hidden nodes in positions marked in Figure 5.4B. These predictions result from the observation that all relations except for R5 could be excluded for the respective pairs of genes. In summary, non confoundable analysis sustains a previous hypothesis on the role of Nanog, Sox2, and Oct4 interactions in stem cell differentiation but also points to possible ambiguities with respect to the role of Tcl1, Tbx3 and Esrrb.

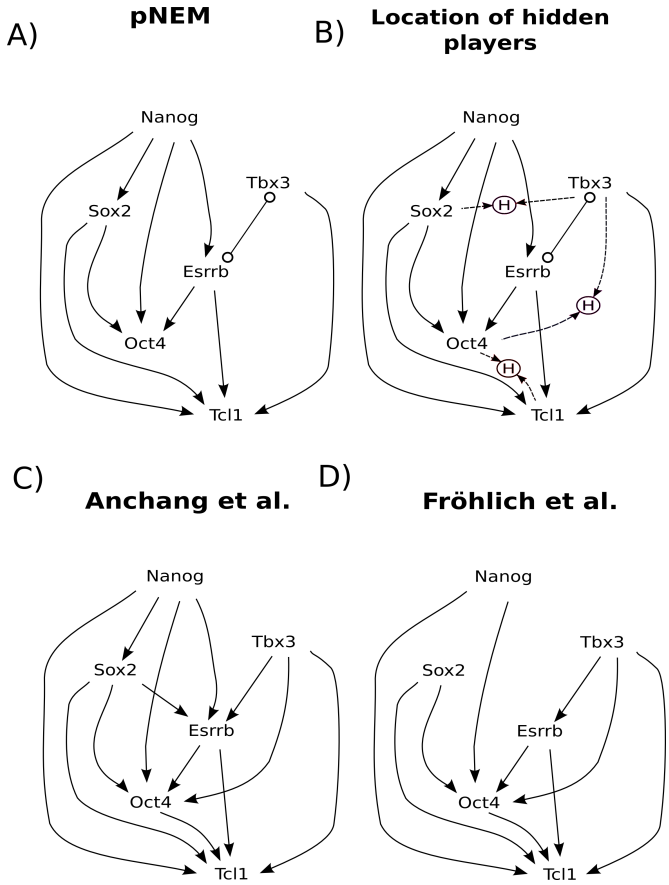

Figure 5.4: Inferred network for murine stem cell development using pNEM - **A** shows the pNEM produced from NoCNA. **B** shows predictions result from the observation that all relations except for R5 could be excluded for the respective pairs of genes. The pNEM predicts the existence of certain hidden nodes in positions marked in **B**. **C** and **D** are the transitive closures of the networks derived in (48) and (50). Many edges of the pNEM are optimally resolved and often agree with those in the two previous models in **C** and **D**.

6

Distorted canonical WNT-signaling in colorectal cancer cells

Aberrant regulation of Wnt signaling pathways plays an important role in the start and progression of colorectal cancer. Mutations in APC or CTNNB1 (β-catenin) genes are found in almost all cases of sporadic colon cancer but the importance of upstream signaling in colon cancer stays with a question mark. This chapter first gives some background on WNT-signaling in colorectal cancer cells (section 6.1). Based on differential gene expression, it is possible to use NEM to calculate and assess potential pathway structures, determining which model represents the best fit for the colon cancer cells (section 6.2). This has already been modeled, but no work has considers unknown unknown mechanisms. I demonstrate the practical use of No-CONAN in the context of a recent study in Wnt signaling pathway in colorectal cancer cells (section 6.3). This project was done in close cooperation with the group of Michael Boutros from University of Heidelberg. Most experiments were done by Gerrit Erdmann.

6.1 Introduction

Cancer is one of the most common causes of death according to the last global survey (86). Cells accumulate mutations that allow them to escape normal homeostatic regulation such as proliferation, differentiation or apoptosis. This occurs constantly throughout the body. The increase of cell number during cell division is called proliferation. Differentiation is the process that assigns a cell a certain biological task. In contrast, cell death occurs in one of two ways. Cells can be killed by the effects of physical, biological, or chemical injury. Additionally, cells are induced to kill themselves. Cell suicide is also referred to as apoptosis. The accumulated mutations in

cancer give cells unlimited division potential, the ability to invade other tissues and to ignore proapoptotic signals that regulate cell number or ensure genomic stability (87, 88). Many different combinations of mutations can lead to a wide variety of cancers. They might require different therapeutic strategies (89, 90). These therapies require drugs that specifically target critical pathways in particular cancer types, which in turn demands detailed understanding of how mutations change signaling pathways to identify specific drug targets.

Colorectal cancer is the third most common cancer world wide with over one million new cases per year (86). Tumorigenesis of sporadic colorectal carcinomas is associated with mutations in two canonical Wnt/β-catenin pathway (91). More than 80% of sporadic colorectal carcinomas are associated with mutations in APC and approximately 10% harbor mutations in β-catenin (91, 92). However, these mutations alone are insufficient to give rise to colorectal cancer. Several additional mutations in other genes are required and their occurrence can in part be associated with distinct stages in colorectal carcinogenesis (93). These mutations are often linked to pathways that have an important role during tissue homeostasis. Despite the clear causative relation between mutations in WNT/β-catenin signaling and colorectal cancer, it remains unclear exactly how these mutations affect the Wnt signaling pathway structurally and mechanistically. Yet, this knowledge is important to identify points for therapeutic intervention (94).

Structure of the intestine In order to understand colorectal carcinogenesis, it is important to decipher the underlying mechanisms and signaling pathways that govern homeostasis in the intestine. Homeostasis is the property of a organism that regulates its internal environment and tends to maintain a stable, relatively constant condition. The intestine can be broadly divided into two different parts, the small and large intestine. The large intestine consists of a flat epithelium called crypts, whereas the small intestine displays similar crypts but also additional finger-like structures called villi, that protrude from the epithelial lining of the intestinal wall. The structure of the large and small intestine are different mainly in their epithelium and this might reflect the different functions. Regardless of structural differences, both the large and the small intestine contain very similar cell types and are regulated by similar mechanisms and pathways that also play a role during carcinogenesis (95, 96).

6.1.1 Signaling pathway in intestinal homeostasis

The intestinal epithelium undergoes constant self-renewal requiring continuous proliferation, differentiation and apoptosis, while it regulates its internal environment and tends to maintain a stable, relatively constant condition. To maintain this intricate balance several pathways such as Wnt signaling control proliferation and cell fate in

the intestine (96, 97). During colorectal cancer development, mutations in the Wnt signaling pathway, frequently occur. In order to understand the roles that these pathways play in colon carcinogenesis, it is important to understand how they interact with the intestinal epithelium in healthy tissue (94).

Wnt signaling The Wnt signaling pathway is a network of proteins that passes signals from receptors on the surface of the cell to DNA expression in the nucleus. It controls cell communications in the embryo and adult, for instance, cell proliferation and differentiation. It frames one of the major signaling pathways for many developmental processes, plays important roles in homeostasis of different tissues and is also linked to carcinogenesis (98, 99). In adult organisms Wnt signaling is often involved in maintenance of stem cell. In a human there are 19 different Wnt ligands that can activate different signaling cascades (100). All 19 Wnt proteins are subdivided into Wnts that trigger canonical and Wnts that trigger non-canonical signaling. Canonical signaling controls β-catenin, a key co-factor of TCF familiy regulating genes involved in development and proliferation. Non-canonical signaling is less well-defined and encompasses several β-catenin independent signaling cascade (94).

β-catenin's role in the Wnt signaling pathway Canonical Wnt signaling is tightly controlled by degradation of β-catenin. When Wnt is not present, β-catenin is constantly degraded by a multi protein complex (100). β-catenin is associated with axin inhibition protein-1 (Axin1) and adenomatosis polyposis coli (APC). The complex recruits the serine/threonine kinase, casein kinase 1 (CK1α). CK1α phosphorylates Axin1 enabling it to bind glycogen synthase kinase β (GSK3β) (101). In order to relieve β-catenin from constant degradation, activation of the Wnt signaling cascade by Wnt ligands is required. When β-catenin is phosphorylated, it is degraded and, thus, will not build up in the cell to a significant level. When Wnt binds to frizzled (Fz), dishevelled (Dsh) is recruited to the membrane. GSK3 is inhibited by the activation of Dsh by Fz. Because of this, β-catenin is permitted to build up in the cytosol and can be subsequently translocated into the nucleus to perform a variety of functions. It can act in conjunction with T-cell specific transcription factors (TCF) such as TCF/LEF1 or TCF4 (TCF7L2) to activate specific target genes involved in different processes. β-catenin turns TCFs into transcriptional activators resulting in the expression of target genes, for instance cyclin D1 (CCND1), Axin2 and cMyc (Figure 6.1). Aside from the respective concentrations of ligands, canonical Wnt signaling is modulated at all levels of the signaling cascade, from secretion of Wnts to activation of target genes in the nucleus (94).

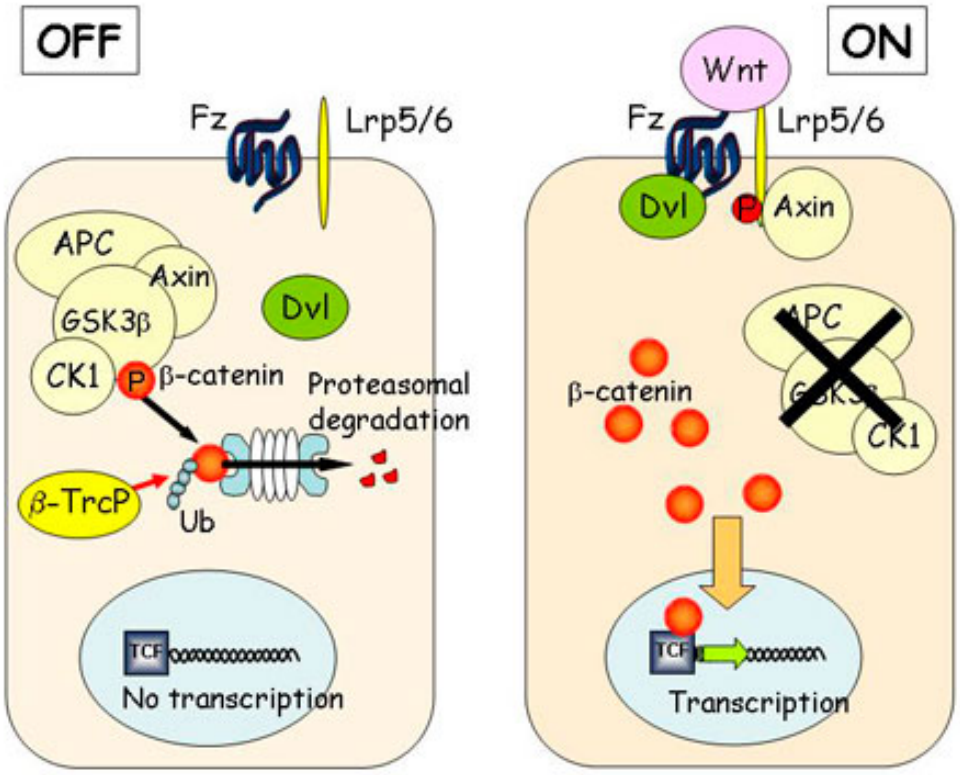

Figure 6.1: The canonical Wnt pathway. - **(OFF)** In the absence of Wnt, cytoplasmic β-catenin is phosphorylated by a protein complex consisting of the scaffolding proteins Axin and APC and the kinases GSK3β and CK1a. Subsequent recognition by the ubiquitin ligase β-TrcP leads to ubiquitin-mediated degradation of β-catenin. **(ON)** Binding of Wnt to Fz leads to recruitment of the cytoplasmic effector protein Dvl. Phosphorylation of the Lrp cytoplasmic tail subsequently provides a docking site for Axin. Taking away Axin from the Axin-APC-GSK3β complex presumably compromises its ability to phosphorylate β-catenin. The figure is adapted from (102).

6.1.2 Mechanism of Wnt pathway mutations in cololectal cancer

Abnormality in Wnt signaling pathway activity is an initial step in colon cancer development but is also important for maintenance of tumors (103). This makes the Wnt pathway an attractive target for therapeutic intervention. In order to find out the main target protein, it is important to understand how mutations in the Wnt signaling pathway affect signal transduction (104).

APC and β-catenin mutation aberrantly activate Wnt signaling Mutation in APC is the most common mutation in colon cancer. When APC does not have an inactivating mutation, β-catenin does. These mutations can be inherited, or arise sporadically, often as the result of mutations in other genes that produce chromosomal instability. A mutation in APC or β-catenin must be followed by other mutations to initiate cancer. APC is a large protein and acts as a scaffold protein binding to β-catenin and Axin1 allowing the phosphorylation of β-catenin by CK1α. Structurally APC contains multiple β-catenin binding sites. β-catenin is found localized at the plasma membrane, in the cytoplasm, as well as in the nucleus. It is unclear how strongly membrane bound and cytoplasmic β-catenin population interchange or if they even represent independent pools. In order to target β-catenin for degradation, it is phosphorylated at specific residues in specific sequential sites. In most colorectal cancers, mutation in β-catenin occur exactly at one of these sites. Loss of one of these phosphorylation sites facilitates ubiquitination required for βcatenin degradation. This cause accumulation of β-catenin and consequently its translocation to the nucleus to aberrant activation of Wnt target genes. In conclusion, both APC and β-catenin mutations aberrantly activate Wnt signaling. This has often led to the assumption that these mutations are dominant over upstream signals and constitutively activate Wnt/β-catenin target genes (94).

Two hypothesis on APC and β-catenin It is not clear yet how mutations in APC or β-catenin change signal transduction events. So far, there are two hypotheses for the pathway topology of the Wnt signaling cascade. The Wnt independent model states that mutations in APC or β-catenin constitutively activate the pathway rendering it insensitive to regulation by upstream factors such as canonical Wnt ligands or antagonists. The Wnt dependent model assumes that APC and β-catenin mutations are not dominant over upstream signaling events. This model would potentially allow therapeutic intervention at all levels of the Wnt/β-catenin signaling cascade, e.g. the targeting of Wnt secretion. In order to develop new therapeutic approaches for colon cancer treatment, it is important to understand the pathway topology of the Wnt signaling cascade (94).

6.2 Wnt secretion is required for Wnt/β-catenin target gene expression

It has been assumed that activating mutations in APC or β-catenin lead to constitutively active Wnt signaling, thus rendering cells insensitive to upstream signaling. Thus drug development efforts for colon cancer mainly focus on downstream targets (91, 105). Our collaboration partners from Boutrous lab at DKFZ have recently shown that pathway activation in colon cancer cells still requires Wnt ligands and that silencing Evi/Wls, a gene controlling the secretion of Wnt molecules impairs the expression of β-catenin dependent transcriptional targets (94). They elucidate the structure of the canonical Wnt pathway in colon cancer and determine the impact of upstream Wnt signaling events on Wnt pathway activity in the presence of mutated β-catenin and APC. In particular, they elucidate whether Wnt/β-catenin signaling in colorectal cancer is dependent on or independent from signals upstream of β-catenin or APC mutations. In order to understand the topology of the Wnt signaling pathway in colon cancer they performed RNAi knockdown of key Wnt pathway components in HCT116 cells for subsequent whole transcriptome sequencing followed by nested effects modeling (NEMs). Since, NEMs allow the prior assumption over the models, they compare Wnt dependent and Wnt independent pathway structures with as little prior assumptions as possible. In particular, depletion of Evi/Wls, a strictly required factor of Wnt secretion, was utilized to determine the dependency of colorectal cancer cells on canonical Wnt proteins.

NEM strongly favors pathways topologies that allow regulation upstream of β-catenin In order to compare different models for the Wnt pathway topology in HCT116 cells NEMs was implemented by Erdmann in collaboration with the Spang group at the University of Regensburg. They compared two hypotheses: (A) network topologies in which constitutive activation of Wnt signaling is solely dependent on mutations of β-catenin or APC. Thus, depletion of upstream pathway components, such as Evi/Wls, has no affect on β-catenin response gene expression in this model; (B) topologies that allow a regulation by upstream components is retained even in the presence of β-catenin or APC mutations. Figure 6.3 shows the schematic of these two hypothesis. They finally implement NEMs to score both competing hypothesis based on the changes in overall gene expression after knockdown. As for the results, they have reported that the expression data did not support any pathway topology assuming Evi/Wls independent activation of β-catenin response genes (hypothesis A) and strongly favoured sustained regulatory input from upstream components (hypothesis B) (Figure 6.3). To avoid overfitting, Bayes Factors were used to compare the hypothesis accounting for the increased complexity of a model that makes β-catenin response gene expression Evi/Wls dependent. The modeling

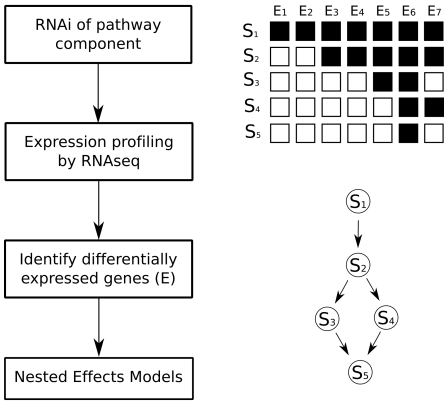

Figure 6.2: Nested effects modeling to determine Wnt pathway structure in colon cancer. - RNAi is used to knockdown several components of a signaling pathway (S-Genes) and expression profiling e.g by RNAseq is used to determine which genes are changed inresponse (E-genes; E1-7). Based on which E-genes are differentially expressed after knockdown of a specific S-gene a hierarchical pathway structure can be inferred. The basic principle assumes that the higher up an Sgene is within a signaling pathway, the more E-genes show an effect because of increasing branching points within a signaling cascade.

results confirm that Wnt pathway activation in HCT116 cells is dependent on Evi/Wls despite the βcatenin mutation without bias (94).

6.3 An application of No-CONAN to WNT-signaling in colorectal cancer cells

The reported results in the last section is important for the development of novel treatment options, since it suggests targeting Wnt signalling upstream of the mutated genes. This work was partially motivated by a nested effect model that predicted cross talk of signaling components upstream and downstream of the mutations. A possible confounding of this model by unobserved signaling molecules was not taken into account. In order to confirm the findings of Erdmann, in this section we show that the dependence on upstream signaling can also be inferred in a non-confoundable analysis. We test No-CONAN in the study on Wnt-signaling in colorectal cancer cells from our collaboration parter in Heidelberg (94).

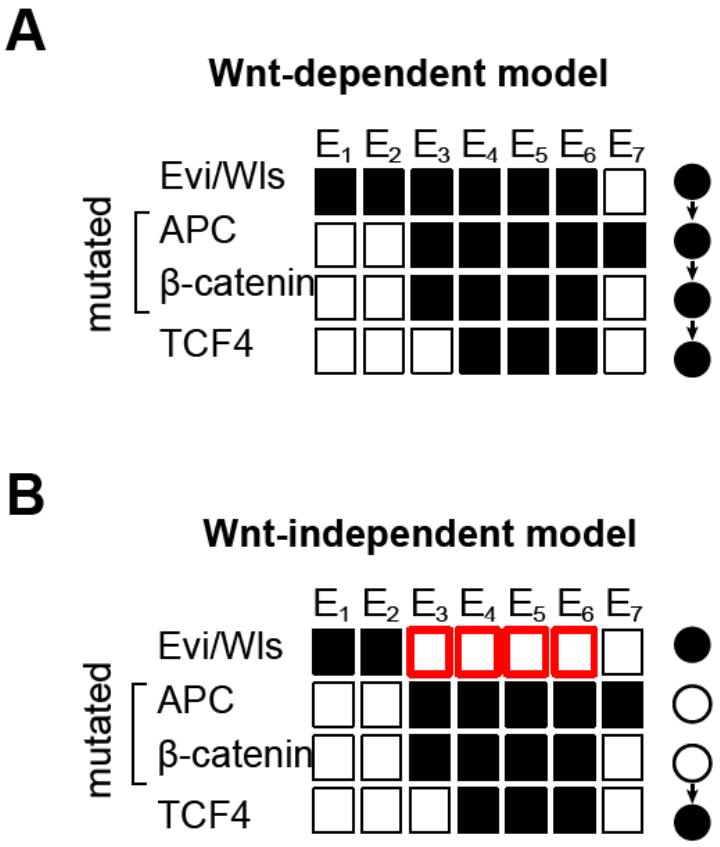

Figure 6.3: Nested effects modeling (NEM) favors Wnt-dependent pathway model. - A) Exemplary scheme of a Wnt-dependent pathway structure. Expected expression changes in effect genes (E1-7) after depletion of a pathway component are displayed as solid boxes. In Wnt dependent models knockdown of Evi/Wls affects expression of Wnt pathway responsive (E3-6)-genes. B) Exemplary scheme of a Wnt-independent pathway structure. In these models knockdown of Evi/Wls does not affect expression of Wnt pathway responsive (E3-6)-genes. The figure is adapted from (94).

6.3.0.1 Data preprocessing

Five pathway components at different levels were selected for RNAi mediated perturbation and subsequent sequencing by Erdmann (94) (Figure 6.4). Evi/Wls, a protein absolutely required for Wnt secretion, was chosen as the most upstream component. In order to find the level of the destruction complex, the negative pathway regulator APC, as well as β-catenin itself, was selected. Finally, BCL9 and TCF/L2 (TCF4), both required for β-catenin mediated target gene expression, were selected to obtain information about the Wnt pathway at the nuclear level. Initially, several single siRNAs against each of these genes were tested for knockdown efficiency. The siRNA with the best knockdown efficiency was used for perturbation and subsequent transcriptome sequencing. Then, knockdown of respective genes was conducted in HCT116 cell using reverse siRNA transfection followed by 72h incubation and subsequent RNA isolation. For each knockdown two biological replicates were generated. For each perturbation of a Wnt pathway component and each target transcript, they calculated the posterior probability that the gene was differentially expressed using Bayesian linear modeling (106).

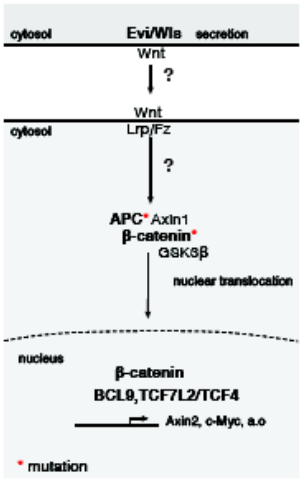

Figure 6.4: Schematic of the Wnt pathway structure in colon cancer - Scheme of the Wnt pathway indicating the genes selected for RNAi and RNAseq (bold). In colon cancer APC or β catenin are frequently mutated (red *) causing aberrant Wnt pathway activity. NEM is used to determine whether signaling components upstream of APC or βcatenin (e.g. Evi/Wls) are still on top of the Wnt signaling cascade. The figure is adapted from (94).

6.3.0.2 Not-confoundable network analysis

For each perturbation of a Wnt pathway component and each target transcript, we calculated the posterior probability that the gene was differentially expressed using Bayesian linear modeling (106). To conduct NEM the posterior probability that an E-gene was differentially expressed after knockdown of a Wnt pathway component was calculated using Bayesian linear modeling (106).

In the language of nested effect models, EVI/Wls, β-catenin, APC, GSK3β, Axin1 and the TCL genes are S-genes of Wnt signaling, while the transciptional targets of the pathway Axin2, SMAD7, EMP1, MGLL are E-genes. The hypothesis of constitutive downstream signaling proclaims that S-genes upstream of the mutation e.g. EVI/Wls are in a disconnected relation, $S_1 \cdot\cdot S_2$, with both β-catenin and TCF transcription factors. Here we use the data from the RNAi silencing assays of EVI/Wls and β-catenin in HCT116 cancer cells. We can show that the disconnect relation can be excluded with high confidence.

In order to test the disconnected relation, we run No-CONAN on silencing assays of EVI/Wls and β-catenin. Since the input of No-CONAN is binary data, we transfer the posterior probability that the gene was differentially expressed to binary values using different cut-offs. Given a disconnected relation, the alien patterns for this relation are at the form (1,1). We test whether the number of alien patterns occur due to noise. To evaluate the reliability of our analysis, we apply No-CONAN using different noise levels and different cut-offs for discretization. Our analysis consists of four different settings:

1. **All β-catenin targets: varying noise** First, we transfer the continuous data to binary data. This continuous data consists of the posterior probability that an E-gene is differentially expressed after knockdown of a Wnt pathway component. We set an E-gene in a certain silencing experiment to 1, if the posterior probability that this E-gene is differentially expressed after knockdown of a Wnt pathway component is bigger than certain a cut-off C, otherwise we set it to 0. We then define individual cut-offs for each E-gene.

   We tried values of C from 0.5 to 0.99 in steps of 0.01. In order to select the E-genes, before discretization we sort all β-catenin targets based on their posterior probability to be differentially expressed. We take the top n and discretize the data. The choice of n depends on the value of the cut-off. We then run No-CONAN on the binary data of all β-catenin target for silencing assays of EVI/Wls and β-catenin using noise levels varying between 0.005 (very low) and 0.32 (very high). The results are organized in Figure 6.5A. The

x-axis corresponds to different cut-offs for the posterior probability that an E-gene is differentially expressed after knockdown of a S-genes, while the y-axis present the P-value of excluding the disconnected model for *Evi* and β-*catenin*. By setting the calibration parameter κ to 0.05, this plot shows that the disconnected model is excluded even for very high noise levels.

2. **TCF7L2 dependent β-catenin targets: varying noise** Since TCF7L2 is a transcription factor influencing the transcription of several genes, it evolves to include a large variety of functions within the cell. For instance, the Wnt signaling pathway leads to the association of β-catenin with BCL9, translocation to the nucleus, and association with TCF7L2, which in turn results in the activation of Wnt target genes (107). In order to investigate the dependency of the relation between EVI/Wls and β-catenin on the TCF7L2 targets, we repeated the first analysis but only for the joint target of β-catenin and TCF7L2. After selecting E-gene, we transfer the continuous data to binary data using different cut-offs, C from 0.5 to 0.99 in steps of 0.01. In the next step we only select the E-genes which show perturbation effect for both β-catenin and TCF7L2. Finally, we run No-CONAN on the binary data of all β-catenin TCF7L2 dependent targets for silencing assays of EVI/Wls and β-catenin using noise levels varying between 0.005 (very low) and 0.32 (very high). The results are organized in Figure 6.5B. The x-axis correspond to different cut-off for the posterior probability that an E-gene is differentially expressed after knockdown of a S-genes, while the y-axis present the P-value of excluding the the disconnected model for *Evi* and β-*catenin*. Similar to the previous analysis, the results show that the disconnected model is excluded with different cut-offs and even for very high noise levels.

3. **All β-catenin targets: varying number of E-genes** In the next analysis, we investigate the influence of different number of selected E-genes for testing the relation between EVI/Wls and β-catenin. Similar to the first analysis, we start by sorting all β-catenin target based on the posterior probability that an E-gene is differentially expressed after knockdown of a Wnt pathway component. We then take the top n E-genes with the high posterior probability, but this time instead of cut-offs we varied n from 10 to 2000 in steps of 10. For discretization only four different cut-offs 0.99, 0.96, 0.92 and 0.68 are used. In the final step, we run No-CONAN for silencing assays of EVI/Wls and β-catenin using four noise levels. Figure 6.6A summarizes the P-value for testing the disconnected model. The x-axis corresponds to different numbers of selected E-genes before discretization, while the y-axis present the P-value of excluding the the disconnected model. Each curve in the plot corresponds to a specific cut-off and noise levels. The results show that the probability that the observed number of alien patterns given the disconnected relations occurs due to the noise fluctuation is smaller than 0.05 for low noise levels. In case of the very high

noise, these probabilities are still lower than calibration parameter for resealable number of E-genes (up to 500 E-genes).

4. **TCF7L2 dependent β-catenin targets: varying number of E-genes** In order to investigate the influence of different number of TCF7L2 dependent target genes for testing the relation between EVI/Wls and β-catenin, we repeat the previous analysis but only for TCF7L2 dependent β-catenin targets. As the result in Figure 6.6B shows, the disconnected model is excluded almost everywhere for different cut-offs and even for very high noise levels.

The partial modeling results confirm that Wnt pathway activation in HCT116 cells is dependent on Evi/Wls despite the βcatenin mutation. Using a combination of RNAi, transcriptome sequencing and No-CONAN, it is possible to show in an non-confoundable way that HCT116 cells are still dependent on Evi/Wls for activation of Wnt/βcatenin dependent target gene activation, despite a mutation in βcatenin. Here, we can not fully resolve the relation between Evi/Wls and β-catenin, but we can exclude the disconnected relation and conclude that they are not independent of each other.

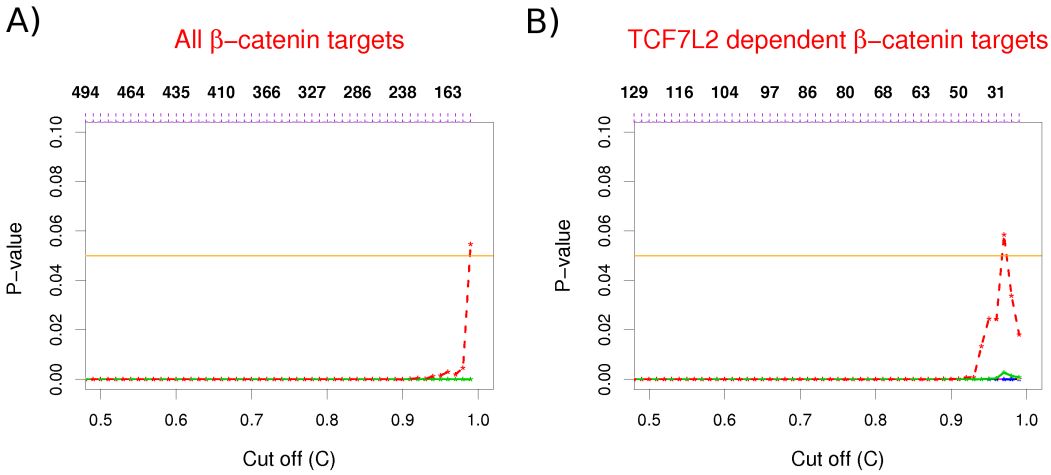

Figure 6.5: No-CONAN inference for the relation between the Evi/Wls and β-catenin: Varying noise - **A** shows the No-CONAN inference given a disconnect relation for the pair of S-genes, Evi/Wls and βcatenin for all βcatenin targets. **B** shows the same inference but only for TCF7L2 dependent β-catenin targets. In each plot, the x-axis corresponds to a different cut-off for the posterior probability that an Egene is differentially expressed after knockdown of a S-gene, while the y-axis presents the P-value of excluding the disconnected relation. The horizontal orange line shows the calibration parameter κ which is set to 0.05. The numbers on top of the plot show the corresponding data size (number of E-genes) for each cut-off on the x-axis. The different lines correspond to different noise levels (red 0.02, green 0.04, blue 0.08 and purple 0.32). The plots show that in both cases, (all β-catenin targets and TCF7L2 dependent β-catenin targets), the disconnected relation is excluded with high confidence.

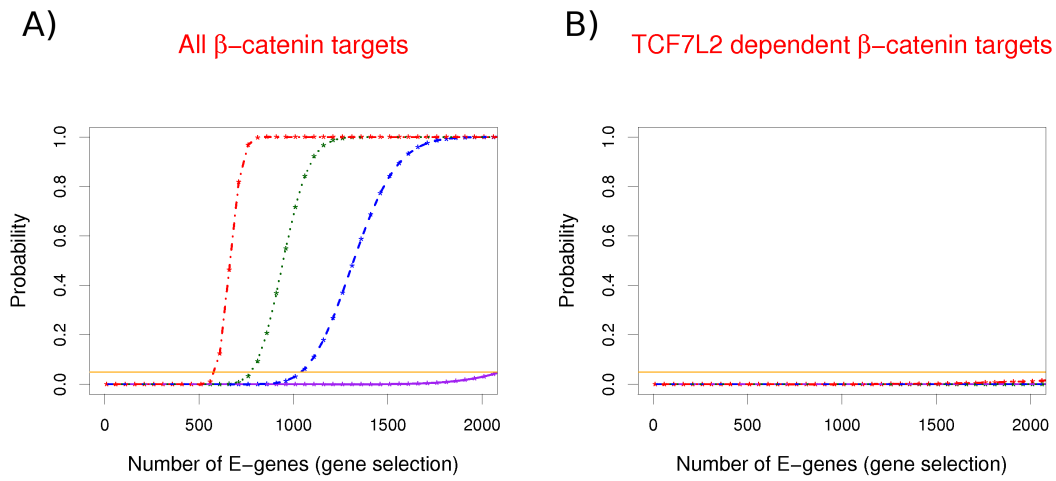

Figure 6.6: No-CONAN inference for the relation between the Evi/Wls and β-catenin: varying number of E-genes - **A)** shows the No-CONAN inference given a disconnect relation while varying number of E-genes for the pair of Evi/Wls and βcatenin only for βcatenin targets. **B** shows the same inference for different data size but only for TCF7L2 dependent β-catenin targets. In each plot, The x-axis correspond to different number of E-genes, while the y-axis present the P-value of excluding the disconnected relation. The horizontal orange line shows the calibration parameter κ which is set to 0.05. The different lines correspond to different noise levels and cut-offs. The plot **A** shows that the probability that the observed number of alien patterns given the disconnected relations occurs due to the noise fluctuation is smaller than 0.05 for small noise levels. In case of the very high noise, these probabilities are still lower than calibration parameter for resealable number of E-genes (up to 500 E-genes). In contrast, the plot **B** show that in both cases, All β-catenin targets and TCF7L2 dependent β-catenin targets, the disconnected relation is excluded almost everywhere with high confidence.

7

Summary and Outlook

Unknown hidden mechanisms in biological networks pose challenges to computational biology. The appearance of methods making use of RNA interference (RNAi) enables researchers to infer the inner workings of complex biological networks by breaking them down using an external stimulus. A cells response to an external stimulus is a complex network. However, due to the complexity of the system, a complete picture with detailed knowledge of the behavior of individual players in the networks is still out of reach. Therefore, our current understanding of virtually all cellular signaling pathways is almost certainly incomplete. We are lacking important but so far unknown players in the pathways and our observation is incomplete. Unknown mechanisms, whose involvement in cellular processes has not yet been determined, can make the interplay of the known mechanisms appear different from what they really are. The effect of unknown players might be mixed up with the effect of known players. Moreover, separating these effects can be difficult and can confound our perspective of the networks. Network reconstruction in the presence of unknown mechanisms may result in confounding. In this dissertation I have addressed the challenge of reconstructing biological networks specific to signaling networks from interventional data in the presence of unknown mechanisms. In doing so I was interested in answering following questions:

- How can the effect of unobserved players be misleading for the network reconstruction?
- What of our current understanding of biological networks can be confounded by hidden mechanisms and what can not?

Conclusion In order to answer these questions, we introduced No-CONAN, a novel method that partially reconstructs the upstream/downstream relations of non-transcriptional signaling networks from interventional data. The method is set in the

framework of nested effects models but has the additional feature that its inference can not be confounded by hidden nodes. The key idea is the definition of alien silencing patterns that can not be confounded by unobserved nodes. The output of No-CONAN is not a fully resolved network but a pNEM: A network of upstream/downstream relations where for some pairs of nodes several relations remain conformable with the data. The information in a pNEM lies in the upstream/downstream relations that it excludes. A pNEM encodes what we know but also what we can not know unless we can be sure that we have observed all nodes of a network. The uncertainties left with certain edges are the price we have to pay to ensure that our results are non-confoundable by mechanisms outside the window of observations.

No-CONAN is reliable in that it does not produce false information by rejecting correct relations. By construction, No-CONAN has two limitations affecting its power in resolving the network. It can never reject the relation $S_1 \rightarrow H \leftarrow S_2$, since this relation has no alien silencing patterns. Moreover, No-CONAN has very little power in resolving a true feedback loop since feedback does not produce the alien patterns of the two directed relations. Nevertheless, No-CONAN is generating new non confoundable insights into network structures by rejecting many though not all incorrect relations.

Partial network reconstruction is a relatively new concept in network analysis. It can be seen as a safeguard against possibly severe confounding effects caused by unobserved mechanisms. Clearly, such a non-confoundable analysis is only valid within the formal context of a network model. I used the framework of nested effect models. The assumptions of nested effect models might be incorrect in certain applications as is true for every modeling framework. I believe that no formal analysis can safeguard against this. However, the concept of unknown mechanisms can be represented in many formal frameworks, and simulations can mimic our partial observation of a true underlying network.

The usefulness of our approach on real data was shown by analyzing two different studies. I demonstrate practical use of No-CONAN in the context of a first application to embryonic stem cell differentiation in mice and a recent study in Wnt signaling pathway in colorectal cancer cells. The results form the first data set in chapter 5 contribute to understanding of how the stem cells carry out differentiation to specialized cells in a non-confoundable way. The reconstructed network by the pNEM differs from the already published models (48, 50) in that it predicts the existence of certain hidden nodes in certain positions. For instance, the non confoundable analysis sustains a previous hypothesis on the role of Nanog, Sox2, and Oct4 interactions in stem cell differentiation but also points to possible ambiguities with respect to the role of Tcl1, Tbx3 and Esrrb. The partial modeling results from second data set in chapter 6 show that HCT116 cells are still dependent on Evi/Wls, an essential factor for Wnt

secretion, for activation of Wnt/β-catenin dependent target gene activation, despite the mutation in β-catenin. Using a combination of RNAi, transcriptome sequencing and No-CONAN, it is possible to show in an non-confoundable way that HCT116 cells are still dependent on Evi/Wls for activation of Wnt/β-catenin dependent target gene activation.

After Donald Rumsfeld gave a new perspective of the concept of unknown-unknowns in 2002 (63), he continued his speech by saying "If I know the answer I'll tell you the answer, and if I don't, I'll just respond, cleverly". I do not know whether a pNEM is a "clever" response but it aims to be a realistic and an honest one. The partial networks aims to encode what we know that we know, but it also encodes what we can not know for certain, unless we are absolutely sure that we have a complete account of all biological mechanisms affecting cell signaling. It is my belief that the methods presented in this dissertation will open a new way to infer the biological networks with effective hidden mechanisms. It is my hope, that it will shed light on new and interesting domains.

Future Prospective The task of reconstructing networks in the presence of hidden variables in probabilistic graphical model in general (Nested effects models in particular) is a central and problematical challenge. This dissertation offers the first step toward methods that treat this problem in the context of binary nested effects models without posing restrictive constraints on the model. The next natural step is to consider the application of the methods presented here for a wider variety of models. This includes extensions to both continuous data and even in more general probabilistic graphical approaches such as Bayesian networks (2) and factor graphs.

Another important avenue of research is to improve our understanding of the method present in this dissertation. Since this method can only partially reconstruct the networks, one might think to combine the idea of alien patterns to remove the confoundable network feature and then reconstruct the networks over the remaining models. This might be still confoundable with hidden variables but using appropriate prior probabilities based on the exclusion-inclusion results might lead to the non-confoundable inferences. Additionally, the method present in this dissertation has to be extended for multiple knockouts experiments. This data was attained by silencing more than one gene at the same time. This will not change the idea of alien patterns, but more sophisticated silencing schemes have to be developed, which encode predictions both from single-gene and multi-gene knockouts. Since the number of possible multiple knockouts increases exponentionally, tools to choose the most informative experiments are needed.

7. SUMMARY AND OUTLOOK

References

[1] C.M. Bishop et al. *Pattern recognition and machine learning*, **4**. springer New York, 2006. 2, 3

[2] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988. 2, 4, 23, 105

[3] F.V. Jensen. *An introduction to Bayesian networks*, **74**. UCL press London, 1996. 2

[4] D. Heckerman, A. Mamdani, and M.P. Wellman. **Real-world applications of Bayesian networks**. *Communications of the ACM*, **38**(3):24–26, 1995. 2

[5] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009. 3, 4, 6, 9, 47

[6] Gal Elidan. *Learning Hidden Variables in Probabilistic Graphical Models*. PhD thesis, 2004. 4, 5

[7] F.V. Jensen, S.L. Lauritzen, and K.G. Olesen. **Bayesian updating in causal probabilistic networks by local computations**. *Computational statistics quarterly*, **4**:269–282, 1990. 4

[8] G.F. Cooper. **The computational complexity of probabilistic inference using Bayesian belief networks**. *Artificial intelligence*, **42**(2):393–405, 1990. 5

[9] D. Heckerman. **A tutorial on learning with Bayesian networks**. *Innovations in Bayesian Networks*, pages 33–82, 2008. 5

[10] D. Heckerman, D. Geiger, and D.M. Chickering. **Learning Bayesian networks: The combination of knowledge and statistical data**. *Machine learning*, **20**(3):197–243, 1995. 5

REFERENCES

[11] RI ACS. **Bayesian Classification (AutoClass): Theory and Results**. 5

[12] N. Friedman et al. **Learning belief networks in the presence of missing values and hidden variables**. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 125–133. MORGAN KAUFMANN PUBLISHERS, INC., 1997. 6

[13] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. **An introduction to variational methods for graphical models**. *Machine learning*, **37**(2):183–233, 1999. 6

[14] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*, **81**. MIT press, 2001. 7

[15] J. Pearl. *Causality: models, reasoning, and inference*, **47**. Cambridge Univ Press, 2000. 7, 13, 44, 56

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society*, **Series B (Methodological) 39 (1)**:1–38, 1977. 7, 8, 30, 44

[17] S.L. Lauritzen. **The EM algorithm for graphical association models with missing data**. *Computational Statistics & Data Analysis*, **19**(2):191–201, 1995. 7, 44

[18] J. Binder, D. Koller, S. Russell, and K. Kanazawa. **Adaptive probabilistic networks with hidden variables**. *Machine Learning*, **29**(2):213–244, 1997. 7, 44

[19] N. Friedman. **The Bayesian structural EM algorithm**. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 129–138. Morgan Kaufmann Publishers Inc., 1998. 7, 8, 9

[20] G. Elidan, N. Lotner, N. Friedman, and D. Koller. **Discovering hidden variables: A structure-based approach**. *Advances in Neural Information Processing Systems*, pages 479–485, 2001. 7, 8, 44

[21] C. Couvreur. *The EM algorithm: A guided tour*. Birkhauser, Boston, 1997. 7

[22] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell. 5th edition*. Garland Science, New York, 2008. 10, 11, 80, 81

[23] Florian Markowetz. *Probabilistic Models for Gene Silencing Data*. PhD thesis, 2006. 11, 12, 15, 17

[24] A. Fire, S.Q. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello. **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. *nature*, **391**(6669):806–811, 1998. 11

[25] Anne E. Carpenter and David M. Sabatini. **Systematic genome-wide screens of gene function.** *Nat Rev Genet*, **5**(1):11–22, Jan 2004. 11

[26] Florian Markowetz and Rainer Spang. **Inferring cellular networks–a review.** *BMC Bioinformatics*, **8 Suppl 6**:S5, 2007. 12

[27] Andreas Wagner. **Reconstructing pathways in large genetic networks from genetic perturbations.** *J Comput Biol*, **11**(1):53–60, 2004. 12

[28] C.D. Novina, P.A. Sharp, et al. **The rnai revolution**. *Nature*, **430**(6996):161–164, 2004. 12

[29] Michael Boutros, Herv Agaisse, and Norbert Perrimon. **Sequential activation of signaling pathways during innate immune responses in Drosophila.** *Dev Cell*, **3**(5):711–722, Nov 2002. 12

[30] A. Wagner. **How to reconstruct a large genetic network from n gene perturbations in fewer than n2 easy steps**. *Bioinformatics*, **17**(12):1183–1197, 2001. 15, 17

[31] Andreas Wagner. **Estimating coarse gene network structure from large-scale gene perturbation data.** *Genome Res*, **12**(2):309–315, Feb 2002. 15

[32] Florian Markowetz, Jacques Bloch, and Rainer Spang. **Non-transcriptional pathway features reconstructed from secondary effects of RNA interference.** *Bioinformatics*, **21**(21):4026–4032, Nov 2005. 16, 20, 24, 25

[33] Holger Froehlich, Mark Fellmann, Holger Sueltmann, Annemarie Poustka, and Tim Beissbarth. **Large scale statistical inference of signaling**

pathways from RNAi and microarray data. *BMC Bioinformatics*, **8**:386, 2007. 20, 28, 29, 30, 39

[34] Florian Markowetz and Rainer Spang. **Inferring cellular networks–a review.** *BMC Bioinformatics*, **8 Suppl 6**:S5, 2007. 20, 28

[35] Achim Tresch and Florian Markowetz. **Structure learning in Nested Effects Models.** *Stat Appl Genet Mol Biol*, **7**(1):Article9, 2008. 21, 23, 24, 28, 29, 30, 37

[36] Cordula Zeller, Holger Froehlich, and Achim Tresch. **A Bayesian network view on nested effects models.** *EURASIP J Bioinform Syst Biol*, page 195272, 2009. 22, 58

[37] R.E. Neapolitan. *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004. 23

[38] Charles J Vaske, Carrie House, Truong Luu, Bryan Frank, Chen-Hsiang Yeang, Norman H Lee, and Joshua M Stuart. **A factor graph nested effects model to identify networks from genetic perturbations.** *PLoS Comput Biol*, **5**(1):e1000274, Jan 2009. 24, 25, 26, 27, 28

[39] Kschischang, Frey, and Loeliger. **Factor graphs and the sum-product algorithm**. *IEEE Transactions on Information Theory*, **47**, 2001. 25, 28

[40] Brendan J. Frey and David J. C. MacKay. **A Revolution: Belief Propagation in Graphs With Cycles**. In *In Neural Information Processing Systems*, pages 479–485. MIT Press, 1997. 28

[41] Brendan J Frey and Delbert Dueck. **Clustering by passing messages between data points.** *Science*, **315**(5814):972–976, Feb 2007. 28

[42] David J.C. MacKay, David J. C. Mackay, Radford M. Neal, and Radford M. Neal. *Good Codes based on Very Sparse Matrices*. Springer, 1995. 28

[43] Holger Froehlich, Achim Tresch, and Tim Beissbarth. **Nested effects models for learning signaling networks from perturbation data.** *Biom J*, **51**(2):304–323, Apr 2009. 28, 30

[44] Juby Jacob, Marcel Jentsch, Dennis Kostka, Stefan Bentink, and Rainer Spang. **Detecting hierarchical structure in molecular characteristics of disease using transitive approximations of directed graphs.** *Bioinformatics*, **24**(7):995–1001, Apr 2008. 29

[45] S.J. Russell, P. Norvig, J.F. Canny, J.M. Malik, and D.D. Edwards. *Artificial intelligence: a modern approach*, **2**. Prentice hall Englewood Cliffs, NJ, 1995. 29

[46] Holger Froehlich, Mark Fellmann, Holger Sltmann, Annemarie Poustka, and Tim Beissbarth. **Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data.** *Bioinformatics*, **24**(22):2650–2656, Nov 2008. 29

[47] Theresa Niederberger, Stefanie Etzold, Michael Lidschreiber, Kerstin C. Maier, Dietmar E. Martin, Holger Frhlich, Patrick Cramer, and Achim Tresch. **MC EMiNEM maps the interaction landscape of the Mediator.** *PLoS Comput Biol*, **8**(6):e1002568, Jun 2012. 30, 31

[48] Benedict Anchang, Mohammad J. Sadeh, Juby Jacob, Achim Tresch, Marcel O. Vlad, Peter J. Oefner, and Rainer Spang. **Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models.** *Proc Natl Acad Sci U S A*, **106**(16):6447–6452, Apr 2009. 32, 35, 83, 86, 88, 104

[49] Florian Markowetz, Dennis Kostka, Olga G Troyanskaya, and Rainer Spang. **Nested effects models for high-dimensional phenotyping screens.** *Bioinformatics*, **23**(13):i305–i312, Jul 2007. 33, 35, 83

[50] Holger Froehlich, Paurush Praveen, and Achim Tresch. **Fast and efficient dynamic nested effects models.** *Bioinformatics*, **27**(2):238–244, Jan 2011. 36, 83, 84, 86, 88, 104

[51] Florian Markowetz, Dennis Kostka, Olga G. Troyanskaya, and Rainer Spang. **Nested effects models for high-dimensional phenotyping screens.** *Bioinformatics*, **23**(13):i305–i312, Jul 2007. 37, 50

[52] Zoubin Ghahramani. **Learning Dynamic Bayesian Networks: Lecture Notes In Computer Science**. **1387**:168–197. 37

[53] Gideon E. Schwarz. **Estimating the dimension of a model**. *Annals of Statistics*, **6 (2)**:461–464, 1978. 40

[54] D. Heckerman et al. **A tutorial on learning with Bayesian networks**. *Nato Asi Series D Behavioural And Social Sciences*, **89**:301–354, 1998. 43

[55] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. **Using Bayesian networks to analyze expression data.** *J Comput Biol*, **7**(3-4):601–620, 2000. 44, 56

[56] Juliane Schfer and Korbinian Strimmer. **An empirical Bayes approach to inferring large-scale gene association networks.** *Bioinformatics*, **21**(6):754–764, Mar 2005. 44, 56

[57] Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Jonathan Epperlein, Regina Samaga, Douglas A. Lauffenburger, Steffen Klamt, and Peter K. Sorger. **Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction.** *Mol Syst Biol*, **5**:331, 2009. 44, 56

[58] Florian Markowetz, Jacques Bloch, and Rainer Spang. **Non-transcriptional pathway features reconstructed from secondary effects of RNA interference.** *Bioinformatics*, **21**(21):4026–4032, Nov 2005. 44, 45, 46, 47, 50, 56, 57, 62, 76, 84

[59] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*, **274**. Wiley New York, 1997. 44

[60] G. Elidan, M. Ninio, N. Friedman, and D. Shuurmans. **Data perturbation for escaping local maxima in learning**. In *Proceedings of the national conference on artificial intelligence*, pages 132–139. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002. 44

[61] D. Colombo, M.H. Maathuis, M. Kalisch, and T.S. Richardson. **Learning high-dimensional directed acyclic graphs with latent and selection variables**. *Arxiv preprint arXiv:1104.5617*, 2011. 47, 57

[62] Chad L. Myers, Daniel R. Barrett, Matthew A. Hibbs, Curtis Huttenhower, and Olga G. Troyanskaya. **Finding function: evaluation methods for functional genomic data.** *BMC Genomics*, **7**:187, 2006. 47

[63] DH Rumsfeld. **DoD News Briefing-Secretary Rumsfeld and Gen. Myers. US Department of Defence**, 2002. 55, 105

[64] T. Richardson and P. Spirtes. **Ancestral graph Markov models**. *The Annals of Statistics*, **30**(4):962–1030, 2002. 57

[65] Florian Markowetz and Rainer Spang. **Inferring cellular networks–a review.** *BMC Bioinformatics*, **8 Suppl 6**:S5, 2007. 76

[66] Arjun Raj and Alexander van Oudenaarden. **Nature, nurture, or chance: stochastic gene expression and its consequences.** *Cell*, **135**(2):216–226, Oct 2008. 79

[67] Vahid Shahrezaei and Peter S. Swain. **The stochastic nature of biochemical networks.** *Curr Opin Biotechnol*, **19**(4):369–374, Aug 2008. 79

[68] D. L. Stocum. **Amphibian regeneration and stem cells.** *Curr Top Microbiol Immunol*, **280**:1–70, 2004. 80

[69] Uri Alon. *An introduction to systems biology : Design Principles of biological circuits*. CHAPMAN & HALL/CRC, 2007. 81

[70] A. K. TARKOWSKI. **Experiments on the development of isolated blastomers of mouse eggs.** *Nature*, **184**:1286–1287, Oct 1959. 81

[71] J.A. Thomson, J. Itskovitz-Eldor, S.S. Shapiro, M.A. Waknitz, J.J. Swiergiel, V.S. Marshall, and J.M. Jones. **Embryonic stem cell lines derived from human blastocysts**. *science*, **282**(5391):1145–1147, 1998. 81

[72] Davor Solter. **From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research.** *Nat Rev Genet*, **7**(4):319–327, Apr 2006. 81

[73] Y. Suda, M. Suzuki, Y. Ikawa, and S. Aizawa. **Mouse embryonic stem cells exhibit indefinite proliferative potential.** *J Cell Physiol*, **133**(1):197–201, Oct 1987. 81

[74] H. Niwa, J. Miyazaki, and A. G. Smith. **Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells.** *Nat Genet*, **24**(4):372–376, Apr 2000. 81

[75] Junji Fujikura, Eiji Yamato, Shigenobu Yonemura, Kiminori Hosoda, Shinji Masui, Kazuwa Nakao, Jun ichi Miyazaki Ji, and Hitoshi Niwa. **Differentiation of embryonic stem cells is induced by GATA factors.** *Genes Dev*, **16**(7):784–789, Apr 2002. 81

[76] Kaoru Mitsui, Yoshimi Tokuzawa, Hiroaki Itoh, Kohichi Segawa, Mirei Murakami, Kazutoshi Takahashi, Masayoshi Maruyama, Mitsuyo Maeda, and Shinya Yamanaka. **The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.** *Cell*, **113**(5):631–642, May 2003. 81

[77] Ian Chambers, Douglas Colby, Morag Robertson, Jennifer Nichols, Sonia Lee, Susan Tweedie, and Austin Smith. **Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells.** *Cell*, **113**(5):643–655, May 2003. 81

[78] Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, David K Gifford, Douglas A Melton, Rudolf Jaenisch, and Richard A Young. **Core transcriptional regulatory circuitry in human embryonic stem cells.** *Cell*, **122**(6):947–956, Sep 2005. 81, 82

[79] Hitoshi Niwa, Yayoi Toyooka, Daisuke Shimosato, Dan Strumpf, Kadue Takahashi, Rika Yagi, and Janet Rossant. **Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation.** *Cell*, **123**(5):917–929, Dec 2005. 81

[80] Natalia Ivanova, Radu Dobrin, Rong Lu, Iulia Kotenko, John Levorse, Christina DeCoste, Xenia Schafer, Yi Lun, and Ihor R. Lemischka. **Dissecting self-renewal in stem cells with RNA interference.** *Nature*, **442**(7102):533–538, Aug 2006. 81, 82, 83, 86

[81] Akira Nishiyama, Li Xin, Alexei A Sharov, Marshall Thomas, Gregory Mowrer, Emily Meyers, Yulan Piao, Samir Mehta, Sarah Yee, Yuhki Nakatake, Carole Stagg, Lioudmila Sharova, Lina S Correa-Cerro, Uwem Bassey, Hien Hoang, Eugene Kim, Richard Tapnio, Yong Qian, Dawood Dudekula, Michal Zalzman, Manxiang Li, Geppino Falco, Hsih-Te Yang, Sung-Lim Lee, Manuela Monti, Ilaria

Stanghellini, Md Nurul Islam, Ramaiah Nagaraja, Ilya Goldberg, Weidong Wang, Dan L Longo, David Schlessinger, and Minoru S H Ko. **Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors.** *Cell Stem Cell*, **5**(4):420–433, Oct 2009. 81, 82

[82] Ian Chambers, Jose Silva, Douglas Colby, Jennifer Nichols, Bianca Nijmeijer, Morag Robertson, Jan Vrana, Ken Jones, Lars Grotewold, and Austin Smith. **Nanog safeguards pluripotency and mediates germline development.** *Nature*, **450**(7173):1230–1234, Dec 2007. 81

[83] Shinji Masui, Yuhki Nakatake, Yayoi Toyooka, Daisuke Shimosato, Rika Yagi, Kazue Takahashi, Hitoshi Okochi, Akihiko Okuda, Ryo Matoba, Alexei A Sharov, Minoru S H Ko, and Hitoshi Niwa. **Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells.** *Nat Cell Biol*, **9**(6):625–635, Jun 2007. 81

[84] Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, David K Gifford, Douglas A Melton, Rudolf Jaenisch, and Richard A Young. **Core transcriptional regulatory circuitry in human embryonic stem cells.** *Cell*, **122**(6):947–956, Sep 2005. 82

[85] David J Rodda, Joon-Lin Chew, Leng-Hiong Lim, Yuin-Han Loh, Bei Wang, Huck-Hui Ng, and Paul Robson. **Transcriptional regulation of nanog by OCT4 and SOX2.** *J Biol Chem*, **280**(26):24731–24737, Jul 2005. 82

[86] Jacques Ferlay, Hai-Rim Shin, Freddie Bray, David Forman, Colin Mathers, and Donald Maxwell Parkin. **Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.** *Int J Cancer*, **127**(12):2893–2917, Dec 2010. 89, 90

[87] Alberto Mantovani. **Cancer: Inflaming metastasis.** *Nature*, **457**(7225):36–37, Jan 2009. 90

[88] Douglas Hanahan and Robert A. Weinberg. **Hallmarks of cancer: the next generation.** *Cell*, **144**(5):646–674, Mar 2011. 90

[89] Gary J. Kelloff and Caroline C. Sigman. **Cancer biomarkers: selecting the right drug for the right patient.** *Nat Rev Drug Discov*, **11**(3):201–214, Mar 2012. 90

[90] Janet E. Dancey, Philippe L. Bedard, Nicole Onetto, and Thomas J. Hudson. **The genetic basis for cancer treatment decisions.** *Cell*, **148**(3):409–420, Feb 2012. 90

[91] P. J. Morin, A. B. Sparks, V. Korinek, N. Barker, H. Clevers, B. Vogelstein, and K. W. Kinzler. **Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC.** *Science*, **275**(5307):1787–1790, Mar 1997. 90, 94

[92] A. B. Sparks, P. J. Morin, B. Vogelstein, and K. W. Kinzler. **Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer.** *Cancer Res*, **58**(6):1130–1134, Mar 1998. 90

[93] E. R. Fearon and B. Vogelstein. **A genetic model for colorectal tumorigenesis.** *Cell*, **61**(5):759–767, Jun 1990. 90

[94] Gerrit Erdmann. *Evi/Wls Mediated Wnt Secretion is Required for Colon Cancer Growth and Survival Despite APC or catenin Mutations*. PhD thesis, 2012. 90, 91, 93, 94, 95, 96, 97

[95] C. S. Potten, C. Booth, and D. M. Pritchard. **The intestinal epithelial stem cell: the mucosal governor.** *Int J Exp Pathol*, **78**(4):219–243, Aug 1997. 90

[96] Elena Sancho, Eduard Batlle, and Hans Clevers. **Signaling pathways in intestinal development and cancer.** *Annu Rev Cell Dev Biol*, **20**:695–723, 2004. 90, 91

[97] Freddy Radtke, Hans Clevers, and Orbicia Riccio. **From gut homeostasis to cancer.** *Curr Mol Med*, **6**(3):275–289, May 2006. 91

[98] Tannishtha Reya and Hans Clevers. **Wnt signalling in stem cells and cancer.** *Nature*, **434**(7035):843–850, Apr 2005. 91

[99] Catriona Y. Logan and Roel Nusse. **The Wnt signaling pathway in development and disease.** *Annu Rev Cell Dev Biol*, **20**:781–810, 2004. 91

[100] Tina Buechling and Michael Boutros. **Wnt signaling signaling at and above the receptor level.** *Curr Top Dev Biol*, **97**:21–53, 2011. 91

[101] T. Nakamura, F. Hamada, T. Ishidate, K. Anai, K. Kawahara, K. Toyoshima, and T. Akiyama. **Axin, an inhibitor of the Wnt signalling pathway, interacts with beta-catenin, GSK-3beta and APC and reduces the beta-catenin level.** *Genes Cells*, **3**(6):395–403, Jun 1998. 91

[102] Madelon Maurice. **Mechanisms of Wnt protein secretion and signal reception in development and cancer**, 2010. 92

[103] Alix Scholer-Dahirel, Michael R. Schlabach, Alice Loo, Linda Bagdasarian, Ronald Meyer, Ribo Guo, Steve Woolfenden, Kristine K. Yu, Judit Markovits, Karen Killary, Dmitry Sonkin, Yung-Mae Yao, Markus Warmuth, William R. Sellers, Robert Schlegel, Frank Stegmeier, Rebecca E. Mosher, and Margaret E. McLaughlin. **Maintenance of adenomatous polyposis coli (APC)-mutant colorectal cancer is dependent on Wnt/beta-catenin signaling.** *Proc Natl Acad Sci U S A*, **108**(41):17135–17140, Oct 2011. 93

[104] Nick Barker and Hans Clevers. **Mining the Wnt pathway for cancer therapeutics.** *Nat Rev Drug Discov*, **5**(12):997–1014, Dec 2006. 93

[105] Baozhi Chen, Michael E. Dodge, Wei Tang, Jianming Lu, Zhiqiang Ma, Chih-Wei Fan, Shuguang Wei, Wayne Hao, Jessica Kilgore, Noelle S. Williams, Michael G. Roth, James F. Amatruda, Chuo Chen, and Lawrence Lum. **Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer.** *Nat Chem Biol*, **5**(2):100–107, Feb 2009. 94

[106] Gordon K. Smyth. **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol*, **3**:Article3, 2004. 97, 98

[107] Tianru Jin and Ling Liu. **The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus.** *Mol Endocrinol*, **22**(11):2383–2392, Nov 2008. 99

Curriculum Vitae

Address

Institute for Functional Genomics
Department of Statistical Bioinformattics
University of Regensburg, Josef Engertstr. 9
93053 Regensburg, Germany.Tel: 0049 (0)941 943 5055
Email:Mohammed.Sadeh@klinik.uni-regensburg.de

Education

September 2007 - Present **University of Regensburg**
PhD Student in Bioinformatics Advisor: Prof. Rainer Spang
Oct 2003 - Nov 2006 **Ferdosi Mashad University, Mashad, Iran**
Master of science in Mathematical Statistics Advisor: Prof. Shahkar
Oct 1998 - July 2003 **Shahid Beheshti University, Tehran, Iran**
Bachelor of science in Statistics Minor: Computer Science

Working Experience

June 2007-present **Institute of Functional Genomics**
Research assistant University of Regensburg, Germany

Sept 2002- Sept 2007 **Institute for Studies in Theoretical Physics and Mathematics**
Student Intern Tehran, Iran

International Conferences

July 15-17th 2011 **ISMB Vienna, Austria**
July 17-19th 2011 **ECCB Vienna, Austria**
Sep 26-29th 2010 **ECCB Ghent, Belgium**

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without using any other than the aids listed. Any thoughts directly or indirectly taken from somebody else's sources are made discernible as such. This dissertation has not previously been presented in identical or similar form to any other German or foreign examination board.

The thesis work was conducted under the supervision of Prof.Dr. Rainer Spang at the Institute of Functional Genomics, University of Regensburg, Germany.

REGENSBURG,