Analysis of state transition in complex systems: statistical procedures in biological networks

Lina Dornelas Thomas

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Estatística

Orientador: Prof. Dr. Anatoli Iambartsev

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES, CNPq e FAPESP Processo 06223-1

São Paulo, março de 2017

Analysis of state transition in complex systems: statistical procedures in biological networks

Esta versão da tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 26/04/2017. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Anatoli Iambartsev (orientador) - IME-USP
- Prof. Dr. Vladimir Belitsky - IME-USP
- Prof. Dr. Ronaldo Fumio Hashimoto - IME-USP
- Prof. Dr. Eugene Abramovish Pechersky - Externo
- Prof. Dr. Marina Vachkovskaia - IMECC-UNIC - Externo

Acknowledgements

During the last three years and a half, I suffered from pain in my neck, shoulder, arm, and hand on the right side of my body. To finish my research and to write this thesis I counted on an army of hands! People who helped me click and type remotely or by my side. I would like to thank all those who helped me deeply. I don't know if I would be able to finish the work if it weren't for you! There were a few people offering to help who I did not contact. I knew I could count on you! Because of you I could relax and focus on my Ph.D. and not fear if my hands would allow me to do the job on time or not!

Henrique Bolfarine, Lucas Altavista, Gustavo Torres, Gustavo Carrijo Duarte, Débora Hasegawa, Humberto H. B. Viglioni, Patricia Viana, Michael Springer, Lilian Luci Lemos, Bharat Rastogi, Daniel de Brito Reis, Carolina Bueno, Daniel Costa Bucher, Paulo Bittencourt Moura, Nathalia Demetrio, Bruna Dornelas Serra, Mila Valle, Rose Mary Ghazal, Omar Martinez, Ana Luisa Losnak, Danielle Bambace, Luís Augusto Magalhães Teles, Frederico Almeida, Leila Lemos, Isabel Guerrero Ochoa, Marcos Virgílio da Silva, Ana Lu Fonseca, Gustavo Henrique Albuquerque, Filó Silva, Leandro Santoro, Renata Stella Kouri, Eduardo Fernandes, Eliete Carvalho, Raphael Pagotto, Laura Rita, Rose Blau Bienemann, Suellen Zatti, Priscila Fulvia Bittencourt, Joana Invernizzi Cunha, Camille Gobbo, Luz Marina Gómez, Milena Egea Marin Guerreiro, Alejandra Rada. Thank you!

I would also like to thank my family and friends for all the support during the tough days when I felt deeply depressed with all the health issues. My mother, father, and aunt Margarida Maria Dornelas who also helped me type. My special thanks to my boyfriend André Belejo who was always there for me, in the good and stressful times, who made me get up and move forward every time I wanted to give up! Thank you for listening to all my complaints with patience! Thank you for taking time from your Ph.D. to type for me, click for me, powerpoint for me! You are my angel!

I would like to thank my adviser with the bottom of my heart for all his support and patience! For not giving up on me when I was very ill and depressed. For respecting my healing time! For helping with the computer when I didn't have any help yet. Without you, I would not be able to do all the work!

Abstract

Thomas, L. D. **Analysis of state transition in complex systems: statistical procedures in biological networks**. 2017. 97 pages. PhD Dissertation - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Complex Systems, a field that studies interactions of large number of components, have been intensely researched over the past recent years due to its ability to represent complex structures that haven't been completely understood. Components can be represented by nodes in a network in which edges illustrate pairwise interactions between nodes. A system faces a state transition when its outcome is altered due to changes in some of its nodes nature and pairwise interactions. New tools to identify the causes of such transitions are in great need. Since biological systems are good examples of complex systems, they are used here to address this question.

Network analysis is a powerful and general approach to investigate systems characterized mainly by pairwise interactions. Correlation coefficients are a statistical measure commonly used to quantify pairwise interactions. Even though this measure is very popular in the research community to define the edges of a network, it has already been proven that partial correlation is a more suitable measure for this purpose in cases where the nodes represent Gaussian random variables. However, its usual calculation, called inverse method, only works on data with large sample size. Since biological data is known for its "Big Data" properties, new partial correlation methods have been already developed and need assessment. Local partial correlation, a new method developed by the author, is herein described and applied to the reconstruction of a complex biological system (cervical cancer). Local partial correlation, graphical lasso and network reconstruction with ridge penalty are examples of such models and are compared for the first time. We have observed in this study that local partial correlation returns similar networks as the method with ridge penalty and both produce good ROC curves.

A tool widely used for analysis of "Big Data" in biology is called co-expression networks. Here, we developed a step-by-step guide on how to reconstruct and analyze co-expression networks using either correlation or local partial correlation approach . Several methods have been proposed to answer biological questions interrogating state transition in these type of networks. Differential co-expression analysis is a recent approach that measures how gene interactions change when a biological system transitions from one state to another. While differentially expressed genes have been extensively investigated, the role of differentially co-expressed genes in gene regulation is not well studied even though the importance of identifying deregulated pathways has already been noted. Here, investigation of differentially co-expressed genes is performed in networks considering only differentially expressed genes as nodes for a relatively simple mono-causal process (B lymphocyte deficiency) and a complex multi-causal system (cervical cancer).

We show that in B cell deficiency the differentially co-expressed genes are highly enriched with immunoglobulin genes (causal genes), whereas, in cervical cancer, differentially co-expressed genes are located close to causal genes and act as "bottlenecks" rather than causal drivers with most flows that come from the key driver genes to the peripheral genes passing through differentially co-expressed genes. Using *in vitro* knockdown experiments for two out of 14 differentially co-expressed genes found in cervical cancer (FGFR2 and CACYBP), we showed that they play regulatory roles in cancer cell growth. Numerical analysis was performed for different graphical structures with different number of nodes and confirmed that, regardless of which nodes suffered knockout, pairs of nodes with high difference of correlation tend to be located close to the perturbation site. Therefore, identifying differentially co-expressed genes in co-expression networks is an important procedure in detecting regulatory genes involved in alterations of phenotype.

Key words: differential correlation, partial correlation, state transition, gene co-expression networks, complex networks

Resumo

Thomas, L. D. **Análise de transição entre estados em sistemas complexos: procedimentos estatísticos em redes biológicas.**. 2017. 97 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Sistemas Complexos, uma área da Ciência que estuda interações de elevado número de componentes, tem sido intensivamente alvo de estudo nos últimos anos devido à sua capacidade de representar estruturas complexas que ainda não foram completamente compreendidas. Componentes podem ser representados por nós em uma rede na qual elos ilustram as interações entre pares de nós . Um sistema está perante a uma transição de estado quando o seu produto é alterado devido a alterações na natureza de alguns dos seus nós ou na interação entre pares de nós. Existe uma grande necessidade de novas ferramentas para identificar as causas destas transições de estado. Uma vez que sistemas biológicos são exemplos de sistemas complexos, são aqui usados para abordar esta questão.

Análise de rede consiste na abordagem geral e poderosa para investigar sistemas que são caracterizados maioritariamente por interação entre pares de nós. Coeficientes de correlação são uma medida estatística frequentemente usada para quantificar interações entre pares de nós. Apesar dessa medida estatística ser muito popular entre a comunidade ciêntifica para definir os elos de uma rede, já foi provado que a correlação parcial é uma medida mais apropriada para casos em que os nós representam variáveis aleatórias Gaussianas. No entanto, o seu cálculo usual, denominado método inverso, funciona apenas em grandes quantidades de dados. Uma vez que dados biológicos são conhecidos pelas suas propriedades "Big Data", novos métodos de correlação parcial já foram desenvolvidos e precisam ser avaliados. A correlação parcial local, novo método desenvolvido pela autora, é aqui descrita e aplicada à reconstrução de um sistema biológico complexo: câncer de colo do útero. Correlação parcial local, graphical lasso e reconstrução de rede com penalidade ridge são exemplos de alguns desses métodos, os quais são comparados pela primeira vez. Foi então observado neste estudo que os métodos de correlação parcial local e com penalidade ridge dão origem a redes similares, e ainda produzem ambos boas curvas ROC.

Redes de co-expressão são uma ferramenta globalmente usada na análise de "*Big Data*" em biologia. Neste trabalho é transmitida uma ideia geral de análise de redes incluindo um guia passo a passo de como construir uma rede de co-expressão usando tanto correlação como correlação parcial local. Vários métodos têm sido propostos para responder a questões relacionadas com transição de estado neste tipo de redes. Análise de co-expressão diferencial é uma abordagem recente que mede como interações entre genes mudam quando o sistema biológico transita de um estado para outro. Enquanto que genes diferencialmente expressos tem sido bastante estudados, a importância de genes diferencialmente co-expressos em regulação gênica não é clara, mesmo sabendo que a

identificação de caminhos desregulados é de elevada importância. Neste trabalho, estudam-se genes diferencialmente co-expressos em redes compostas apenas por genes diferencialmente expressos para um processo monocausal simples (deficiência de linfócito B) e um sistema multi-causal complexo (câncer de colo do útero).

Resultados indicam que muitos dos genes diferencialmente expressos são genes imunoglobulina (genes causais) nas redes de deficiência de células B, enquanto que, nas redes de câncer de colo do útero, os genes diferencialmente co-expressos estão localizados próximo aos genes causais, atuando como "bottlenecks" e não como *drivers* causais. Neste caso, um maior número de correntes originadas no gene condutor seguindo para os genes periféricos passam pelos genes diferencialmente co-expressos. Ao realizar ensaios experimentais *knockdown in vitro* em 14 genes diferencialmente co-expressos outrora encontrados em cervical cancer(FGFR2 e CACYBP), foi verificado que os genes diferencialmente co-expressos têm um papel fundamental na regulação do crescimento de células cancerígenas. Foram ainda efetuadas análises numérica para tipos de estruturas gráficas distintas com diferente número de nós tendo sido confirmado que os pares de nós com maior diferença de correlação têm tendência para estar localizados perto da região da perturbação, independentemente dos nós que sofreram *knockout*. Portanto, identificar genes diferencialmente co expressos nas redes de co-expressão é um importante procedimento na detecção de regulação de genes envolvidos nas alterações de fenótipo.

Palavras-chave: correlação diferencial, correlação parcial, transição entre estados, redes de co-expressão gênica, redes complexas

Contents

List of abbreviations

BcKO	Bcell knockout
DC	Differentially Co-expressed
DCP	Differentially Correlated Pair
DEG	Differentially Expressed Gene
FDR	False Discovery Rate
GGM	Gaussian Graphical Models
GNW	GeneNetWeaver
Ig	Immunoglobuline
MLE	Maximum Likelihood Estimation
ODE	Ordinary Differential Equation
ROC	Receiver Operating Characteristic
SDE	Stochastic Differential Equation

List of symbols

Ω	Precision Matrix
S	Sample Covariance Matrix
R	Correlation Matrix
n	Number of Samples
p	Number of Variables/Nodes
k	Number of datasets
α	Threshold for p-value
p_{Fisher}	Fisher p-value
χ^2_k	chi squared distribution with k degrees of freedom
$\mathcal{V}$	Set of nodes
$\mathcal{E}$	Set of edges
$G(\mathcal{V}, \mathcal{E})$	A graph with set of nodes $\mathcal{V}$ and set of edges $\mathcal{E}$
X_i	A random variable
$N_p(\mu, \Sigma)$	p-multivariate normal distribution with mean vector μ and covariance matrix Σ
$\rho_(i, j)$	Correlation between X_i and X_j
$\rho_{ij.Y}$	Partial correlation between X_i and X_j given remaining variables $\mathbf{Y}$

List of Figures

List of Tables

Chapter 1

Introduction

Complex Systems is a field that studies systems with large number of interacting components, [BY97, CS99]. Researchers have been seriously interested in its understanding throughout the recent years due to its ability to represent complex structures which are still a challenge to fully comprehend ([HSA06, HJG17]). Examples of complex systems are the brain, a living cell, the stock market, genetics, social organization. A common representation is network reconstruction where the nodes are its components and the edges are interactions between components([Str01, BS09]). Sometimes there can be a change in the nature of a few components leading to changes in the behavior of other components which will change the entire system outcome. This situation can be viewed as a state transition and can be exemplified by drops in the stock market, diseases, strikes, among others.

This project aims to develop new tools that analyze changes in complex networks when a state transition occurs. We would like to be able to identify the components that trigger the transition as well as to comprehend how the structure changes in different complexities. For this reason we chose to work with biological systems, more explicitly co-expression networks reconstructed from microarray gene expression data, and with simulations for different graph structures. In real data approach, the nodes represent genes and the edges show if there is an association between each pair of genes. In both data, when one or more nodes are disturbed, perturbations arise in other nodes changing the network structure.

General changes in a biological system after a state transition have already been extensively investigated, [Li02, NHDQ11, SYAP11, PSS+13]. The most popular one is when several genes present statistically significant changes in their expression means. These genes are called Differentially Expressed Genes (DEGs), [DYCS02, RYB03, XFG+04]. A structural change that has been recently getting attention is when the interactions within one or more pairs suffer severe alterations. Two genes can either lose or gain interaction or can start interacting in a different direction, for example, a gene acts as an inhibitor in one state and then becomes an enhancer in another state or the other way around. Different approaches to calculate structural changes in biological systems have already been suggested in previous publications, [LWCZ04, Wat06, SKV+11, DYK12, Fuk13].

Co-expression networks are herein reconstructed setting DEGs as nodes and the interactions between two DEGs are established through significant Pearson correlation coefficient. Sometimes correlation coefficients are not ideal to define the edges of a network since it allows the appearance of indirect links. In studies with enough sample sizes, it is interesting to represent only direct connections between nodes which can be done through partial correlation. However, gene expression data is often composed by thousands of genes vs. tens of samples which makes partial correlation inapplicable. In order to enable the use of partial correlation to high dimensional data the author created a new method called local partial correlation (LPC) which basically applies the inverse method - usual method for partial correlation - to the neighborhood of each pair of correlated nodes, [Tho12]. The efficiency of LPC in network reconstruction is compared for the first time to two partial correlation methods with regularization: Graphical Lasso [FHT08] and GGM Ridge [HS14]. Both of them can also be applied to high dimensional data. The comparison was made using simulated and real data.

We say that there is a change in interaction when a pair of genes goes through a statistically significant change in Pearson correlation after a state transition. We call the pairs that go through these changes Differentially Correlated Pairs (DCPs). We denominate the genes forming the DCPs as Differentially Co-expressed (DC) genes. We already know from literature that there are strong indications of some kind of relationship between DEGs and DCPs, [dlF10]. However, no further investigation about such relationship was found leaving the role of DC genes in gene regulation unknown. This project also aims to reveal how DCPs influence DEGs networks out of two different conditions. The main goal of this part is to aid the development of tools that are able to detect a few candidate causal genes that can later be experimentally tested.

Our biologist collaborators from Oregon State University (OSU) provided us with insights and data from two of their previously published papers. The first one [SMH+11] studies the global gene expression in the jejunum of deficient B cells mice (Bcell knockout, a homogenous one causal factor system) and the second one [MSY+13] seeks to understand the process of cervical cancer, a heterogeneous multi-clausal system. The choice of these two biological models is due to the already existing knowledge about the causes of the state transition. All datasets can be obtained from the two following datasets resource: NCBI GEO and Array Express at the European Bioinfomatics Institute.

Numerical analysis is also performed in different graph structures in order to confirm the results from real data analysis. Simulations of state transition are based on Gaussian Graphical Models and GeneNetWeaver, a network reconstruction software from DREAM Project.

1.1 General Objectives

General objectives and motivations of this dissertation are listed below:

- Compare the efficiency of local partial correlation with other partial correlation methods.
- Describe local partial correlation properties.
- Understand how network changes after a state transition.
- Evaluate the importance of differential correlated pairs in a state transition.
- Develop mechanisms to detect causal-gene candidates to be experimentally tested.
- Confirm the results from Biological Networks through numerical analysis and find out if they can be generalized to any kind of network structure.

1.2 Dissertation Outline

Chapter 2 describes several basic concepts about graph theory and biological networks necessary for the comprehension of the entire document.

Chapter 3 defined local partial correlation along with some of its properties and simulation studies developed in [Tho12].

Chapter 4 describes the steps involved in biological network reconstruction : 1) data normalization, 2) discovery of DEGs, 3) correlation analysis for network reconstruction, 4) discrimination between direct and indirect links, 5) proportional unexpected correlations, 6) meta-analysis, 7) identification of differentially co-expressed gene pairs ([DYR+15]).

In Chapter 5 co-expression networks of B cell deficiency (Control and BcKO) were reconstructed using Pearson correlation coefficient for mus musculus datasets. Co-expression networks of cervical cancer data (normal and cancer) were reconstructed using local partial correlation method. Differentially correlated pairs were identified along with the location of their genes in BcKO and in cancer networks. Minimum Shortest Path and Bi-partite Betweenness Centrality where statistically evaluated for differentially co-expressed genes in corresponding networks. Experimental and numerical

analysis are performed to support the results obtained from real data analysis. This chapter is a product of [TVS+16].

Chapter 6 describes two partial correlation methods with regularization: graphical lasso [FHT08] and network reconstruction with ridge penalty [HS14], then compares local partial correlation with these two other partial correlation. This comparison was made using GGM and GNW simulations and real data from cervical cancer studies.

Finally in Chapter 7 some general conclusions are presented along with suggestions for future work.

Chapter 2

Basic Concepts

In this chapter, we introduce some basic concepts one must know in order to be able to fully comprehend the work developed in the scope of this research. Graph theory and graphical modeling are herein described. Several properties of Biological networks are also explained.

2.1 Graph Theory

In mathematics, graph theory is the study of graphs, which are also known as networks and represent mathematical pairwise relations between objects. A graph is an ordered pair $G = (\mathcal{V}, \mathcal{E})$ constituted by a set of nodes $\mathcal{V} = \{1, 2, ..., p\}$ and a set $\mathcal{E}$ of edges, which are 2-element subsets of $\mathcal{V}$, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. It may be undirected, where (i, j) and (j, i) mean the same, or directed. In the latter case (i, j) represents an edge coming from i and pointing to j and (j, i) means the other way around. This research only considers undirected edges.

Each node can be a component of one or more edges. The amount of edges connected to a certain node is called degree. Nodes with the degree 1 are known as peripheral or leaf nodes while nodes with unusually high degree are known as hubs. The probability distribution of a measure of edges connected to each node is the degree distribution. From now on, we will consider hubs the nodes whose degree is higher than the 80^{th} percentile of the degree distribution.

A path in a graph is a sequence of edges - either finite or infinite - necessary to connect two nodes. Sometimes there is more than one path between two nodes, and the path formed by less edges is called shortest path, which is an important measure to be considered in many studies. Another relevant graph measure is betweenness centrality, which is the number of shortest paths between any pairs of nodes passing through a certain node.

$$g(\nu) = \sum_{s \neq \nu \neq t} \frac{\sigma_{st}(\nu)}{\sigma_{st}}, \tag{2.1}$$

where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those shortest paths that pass through vertex v (node for which the metric is calculated).

2.1.1 Graph Structure Models

There are several different probabilistic models to structure random graphs, that is, for a fixed set of nodes $\mathcal{V}$ with size p, the set of edges $\mathcal{E}$ is randomly generated. In most models, the existence of each edge is decided upon a pre-established probability. These types of models are also known as random graphs. Here we describe a few graph structures which are taken into consideration when numerical analysis is concerned. There are already functions for most of them in the igraph R package. R scripts for three graph structures can be found in Appendix B.1: edge percent, Galton Watson trees and lattice with leaves.

Erdõs-Rényi

The most common model is called Erdõs-Rényi [ER59] and comes with two versions: $G(p, e)$, which is uniformly drawn out of all possible structures with p nodes and e edges, and $G(p, q)$, which is generated by assigning probability q to each possible edge. It is worth noting that both versions result in graphs with unconnected nodes, specially when e or q are small. In order to avoid such outcome we set $q \cong \frac{6}{p}$ leading to graphs with $|\mathcal{E}| \approx 3p$. It is known that if $q > \frac{(1+\epsilon)\ln p}{p}$, then a graph $G(p, q)$ will almost surely be connected. Since $6 > \ln p$ for $p < 400$, we almost surely guarantee a connected graph for our simulations of graphs with 50, 100 and 200 nodes. Besides that, a few minor modifications were made to $G(p, e)$ algorithm to assure at least one connection to all nodes. From now on, we will call Erdõs-Rényi only structures generated by $G(p, q = \frac{6}{p})$ with function sample_gnp from igraph R package while our modified version will be referred to as edge percent (EP) and can be found in Appendix B.1.

Tree

Another popular graphical structure among mathematicians is the tree graph, which is defined by the existence of only one path connecting any pair of nodes. There are several ways to generate a tree. In this work we will consider two types of tree models: a) 2-regular tree starting with one root node and each node with two offspring nodes until the intended number of nodes is reached, b) Galton-Watson tree also starting with one root node and offspring distribution lognormal$(0, 1)$, that is, a Galton-Watson branching process. Since the chosen distribution is continuous, the number of offspring of each node i will be rounded up to the next integer, that is, $ceiling(Z_i)$, where $Z_i \sim$ lognormal$(0, 1)$ and ceiling is a function of R. The choice of this distribution was due to observing such offspring distribution in most of the reconstructed co-expression networks from our lab.

Scale free

It is already known that the degree distribution of large networks, such as several natural and human-made systems, is frequently a scale-free power-law distribution, that is, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a node has degree k (see [BA99, JTA+00]). Since our goal involves the investigation of biological network with hundreds of nodes, it is of our interest to comprehend this type of graph structure. A popular algorithm used to generate random Scale-free graphs is the Barabasi-Albert [BA99] model, in which new vertices attach preferentially to high-degree nodes enabling the presence of hubs. We simulated the Scale-free network using the preferential attachment principle as a random process. For a network with p nodes, we start with two connected nodes in a graph $G_{SF} = (\mathcal{V}, \mathcal{E})$, starting with $\mathcal{V} = \{1, 2\}$ and $(1, 2) \in \mathcal{E}$). The next node 3 is added, which connects randomly to one of the nodes $j \in V$ with probability distribution

$$p_j = P\frac{deg(j)}{\sum_{l\in\mathcal{V}} deg(l)},$$

where $deg(j)$ is the number of connections of node j. The process goes on, until $|\mathcal{V}| = p$, generating a graph with Scale-free properties.

Small world

In addition, [CHBA03] has shown analytically that Scale-free networks are ultra-small worlds which drew our attention to such type of networks. Small-world networks are a type of mathematical graph in which the shortest path between any two nodes is rather small and its growth is proportional to $\log p$, [WS98]. It is based on the idea of modeling social interactions and other networks in real world applications. To generate small world graph structures, we use the function sample_smallworld from igraph R package.

Lattice

Another graph structure often studied in probability models is lattice graph. We will focus on square grid graph, whose vertices represent points in plane grid formed by the Cartesian product $\{1, 2, \cdots, n\} \times \{1, 2, \cdots, m\}$, where $n, m \in \mathbb{N}$, and the edges exist only for pairs of nodes that are located in the plane at distance one. Our interest in studying the effect of leaf nodes on DCP studies (see Section 2.2.2) led us to add a few extra nodes to a parallel one-distance plane only connected to one node in the lattice plane.

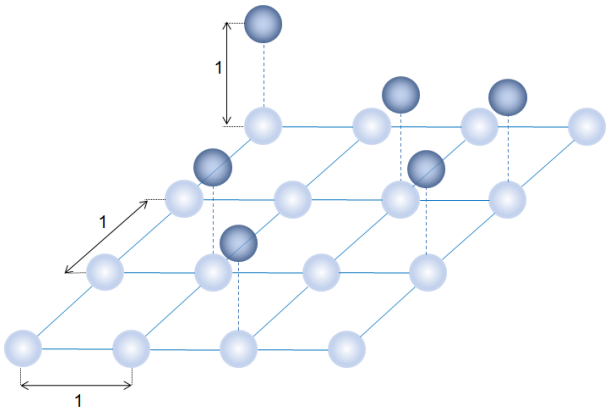


Figure 2.1: *Example of leaves in a lattice graph. The light blue nodes are vertices in the square grid while dark blue nodes are leaves added at a distance of* 1.

2.1.2 Graphical Models

Gaussian Graphical Modeling

Gaussian graphical model (GGM) is characterized by a graph $G = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, \cdots, p\}$ is the set of nodes representing a p-dimensional multivariate normal distributed random vector $\mathbf{X} = \{X_1, \cdots, X_p\} \sim N_p(\mu, \mathbf{\Sigma})$ and $\mathcal{E}$ is defined by pairwise conditional dependence, that is, $(i, j) \in \mathcal{E}$ if and only if X_i and X_j are conditional dependent given $\{X_k, k \neq i, j\}$. Consequently $(i, j) \notin \mathcal{E} \Leftrightarrow X_i$ and X_j are conditionally independent given $\{X_k, k \neq i, j\}$. For more details, see [MKB79, Edw95, Lau96]. However, calculation of conditional independence can be very complex. In [BSS04], it is shown that conditional independence can be replaced by partial correlation 0 when the set of variables follows a multivariate normal distribution.

It is common knowledge that each non-zero entry ($\omega_{ij} \neq 0$) of the inverse correlation matrix $\Omega = \Sigma^{-1}$ (also called precision matrix) indicates the existence of conditional dependence between the corresponding pair, that is, $\mathcal{E} = \{(i, j) \in \mathcal{V} \times \mathcal{V} : \omega_{ij} \neq 0\}$. More details can be found in Chapter 3.

Partial Correlation

In a model with $p > 2$ interacting variables, correlation between 2 variables is actually a result of the interaction between those 2 variables combined with interactions of those 2 variables with other variables and so on. For example, lets suppose it is already known that X is correlated to Y and to Z. Therefore, when the value of X goes up or down, then the values of Y and Z will also go up or down depending on the direction (sign) of the correlation. It is easy to realize that when X varies, a correlation between Y and Z will probably be observed. In this case, if we keep X constant, how do we know if Y and Z are really correlated? In the case of network reconstruction a link is called indirect when a correlation between a pair of vertices is only the result of correlations from other pairs of vertices. In a graphical model, the best situation is to look at a network consisting of only direct links. This allows researchers to actually observe how nodes really interact with one another.

Partial Correlation is a common way to measure conditional independence of each pair of variables in a model. The main idea is to remove the effect of the remaining variables from the considered pair. The original definition is based on linear regression as follows:

Definition 2.1.1 (Partial Correlation)**.** Let $\mathbf{X} = \{X_1, \cdots, X_p\}$ be a random vector and suppose we want to find if X_i is really correlated to X_j. Denote by $\mathbf{Y}$ all variables in $\mathbf{X}$ but X_i and X_j, that is, $\mathbf{Y} = \mathbf{X} \backslash \{X_i, X_j\}$.

The effect of $\mathbf{Y}$ on X_i and on X_j is removed by assessing the correlations between the residues of the projections of X_i and X_j on the linear space generated by $\mathbf{Y}$, that is, subtracting part of the linear relation due to $\mathbf{Y}$. Mathematically:

$$X_i = \alpha_i + \mathbf{Y}\boldsymbol{\beta}_i + \epsilon_i$$
$$X_j = \alpha_j + \mathbf{Y}\boldsymbol{\beta}_j + \epsilon_j,$$

where $\mathbf{Y}$ is a horizontal vector of size $p-2$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \cdots, \beta_{i(p-2)})^T$ are the coefficients of the linear regression. Through the regression method of least squares, we compute $\hat{X}_i$ and $\hat{X}_j$, which are the estimators of X_i and X_j respectively. The residues are given by:

$$Res_i = X_i - \hat{X}_i$$
$$Res_j = X_j - \hat{X}_j.$$

Note that, Res_i and Res_j are orthogonal to $\boldsymbol{Y}$ and therefore the correlation $\rho(Res_i, Res_j)$ represent the correlation between the components of X_i and X_j that don't show linear dependency with $\boldsymbol{Y}$. Therefore, the partial correlation is given by

$$\rho_{ij.Y} = \rho(Res_i, Res_j) = \rho(X_i - \hat{X}_i, X_j - \hat{X}_j),$$

where $\rho(Res_i, Res_j)$ is the correlation between variables Res_i and Res_j.

Here, $\hat{X}_i = E(X_i) + \Sigma_{X_iY}\Sigma_{YY}^{-1}(\mathbf{Y} - E(\mathbf{Y}))$ is the projection of X_i, that is, the conditional expectation of X_i given $\mathbf{Y}$, where Σ_{X_iY} is the covariance matrix of X_i with $\mathbf{Y}$ and Σ_{YY} is the covariance matrix of $\mathbf{Y}$. The conception of $\hat{X}_j$ is analogous to $\hat{X}_i$

The interpretation of $\rho_{ij.Y}$ is the following: $\rho_{ij.Y} = 0 \Rightarrow X_i$ and X_j are not correlated when the effect of $\mathbf{Y}$ on X_i and X_j is removed. On GGM, the latter implication runs both ways, that is, the statements are necessary and sufficient to one another.

Another approach that uses matrices to calculate all partial correlations between all pairs of random variables X_i and X_j, given all remaining variables, is called Inverse Method.

Definition 2.1.2 (Inverse Method)**.** Let R be the correlation matrix, where $\rho_{ij} = \rho(X_i, X_j)$ are the elements of R. If R is invertible, define $Q = R^{-1}$. The partial correlation is then defined by:

$$P = -\text{scale}(R^{-1}), \tag{2.2}$$

where $\text{scale}(Q) = D_Q^{-\frac{1}{2}} Q D_Q^{-\frac{1}{2}}$ in which D_Q is a diagonal matrix with $d_{ii} = q_{ii}$. Each entry of P will be given by

$$\rho_{ij.Y} = -\frac{q_{ij}}{\sqrt{q_{ii}q_{jj}}}, \tag{2.3}$$

where q_{ij} are the elements of Q.

Note that, if the number of observations n is lower than the number of random variables, then R will be singular and we cannot calculate P. Options to solve this problem can be found in Chapter 3.

The demonstration of the equivalence of the inverse method with the linear regression definition can be found in [Tho12].

2.1.3 GeneNetWeaver

Gene Net Weaver (GNW) is a a popular software among biologists which generates data based on real biological systems that have been extensively investigated. It is used in the DREAM Project which aims to challenge researchers to infer simulated and *in-vivo* gene regulation networks. The DREAM Challenges are crowdsourcing challenges examining questions in biology and medicine. They are a non-profit, collaborative community effort and are created and managed by experts in systems biology, statistics, and challenge design so that the results will be solvable and reproducible in a meaningful way.

GNW generates network structures as subnetworks (modules) from known network systems: E.coli [GCJJPG+08] and Yeast S.cerevicie [BBI+06]. For a chosen structure the expression data is generated using ordinary differential equations (ODEs) adding a molecular noise into the dynamics or adding experimental noise observed in microarrays [MSMF09, MPS+10]. Stochastic kinetics (SDEs) prevents a gene that suffered knockdown to suddenly reach a very high transcription rate caused by noise. Therefore this approach allows to perform perturbations on one or more genes.

Both transcription and translation are modeled using a standard thermodynamic approach [AJS82] allowing both independent regulatory interactions ("additive") and synergistic ('multiplier'). For each gene i in a network, the rate of change of the concentration of mRNA, f_i^{RNA}, and the rate of change of protein concentration, f_i^{Prot} , are described by

$$F_i^{RNA} = \frac{dx_i}{dt} = m_i \cdot f_i(\mathbf{y}) - \lambda_i^{RNA} \cdot x_i \tag{2.4}$$

$$F_i^{Prot} = \frac{dy_i}{dt} = r_i \cdot x_i - \lambda_i^{Prot} \cdot y_i \tag{2.5}$$

The integration of the equations defined by (2.4) and (2.5) results in mRNA levels and noise-free protein concentration, respectively $x_i(t)$ and $y_i(t)$ to the gene i.

2.2 Changes in Biological Networks

Biological regulatory mechanisms are a good example of complex systems as already defined. Hence, investigation of state transition is herein performed using gene expression data acquired through mircroarray measurements. This data is organized in matrices where each column contains the expression intensities of different genes in a cell under certain experimental conditions. Data for different conditions are kept in different matrices which contain tremendous amounts of information about the complex interactions between genes, and ultimately characterize a cell's behavior.

Biological Networks are seldom built using correlation coefficients to define the set of edges (pais of interacting genes). When a state transition occurs the gene expression network goes through several types of alterations that affects the whole system. Two types of changes in Biological Networks have already been already researched: Differentially expressed genes and Differentially correlated pairs.

2.2.1 Differentially expressed genes (DEGs)

Genes that have their gene expression mean changed when transitioning between two different states are called differentially expressed genes and can be viewed as transition indicators. The DEGs are usually found through the t-Student test (with random variance model). Figure 2.2(a) shows an example of DEGs, with a gene expression mean increase from state 1 to state 2.

2.2.2 Differentially correlated pairs (DCPs)

As already described above, DCPs are pairs of genes that suffered correlation change between two states, Figure 2.2(b). They can either gain or lose correlation and also change correlation di-

rection (sign). Therefore the identification of DCPs can be done in two parts:

1. A pair of genes can only be a DCP if it is correlated in at least one state, so, for each pair of genes A and B, the following filters will be used:

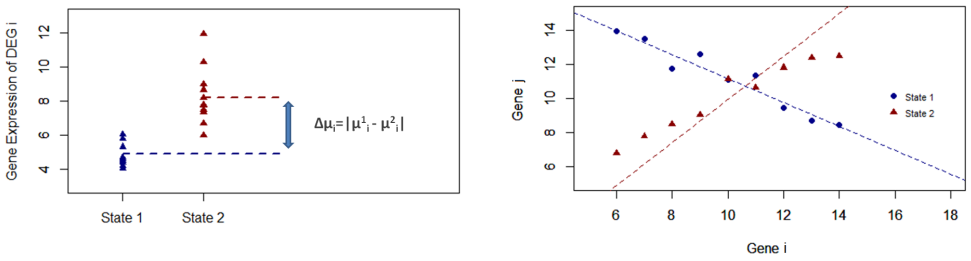
$$\begin{array}{ll} \rho_{ij}^{(1)} \neq 0 \text{ e } \rho_{ij}^{(2)} = 0 & \text{correlation loss} \\ \rho_{ij}^{(1)} = 0 \text{ e } \rho_{ij}^{(2)} \neq 0 & \text{correlation gain} \\ \rho_{ij}^{(1)} > 0 \text{ e } \rho_{ij}^{(2)} < 0 & \text{change of sign} \\ \rho_{ij}^{(1)} < 0 \text{ e } \rho_{ij}^{(2)} > 0 & \text{change of sign} \end{array}$$

where ρ_{ij}^{k} is the correlation of the pair (i, j) in the state k

2. Then, for each pair of genes i and j that passed the filter in 1., we will statistically test the difference of Pearson correlation between states 1 and 2 using the following null hypothesis:

$$H_0 : \rho_{ij}^{(1)} - \rho_{ij}^{(2)} = 0$$

The Fisher z-transformation will be used to test the pairwise differences of correlations between two states. More details about this calculation can be found in [SKV+11].



(a) Example of a DEG.

(b) Example of a DCP.

Figure 2.2: *(a) Graph showing the change of gene expression mean of gene i when the system goes from state 1 to 2. $\Delta\mu_i$ is the difference of means of gene i between these two states. (b) Graph representing a possible change in the correlation of gene expression of any two genes i and j from state 1 to 2.*

Chapter 3

Local Partial Correlation - LPC

Conditional independence is essential to build networks where edges represent direct regulatory relations [MKB79, Edw95], as already mentioned in Section 2.1.2. However, its calculation can be very complex. In [BSS04] it is shown that conditional independence can be replaced by partial correlation 0 when the set of variables follows a multivariate normal distribution. You can find the definition of partial correlation and a few analyses thereof, in Section 2.1.2. Just as a reminder, the main idea of partial correlation is to remove the effect of the remaining variables on the model on each pair of variables through MLE. Nonetheless, it is common knowledge that the existence of the MLE is not guaranteed in general for high-dimensional data [Buh93].

Local partial correlation is an adaptation of the usual partial correlation method that serves as an alternative to enable the calculation of partial correlation when dealing with more variables than samples. The new method, developed by the author [Tho12] applies the inverse method (described in section 2.1.2) considering only the neighborhood of each pair of variables separately.

3.1 Definitions

3.1.1 Partial Correlation

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^T$ be a vector of random variables following a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ_p, that is, $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. In GGM, each node (or vertex) $i \in \mathcal{V} = \{1, 2, \cdots, p\}$ represents a random variable $X_i \in \boldsymbol{X}$. Let $\Omega = \Sigma^{-1}$ be the precision matrix and $\mathbf{Y} = \mathbf{X} \backslash \{X_i, X_j\}$ be the remaining variables in the model apart from X_i and X_j. The partial correlation coefficient for (X_i, X_j) given $\mathbf{Y}$ is as follows:

$$\rho_{ij \cdot \boldsymbol{Y}} = -\frac{w_{ij}}{\sqrt{w_{ii} w_{jj}}},$$

where $w_{ij}, i, j = 1, \cdots, p$ are the elements of Ω. Note that

$$\rho_{ij \cdot \boldsymbol{Y}} = 0 \Leftrightarrow w_{ij} = 0.$$

Therefore, the elements of the precision matrix also indicate the existence of an edge between two vertices.

3.1.2 Adjacency Matrix

The network structure can be represented by the adjacency matrix $A = (a_{ij})$, where

$$a_{ij} = \begin{cases} 1, & \text{if } \rho_{ij \cdot \boldsymbol{Y}} \neq 0 \\ 0, & \text{if } \rho_{ij \cdot \boldsymbol{Y}} = 0 \end{cases}$$

and $a_{ij} = 1$ determines the existence of an edge between i and j.

In this project, we are interested in network reconstruction methods that are able to predict the structure closest to the original one as possible based on observations, rather then re-estimating the partial correlation coefficients. Therefore our goal is to estimate the adjacency matrix A.

3.1.3 Neighborhood

Neighborhood of vertex i is composed by the vertices connected to i, that is, $ne(i) = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\} = \{j \in \mathcal{V} : a_{ij} = 1\}$ and the neighborhood of the pair of vertices (i,j) as $ne(ij) = ne(i) \cup ne(j)$. Neighborhood selection is a problem that has been explored. For example, in [MB06] the authors select the neighborhood by solving a regression problem with Lasso penalty.

3.1.4 Local Partial Correlation

The starting point of this method is neighborhood selection, which is done through Pearson correlation coefficient. For each pair of variables (X_i, X_j) the following new hypothesis is tested: $H_0 : r_{ij} = 0$. The edges of a reconstructed network are pre-determined according to a fixed level of significance α_1. Consider r_{ij} the sample correlation. The estimated adjacency matrix is then given by $\hat{A}_{\alpha_1} = (\hat{a}_{ij}^{(\alpha_1)})$ where

$$\hat{a}_{ij}^{(\alpha_1)} = \begin{cases} 1, & \text{if p-value of } r_{ij} < \alpha_1 \\ 0, & \text{if p-value of } r_{ij} > \alpha_1 \end{cases}$$

The estimated neighborhood of X_i given α_1 is then given by $\hat{ne}_{\alpha_1}(i) = \{j \in \mathcal{V} : \hat{a}_{ij}^{(\alpha_1)} = 1\}$. The estimated neighborhood of (i,j) is then defined by $\hat{ne}_{\alpha_1}(i,j) = \hat{ne}_{\alpha_1}(i) \cup \hat{ne}_{\alpha_1}(j)$.

For each pair of variables (X_i, X_j), the partial correlation coefficient is calculated through the inverse of the local covariance matrix $S_{i,j}$, which is a sub-matrix (partition) of the sample covariance matrix S with rows and columns representing only the nodes in $\hat{ne}_{\alpha_1}(i,j)$.

We can see in Figure 3.1 an example of a correlation network with nine nodes, built based on α_1. In this case we want to calculate the local partial correlation between X_2 and X_5. Note that $\hat{N}_{\alpha_1}(X_2) = \{X_3, X_5, X_8, X_9\}$ and $\hat{N}_{\alpha_1}(X_5) = \{X_2, X_6, X_9\}$. Therefore, $\hat{N}_{\alpha_1}(X_2, X_5) = \{X_2, X_3, X_5, X_6, X_8, X_9\}$. Furthermore, X_1, X_4 and X_7 are left out of the local covariance matrix $S_{2,5}$ and the inverse method considers all correlations regarding the variables in $\hat{N}_{\alpha_1}(X_2, X_5)$. In [Tho12], it was numerically shown that $\exists\ n_0 : \forall n > n_0,\ |ne(i,j)| < \frac{n}{2}$. In cases where there are still more neighbors than samples we select the $\frac{n}{2}$ neighbors with lower p-values, which in most cases, mean higher LPC coefficients.

The next step is to build the adjacency matrix $\hat{A}_{\alpha_2} = (\hat{a}_{ij}^{(\alpha_2)})$ considering the level of significance α_2 for local partial correlation, that is,

$$\hat{a}_{ij}^{(\alpha_2)} = \begin{cases} 1, & \text{if p-value of } \hat{\rho}_{ij \cdot \mathbf{Y}} < \alpha_2 \\ 0, & \text{if p-value of } \hat{\rho}_{ij \cdot \mathbf{Y}} > \alpha_2 \text{ or } \hat{a}_{ij}^{(\alpha_1)} = 0 \end{cases}$$

The hypothesis test for LPC is done using Fisher's z-transform of the partial correlation:

$$z(\hat{\rho}_{ij \cdot \mathbf{Y}}) = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{ij \cdot \mathbf{Y}}}{1 - \hat{\rho}_{ij \cdot \mathbf{Y}}} \right).$$

The null hypothesis is $H_0 : \rho_{ij \cdot \mathbf{Y}} = 0$, to be tested against the alternative hypothesis $H_A : \rho_{ij \cdot \mathbf{Y}} \neq 0$. We reject H_0 with significance level α_2 if:

$$\sqrt{n - |\hat{ne}_{\alpha_1}(i,j)| - 3} \cdot |z(\hat{\rho}_{ij \cdot \mathbf{Y}})| > \Phi^{-1}(1 - \alpha_2/2),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a Gaussian distribution with zero mean and unit standard deviation, and n is the sample size. The distribution of the sample partial correlation

was described by Fisher, see [Fis24].

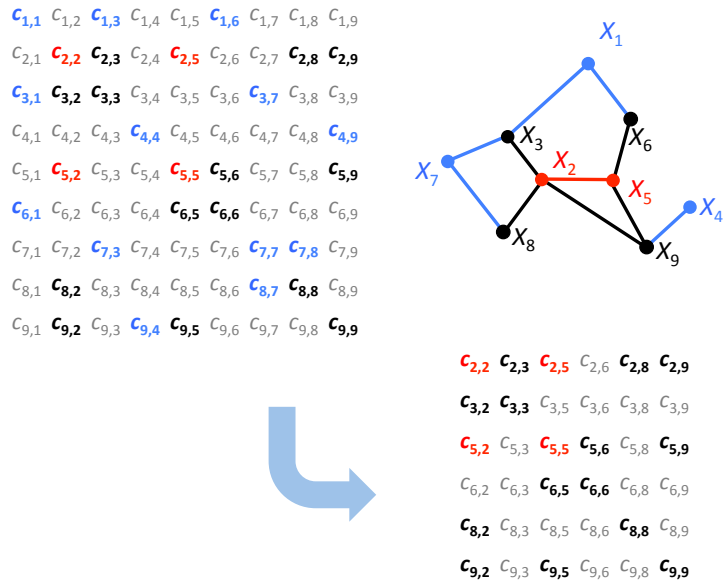


Figure 3.1: *Local partial correlation scheme: we calculate the LPC for pair X_2 , X_5, (red nodes/edge). The neighborhood of this pair is the set of nodes X_3, X_6, X_8, X_9 (black nodes/edges). X_1, X_4, X_7 (blue nodes) are significantly correlated with the black nodes (blue edges), but not with the red nodes. Thus the inverse method is applied exclusively to the correlation sub-matrix formed only by the genes X_2, X_5, X_3, X_6, X_8, X_9. In correlation matrices the gray entries are statistically non-significant empirical correlations.*

3.2 Simulation Studies

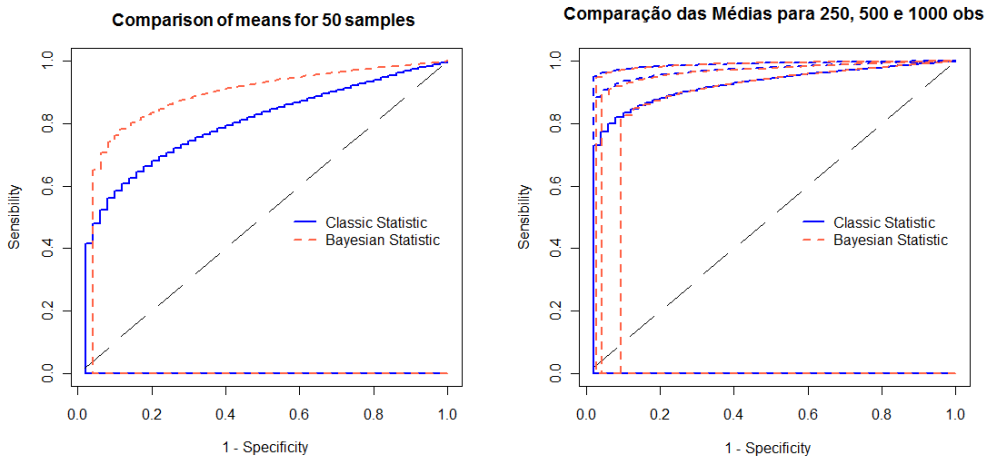
In my masters dissertation, [Tho12], we simulated networks with 100 variables through 2 different network generation methods. The first one is a tree with offspring distribution following a Lognormal(1, 1). That allows the degree distribution to have heavy tail, which means that only a few nodes will have high degree. The second one is a mixed network, that is, the offspring distribution follows a Lognormal(1, 1) and the parents distribution follows a Geometric(0.6). The offspring distribution adopted in this work leads the degree distribution to produce heavy tail and the nodes to have more than one parent but not too many parents.

Two approaches were adopted: classical and Bayesian (local) partial correlation. Classical partial correlation is exactly what we have already defined previously. In contrast, Bayesian approach uses partial covariance (see [Edw95]) to estimate partial correlation. In my masters dissertation we considered that the data followed a multivariate normal distribution, where, a priori, its covariance followed an inverse Wishart distribution while its mean given the covariance followed a normal distribution. Note that this is a conjugate priori, therefore, a posteriori, the distributions are the same but the parameters are updated based on the data (see [DS02]). Another important observation is that, instead of p-values, we used the evidence value, also referred to by e-value, from FBST (Full Bayesian Statistics Test - see [PS99]) in order to statistically test the (local) partial correlations.

When investigating through classical approach we simulated 500 networks and assessed ROC curves for 50, 250, 500 and 1000 samples. Due to high computational costs, for Bayesian approach in mixed networks we simulated 50 networks and performed ROC curve analysis considering 50, 250 and 500 samples. Just as a remark, local partial correlation is only necessary when there are more variables than samples - in this case, it was applied only to simulations with 50 samples. In order to compare the ROC curves, for each group of networks - defined by sample size and generation method - we calculated the mean of sensibility and of specificity. You can see on Figures 3.2 and 3.3 that for most of the time, either for trees or mixed networks, the mean ROC curves overlap

showing similar results regarding classical and Bayesian statistics and the number of samples. The only exception is when there are less samples (50) than variables (100). Figure 3.2(a) shows that the Bayesian approach is able to predict the original networks slightly better than the Classical approach.

Due to high computational costs and not observing a big difference in the mixed network ROC curves when $n << p$, we have decided to work with local partial correlation in the Classical approach throughout this project.



(a) (b)

Figure 3.2: *Comparison between ROC curves means with classical and Bayesian statistics for trees built from (a) 50 samples and (b) 250, 500 and 1000 samples*

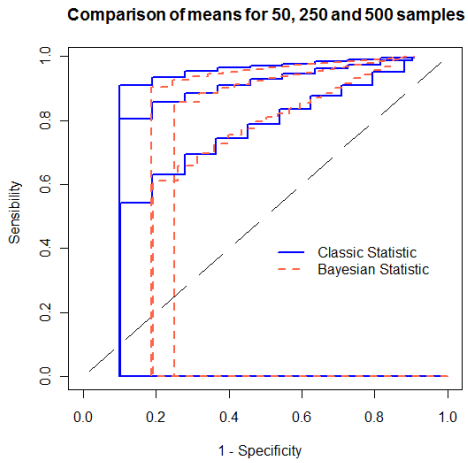


Figure 3.3: *Comparison between ROC curves means with classical and Bayesian statistics for mixed networks built from 50, 250 and 500 samples*

Chapter 4

Biological Co-expression Networks

4.1 Introduction

In saying that we understand a biological process, we usually mean that we are able to predict future events and manipulate the process into a desired direction. Thus, biological inquiry could be viewed as an attempt to understand how a biological system transits from one state to another. Such transitions underlie a wide range of biological phenomena from cell differentiation to recovery from disease. In attempting to understand these transitions, a simple and frequently used approach is to compare two states of a system (eg, before and after stimulus, with and without mutation, or healthy and diseased). Although more sophisticated approaches with timeseries data, dose-effect data, or three or more sample groups can be also used, here we discuss analysis of data from a two-class study design. Furthermore, most of the methods that we describe can, with slight modifications, be used for other study designs. Today, omics technologies enable unbiased investigation of biological systems through massively parallel sequence acquisition or molecular measurements, bringing the life sciences into the era of Big Data. A central challenge posed by such omics datasets is how to navigate through the haystack of measurements (eg, differential expression between two states) to identify the needles comprised of the critical causal factors.

Network analysis is a powerful and general approach to this problem, in which the biological system is modeled as a network whose nodes represent dynamical units (eg, genes, proteins, metabolites, etc) and edges stand for links between them. Network analysis consists of two fundamental stages: network reconstruction and network interrogation. For omics molecular measurements such as gene expression, a particular type of network analysis called co-variation network analysis has become a dominant approach. In such networks, a node represents the expression of the gene being measured, and an edge indicates that the expressions of two genes are correlated. Multiple groups including ours have been successfully using such methods to gain a systems-level understanding of biological processes and to reveal mechanisms of different diseases [AGC+09, SYC+11, YSH+13]. Several recent discoveries ranging from genes that drive progression of different cancers [MSY+13, CAT+14] to microbes and microbial genes that cause a human illness [MDD+15] became possible because of the predictive power of network analysis. In particular, such insights would be very difficult to achieve if analysis is limited to finding differentially expressed genes and follow-up data mining of those genes. Due to the rapid pace of evolution of techniques and omics technologies, the practical application of network analysis has usually required a dedicated computational biologist or statistician. This requirement has limited the extent to which the larger biological sciences community has benefited from network analysis. Here we provide an overview of co-variation network reconstruction and interrogation, including a step-by-step guide on how to perform and use network analysis to investigate a biological question (Fig. 4.1). In this guide, we include the software packages that we employ (and specific pointers to the methods or software used by other groups) for each of the steps of a model network analysis work flow.

In general, the types of omics measurements that are amenable to network analysis include microarrays, next-generation sequencing (for genotyping, transcriptome profiling, or microbiome

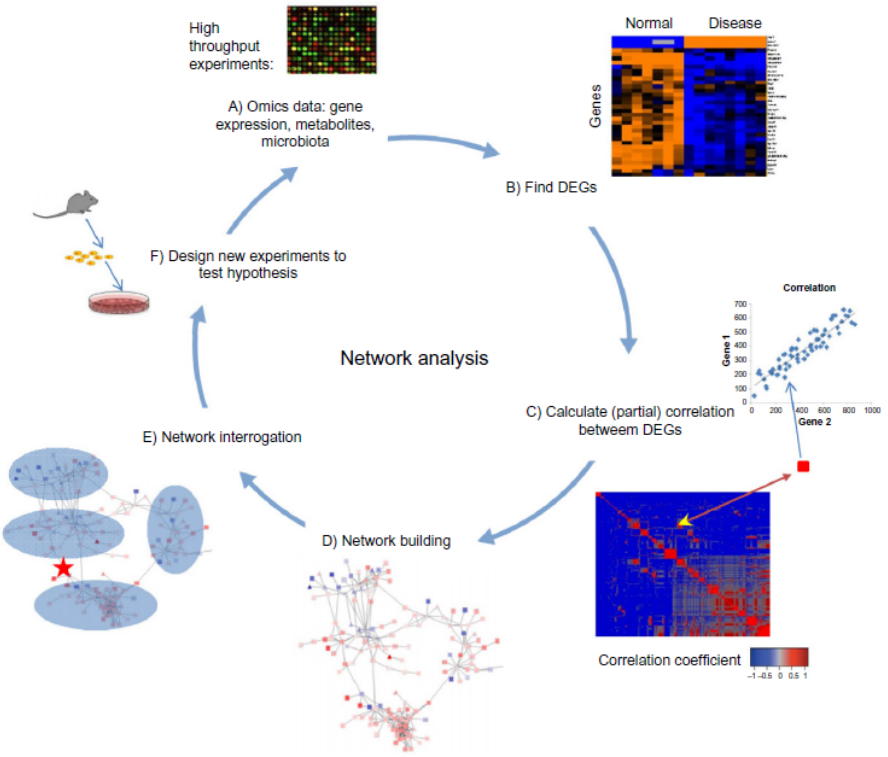


Figure 4.1: *Work flow of network analysis. (A) Network analysis starts from data obtained from high-throughput experiments such as microarray experiments detecting expression of genes in samples. (B) Differentially expressed genes are found between two states of a system (eg, normal vs disease). (C) Correlations of DEGs based on their expression values are calculated to detect regulatory relationship among them. (D) Significant correlations suggest connections between differentially expressed genes (DEGs) and are used to generate a network of DEGs. (E) Network interrogation is performed to detect modules, key regulators, and functional pathways that are important for state transitions. (F) Based on the findings from network interrogation, new hypotheses are generated, which can be tested in newly designed experiments. Data from new experiments could also be subject to further analysis.*

analysis), and mass spectrometry-based proteomics and metabolomics data. In this guide, we use gene expression data to illustrate the process of network reconstruction and interrogation.

4.2 Network reconstruction

The first stage of network analysis is network reconstruction, which is the data-driven discovery or inference of the entities/nodes (transcripts, proteins, genes, metabolites, or microbes) and relationships or edges between these entities that together constitute the biological network. Here, we describe the steps involved in network reconstruction starting from entity abundance or frequency data.

4.2.1 Normalization (data preprocessing)

Customarily, abundance data are normalized in order to correct for sample-to-sample variation in the overall distribution of abundance values (or more generally, to normalize specific quantities that depend on the distribution). Measurements of gene expression levels (as well as other types of omics data) can be affected by a variety of non-biological factors including unequal amount of starting RNA, different extents of labeling, or different efficiencies of detection between samples. Before normalization, data are often log_2-transformed in order to stabilize variances when measurements span orders of magnitude. Frequently used normalization schemes include median normalization,

quantile normalization, LOWESS normalization [BHJ+04] for RNA microarray data, reads per kilobase per million mapped reads (RPKM), [MWM+08] and trimmed mean of M-values [RO10] for RNA-seq data. In practice, we use normalization procedures available in the software package BRB Array Tools [SLL+07] for normalization of microarray data (Table 1). In addition, most normalization procedures are available as software packages in the Bioconductor toolkit [GCB+04]. Systematic evaluations of transcriptome normalization methods have been reported for both microarrays [LWLC07] and RNA-seq [DRA+13]; however, evaluations using large numbers of sample groups are needed in order to determine which normalization method is most appropriate for covariance network inference. Selection of an appropriate normalization method is clearly important, given that selection of a suboptimal normalization scheme can lead to overestimation of gene-gene correlation coefficients [LWLC07]. Beyond transcriptome profiling, different omics data types may benefit from different types of normalization. For example, new methods have been proposed for normalization of metabolomics [JBN+14] and microbiome [MH14] data. Although there is no consensus about the best methods for many types of data, in the experience of the authors, [MSY+13, BAS+04, PDMS07, SMH+11, SYGP+05, SKV+11] simple methods such as quantile, LOWESS, or even median normalization perform reasonably well for class comparison and correlation if there are no major biases in the data such as batch effects.

4.2.2 Discovery of differentially expressed genes (selecting nodes)

A crucial step in network reconstruction is the identification of the relevant subset of variables/-genes that will constitute the nodes in the network; for a transcriptome profiling study, these would be genes for which there is significant differential expression between the sample groups. A variety of statistical tests are commonly used for the identification of differentially expressed genes (DEGs), including Welch's t-test, moderated t-test, and permutation tests. For parametric tests, accurate estimation of intra-sample-group variance is a critical issue; two improved variance estimation techniques are the locally pooled error [JTB+03] and empirical Bayes methods [SMS05]. To find DEGs, we usually use the t-test with the ordered set of p-values converted to cumulative false discovery rate (FDR) estimates, for which a typical cutoff would be 10%. Both statistical functions are implemented in BRB Array Tools [ZS08]. During the last two decades, multiple statistical approaches have been proposed for differential expression testing [Pan02]. Overall, they provide similar results with small differences [Pan02]. Thus, careful study design (rather than trash in, trash out) and the use of meta-analysis techniques to integrate multiple datasets are likely to be more important for reliable DEG discovery than a choice of one or another statistical test. Because omics data analysis typically involves tens of thousands of statistical tests, the correction for multiple hypotheses is essential [DSB03].

4.2.3 Correlation analysis for network reconstruction (finding links between nodes)

The central biological principles underlying correlation network analysis are 1) that DEGs reflect functional changes, and 2) that DEGs do not work individually but interact (eg, at the protein or pathway level) to functionally alter the biological system. In gene expression networks, nodes represent genes and edges represent significant pairwise associations between gene expression profiles. The central mathematical/statistical principle that allows us to use correlation networks for analysis of biological systems is that the correlation between two variables, if statistically significant, is always a result of causation. Specifically, correlation results from regulatory relations between the two variables, or from a common causal regulator to the two variables, or both, as in the case of a feed-forward loop [Pea01]. To reconstruct the network, the Pearson or Spearman correlation coefficient can be used to obtain an association (similarity) measure for each possible pair of DEGs, with a cutoff for statistical significance (for example an FDR cutoff of 10% for the $\frac{p(p-1)}{2}$ possible pairwise associations tested) and for a minimum correlation level. Together with the nodes, the edges whose similarity measures exceed this cutoff constitute a network. In practice, normalized expres-

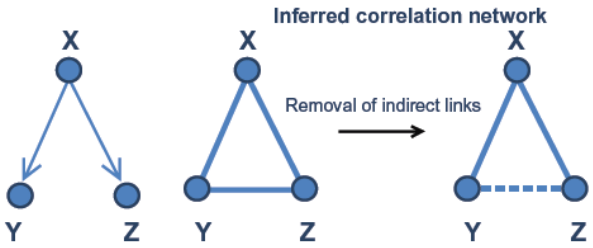


Figure 4.2: *Removal of indirect links. As a demonstration, gene X can regulate the expression of both gene Y and Z. But there is no direct regulatory relationship between gene Y and Z. From the calculation of correlation of expression levels of three genes, correlations between gene X and Y, Z are observed as expected. However, genes Y and Z are also significantly correlated since they are both directly regulated by gene X. This correlation from common cause is called indirect link and can be removed by techniques, such as partial correlation, generating a network reflecting regulatory relationships.*

sion data for DEGs are retrieved and pairwise correlations along with the corresponding p-values are calculated for each class (biological state) separately using the R statistical analysis software, with the function cor.test; FDR is calculated using the function p.adjust(<pvalue vector>, method = "fdr"). Several other software programs that can be used for calculating gene-gene associations (correlations, mutual information and others) are listed in Table 1. Note that correlations should be calculated within a group of samples that belong to one class/biological state (pooling samples from different states/classes to compute the correlation coefficient leads to significant bias).

4.2.4 Discriminating between direct and indirect links

Co-variation gene networks in general consist of connections that result from a combination of direct and indirect effects between genes. For example, if a gene Y strongly depends on gene X and gene Z also depends on X, it is likely that a high association (eg, correlation) will exist between Y and Z even if there is no direct dependence between them (Fig. 4.2). Moreover, even if a true dependence exists between a pair of genes/nodes, its strength estimation can be biased by additional indirect relationships [Pea01]. For this reason, correlation networks in general have many edges that reflect indirect relationships between pairs of genes, where no direct relationship exists. Finding direct relationships between genes is important when one attempts to identify causal gene regulators of a given biological process.

Mathematically, direct effects can be defined as the association between two genes, holding the remaining genes constant [Pea10]. An effect that is not direct is called an indirect effect. The identification of direct links is an important goal of network reverse engineering.

To infer direct links between DEGs, we have been using the partial correlation coefficient [DLFBHM04, MCK+12]. To calculate partial correlations, we use a method called the *inverse method* [Whi09]. Its implementation is straightforward in R using the function cor2pcor from the package corpcor. The detailed algorithm is described in Apendix B.2.1. After calculation of partial correlation, the network can be built using links with absolute value of the partial correlation larger than a user-defined threshold or p-value of partial correlation higher than a specific α.

Several other methods have been proposed to discriminate between direct and indirect links in co-variation networks [MNB+06, FHT08, JMC13, BB13, FMMK13, HS14]. For example, a variant of the partial correlation, which we call the *local partial correlation*, can be used in order to overcome the limitations of other methods.[TFY12]

4.2.5 Proportion of unexpected correlations (improvement of reconstruction and error evaluation)

A fundamental problem of the standard correlation network approach is that practical limitations in the numbers of sample measurements can lead to an unacceptably high error rate. Recently,

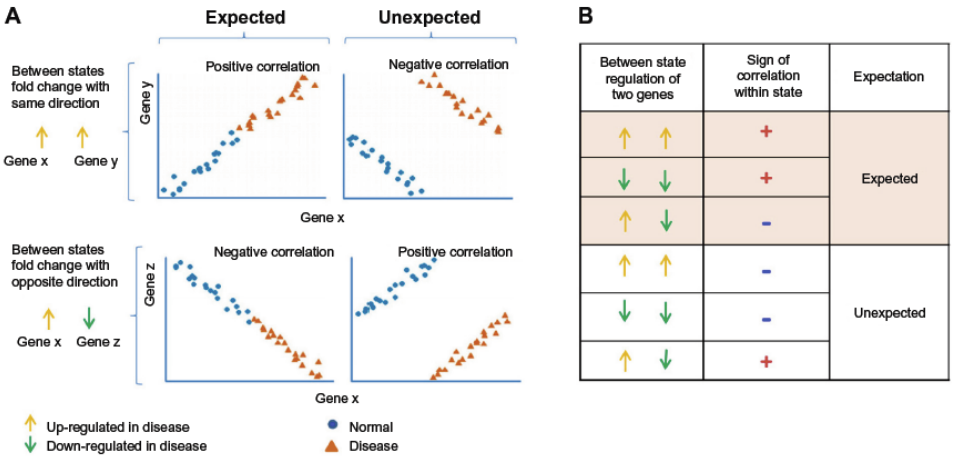


Figure 4.3: *Illustration of expected and unexpected correlations. (A) When expression of two genes (gene x and gene y) are regulated toward the same direction when comparing two states, eg, both up-regulated in disease (upper two panels), we should expect their expression levels to be positively correlated within each state if there exists regulatory relationship between gene x and gene y. When two genes are oppositely regulated when transiting from normal to disease (in the lower two panels, gene x is up-regulated while gene z is down regulated), we should expect negative correlation between those two genes in each state. (B) Different combinations of between states and sign of correlations used to define expected or unexpected correlation.*

our group has proposed a method called *proportion of unexpected correlations* (PUC), which allows identifying and removing approximately half of false positive edges from a co-variation network with no reduction in statistical power, see [YPK+16]. The method takes into account a relation between the direction of regulation of two DEGs and the sign of correlation between the two genes. Thus, two up and two down-regulated genes must correlate positively; and a pair of oppositely regulated genes (one up-regulated and one down-regulated) should have negative correlation, that is, $\Delta_X \Delta_Y \rho(X, Y) \geq 0$, where Δ_X and Δ_Y are the variation of gene expression mean of genes X and Y between two states. Any deviation from this rule represents unexpected/erroneous edges and is removed from the network (Fig. 4.3). The proportion of these unexpected edges provides an error estimate for the whole network. For network reconstruction, each edge in a network can be evaluated and removed if it is unexpected.

4.2.6 Meta-analysis (improvement of reconstruction and error evaluation)

In omics-based network reconstruction, because of the large number of genes or variables measured (up to tens of thousands) and the limited number of samples (typically tens or hundreds), it is critical to assess the reproducibility of results. Although widely used methods (eg, FDR [RYB03]) enable accounting for multiple hypothesis tests, the discrepancy between the number of samples and variables inherent to omics datasets limits the sensitivity and specificity for detecting edges through network reconstruction.

In order to overcome this problem and to augment the statistical significance for the nodes and links in a network, meta-analysis can be employed. This statistical approach combines results from different studies in order to achieve reproducibility.

The studies can be obtained from standardized omics data repositories. Good examples of such repositories are the Gene Expression Omnibus (GEO) [EDL02] and Array Express [BPS+03] (for transcriptomics and epigenomics datasets); PRIDE [MHJ+05] (for proteomics datasets), the Human Metabolome Database [WTK+07] (for metabolomics datasets), and lipid MAPS [FSM+09] (for lipidomics datasets). Additionally, molecular interaction data from the BioGRID [SBR+06] or BioCyc databases [CFF+08] can be used as a prior for edge reconstruction.

In meta-analysis of multiple datasets - whether from publicly available datasets or experiments produced in the same lab - the strategy is usually the same. The datasets to be co-analyzed in a

meta-analysis should be selected on the basis of their congruence with the central biological question of interest, and they should pass some predefined sample size and quality requirements (eg, number of measured/ detected genes). After choosing the datasets, as a first step for meta-analysis we apply two filters:

1. The same sign of statistic (mean, covariance, or correlation) throughout all datasets (ie, if gene A is up-regulated in case over control in dataset 1, it should have the same direction of regulation in all other datasets to pass the filter);

2. P-value thresholds across all datasets. These filters provide consistency and control for heterogeneity across datasets for a given gene (or gene pair in case of correlation). The next step is an actual statistical evaluation. In this step, meta-analysis combines common statistical measures, such as p-values, and calculate a *weighted average* for such measures. As a *weighted average*, we frequently use the Fisher's p-value calculation. Let $p_1, \cdots, p_k$ be the p-values of one measure into k datasets (studies). For example, p_i can be the t-Student test p value for gene A to be differentially expressed in study i for all $i = 1, \cdots, k$. Then the Fisher's p-value p_{Fisher} summarizes all these p-values $p_1, \cdots, p_k$ into one average p-value by the formula

$$p_{Fisher} = P\left(\chi^2_{2k} \geq -2\sum_{i=1}^{k} ln(p_i)\right)$$

   where χ^2_{2k} is a random variable with chi-square distribution with $2k$ degrees of freedom. After calculating Fisher's p-values for all genes, the standard FDR procedure can be used to adjust for multiple hypothesis testing. Several other approaches have been proposed for meta-analysis of gene expression data (Table 1) [RYS+04, HRR+05]. In Appendix B.2.2 we describe in more detail the algorithm that we have employed for integrating differential expression, correlations, and differential associations/correlations[MSY+13].

4.2.7 Differentially co-expressed gene pairs (evaluating network changes)

The networks discussed above model static correlations between genes that change their expression when the biological system transits from one state to another. However, the sets of edges within a gene co-variation network can themselves vary from state to state, for example, when two genes are highly correlated in a subset of conditions but not across all conditions [SS+05a]. Such a gene pair is called a *differentially co-expressed* gene pair (Fig. 4.4). It has been shown that differentially co-expressed gene pairs frequently play critical roles in pathogenesis. Several studies have explored gene co-expression changes in cancer, revealing known cancer genes that were top-ranked among co-expression changes but not necessary (separately) among differentially expressed genes [SKV+11, KS04].

In order to search for differentially co-expressed gene pairs, our group adapted a simple approach called differentially correlated pairs (DCPs) [SKV+11]. The DCPs algorithm is described in Appendix B.3. In addition to DCPs, multiple methods/ software have been developed to find the changing edges in gene expression networks (Table 1) [SKV+11, Wat06].

4.3 Conclusion

In this chapter, we have described how network analysis can help us to answer different questions commonly asked in biological research. We have also provided a detailed methodology for this analysis, including approaches employed by our group as well as frequently used by the network-biology community.

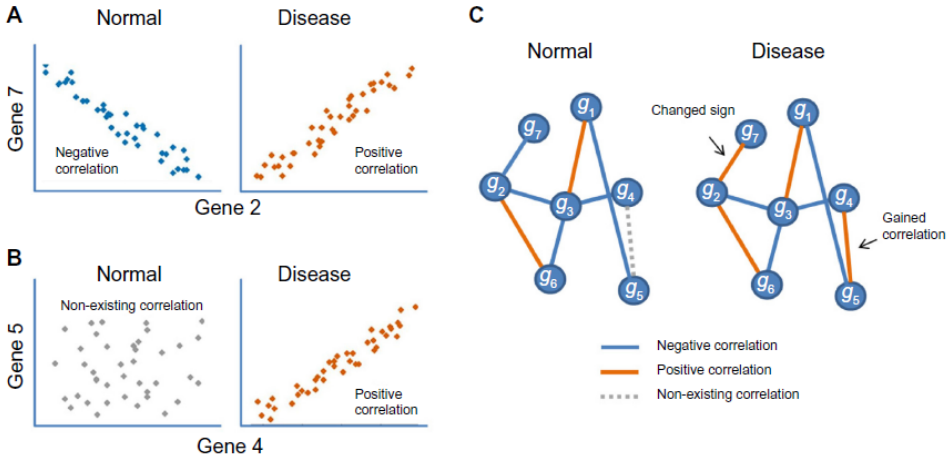


Figure 4.4: *(A) Gene 2 and gene 7 correlate with each other in both normal and disease conditions, but the signs of the correlation coefficient are opposite. (B) In normal condition, there is no correlation between gene 4 and gene 5, but they gain positive correlation when the biological system transitioned to disease. (C) Example of visualization of a network transitioning between normal and disease conditions. Red lines represent positive correlation, blue line represent negative correlation, and dotted gray lines represent non-existing correlations in one condition that strongly appear in the other condition (on this case, becomes positively correlated).*

Chapter 5

Differential Correlation Analysis

5.1 Introduction

Recent technological advances have moved the focus of biologists from how to measure biological parameters to how to analyze and interpret tens of thousands of measurements, frequently called omics data. The first solutions for such a problem were limited to hierarchical clustering [KR09, PTVF07, HTF09] and simple comparisons between classes of data through the identification of differentially expressed genes (DEGs) [DYCS02, RYB03]. Nowadays, reconstruction and interrogation of biological networks have become a widely used approach to get insights from different types of omics data [DYR+15, MDD+15].

After establishing co-expression networks for different states of one biological system, differential co-expression analysis investigates their structural changes when a system goes through a state transition. This analysis, first proposed more than a decade ago [KS04, XFG+04], identifies the pairs of genes that have their interaction changed during such transition. Several later publications have suggested different algorithms and statistics to determine differential gene co-expression [SYAP11, NHDQ11, ASS13, dlF10, LWCZ04, Li02, DGP05, Wat06, MLW+08, HQG+09, CKK09, SKV+11, DYK12, Fuk13, LPS+12, CYYK05, PSS+13, CKP12]. Fewer studies, however, attempted to evaluate the biological significance of these changes [MLW+08, SKV+11]. Also, to the best of our knowledge, there have been no studies that would investigate how this approach performs depending on the type and complexity of the biological system analyzed.

Commonly, a state transition of a biological system is related to perturbation of a set of genes, which propagates through network interactions and affects other genes. Thus, there is a possibility that differentially co-expressed (DC) genes (directly or indirectly) contribute to the propagation of perturbations. In order to investigate the role of DC genes in a state transition of a biological system, we considered two biological processes [SMH+11, MSY+13] previously analyzed by our group. The first one (B cell deficiency in mice) is a homogenous, one-causal-factor process, while the second one (cervical cancer) represents a heterogeneous multi-causal system.

In this work, a co-expression network is an undirected graph, where the set of nodes consists of a set of DEGs, and a pair of nodes is connected if there is a significant correlation between them. Differential co-expression analysis is done by identifying the pairs of genes that suffer significant changes in correlation between two states. Throughout this paper such pairs are called differentially correlated pairs (DCPs) and the genes forming these pairs are considered DC genes.

5.2 Material and methods

5.2.1 Preparation of microarray data

- **BcKO.** All microarray data were analyzed using BRB Array-Tools developed by the Biometric Research Branch of the National Cancer Institute under the direction of R. Simon (http://linus.nci. nih.gov/BRB-ArrayTools.html). Array data were filtered to limit analysis

to probes with greater than 50% of samples showing spot intensities of > 10 and spot sizes > 10 pixels, and a median normalization was applied.

- **Cervical cancer.** Same as in cervical cancer[MSY+13]. The data were analyzed using BRB Array-Tools using the original normalization used in three studies [BBM+08, PNL+07, ZKN+07] and median normalization over entire the array for the fourth study [SNN+08]. For all studies, we only considered genes found in at least 70% of arrays.

5.2.2 Finding Differentially Correlated Pairs

For both biological systems studied in this paper we identified the Differentially Co-expressed Pairs using the same procedure. We start considering all genes in the dataset and filter out the genes presenting more than 30% missing data. Next, we calculate the Pearson correlation for each possible pair of genes in 2 different states separately and then the difference of correlation between those states and filtered out pairs that are not present in at least a fixed number of datasets (BcKO: 2 (all studies), cervical cancer: 3 out of 5). In all datasets the difference between correlations in two states must have the same direction (sign). To assure similarities between datasets we select the pairs that have the same direction (sign) of correlation at a significance level of 20% in at least one state. This way we ascertain that the pair is correlated in at least one state and has the same behavior in the state which the correlation occurs. We then proceed to the computation of the p-value for the difference of correlation [SKV+11] and only keep the pairs with p-value lower than 20% in all studies. Now meta-analysis is done through Fisher's method and then FDR. Next we eliminate the pairs that show FDR higher than a threshold (Tables A.5 and A.6). The final step is to identify the pairs that passed the FDR filter and were considered significantly correlated in the final reconstructed network (correlation network for BcKO and local partial correlation network for cervical cancer). Differentially Co-expressed Genes for BcKO and cervical cancer can be found in Table A.2 in Appendix and Table 5.1 respectively.

5.2.3 Filtering and meta-analysis of microarray data

In every analysis (DEGs, DCPs and networks), filter of direction (same sign of correspondent parameter - difference of mean, difference of correlation, correlation and partial correlation) was required in a fixed number of datasets (2 out of 2 in BcKO and 3 out of 5 in cervical cancer). Then meta-analysis was done through Fisher combined probability test [Fis25]. Next, the pairs with false discovery rate (FDR) [BH95] lower than a threshold are chosen. At last, only the pairs that pass PUC [YPK+16] are considered correlated and therefore represent edges in the network.

5.2.4 Analysis of microarray data

- **BcKO.** DEGs between groups of samples were identified by random variance paired t-test p-value lower than 5% with adjustment for multiple hypotheses by setting the FDR below 10% in BRB Array-Tools. Co-expression networks (BcKO and Control) were inferred through Pearson correlation with p-value $< 20\%$ and FDR adjustment below 2.5%. DCPs were calculated for pairs that were initially correlated (p-value $< 20\%$) in at least one state. Then differences of Pearson correlation were tested following [SKV+11] with a p-value below 10% and FDR $< 2\%$. At last only the DCPs that showed up in one of the networks were selected.

- **Cervical cancer.** DEGs were retrieved from a cervical cancer paper[MSY+13]. Correlation networks and DCPs followed the same procedure and in BcKO but with different p-values (correlation p-value $< 10\%$ with FDR $< 10-8$ and difference of correlation p-value $< 10\%$ with FDR $< 0.25\%$). Partial correlation was computed using local partial correlation method30. The initial significance was p-value lower than 40% and then FDR $< 5\%$. For more details about the thresholds used, see Tables A.5 and A.6.

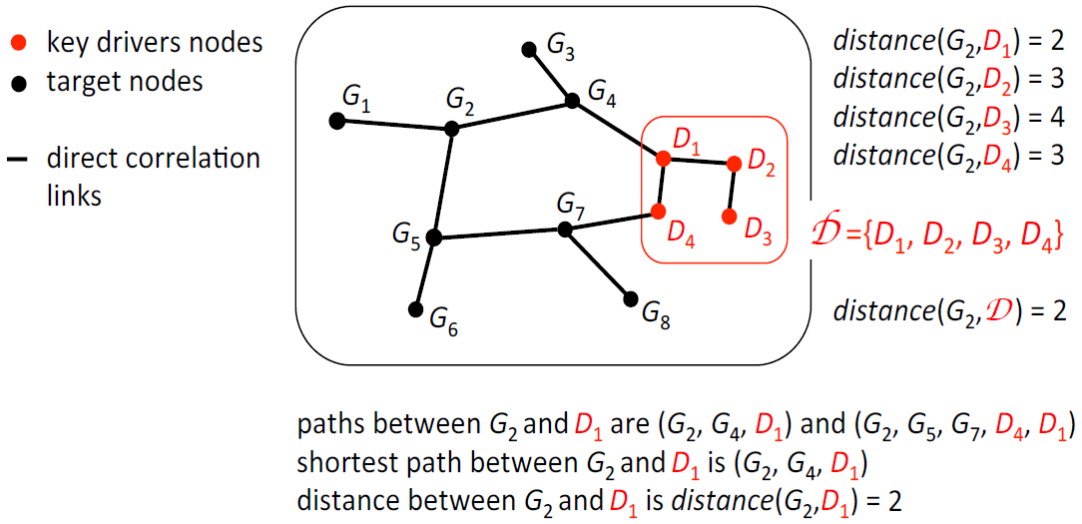


Figure 5.1: *In this example we show how to calculate the distance (length of shortest path) between the gene G_2 and group of genes D_1, D_2, D_3, D_4 (nodes in red).*

5.2.5 Local partial correlation network

Two aspects of cervical cancer data led us to use local partial correlation for this system. First of all, we have more samples throughout five datasets (see Tables A.7 and A.8 in Appendix) which allows us to have more confidence in our results and second we already know that tumors in general present heterogeneous causal factors. The partial correlation approach gives us the alternative to only consider edges that represent direct regulatory relations.

In this paper we used the new approach developed in [TFY12] called local partial correlation. This approach was elaborated specially for cases when there are more variables than samples, which happens regularly in genetics and is a serious problem in classical statistics. First we calculate the correlation network. Then for each significantly correlated pair the inverse method is applied exclusively to the correlation sub-matrix formed only by the closest neighbors of the pair along with the genes forming the pair. If the number of closest neighbors is still higher than the number of samples n, then we decreasingly rank the correlations of the neighbors to either genes in the pair and select the first $\frac{n}{2}$ neighbors. For each sub-matrix, we only keep the partial correlation value regarding the pair that formed that sub-matrix and then calculate its p-value also based on the sub- matrix. R script for calculation is available in Section 3 along with more detailed information.

Partial correlations were estimated only for the significant (Pearson) correlations in co-expression network. Thus the same definition of DCPs (by Pearson correlation) can still represent structural changes as long as it remains present in one of the two networks.

Figure 5.5 illustrates the local partial correlation network for cervical cancer using only tumor data. It has 578 connected nodes and 824 edges.

5.2.6 Minimum shortest path

The shortest path is a distance between 2 nodes in a network. It consists of the minimum number of edges connecting 2 nodes. We define want to define a distance between one node and some subset of nodes in a network, see Figure 5.1. For each gene we calculate the shortest path to all key drivers and get the minimum value. Then we compare the minimum shortest path to key drivers coming from DCP genes and the remaining genes.

5.2.7 Bi-partite betweenness centrality

Once a relationship between genes has been established, the next question is which nodes or genes are responsible for the interaction. Although multiple genes could act as mediators of interaction between two pathways, their relative importance can be different. Few approaches have been developed to find which nodes are critical for crosstalk between different groups of genes in a network.

We have developed an approach that identifies nodes in a network responsible for interactions between modules that potentially correspond to genes regulating crosstalk between pathways represented by these modules. The approach is based on the idea that the genes that are in the shortest paths between modules should be more important in controlling perturbation from one pathway to another, mediating inter-module signaling or regulation. Several centrality measures have been proposed to evaluate the importance of nodes in a network, see [Fre78]. [DSJ+10] shows that betweenness centrality measures the importance of a node in acting as a bridge between any nodes within a network. We modified standard betweenness centrality ([New04]) to adapt to the case of interaction between two defined subnetworks and to specifically address the question of which nodes have a higher probability to be bottlenecks in the transfer of signal from nodes belonging to subnetwork $\mathcal{A}$ to the nodes in subnetwork $\mathcal{B}$, and vice versa. For this metric, the shortest paths are calculated only between nodes of two subnetworks and not between any nodes within a network. This bi-partite betweenness centrality can be calculated just as in equation **??** but with a slight modification:

$$g(\nu) = \sum_{s \neq \nu \neq t} \frac{\sigma_{st}(\nu)}{\sigma_{st}},$$

where $s \in \mathcal{V}(\mathcal{A})$, $t \in \mathcal{V}(\mathcal{B})$, σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those shortest paths that pass through vertex v (node for which the metric is calculated). Thus, this measurement represents the importance of a node in mediating information flow between two connected modules in a network. A gene with high betweenness centrality has a great influence on the transfer of signal through the network. Figure 5.2 illustrates the calculation of bi-partite betweenness centrality for nodes D and C located between subnetworks $\mathcal{A}$ and $\mathcal{B}$. Note that all paths connecting $\mathcal{A}$ to $\mathcal{B}$ must path through D and not necessarily through C, therefore the bi-partite betweenness centrality of D is higher than C. In this study, we are interested in the signal passing from key drivers throughout the network, therefore the two subnetworks are formed by either key drivers or peripheral genes.

5.3 Results

5.3.1 B cell deficiency

We started by analyzing the B cell knockout (BcKO) data [SMH+11], which represents a relatively simple experimental model with only one causal factor (B lymphocytes) and homogenous subject groups since this experiment was performed in highly inbred strains of mice.

In order to select the nodes to reconstruct the co-expression networks (BcKO and Control) we compared gene expression in jejunum between BcKO and control mice and found 509 DEGs ([TVS+16]). Next, the edges for each network were determined using significantly correlated pairs of DEGs (Figure 5.3). To identify DCPs we used the method introduced in [MH14] which compares correlations in the BcKO group and in the Control group. Eighty DCPs were found (Table A.2 in Appendix), of which 56 represent correlation gains (edges which were not present in Control network but showed up in BcKO) and 24 represent losses.

Now we investigate whether network structural changes, herein represented by DCPs, are related to actual causes of global change in gene expression. In the previous study [SMH+11], it was shown that intestinal gene expression alterations in BcKO mice are mostly dependent on the ability of B

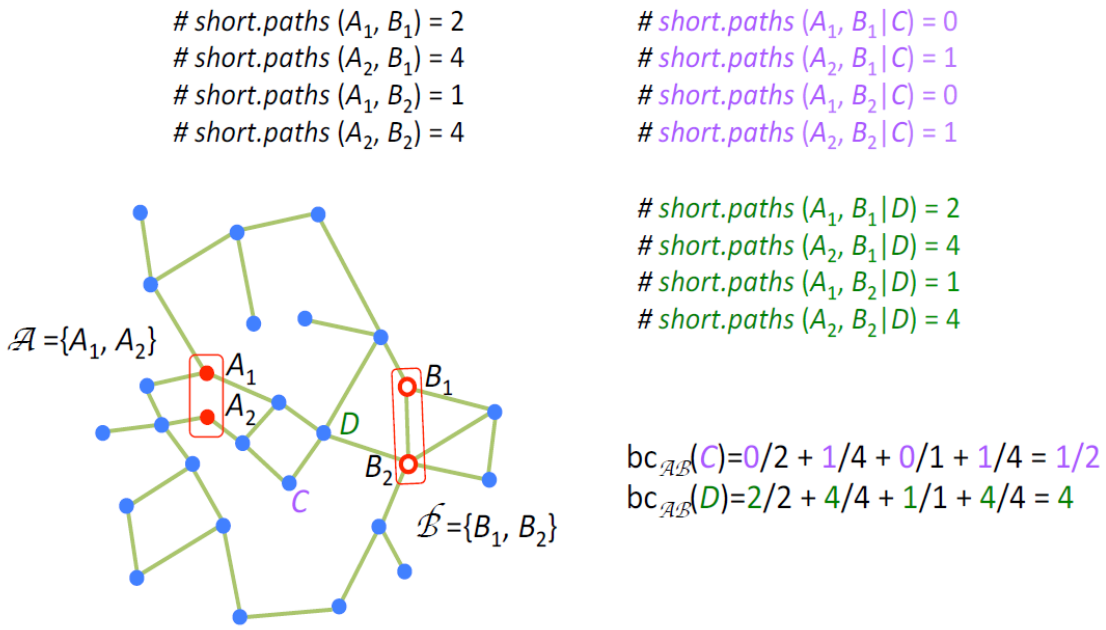


Figure 5.2: *Here we explain how to calculate bi-partite betweenness centrality (bc) between groups A and B. Note that node D has bigger bi-partite bc because all shortest paths connecting nodes in group A to nodes in group B pass through the node D.*

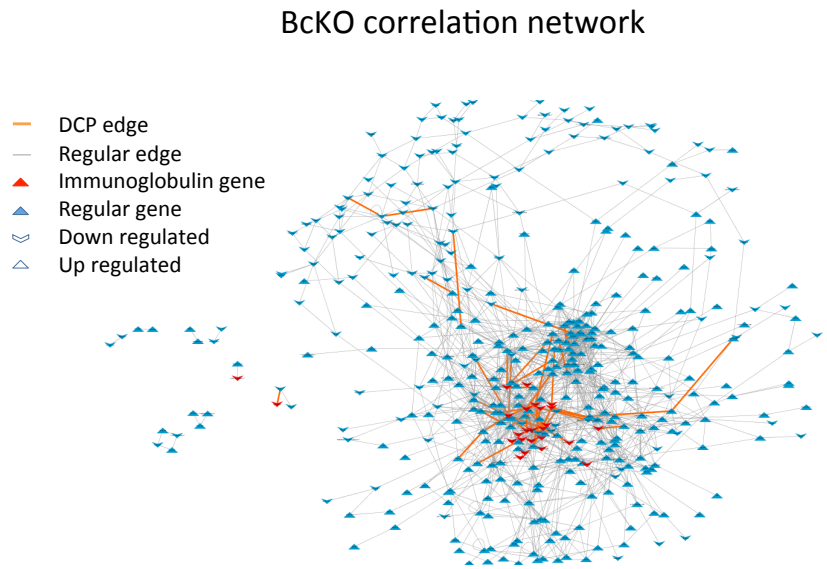
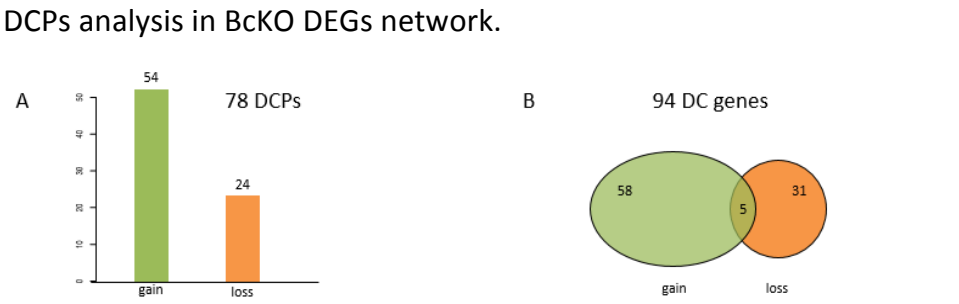


Figure 5.3: *Co-expression networks for BcKO data. The nodes are composed by DEGs and the edges represent significant correlations between nodes. The causal genes (immunoglobulin genes) and the DCP edges are concentrated in the high connectivity region with several causal genes forming DCPs.*



	gain pairs	loss pairs
DCP without Ig genes	15	22
DCP with one Ig gene	31	2
DCP with two Ig genes	8	0
total	**54**	**24**

	genes from gain pairs	genes from loss pairs	non-DC genes
Ig-YES	15	1	11
Ig-NO	48	35	404
total	**63**	**36**	**415**

Figure 5.4: *A) 78 Differentially Correlated Pairs (DCPs) were found, of which 54 represent correlation gains (edges which were not present in Control network but showed up in BcKO) and 24 represent correlation losses. The table stratifies the set of pairs representing correlation gains and losses according to the amount of Ig genes (0, 1 or 2) present in a pair. Note that 39 out of 54 of correlation gain DCPs are formed by at least one Ig gene while only 2 out of 22 correlation losses have at least one Ig gene. B) The 78 DCPs are formed by a total of 94 Differentially Co-expressed genes (DC genes). 58 DC genes participate only in correlation gain DCPs, 31 only in correlation loss DCPs and 5 of them participate in both correlation gain and loss DCPs. The results show enrichment for Ig genes among DC genes in correlation gain: 24% (15 out of 63(= 58 + 5)) of DC genes are Ig genes vs 2.7% (11 out of 415) of other DEGs are Ig genes (p value < 0.001). Meanwhile no enrichment was observed for correlation loss as a result of B cell deficiency: 3% (1 out of 36(= 31 + 5)) of DC genes are Ig genes vs 2.7% (11 out of 415) of other DEGs are Ig genes.*

lymphocytes to produce antibodies. Therefore, we analyzed the presence of immunoglobulin coding genes (Ig genes, see Table A.3 in Appendix) among differentially expressed genes (26 Ig genes among 509 DEGs) in DCPs. We observed that 72% (39 out of 54) of correlation gain DCPs are formed by at least one Ig gene, (Figure 5.4A). Moreover, we found strong enrichment for Ig genes among DC genes in correlation gain (24% (15 out of 63) of DC genes are Ig genes vs 2.7% (11 out of 415) of other DEGs are Ig genes), while no enrichment was observed for correlation lost as a result of B cell deficiency (Figure 5.4B). Thus, these results support the idea that differentially expressed genes that acquire correlations during transition from one biological state to another have a high chance to play causal roles in such transition.

5.3.2 Cervical cancer

Analysis of gene expression data.

In order to study differentially co-expressed genes in a more complex biological model we turned to cancer. It is well known that cancers of the same clinically/ morphological type can be very different on molecular levels. One of the most studied causes for such diversity is the different sets of chromosomal aberrations and mutations harbored by tumors otherwise defined as the same cancer. In previous study [MSY+13], we have found 36 cervical cancer driver genes located in multiple chromosomal aberrations (A.4 in Appendix). Thus we decided to use cervical cancer data from [MSY+13] for investigation of the role of DCPs in complex biological processes due to its heterogeneity and previously acquired knowledge of essential causal genes.

We used the DEGs between tumor and normal tissue as the nodes of the co-expression networks. Since the number of samples (five datasets, 148 tumor samples and 67 normal samples) was larger than in BcKO study (two datasets, 22 paired samples), we used the partial correlation coefficient as a measure of co-expression (Figure 5.5). The potential advantage of using partial correlation is that it aims to infer edges that are a result of direct regulatory relations [MDD+15]. Partial correlations

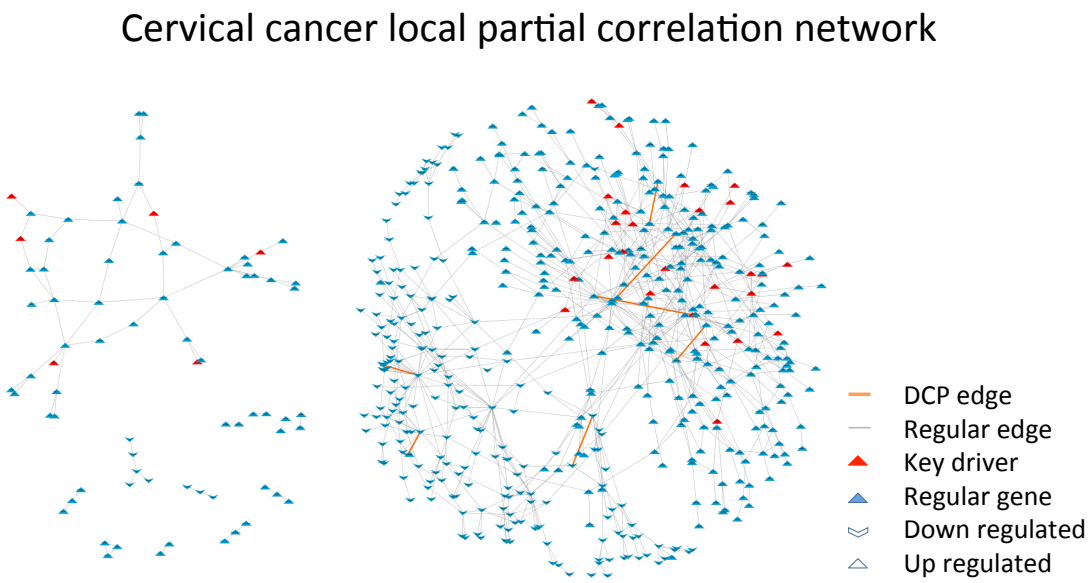


Figure 5.5: *Co-expression networks for cervical cancer data. The nodes are composed by DEGs and the edges represent significant local partial correlation between nodes. A few causal genes (key drivers) and DCP edges are located in the high connectivity region, but scattered throughout the network. Only one key driver is amongst the genes in DCPs.*

were calculated through the Local Partial Correlation (LCP) method [TFY12] (Section 3).

In this study seven DCPs composed of 14 DC genes were found. Interestingly, all DCPs were differential correlations gained in tumors (Table 5.1). Only one of the 36 key drivers (CEP70) was identified among the 14 DC genes. Accordingly, no enrichment of key driver genes among DC genes was detected in this analysis.

Table 5.1: *DCPs - cancer (* key drivers)*

Gene 1	Gene 2	Change direction	Sign of LPC in tumor	Regul. 1	Regul. 2
ANP32E	CACYBP	Gained edge	> 0	UP	UP
CENPN	DHFR	Gained edge	> 0	UP	UP
C10orf68	FGFR2	Gained edge	> 0	DN	DN
AK2	HNRNPR	Gained edge	> 0	UP	UP
CEP70*	SEPHS1	Gained edge	> 0	UP	UP
NIPAL2	TRPM3	Gained edge	> 0	DN	DN
ARHGEF12	ZSCAN18	Gained edge	> 0	DN	DN

Even though we observed that DCPs are not necessarily formed by key drivers, it is known from literature that most of the DC genes found play regulatory roles in other types of cancer [GLT+13, CCK15, HCH+14, LPO+07, WLC+10, BEL+05, LJB+08, NYW+10, HLN+10, HKJ+07, Kat08, JSP01, VAE+15, BW91, KCHS15, NSH+07, SNL+07, MRG+11]. Thus we hypothesized that DCPs are located downstream of key drivers and can be responsible for changes of regulatory chain events coming from the key drivers and spreading throughout the network. In order to verify this hypothesis, we investigated how close DC genes are to key drivers and whether their "signal flow" [FSM+09] in the tumor co-expression network is stronger than that of the other genes. In order to verify this hypothesis we investigated two network measures: Minimum Shortest Path and Bi-partite Betweenness Centrality.

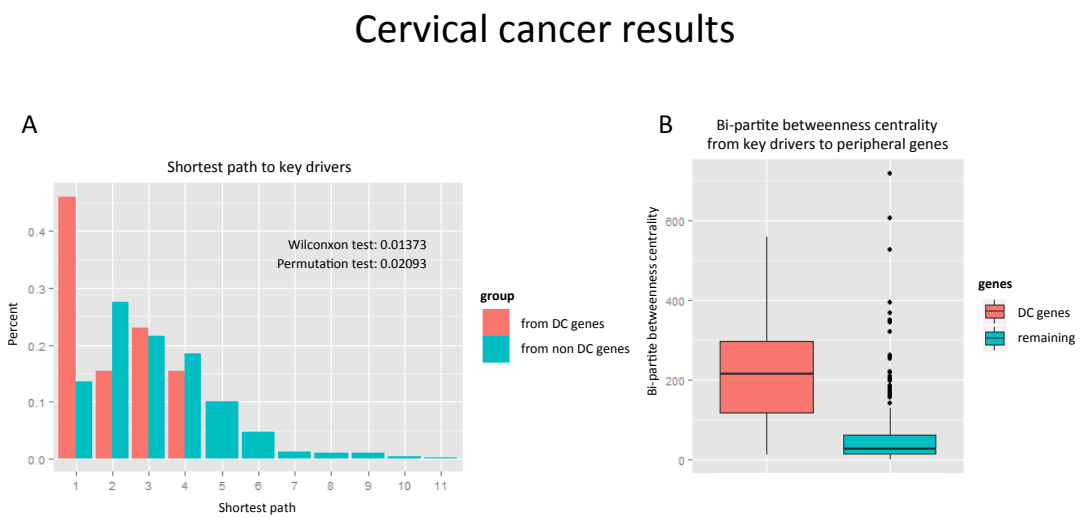


Figure 5.6: *Topological properties of Differentially Correlated Genes (DCGs). A) Barplot of the shortest path to the causal genes and originated in either the genes in DCPs (in orange) or the non DCP genes (in blue). The distribution in orange is concentrated in lower values. B) Boxplot comparing the values of Bipartite Betweenness Centrality of the genes in DCPs (in orange) and the non-DCP genes (in blue). The boxplot on the left is concentrated in higher values.*

First we compared the shortest paths from key driver genes to DC genes and to all other DEGs in the network. We found that DC genes are located statistically closer than the rest of genes in the network to key drivers (Figure 5.6A, Wilcoxon test < 0.014 and Permutation test < 0.021). Then we used Bi-partite Betweenness Centrality as a measure of the signal flow from key drivers to peripheral genes (genes with only one edge) [DYR+15]. We evaluated this measure for DC genes and remaining DEGs and observed that DC genes had much higher values than other genes in the network. Figure 5.6B illustrates a comparison of boxplots of bi-partite betweenness centrality between these two groups concerning DCPs and the rest (non DCPs, non-key drivers, non-peripheral). We can observe that the bi-partite betweenness centralities of DCPs are concentrated in higher values than the rest. Mann-Whitney test gave us a p-value of 7.868×10^{-5}, which gives us evidence that the distribution of Bi-Partite Betweenness Centrality in DCP genes is higher. For more details see Figure S2 in appendix. Thus, altogether these results suggest that DC genes might be "bottlenecks", that is, required to transmit a signal from key drivers to other genes in the network, therefore, supplement the hypothesis of a regulatory role of DC genes (Figure S1 in appendix).

Knockdown experiments

In addition, data from other cancers provide indirect support for the idea of regulatory role of DC genes in cervical cancer [GLT+13, CCK15, HCH+14, LPO+07, WLC+10, BEL+05, LJB+08, NYW+10, HLN+10, HKJ+07, Kat08, JSP01, VAE+15, BW91, KCHS15, NSH+07, SNL+07, MRG+11]. However, since a role of these DC genes in carcinogenesis was not as straightforward as for immunoglobulin genes in B cell deficiency, we decided to perform experimental tests. Among the DC genes found for cervical cancer, there were seven up-regulated and seven down-regulated in cancer.

Therefore, for validation experiments we chose one down-regulated (FGFR2) and one up-regulated (CACYBP) gene that have not been previously studied in cervical cancer for regulatory properties, but have a potential connection with cell death or proliferation based on their Gene Ontology annotations. In order to test if FGFR2 and CACYBP play critical regulatory roles in cancer pathogenesis, we evaluated the effect on in vitro knockdown of these genes on cell proliferation in a cervical carcinoma cell line.

First, we tested two cervical cancer cell lines (Hela and ME180) and found that only ME180 had detectable expression levels of both genes. In order to perform these tests, we evaluated siR-

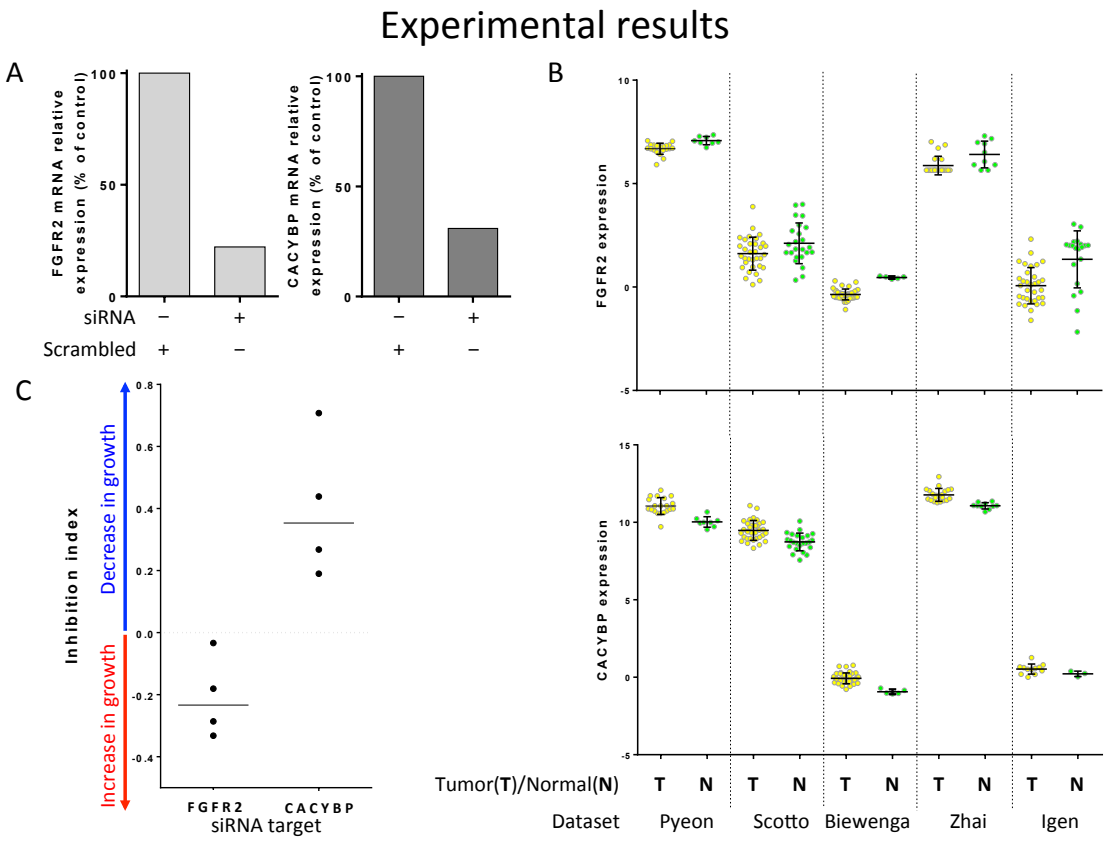


Figure 5.7: *Experimental evaluation of DCGs in cervical cancer. A) Efficacy of FGFR2 and CACYBP siRNA knockdown. qRT-PCR with primers for GAPDH as the internal control was used to determine expression and efficacy of FGFR2 and CACYBP specific siRNA knockdown in endothelial cells (ME180). ME180 cells were harvested* 72 *h after transfection with vehicle (Lipofectamine) and either scrambled control or targeting siRNA. B) Gene expression of FGFR2 and CACYBP (mean +/− standard deviation) for tumor and normal samples from five datasets used in the analysis. Since FGFR2 was found down-regulated in tumor tissue, its potential regulatory role would be as a tumor suppressor. However, CACYBP is up-regulated, thus CACYBP should function as an oncogene promoting cell proliferation. C) Evaluation of cell proliferation inhibition using xCelligence System. Proliferation data (cell index) was obtained at* 72 *h after transfection with Lipofectamine and either scrambled control or targeting siRNA. Inhibition index was calculated (two step normalization of cell index): inhibition index* > 0 *- cells transfected with targeting siRNA showed decrease in proliferation;* < 0 *- showed increase in proliferation;* $= 0$ *- no difference from control was found. One sided T test for mean (*< 0 *for FGFR2 and* > 0 *for CACYBP) was applied and returned statistically significant p-values for both of them (*0.0258 *for FGFR2 and* 0.01978 *for CACYBP).*

NAs and observed that they were able to knock down expression of both genes by at least 70% (Figure 5.7A). CACYBP is up-regulated in tumor tissue, as compared to normal tissue (Figure 5.7B). Consequently, if CACYBP has regulatory potential, as predicted by our analysis, it should function as an oncogene promoting cell proliferation. Therefore, the knockdown of this gene should result in a decrease of cell growth/survival. Since FGFR2 was found down-regulated in cervical carcinomas (Figure 5.7B) its potential regulatory role would be as a tumor suppressor. Therefore, the knockdown of this gene is expected to increase cell growth. The subsequent analysis of cell proliferation confirmed our predictions for both genes: knockdown of CACYBP led to a decrease of cell growth, while knockdown of FGFR2 induced higher cell proliferation (Figure 5.7C). Thus, these results provide additional support to our *in silico* prediction that DC genes may play a regulatory role in cell proliferation related to tumor growth.

5.4 Experimental design

5.4.1 FGFR2 and CACYBP knockdown experiment

ME180 cells were transfected with FGFR2-, CACYBP-specific siRNA or control siRNA using Lipofectamine RNAiMAX Transfection Reagent. Cell growth rate during 72h after siRNA transfection was measured using xCelligence system as described below.

Evaluation of siRNA efficacy in knocking down the gene targets.

ME180 cell line was obtained from Dr. Pulivarthi H. Rao. It was cultured in RPMI medium with 10% FBS and 1added. The cells were seeded at density 4000 cells per well in 96 F-bottom plates (seeding procedure was done according to ATCC protocol for ME180 cell line) and with cell culture media 200 ul per well. 24 hours after seeding, cells were transfected with one of the three siRNA (Table 5.2).

Before transfection, 100 uL of media was taken from each well. Transfection procedure was done according to Lipofectamine RNAiMAX Reagent protocol (Protocol Pub. No. MAN0007825 Rev. 1.0). 3pM of siRNA per well and Lipofectamine 0.6 uL per well were delivered in 20uL. 80 uL of fresh cell culture media was added to each well.

Cells were collected 72 h after transfection using Lysis buffer from RNeasy Mini Kit (QIAGEN). RNA extraction was done using RNeasy Mini Kit (QIAGEN) according to the manufacturer's protocol (no Dnase treatment step was done). Concentrations of RNA measured with Qubit RNA BR Assay Kit. cDNA was done using Bio-Rad iScript cDNA Kit according to the manufacturer's protocol.

Quantitative Real-Time PCR was done for the samples using QuantiFast SYBR Green PCR Kit and GAPDH as a control gene. Primers for the targets you can see in the Table 5.3.

qRT PCR set up: sample was heated to 95Â°C, followed by 40 cycles of 95Â°C for 10 sec and 60Â°C for 30 sec.

Table 5.2: *Suppliers*

Target	Supplier	Supplier ID
FGFR2	ThermoFisher	s5173
CACYBP	ThermoFisher	s25819
Non-targeting siRNA	Dharmacon	D-001810-01-05

Table 5.3: *Primers and Targets*

Target	Forward/ Reverse	Primer sequence (5' → 3')
FGFR2	Forward	AACAGTTTCGGCTGAGTCCAG
FGFR2	Reverse	GCCCAGTGTCAGCTTATCTCTT
CACYBP	Forward	CTCTGTGGAAGGCAGTTCAAA
CACYBP	Reverse	TCAGGTAATCCCACCTTGTGTT
GAPDH	Forward	GGAGCGAGATCCCTCCAAAAT
GAPDH	Reverse	GGCTGTTGTCATACTTCTCATGG

Evaluation of cell growth after knock down of gene targets.

CACYBP is up-regulated in tumor tissue, as compared to normal tissue (Figure 5.7B). Consequently, if CACYBP has regulatory potential, as predicted by our analysis, it should function as an oncogene promoting cell proliferation. Therefore, the knockdown of this gene should result

in a decrease of cell growth/survival. Since FGFR2 was found down-regulated in cervical carcinomas (Figure 5.7B) its potential regulatory role would be as a tumor suppressor. Therefore, the knockdown of this gene is expected to increase cell growth.

Cell growth was evaluated using xCelligence system (The RTCA DP Instrument) using manufacturer's protocol. ME180 was cultured in RPMI media with 10% FBS and 1% PenicillinStreptomycin added. The cells were seeded at density 4000 cells per well (E-Plate 16) in 200 uL of cell culture media.

Twenty four hours after seeding, the experiment was paused for transfecton. Before transfection, 100 uL of media was taken from each well. Transfection procedure was done according to Lipofectamine RNAiMAX Reagent protocol (Protocol Pub. No. MAN0007825 Rev. 1.0). 3pM of siRNA per well and Lipofectamine 0.6 uL per well were delivered in 20uL; 80 uL of fresh cell culture media was added to each well. Plate was placed back in the slot and cell growth was evaluated for another 72 h.

Cell index normalization.

To evaluate cell growth rate cell index was transformed into Inhibition index in two steps:

1. Cell indexes for all wells were exported to the excel file. For each treatment (including non-targeting siRNA transfected wells) we extracted cell index average for all wells at 20 h after seeding (Cell Index Before Treatment) and at 96 h after seeding (Cell Index After Treatment). To normalize cell index to initial cell number differences for each of the treatments we used the following formula:

$$After/BeforeTreatmentNormalizedCellIndex(A/BIndex) = \frac{CellindexAfterTreatment}{CellindexBeforeTreatment}$$

2. In next step we normalized each treatment with targeting siRNA to treatment with non-targeting siRNA. For this purpose in each experiment A/B Index from treatment (siRNA targeting either FGFR2 or CACYBP) was normalized to A/B Index from control treatment using the following formula:

$$InhibitionIndex = \frac{ControlA/BIndex - TreatmentA/BIndex}{ControlA/BIndex}$$

Final evaluation of growth was done according to the value of Inhibition Index: > 0 - there is a decrease in growth; 0 - no difference between treated with targeting and treated with non-targeting siRNA; < 0 - there is a growth after treating with targeting siRNA.

5.4.2 Data availability

BcKO: Gene expression files containing array data from [SMH+11] are available under the GSE23934 superseries in the Gene Expression Omnibus (GEO) data repository. We worked with two groups of samples: B10.A litter-mates and BALB/C (Table A.7 in Appendix). Cervical cancer: We have used the same datasets as in previous study[MSY+13] available at GEO: GSE741050, GSE679151, GSE780352, GSE975053, GSE26342[MSY+13] (Table A.8 in Appendix).

5.5 Discussion regarding real data analysis

In the current study, the differential co-expression analysis [SKV+11] was applied to two relatively well-investigated biological systems[SMH+11, MSY+13] in order to evaluate the potential importance of genes found using differential correlation analyses. Overall, the obtained results support

the idea that DC genes play a regulatory role. While in B cell deficiency DCPs were found highly enriched with immunoglobulin genes (i.e. causal genes for alterations in the gut) we did not observe enrichment for key driver genes in cervical cancers. Rather, DCPs of cervical cancer seem to be located downstream of causal genes. Indeed, those DCPs have been found closer to key regulators than other genes in the network, actually representing "bottlenecks" for communication between driver genes previously published in [MSY+13] and the rest of the network (Figure 5.6). Furthermore, some differentially co-expressed genes in cervical cancer have been previously implicated in processes such as metastasis, oncogenic autophagy and apoptosis. For example, CACYBP has been shown to promote colorectal cancer metastasis [GLT+13], TRPM3 was observed to play a role in oncogenic autophagy in clear cell renal cell carcinoma [CCK15, HCH+14], and AK2 was reported to activate apoptotic pathway [LPO+07]. Several genes are investigated for prognostic value for cancers such as myeloma [WLC+10], lymphoma [BEL+05], breast [LJB+08, NYW+10, HLN+10, HKJ+07, Kat08] and gastrointestinal cancers [JSP01, VAE+15]. At least two genes were previously proposed as targets for anti-cancer agents: DHFR [BW91] and FGFR2 [KCHS15]. Moreover, CACYBP and ZSCAN18 were also reported as putative tumor suppressor genes in renal cell carcinoma [NSH+07, SNL+07]. In addition, we have tested two DC genes and confirmed their regulatory role (FGFR2 as a tumor suppressor and CACYBP as a potential oncogene in cervical cancer) by manipulating their expression in vitro. Altogether, published observations and our experimental validation for these two genes support the idea that DC genes revealed in the current study play a regulatory role and can be candidate targets for cervical cancer treatment.

Interestingly, while in the model of B cell deficiency, the DC genes are highly enriched with causal regulatory genes, there was only one key driver in cervical cancer (CEP70), despite the DC genes in this system still seeming to play a regulatory role overall. Such a difference is potentially related to the fact that the mouse system studied in [SMH+11] is highly homogeneous (inbred mice) with only one cause of alterations (i.e. absence of B lymphocytes). Cervical cancer, however, is a heterogeneous system with different chromosomal aberrations and consequently turned-on expression of different driver genes in different patients. Therefore, we can speculate that differential correlations point to regulatory genes that are shared by majority of samples. This hypothesis warrants further investigation, especially considering that DCPs could represent common therapeutic targets for tumors that originated as a result of different genomic or epi-genomic events.

In conclusion, this study provided additional evidence for the previously suggested idea ([KS04, XFG+04, SYAP11, NHDQ11, ASS13, dlF10, LWCZ04, Li02, DGP05, Wat06, MLW+08, HQG+09, CKK09, SKV+11, DYK12, Fuk13, LPS+12, CYYK05, PSS+13, CKP12]) that genes presenting alterations in correlation patterns between different phenotypes (i.e. states of biological system) play a critical regulatory role in transitions from one state to another. Furthermore, although our results do not allow for full generalization, they indicate that gain and not loss of correlations connects critical genes involved in transitions to new phenotypes. However, further studies are required to understand how changes in correlation patterns can point to genes with critical capacity to guide a biological system into certain state/ phenotype.

5.6 GGM Numerical Analysis

As already introduced in Section 2.1.2, in cases where variables follow a normal distribution, conditional independence can be replaced by partial correlation. Thereafter, as can be seen in Section 3.1.1, the precision matrix $\Omega = \Sigma^{-1}$ can be also interpreted as a representation of edges in a network.

5.6.1 Covariance Matrix Simulation

For a determined graph structure G we want to develop a GGM model. Different GGM simulators were analyzed such as the simulator in GGMridge [HS14] and Igraph [CN06] R package among others. We chose to work with the simulator in [CLZ+16].The first step is to build the precision

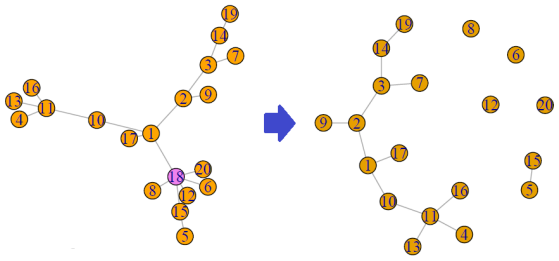


Figure 5.8: *Example of a knockout in a scale-free graph structure. The purple node in the network to the left is the knockout gene. The network to the right is the structure after perturbation.*

matrix in the following way: for every non-zero entry of the adjacency matrix we draw a value from an $Uniform[0.25, 0.75]$ distribution with 70% chance of being a positive value. Then, to guarantee the positive definite property, $\Omega > 0$, the diagonal entries are equal to the minimum eigenvalue of Ω times 0.05, that is,

$$diag(\Omega) = \Omega + 0.05min(eigen(\Omega)). \tag{5.1}$$

After calculating the precision matrix with non-zero entries only for the pair of variables corresponding to an edge and for the diagonal entries, the next step is to calculate the covariance matrix. However, note that equation 5.1 produces the same value for all diagonal entries, which will force all variables to have the same variance. In order to allow the variables to have different variances, let D be a diagonal matrix with entries drawn from an $Uniform[1, 5]$. The simulated covariance matrix is as follows:

$$\Sigma_{sim} = D\Omega^{-1}D. \tag{5.2}$$

5.6.2 Perturbations

Once a graphical structure is generated we have the original/wildtype network. But that is not enough. Network perturbations are necessary for a state transition to occur. The analysis of effects of an "artificial" perturbation can provide more evidence to support the conclusion that DCPs tend to be close to the perturbed nodes as observed in Section 5.3. It will also determine if this is a result of a specific state transition or a network property. We have studied several possibilities of disturbing a networkand have decided to start with the most simple way: knockout of a gene or an edge. This way we have more control of what is happening in the network and we know exactly where the change flow is coming from.

Knockout of a node

The first type of perturbation applied is the **knockout** of a node, which, in the GGM approach, basically means that a node will not be a random variable anymore. Instead, it becomes a constant value and the covariance matrix changes accordingly.

Let $\mathbf{X} = \{X_1, \cdots, X_p\}$ be a random variable vector representing nodes $1, \cdots, p$ respectively and following a Normal distribution $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Suppose the variable X_k assumes a constant value c. Then, we have $\mathbf{X}$ conditional on $X_k = c$, that is, $\mathbf{X}|X_k = c$ (a graphical example is shown in Figure 5.8). It is already known that $\mathbf{X}|X_k \sim N(\boldsymbol{\mu}^*, \Sigma^*)$. To get to the new values of the conditional expectation and conditional covariance matrix, let

$$Y = \mathbf{X} \backslash X_k = \{X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_p\}$$

and lets partition $\boldsymbol{\mu}$ and Σ in the following way:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_Y \\ \mu_k \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{YY} & \boldsymbol{\sigma}_k^T \\ \boldsymbol{\sigma}_k & \sigma_{kk} \end{bmatrix}$$

where μ_k is the expectation of X_k, $\boldsymbol{\mu}_Y$ is the expectation vector of all variables in $\mathbf{X}$ except for X_k, σ_{kk} is the variance of X_k, $\boldsymbol{\sigma}_k$ is the covariance vector between X_k and $\mathbf{Y}$, $\boldsymbol{\sigma_k}^T$ is $\boldsymbol{\sigma}_k$ transposed and Σ_{YY} is the covariance matrix of $\mathbf{Y}$.

The conditional expectation $\boldsymbol{\mu}^*$ and the conditional covariance Σ^* are then given by

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y + \boldsymbol{\sigma}_k^T \sigma_{kk}(c - \mu_k)$$

and

$$\Sigma^* = \Sigma_{YY} - \boldsymbol{\sigma}_k^T \sigma_{kk} \boldsymbol{\sigma}_k$$

respectively.

From the conditional covariance matrix, we get the conditional correlation matrix ρ^* through Pearson correlation coefficients and the difference of correlation matrix $\Delta\rho$ as follows:

$$\Delta\rho = |\rho_Y - \rho^*|.$$

5.6.3 Edge Removal

The knockout of a node in GGM causes the disappearance of all edges connected to such node. Since the occurrence of correlation in only one state is enough to call the edge a DCP, these edges vanished by a knockout are automatically considered DCP and are the closest to a perturbed node as edges can be. To make it a little less straight forward, we proceed to a different network perturbation: edge removal.

After simulating the covariance matrix, an edge connecting two highly connected genes (for example i and j) is removed - KO edge. Afterwords, a new edge is added by connecting the node in the KO edge with higher degree to a leaf (node with degree 1). Figure 5.9 illustrates a scale free graph with 20 nodes where the edge $(1, 3)$ is removed and the edge $(1, 9)$ is included. Mathematically this is done by setting to 0 the entry σ_{ij} in the simulated covariance matrix Σ correspondent to the KO edge while the covariance entry representing the included edge assumes a value from an $Uniform[0.25, 0.75]$ distribution with 70% chance of being a positive value. The equations 5.1 and 5.2 are applied again for the sake of coherence.

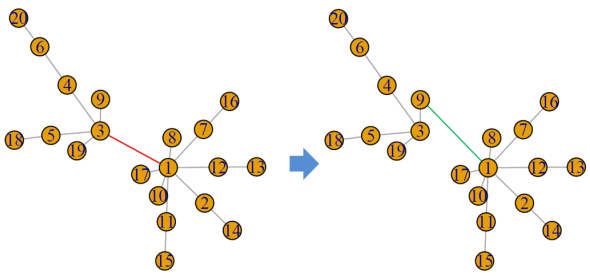
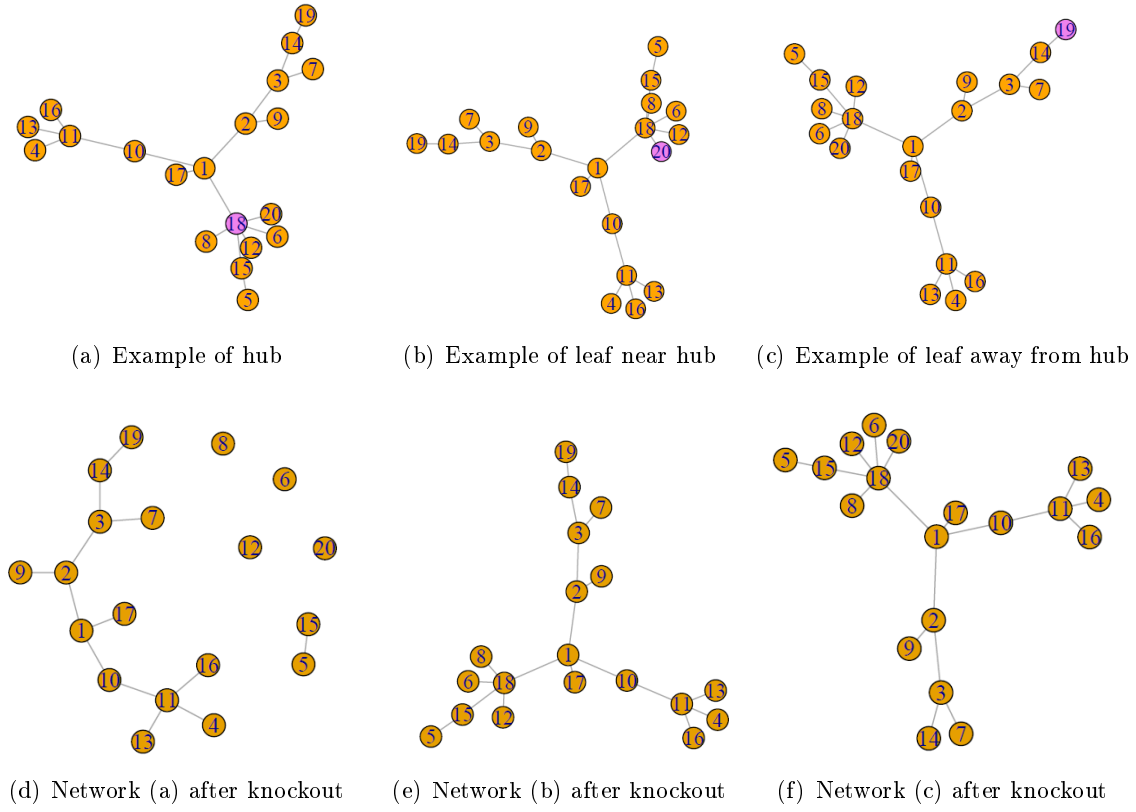


Figure 5.9: *Example of a knockout of an edge in a scale-free graph structure. The red edge in the network to the left is the knockout edge. The network to the right is the structure after perturbation with a new edge (green).*

5.6.4 Analysis

Different graphical structures were generated for 20, 50, and 100 nodes. The selected structure models to be analyzed in this study are the following: random graph (Erdos-Renyi and edge percent - both with 3 edges per node in average), tree (regular with two leaves and Galton Watson with



(a) Example of hub (b) Example of leaf near hub (c) Example of leaf away from hub

(d) Network (a) after knockout (e) Network (b) after knockout (f) Network (c) after knockout

Figure 5.10: *Examples of a scale free graphical structure with 20 nodes. The purple nodes in (a), (b) and (c) are being knocked out. (d), (e) and (f) represent the networks right above them after knockout*

number of offspring following a log-normal distribution with $\mu = 0, \sigma^2 = 1$, which correspond to $meanlog = 0$ and $sdlog = 1$ in R), scale free as in Section 2.1.1, small world with neighborhood equals 2 and $p = 0.05$, and brush-like lattice with length = ceiling($\sqrt{p}$), width = floor($\frac{p}{length}$), generating a lattice with approximate p nodes and values = $\frac{p}{4}$.

As a starting point, after covariance matrices are simulated, three types of nodes were knocked out separately for each structure: hubs (genes whose degree is above the 80% percentile), leaves near hubs (shortest distance = 1) and leaves as far away from hubs as possible. Note that leaves are genes connected to only one other node (degree = 1). Illustrations of the different type of knockouts in a scale-free graphical structure with 20 nodes can be found in Figure 5.10 along with their new structures after knockout.

For each scenario (structure, number of nodes and KO node), we checked the existence of any signs of relationship between difference of correlation and several edge measures in a graph: shortest path between the knockout node and one of the nodes corresponding to a component of a pair of variables, sum of shortest paths between the KO node and both edge components (shortest distance from KO), number of leaves and hubs in each edge, among others.

Shortest distance from KO and number of leaves were the only measures that have shown any influence in the identification of high difference of correlation values. Scatterplots of $\log \Delta\rho$ versus shortest distance from KO were investigated and in all scenarios described previously we could observe that there is a negative linear relationship between $\log \Delta\rho$ and shortest distance from KO node, which means that there is an exponential increase of $\Delta\rho$ as edges get closer to the perturbation. These scatterplots were also performed separately according to the amount of leaves in a pair - either no leaves, 1 leaf or 2 leaves. Unexpectedly, for some graph structures, such as scale free and lattice with leaves, the amount of leaves in a pair can also indicate higher difference of correlation among pairs with the same shortest distance from KO, which can be seen in Figure 5.11 for networks with 100 nodes. Scatterplots for the three knockouts described above (hub, leaf

near hub and leaf away from hub) are illustrated in Figures 5.11(a), 5.11(b) and 5.11(c) while only hub knockout in lattice structure is represented in Figure 5.11(d). We can see that linear regression curves of $\log \Delta\rho$ versus shortest distance from KO gets steeper as the amount of leaves in a pair increases, which means that when the pair is really close to the perturbation - distance of 2, 3, 4 - the chances of higher $\Delta\rho$ increase as the number of leaves in that pair also increases.

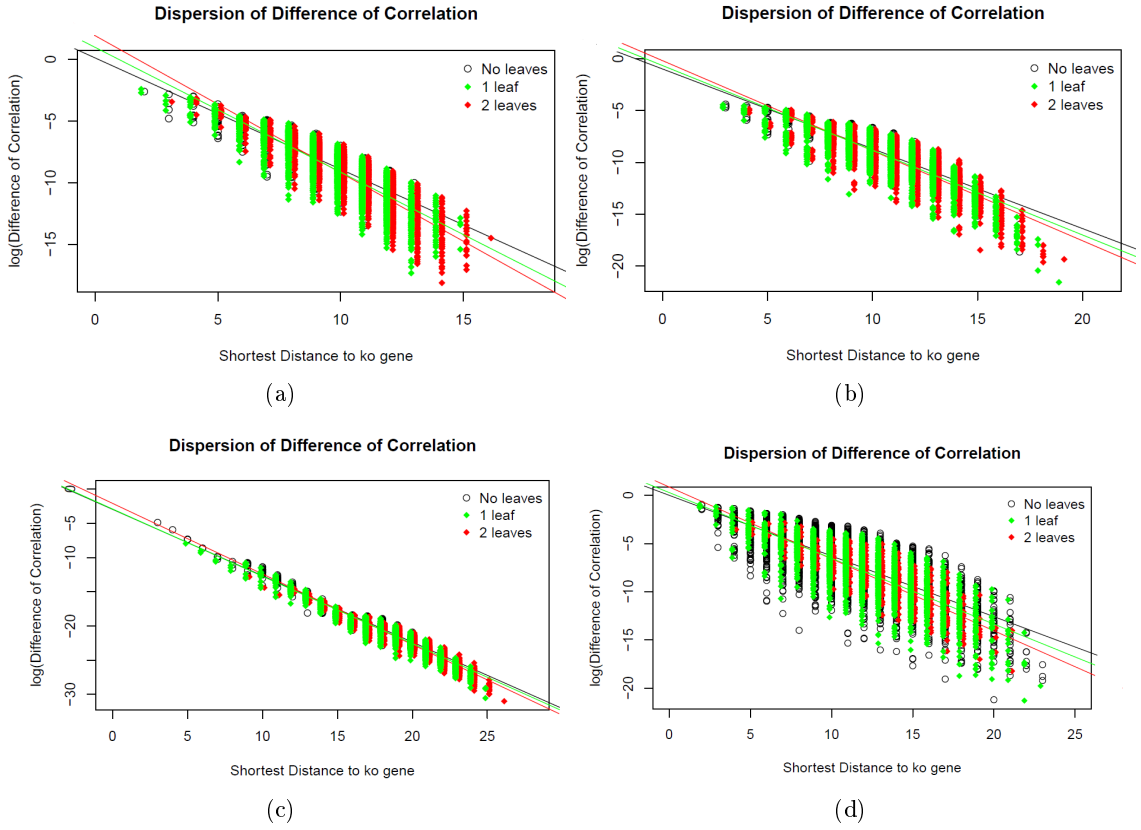


Figure 5.11: *Scatterplots of* $\log \Delta\rho$ *versus shortest distance from KO: a) hub knockout in a scale free structure, b) leaf near hub knockout in a scale free structure, c) leaf away from hub knockout in a scale free structure, and d) hub knockout in a lattice structure with leaves*

In addition, for each graph structure, knockouts of all p nodes were performed separately one at a time. Linear regression coefficients of all p knockouts were used to generate boxplots for all structures which are illustrated in Figure 5.12. As you can see, all boxplots are entirely located below 0, that is, all knockouts in every graph structure produced negative coefficient values regardless of graph structure and which node suffered knockout. Moreover, Galton-Watson and regular trees present small variance in the coefficient values, as opposed to the other structures with bigger variance. More interestingly, it is clear that scale free structures present higher absolute coefficient values, showing that in scale free structures the shortest distance has more influence over the difference of correlation: the closest the highest. These results support the conclusion observed in Section 5.3 that DCPs are located near the perturbed nodes specially in Genetics since the co-expression networks seldom presents scale free structures.

Further assessment of scale free structures was performed in order to address the influence of the amount of leaves - two leaves, one leaf or no leaves - on the greatness of difference of correlation. For each node knockout, linear regression coefficients were calculated using only the correlation coefficients corresponding to pairs of nodes formed by 2 leaves, 1 leaf and no leaves separately. Figure 5.13 shows the boxplots of linear regression coefficients of scatterplots of $\log \Delta\rho$ versus shortest distance from KO for scale free structure for all pairs of nodes (black), pairs composed by 2 leaves (red), 1 leaf (green) and no leaves (blue), considering 20, 50, and 100 nodes. Regardless of the number of nodes on scale free structures, the median value of all coefficients decreases as

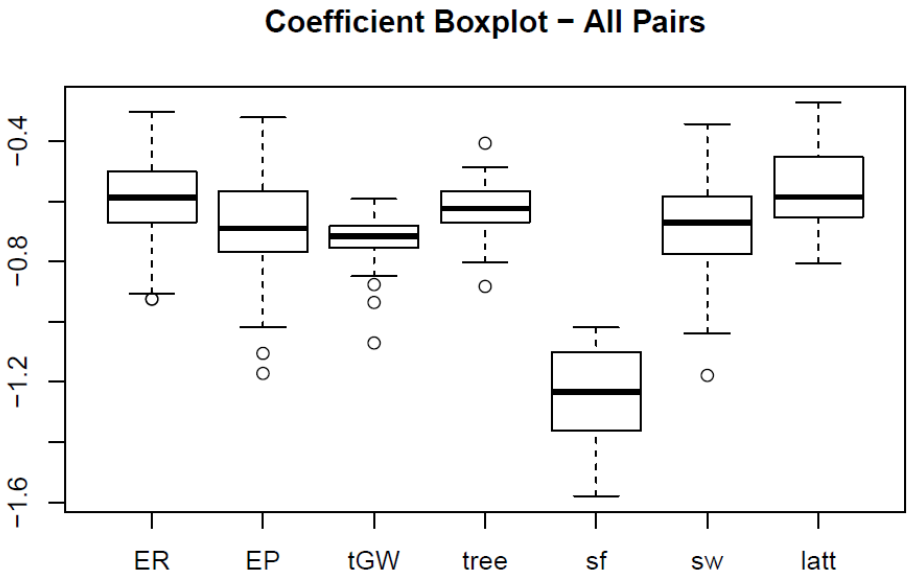
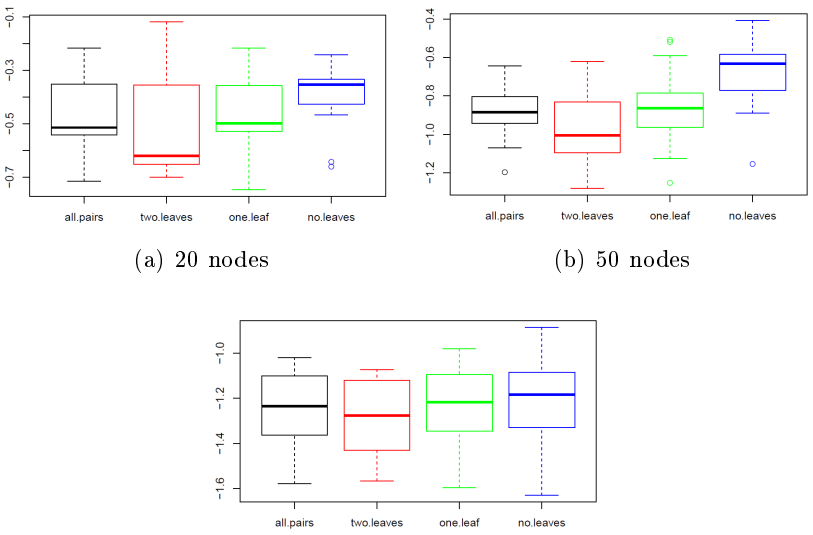


Figure 5.12: *Boxplots of linear regression coefficients of scatterplots of* $\log \Delta\rho$ *versus shortest distance from KO for all structures: Erdos Renyi (ER), edge percent (EP), Galton Watson tree (tGW), regular tree with two offspring (tree), scale free (sf), small world (sw), lattice with leaves (latt). Note that the black boxplots are a result of correlations of all pairs of variables in the model, not a stack of the colored boxplots*



(a) 20 nodes

(b) 50 nodes

(c) 100 nodes

Figure 5.13: *Boxplots of linear regression coefficients of scatterplots of* $\log \Delta\rho$ *versus shortest distance from KO for scale free structure with: (a) 20 nodes, (b) 50 nodes, and (c) 100 nodes. The black boxplots consider all pairs, the red boxplots only consider pairs composed by 2 leaves, the green boxplots only consider pairs composed by only 1 leaf, and the blue boxplots only consider pairs with no leaves.*

the number of leaves in a pair increases, which supports the result that linear regression curves of $\log \Delta\rho$ versus shortest distance from KO gets steeper as the amount of leaves in a pair increases. Therefore, when the pair is really close to the perturbation, the number of leaves in the pair impacts $\Delta\rho$: pairs of nodes composed by two leaves tend to have higher $\Delta\rho$ than pairs with one leaf which in turn tend to have higher $\Delta\rho$ than pairs with no leaves.

Edge removal or edge knockout was also performed for a scale-free structure with 100 nodes. The edge removed was the one connecting the node with the maximum degree in the network

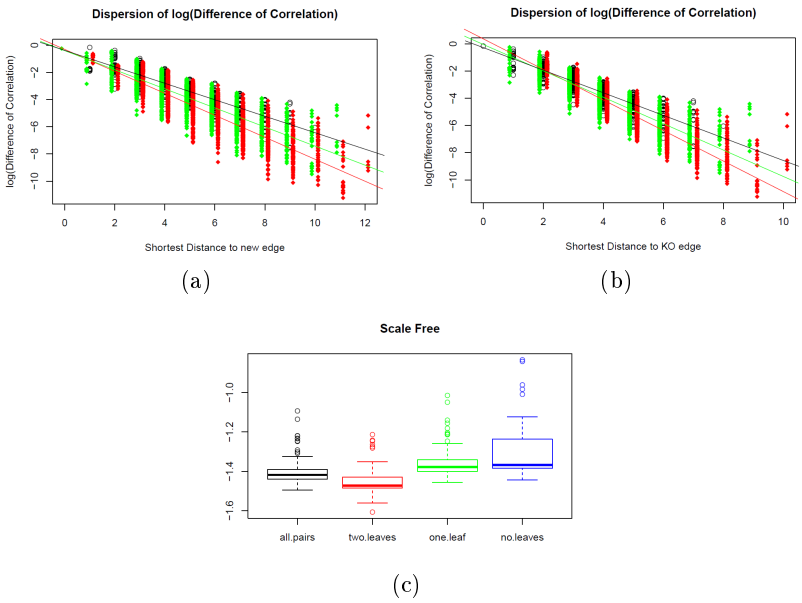


Figure 5.14: *Scatterplots of* $\log \Delta\rho$ *versus shortest distance from: (a) new edge, (b) KO edge, (c) Boxplots of linear regression coefficients of* $\log \Delta\rho$ *versus shortest distance from KO for scale free structure with 100 nodes. The black boxplots consider all pairs, the red boxplots only consider pairs composed by 2 leaves, the green boxplots only consider pairs composed by only 1 leaf, and the blue boxplots only consider pairs with no leaves.*

N_{max} with its direct neighbor with higher degree. Then an edge connecting N_{max} is added. Figures 5.14(a) and 5.14(b) show a scatterplot of $\log \Delta\rho$ along with the shortest distance to the new edge and the KO edge respectively. Note that the three curves cross in the region of low distance just like when perturbing a node. To expand the study, the same KO edge is kept but we vary the new edge separately. Each time connecting N_{max} to a different node, that is, $(N_{max}, j) \notin \mathcal{E}$ and calculating the linear regression coefficient for $\log \Delta\rho$ versus shortest distance. Figure 5.14(c) shows the boxplots for such coefficients when considering all pairs of variables (black), only pairs representing two leaf nodes (red), one (green) and none (blue). This possible to see and ascension of boxplots as the number of leaves in a pair decreases. Since the boxplots are composed by only negative values, we can conclude that the linear regression curves tend to be steeper as the number of leaves in a pair increases. Therefore, giving support to the results for node knockout in scale-free networks.

This study shows that the distance from the network perturbation has direct effect on the changes after a network transition. The closest the pair of nodes are to the perturbation site, the bigger difference of correlation. Besides distance, another feature that seems to affect the changes is the amount of leaves in the pair uncertain network structures. If we compare pairs with same distance from knockout, the pairs with more leaves seem to present higher difference of correlation. The reason why this happens could be the fact that there is only one last edge to get to the leaves. Therefore, if changes come through that edge, there are no other edges to " fight" them. More research is necessary to confirm this observation such as more general simulations and mathematical demonstrations.

Chapter 6

Comparison of Partial Correlation Methods

Pearson and Spearman correlation are often used in Biology to identify associations between selected genes and determine the edges of relevant networks [SS+05a]. However, undesirable spurious associations may be detected as a result of signal flow passing through several true network interactions. In order to overcome this problem, several methods of network reconstruction have been recently developed. Most of these methods are based on Gaussian Graphical Models (GGM), where variables are assumed to follow a Gaussian distribution allowing the use of partial correlation to determine direct associations. Some of these methods are also a solution for high dimensional data, which is still an issue in Classical Statistics characterized by sample sizes much smaller than the number of variables.

In this project, one of the goals is to assess the efficiency of local partial correlation, which estimates partial correlation considering only the neighborhood of each correlated pair of variables. We would like to ascertain that the new method developed by the Author ([Tho12]) is as efficient as other methods recently developed. In this chapter, we compare LPC - explained in Section 3 - to partial correlation methods based on methodologies of regularization: graphical lasso [FHT08] and GGM Ridge - network reconstruction using ridge penalty [HS14]. Both methods assume that the nodes represent variables following a Normal distribution and already have R packages available: glasso and GGMridge.

Graphical Lasso, defined in [FHT08] and the most popular among the three methods mentioned above, is an adaptation of the optimization developed by [BEGd08], which in turn improved the method based on neighborhood selection with the lasso described in [MB06]. Graphical Ridge, even though has not been used much, consists of an interesting alternative to Graphical Lasso using Ridge penalty instead. Besides providing partial correlation coefficients, the procedure developed by [HS14] and based on [SS05b] also returns p-values for each partial correlation estimates from empirical distributions. Both methods are adaptations of covariance estimation through MLE.

The assessment of the methods described above is performed using simulations and real-life data. GGM data was generated for scale-free, tree and small world graph structures while GeneNetWeaver, a software based on differential and stochastic models, provided data from gold standards E-coli subnetworks. ROC curves are employed to compare the performance of each method in the simulated data where there are original networks to liken. The methods are then applied to data from cervical cancer studies in Chapter 5 and are evaluated through edges and nodes similarities.

6.1 Partial correlation methods with regularization

Assume $\boldsymbol{X} \in \mathbb{R}^p$ a random vector following a multivariate normal distribution with expected vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. The probability density function on each component X_i of $\boldsymbol{X}$ is

$$f(x) = (2\pi)^{-\frac{p}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-\boldsymbol{\mu})^T\Sigma^{-1}(x-\boldsymbol{\mu})), \tag{6.1}$$

where $\Sigma > 0$ is a positive-definite (nonsingular) matrix. Suppose that $X_{n\times p} = (x_{ij})$ is the data matrix where each column is composed by independent and identically distributed (iid) samples from 6.1 with size n. The estimation of Σ is based on the observed values $x_1, \cdots, x_n, x_i = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}$. The likelihood function is then given by

$$L(\boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{np}{2}} \prod_{i=1}^{n} \det(\Sigma)^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-\boldsymbol{\mu})^T\Sigma^{-1}(x-\boldsymbol{\mu})) \tag{6.2}$$

It can be shown that

$$l(\boldsymbol{\mu}, \Sigma) \propto \log\det(\Sigma^{-1}) - tr(S\Sigma^{-1}), \tag{6.3}$$

where $l(\boldsymbol{\mu}, \Sigma)$ is the loglikelihood function and $S = \frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})^T$. Therefore, MLE of Σ is

$$\hat{\Sigma}_{MLE} = \underset{\Sigma>0}{\operatorname{argmax}} \log\det(\Sigma^{-1}) - tr(S\Sigma^{-1}). \tag{6.4}$$

However datasets with much more variables than samples, $n << p$ lead to a sample covariance matrix S that is not of full rank, so its inverse does not exist. This makes equation 6.4 very complex to be solved, and sometimes even inapplicable. In this section we present two methods that deal with this high dimensional problem using regularization penalties and compare them with LPC, described in Chapter 3, which deals with this problem with neighborhood selection.

6.1.1 Graphical lasso

This method was described in [BEGd08, FHT08] and is also known as Graphical Lasso or Glasso. It is an adaptation of neighborhood selection with the lasso [MB06], where the lasso penalty was applied to a standard regression problem to each variable individually and each time using the others as predictors, such as

$$\beta^{i,\lambda} = \underset{\beta}{\operatorname{argmin}}\{n^{-1}\|X_i - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1\}, i \in \mathcal{V}, \tag{6.5}$$

where $\beta = (\beta_1, \cdots, \beta_p)$, $\beta_i = 0$ and λ is the tuning parameter. The neighborhood of each node i is then given by

$$\hat{ne}_i^{\lambda} = \{j \in \mathcal{V} : \hat{\beta}_j^{i,\lambda} \neq 0\} \tag{6.6}$$

This method leads to a sparse graphical model and consistently estimates the set of nonzero elements of Σ^{-1}.

Since the existence of an edge between two vertices is determined by the non-zero elements of the inverse covariance matrix $\Omega = \Sigma^{-1}$, this idea was then implemented more generally in [BEGd08] by applying the l_1-norm penalty to solve the maximum likelihood problem to estimate $\Omega = \Sigma^{-1}$ and reach a graph as sparse as possible.

$$\hat{\Omega} = \underset{\Omega>0}{\operatorname{argmax}} \log\det(\Omega) - tr(S\Omega) - \lambda\|\Omega\|_1. \tag{6.7}$$

It is also shown a dual problem to 6.7

$$\hat{\Omega} = \max\{\log\det W : \|W - S\|_\infty \leq \lambda\}, \tag{6.8}$$

where $W = S + U$ with U a symmetric matrix and $S + U > 0$. To solve 6.8 it was developed also in [BEGd08] a block coordinate algorithm which is slightly changed in [FHT08] to ascertain the equivalence to solving the lasso equations 6.5. The estimation of the covariance matrix using lasso

penalty with the purpose of increasing its sparsity is intuitive and logical. The algorithm can be found in [FHT08] and is implemented in glasso R package.

6.1.2 Network reconstruction using ridge penalty

In [HS14], they propose a new three-step approach to estimate a high-dimensional sparse graph. The first step of this method is to use Ridge penalty to obtain a penalized partial correlation matrix. Then a hypothesis test is applied on a mixture distribution and a few entries are set to zero according to a p-value threshold. Finally, the non-zero partial correlation coefficients are re-estimated. Since this research is focused on reproducing the graph structure, we will only considered the first and second steps. This method will be referred to as GGMridge or Graphical Ridge.

Step 1

Let's assume $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma_p)$. As already seen in Equation 2.2, the partial correlation estimate following the inverse method is given by

$$\hat{P} = -\text{scale}(R^{-1}),$$

where R is the sample correlation matrix and the function scale(.) was defined in Section 2.1.2. Now assume that the data matrix X is standardized to have mean 0 and standard deviation 1 so that $S = \frac{XX^T}{n}$ is the sample correlation matrix. To solve the singularity problem of S when dealing with $n < p$, a positive constant λ is added to the diagonal elements of S, that is,

$$\hat{P_\lambda} = -\text{scale}((S + \lambda I_p)^{-1}), \tag{6.9}$$

where λ is the tuning parameter and I_p is the $p \times p$ identity matrix. It is shown that the partial correlation matrix shrinks towards the identity matrix as λ goes to infinity.

Step 2

The hypothesis testing is done as follows:

1. Fisher's Z-transformation is applied on $\hat{P_\lambda} = (\hat{p}_{ij})$, denoted by $\{\Psi(\hat{p}_{ij})\}$
2. $\{\Psi(\hat{p}_{ij})\}$ are assumed to follow a mixture of null and alternative distributions

$$f(\Psi) = \eta f_0(\Psi) + (1 - \eta) f_a(\Psi) \tag{6.10}$$

   where $f_0(\Psi) \sim N(\mu_0, \sigma_0^2)$ is the null distribution related to $H_0 : \rho_{ij} = 0$, $f_a(\Psi)$ is the alternative distribution related to $H_a : \rho_{ij} \neq 0$ and is left unspecified, and η is the proportion of non-rejected null hypotheses.

3. Efron's central matching method is used to estimate $f_0(\Psi)$ while $f(\Psi)$ is estimated using polynomial Poisson regression.
4. P-values are then calculated for each $\hat{p_{ij}}$.

The non-zero partial correlation estimates are selected according to a threshold α_{Ridge} which is calculated using cross-validation alongside with λ.
$\mathcal{E}_{Ridge} = \{(i,j) \in \mathcal{V} \times \mathcal{V} : \text{p-value}_{Ridge} < \alpha_{Ridge}\}$

6.2 Average of ROC Curves

The Receiver Operating Characteristic curve, commonly called ROC curve, is a plot that indicates the accuracy of a model when compared with the real case. The vertical axis represents the

model sensitivity, and the horizontal axis corresponds to 1− specificity as the p-value threshold $\boldsymbol{\alpha}$ varies from 0 to 1. Sensibility is the percentage of true positives (TP) out of all real positives, that is the sum of true positives and false negatives (TP+FN), while specificity is the percentage of true negatives (TN) out of all real negatives, that is, the sum of true negatives and false positives (TN+FP).

In network reconstruction, positives are existing edges and negatives are non-existing edges. When comparing the reconstructed network with the original structure used to generate data, we can tell which edges have been correctly identified (TP) or not (FP). The same comparison can be done regarding non-existing edges (TN) and (FN). The best models are the ones that combine both high sensitivity and specificity, since models with high sensibility present low type II error and models with high specificity have low type I error. Therefore, the higher the curve is (the higher the area under the curve is) the better as can be seen in Figure 6.1

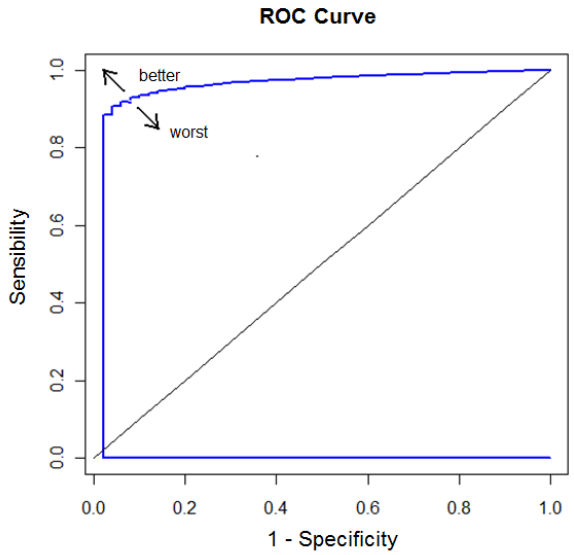


Figure 6.1: *Example of ROC curve*

Assume that $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$, where $\alpha_1 = 0$ and $\alpha_m = 1$, and knowing that $\alpha_{i+1} = \alpha i + \frac{1}{m}$. For each α_i we have a value for specificity and for sensitivity forming corresponding vectors. Each simulation produces a specificity and a sensitivity vector. Lets stack these vectors in a such a way that each column i will correspond to the specificity or sensitivity values produced by α_i. Figure 6.2 shows two matrices where each row is the vector for sensitivity (left matrix) or specificity (right matrix) and the columns are composed by entries of each vector corresponding to a specific α_i. The average of ROC curves is the plot of the mean vectors of specificity and sensibility where each entry i is the mean of column i, that is, mean of sensibility or specificity values produced by α_i in each simulation. Then the standard deviation is added and subtracted from the mean to form a variation range.

	Sensivity					Specificity				
sim	α_1	α_2	α_3	...	α_m	α_1	α_2	α_3	...	α_m
1	*0*		...		*1*	*0*		...		*1*
2	*0*		...		*1*	*0*		...		*1*
3	*0*		...		*1*	*0*		...		*1*
⋮	⋮		⋱		⋮	⋮		⋱		⋮
total sim.	*0*				*1*	*0*				*1*

Figure 6.2: *Example of how sensitivity (left) and specificity (right) vectors are stacked into matrices. The columns are composed by entries of each vector corresponding to a specific α_i.*

6.3 Simulations

Local partial correlation, graphical lasso and GGMridge methods are first applied to data from two different simulation methods: GGM and GNW. Since these simulation methods provide the original graph structure used to generate the data, we can use ROC curves to compare the efficiency of each partial correlation method.

6.3.1 Graphical Gaussian Models

Data was generated from a normal distribution with mean $\boldsymbol{\mu}_p = 0_p$ and covariance matrix Σ_{sim} simulated according to the steps described in 5.6.1. The simulation process was repeated 300 times for each scenario (graph structure, number of nodes and sample size). Erdos Renyi, scale free and small world are the three graph structures selected to perform this study. We started with a small size of network: $p = 50$, $n = 20$. Then we proceed to a slight bigger size: $p = 100$, $n = 50$, and then $p = 200$, $n = 50$. Note that the sample size is always smaller than the number of nodes due to our goal to assess partial correlation methods on high-dimensional data. For each generated dataset, we apply local partial correlation, graphical lasso and GGMridge methods separately leading to three reconstructed networks.

The three methods are compared using ROC curves. The p-value threshold range used to generate the specificity and sensitivity vectors were $\{0.001, \cdots, 1;\ \text{by } 0.005\}$, for GGMridge and $(0.0001, \cdots, 0.4;\ \text{by } 0, 1\} \cup \{0.6, 0.7, 0.8, 0.9, 1\}$ for LPC. Since GLasso does not use p-values to build the sparse graph, the ROC curves were built based on the following range for the lasso tuning parameter λ: $\{0.0001, \cdots, 40;\ \text{by } 0.1\}$. For each method, the average of the ROC curves obtained from the 300 simulations are plotted in Figure 6.3 along with the 1 standard deviation range. We can see that in all graphs the GGMridge curves are higher than the curves corresponding to LPC and GLasso, while these last two curves are very close to each other and sometimes overlap, which means that, although LPC and GLasso methods present similar results, GGMridge has proven to be a slightly better method to be used on GGM simulations.

6.3.2 GNW

Choice of parameters

To avoid incorrect identification of DCPs, we should work with simulated data that adjusts to the methods we use to reconstruct networks. In other words, it would be ideal if the methods we used in this research could reproduce networks similar to the original structures that generated them. As a starting point, a detailed study was performed on wildtype data - gene expression without perturbations - generated by GNW from Ecoli goldstandards. We checked how much the significant correlation coefficients are able to correctly identify the edges present in the original network, also called true positives (TP).

Wildtype data was generated from two different sub-networks for different numbers of variables and samples using two GNW models and several types of noise, which include SDE, microarray, Gaussian and log-normal noise. Table A.1 shows the models, noise and parameters used along with their index number. Next, gene coexpression networks based on Pearson correlation were reconstructed and the percentage of TP were compared.

We can see on Figure 6.4 that there are two groups of parameters that return a higher percentage of true positives in both networks and that, according to Table A.1, those circled groups are modeled with only SDE noise. In addition to this finding, by comparing Figure 6.4(a) with 6.4(b) and Figure 6.4(c) with 6.4(d) we can observe that - with a fixed p - comparing $n < p$ with $n >> p$ shows that the percentage of TP is higher for $n >> p$, as expected. Figure 6.5 illustrates this latter observation further with a bar plot showing the growth of percentage of true positives (TP) as the number of samples n increases from 25 to 1000. Based on the information provided by this study, we have decided to work only with SDE noise of 0.4 - index 6 and 27 in Table A.1.

Comparison of Partial Correlation Methods

The results from GGM simulations in Section 6.3.1 led us to question how the partial correlation methods would perform on these types of data. Considering both percentage of TP and TN (true negatives: non-existing edges in original network).To this purpose the ROC curves for GLasso, GGMridge and lcp methods were calculated and can be found on Figure 6.6. For LPC method we used correlation p-value threshold of 0.1 and 1. The latter was done so that the percentage of TP was not pre-determined by the correlation p-value. The parameters chosen this time was SDE noise of 0.4 for SDE (index 6 in Table A.1) and ODE with SDE (index 27 in Table A.1) models. This choice of parameter was based on Figure 6.4(b) and 6.4(d) where both networks return similar high percentage of TP when $n >> p$. All methods in all charts in Figure 6.6 show poor ROC curves. In Figure 6.6(b) and 6.6(d), where we have higher p and higher n, we can see that the method GLasso has a much worse performance comparing to the other methods.

6.4 Real Data

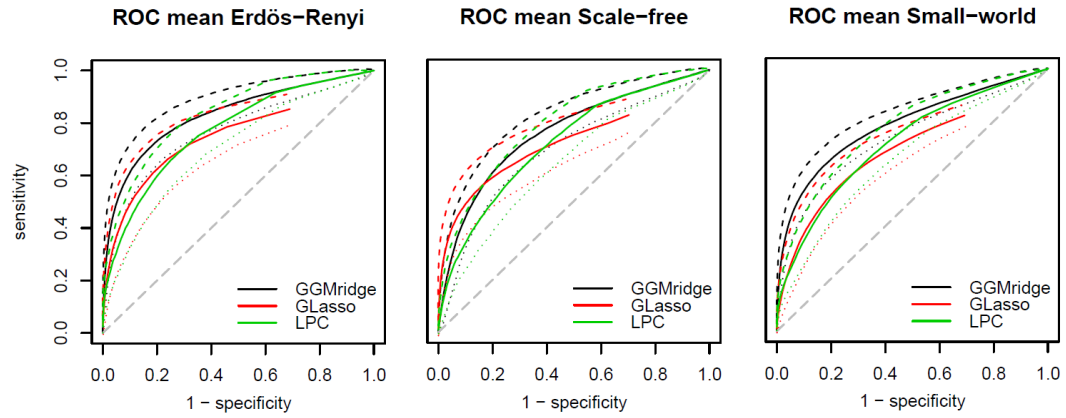
The three partial correlation methods analyzed previously were applied to one of the datasets used in cervical cancer studies in section 5 and collected in [ZKN+07]. This data can be found in "GEO repository" and is already processed by the authors.

This dataset has tens of thousands of genes. However, only the 1268 Differentially Expressed Genes (DEGs) identified by [MSY+13] were considered as starting nodes. Dependencies are checked between every pair of DEGs. We focused on reconstructing the tumor network due to its bigger sample size (10 normal and 21 tumor samples, see Table A.8). All methods returned different networks, that is, the nodes connected to one or more edges are not necessarily the same (nodes with degree 0 are not taken into consideration), and edges may connect different pairs of nodes.

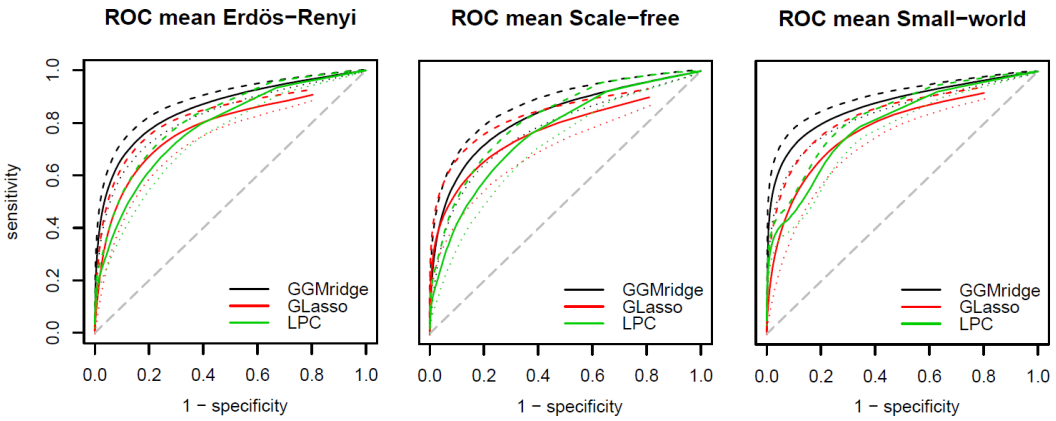
A visual support can be found in Figure 6.7. Figure 6.7(a) illustrates the networks reconstructed using LPC. Next, in Figure 6.7(b), we add the edges and nodes produced by GGMridge that were not present in LPC network. Even though the number of edges increases considerably, we can see that only a few different nodes were added. Finally, we also add the nodes and edges identified by Graphical lasso that are not common neither to LPC or GGMridge networks, Figure 6.7(c). It is visually clear how GLasso generates a network completely different form the other two methods.

The number of connected nodes and edges estimated by all three methods can be found in Venn diagrams in Figures 6.8(a) and 6.8(b) respectively. Besides the 217 nodes present in all three networks, LPC and GGMridge have a lot more nodes in common (343) when compared to GLasso and LPC (54) or GLasso and GGMridge (33). This result indicates that GLasso estimates a network very different from LPC and GGMridge. When considering the intersection of edges, the difference is even more clear. The number of edges present in all three methods is really low (only 14) while t he intersection between LPC and GGMridge shows a much higher number (247) than GLasso and LPC or GLasso and GGMridge (64). The illustration of the union of all three networks can be found in Figure 6.8(c). Here we can see the nodes organized by the intersections observed in Figure 6.8(a).

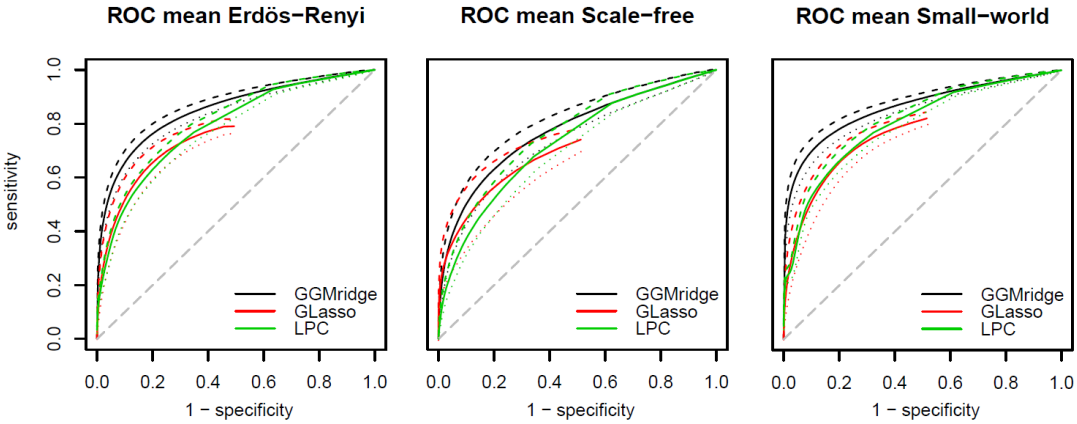
This study shows that there is a greater similarity between the networks produced by LPC and GGMridge than with GLasso. This result gives us support to believe that LPC is an efficient method and that LPC and GGMridge are more trustful methods than GLasso. Much more still need to be done to prove these assumptions, such as applying these on other datasets or comparing networks estimated by the same methods using different datasets.



(a) 50 variables and 20 samples

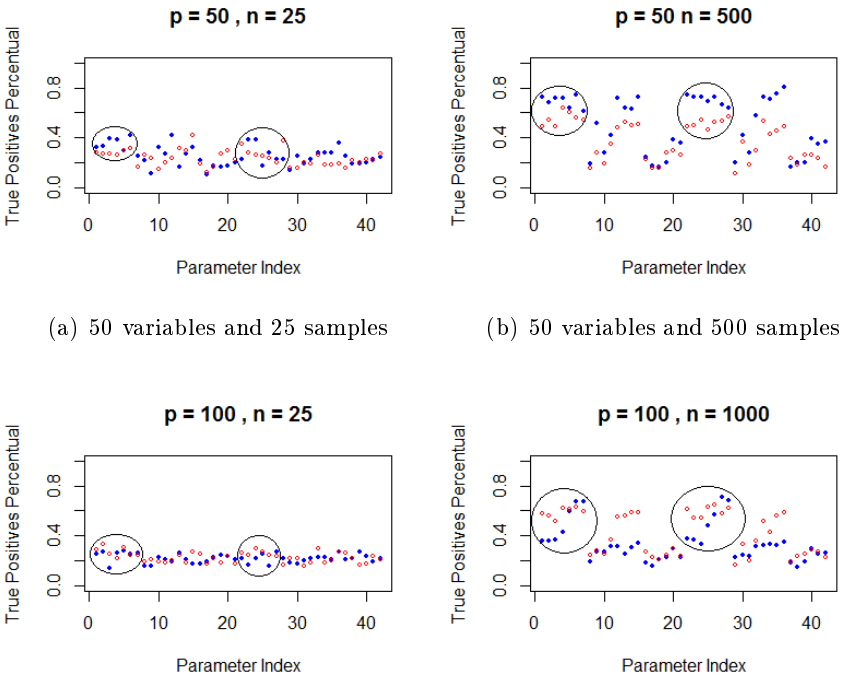


(b) 100 variables and 50 samples



(c) 200 variables and 50 samples

Figure 6.3: *ROC curves from three different partial correlation methods (G-rigde in black, GLasso in red, LPC in in green) applied on data generated from Erdos-Renyi, Scale free and Small world graph structures with: (a) 50 nodes and 20 samples; (b) 100 nodes and 50 samples; and (c) 200 nodes and 50 samples. The straight lines are average ROC curves, while the dashed and dotted lines refer to average plus standard deviation and average minus standard deviation respectively.*



(a) 50 variables and 25 samples

(b) 50 variables and 500 samples

(c) 100 variables and 25 samples

(d) 100 variables and 1000 samples

Figure 6.4: *Study of the percentage of true positives in the reconstruction of two different sub-networks (red and blue) for 42 different GNW parameters. Note that the parameters are indexed as natural numbers and are described in Table A.1. The circled regions of all 4 plots correspond to the data generated with only SDE noise.*

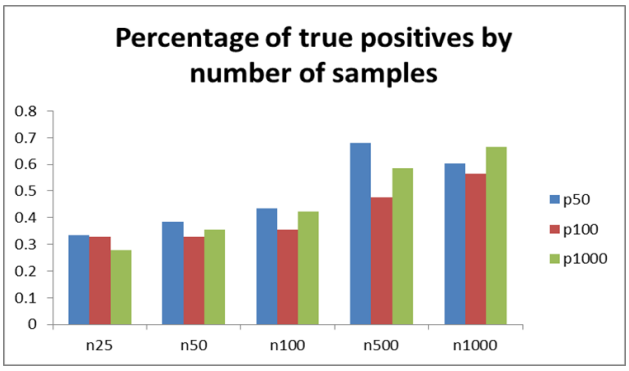
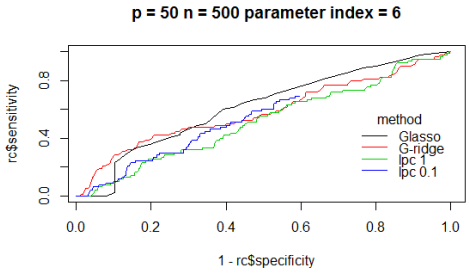
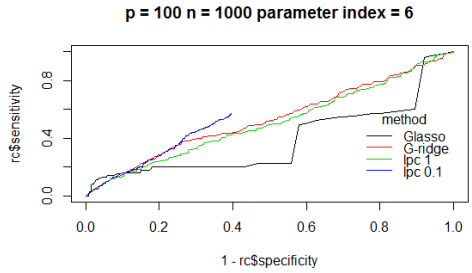


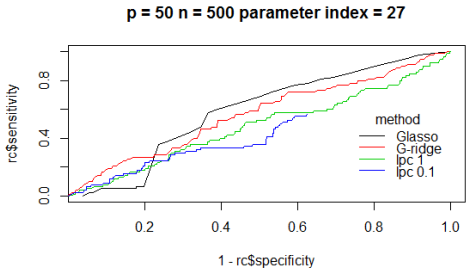
Figure 6.5: *Bar plot showing the growth of percentage of true positives (TP) as the number of samples n increases from 25 to 1000 for three networks with different number of vertices p: 50 (in blue), 100 (in red) and 1000 (in green).*



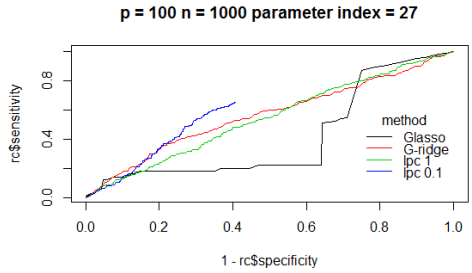
(a) 50 variables and 500 samples: SDE model with SDE noise of 0.4



(b) 100 variables and 1000 samples: SDE model with SDE noise of 0.4

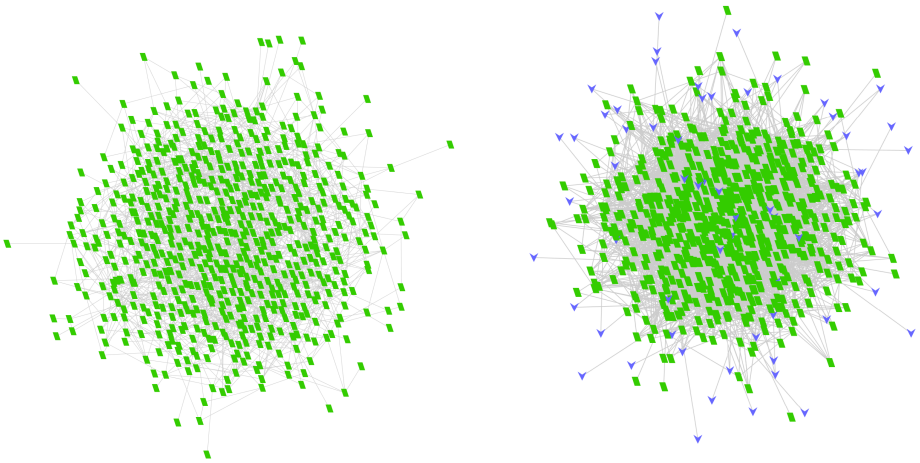


(c) 50 variables and 500 samples: ODE + SDE model with SDE noise of 0.4



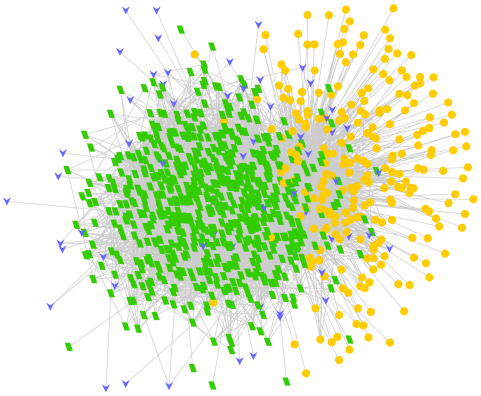
(d) 100 variables and 1000 samples: ODE + SDE model with SDE noise of 0.4

Figure 6.6: *ROC curves of different partial correlation methods (GLasso in black, G-rigde in red, LPC with correlation p-value threshold 1 in green and LPC with correlation p-value threshold 0.1 in blue) in the reconstruction of networks using two different number of variables (50 and 100) and two different number of samples (500 and 1000). Note that in (a) and (c) all the curves overlap close to the identity curve while in (b) and (d) the GLasso ROC curve is much lower than the other curves. Regardless of the overlapping results, all the ROC curves show poor performance.*



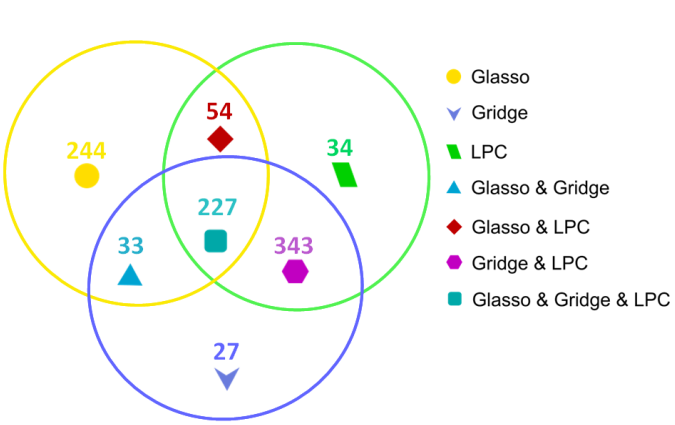
(a) LPC

(b) LPC + GGMridge

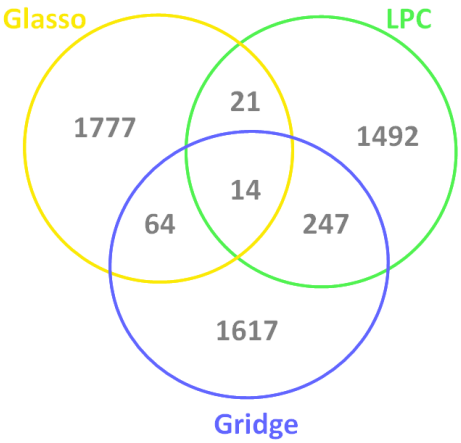


(c) LPC + GGMridge + Lasso

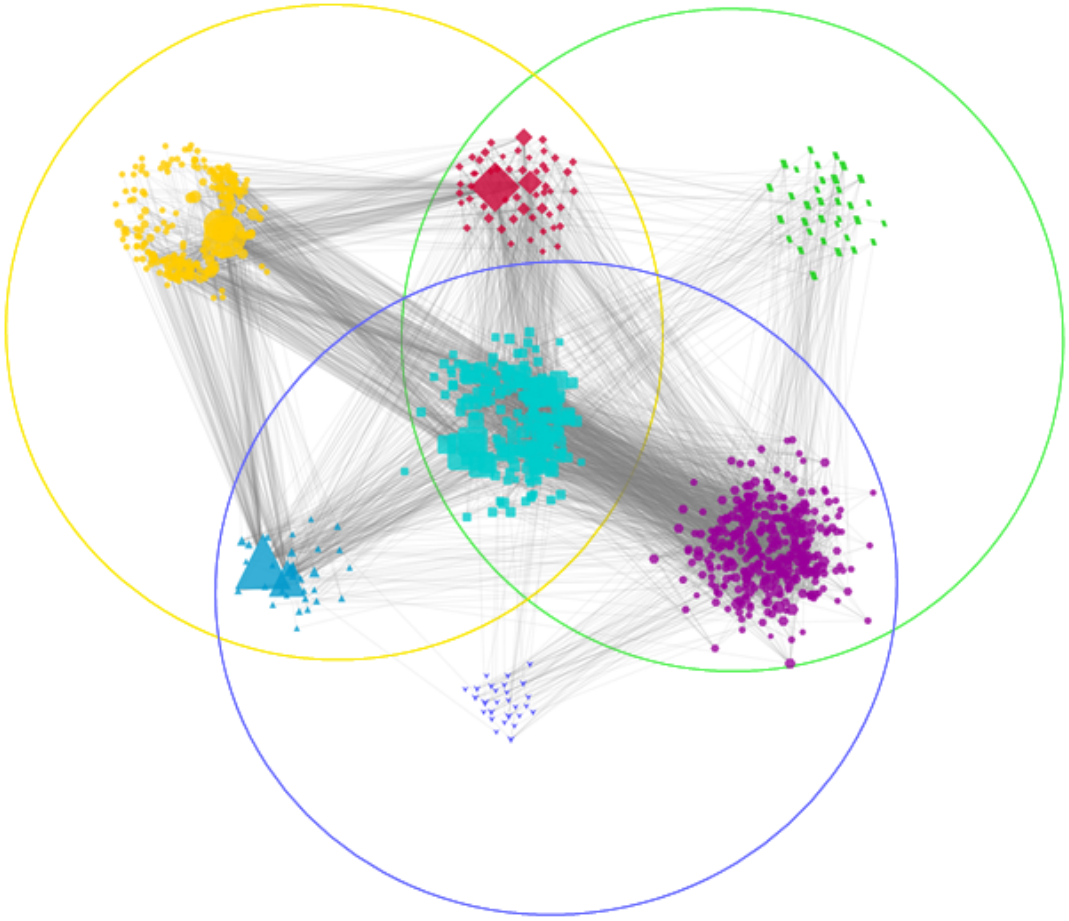
Figure 6.7: *(a) Network built with LPC method; (b) Union of networks built with LPC method and GGMridge: the blue nodes in the network appear after the union of GGMridge network with (a); (c) Union of networks built with LPC method, GGMridge and Graphical Lasso: the orange nodes in the network appear after the union of Graphical Lasso network with (b)*



(a) Number of nodes



(b) Number of edges



(c) Union of all networks

Figure 6.8: *Venn Diagrams. (a) Number of connected nodes in networks reconstructed through the three partial correlation methods herein analyzed: Graphical lasso (yellow circle), LPC (green circle) and GGMridge (blue circle). We can see that the number of nodes in the intersection of LPC and GGMridge (purple hex) is considerably bigger that the other pairwise intersections (red diamond and blue square). (b) Number of edges in common in all three networks. The circle colors are kept the same. The number of similar edges in the intersection of LPC and GGMridge are also higher than the other pairwise intersections. (c) Union of all networks organized in a Venn Diagram. The circle colors are once more kept the same. The nodes in each region represent the numbers in (a).*

Chapter 7

Conclusions

In this work, a methodology for biological network reconstruction was developed in order to answer commonly asked questions made by biological research community. It was applied to two previously published biological data - Bcell knockout and cervical cancer - in order to evaluate the importance of differentially correlated pairs in detecting nodes responsible for state transition. Overall, the obtained results support the idea that differentially co-expressed genes play a regulatory role. While an enrichment of causal genes among DC genes was observed in B cell knockout system, only one causal gene (key driver) was found among DC genes in cervical cancers. On the other hand, the DCPs were located close to genes involved in chromosomal-aberration (perturbations). Moreover, the DCPs identified in cervical cancer have also been found closer to key regulators than other genes in the network, representing "bottlenecks" for communication between driver genes and the rest of the network. Therefore, the idea that genes presenting alterations in correlation patterns between different phenotypes play a critical regulatory role in transitions from one state to another was confirmed.

Difference of correlation after a node or edge knockout was evaluated through a numerical analysis using Graphical Gaussian Models on several types of graph structures. This study demonstrated that the distance from the network perturbation has direct effect on the changes after a network transition. It confirmed that DCPs tend to be close to the perturbation site, specially on scale free graphs. It is known that this type of structure is a good representation of biological networks. Another observed interesting feature of scale free structures is the effect of number of leaves in a pair of nodes on difference of correlation: when comparing pairs with same distance from knockout, the pairs with more leaves seem to present higher difference of correlation. On the other hand, no evidence was found regarding bottlenecks on GGM simulations. Further investigation is necessary to confirm the aforementioned conclusions such as more general simulations and mathematical demonstrations.

Comparison of partial correlation methods was performed in order to evaluate the use of Local Partial Correlation, a new method developed by the Author. GGMridge presented better ROC curves on GGM simulated data, besides producing similar networks reconstructed from real biological data when comparing to LPC methodology. However, since gene expression data is not entirely Gaussian due to its complexity and the existence of external factors, the obtained results support the idea that LPC is an efficient method and that both LPC and GGMridge are more reliable methods than Glasso. Also here more work need to be done to prove these assumptions, such as applying these partial correlation procedures on other datasets, and also comparing networks estimated by the same methods using different datasets.

The overall conclusions of this work are:

- DCPs can be used to identify a group of genes containing nodes responsible for network changes. However, further experimental analysis is necessary to correctly identify some of the real causal genes.

- Local Partial Correlation was proven to be a reliable method to be applied on gene expression

data, and due to the fact that it follows a straightforward methodology, its application is strongly recommended when dealing with small number of variables/genes (e.g. $p \leq 200$). Otherwise, for larger number of variables, the GGMridge method seems to be more appropriate due to its smaller processing time.

Appendix A

Tables

Parameter Index				
Model	Noise			Index
SDE	SDE Noise 0.01			1
SDE	SDE Noise 0.05			2
SDE	SDE Noise 0.1			3
SDE	SDE Noise 0.2			4
SDE	SDE Noise 0.3			5
SDE	SDE Noise 0.4			**6**
SDE	SDE Noise 0.5			7
SDE	SDE Noise 0.01	+	microarray noise	8
SDE	SDE Noise 0.05	+	microarray noise	9
SDE	SDE Noise 0.01	+	Gaussian noise 0.025	10
SDE	SDE Noise 0.01	+	Gaussian noise 0.01	11
SDE	SDE Noise 0.01	+	Gaussian noise 0.001	12
SDE	SDE Noise 0.05	+	Gaussian noise 0.025	13
SDE	SDE Noise 0.05	+	Gaussian noise 0.01	14
SDE	SDE Noise 0.05	+	Gaussian noise 0.001	15
SDE	SDE Noise 0.01	+	log normal noise 0.075	16
SDE	SDE Noise 0.01	+	log normal noise 0.05	17
SDE	SDE Noise 0.01	+	log normal noise 0.025	18
SDE	SDE Noise 0.05	+	log normal noise 0.075	19
SDE	SDE Noise 0.05	+	log normal noise 0.05	20
SDE	SDE Noise 0.05	+	log normal noise 0.025	21
ODE + SDE	SDE Noise 0.01			22
ODE + SDE	SDE Noise 0.05			23
ODE + SDE	SDE Noise 0.1			24
ODE + SDE	SDE Noise 0.2			25
ODE + SDE	SDE Noise 0.3			26
ODE + SDE	SDE Noise 0.4			**27**
ODE + SDE	SDE Noise 0.5			28
ODE + SDE	SDE Noise 0.01	+	microarray noise	29
ODE + SDE	SDE Noise 0.05	+	microarray noise	30
ODE + SDE	SDE Noise 0.01	+	Gaussian noise 0.025	31
ODE + SDE	SDE Noise 0.01	+	Gaussian noise 0.01	32
ODE + SDE	SDE Noise 0.01	+	Gaussian noise 0.001	33
ODE + SDE	SDE Noise 0.05	+	Gaussian noise 0.025	34
ODE + SDE	SDE Noise 0.05	+	Gaussian noise 0.01	35
ODE + SDE	SDE Noise 0.05	+	Gaussian noise 0.001	36
ODE + SDE	SDE Noise 0.01	+	log normal noise 0.075	37
ODE + SDE	SDE Noise 0.01	+	log normal noise 0.05	38
ODE + SDE	SDE Noise 0.01	+	log normal noise 0.025	39
ODE + SDE	SDE Noise 0.05	+	log normal noise 0.075	40
ODE + SDE	SDE Noise 0.05	+	log normal noise 0.05	41
ODE + SDE	SDE Noise 0.05	+	log normal noise 0.025	42

Table A.1: *Index table to indicate all parameters used in the generation of gene expression through GNW. The cells in green correspond to the circled groups in Figure 6.4 while the indices in red are the parameters used to reconstruct the networks in Figure 6.6.*

Gene 1	Gene 2	Change direction	Sign of LPC in BcKO	Sign of LPC in control	Regul. 1	Regul. 2	Nr. of lg genes
HV44_MOUSE	H2-T10	Gained edge	< 0	0	DN	UP	1
Rad54l	Igh-VJ558	Gained edge	< 0	0	UP	DN	1
Paox	Igh-VJ558	Gained edge	< 0	0	UP	DN	1
Parp12	Igh-VJ558	Gained edge	< 0	0	UP	DN	1
Mocos	Igj	Gained edge	< 0	0	UP	DN	1
Trim34	Igj	Gained edge	< 0	0	UP	DN	1
Eif1a	Igk-C	Gained edge	< 0	0	UP	DN	1
Psmb9	Igk-C	Gained edge	< 0	0	UP	DN	1
Mybl2	Igk-C	Gained edge	< 0	0	UP	DN	1
Rnf31	Igk-V1	Gained edge	< 0	0	UP	DN	1
G3bp2	Igk-V1	Gained edge	< 0	0	UP	DN	1
Irf7	Igk-V1	Gained edge	< 0	0	UP	DN	1
Rhof	Igl-V1	Gained edge	< 0	0	UP	DN	1
Batf2	KV2G_MOUSE	Gained edge	< 0	0	UP	DN	1
Batf2	LOC100046894	Gained edge	< 0	0	UP	DN	1
Il15ra	LOC636126	Gained edge	< 0	0	UP	DN	1
Trim15	LOC676193	Gained edge	< 0	0	UP	DN	1
BC006779	LOC676193	Gained edge	< 0	0	UP	DN	1
Muc4	LOC676193	Gained edge	< 0	0	UP	DN	1
Ube2l6	LOC676193	Gained edge	< 0	0	UP	DN	1
Igj	Ppm1k	Gained edge	< 0	0	DN	UP	1
Igl-V1	Ppm1k	Gained edge	< 0	0	DN	UP	1
Chrna6	Serpina1e	Gained edge	< 0	0	UP	DN	0
LOC56304	Trim34	Gained edge	< 0	0	DN	UP	0
Trp53inp1	Zadh1	Gained edge	< 0	0	DN	UP	0
Irf8	9130017N09Rik	Gained edge	> 0	0	UP	UP	0
OTTMUSG00000016644	BC020489	Gained edge	> 0	0	UP	UP	0
Usp18	BC020489	Gained edge	> 0	0	UP	UP	0
Pex14	Eif1a	Gained edge	> 0	0	UP	UP	0
Cdadc1	Ghr	Gained edge	> 0	0	DN	DN	0
Slc6a20a	Ghr	Gained edge	> 0	0	DN	DN	0
Igk-V38	Gm1499	Gained edge	> 0	0	DN	DN	1
Nkg7	Hspa12a	Gained edge	> 0	0	UP	UP	0
Igk-V38	Igh	Gained edge	> 0	0	DN	DN	2
LOC100046894	Igh-VJ558	Gained edge	> 0	0	DN	DN	2
LOC100047628	Igh-VJ558	Gained edge	> 0	0	DN	DN	1
Igh-VJ558	Igk-C	Gained edge	> 0	0	DN	DN	2
LOC100046894	Igk-C	Gained edge	> 0	0	DN	DN	2
Sypl2	Igk-C	Gained edge	> 0	0	DN	DN	1
Gm1420	Igk-C	Gained edge	> 0	0	DN	DN	1

Gm1499	Igk-V1	Gained edge	> 0	0	DN	DN	1
Igh-VJ558	Igk-V1	Gained edge	> 0	0	DN	DN	2
LOC100046894	Igk-V1	Gained edge	> 0	0	DN	DN	2
KV2G_MOUSE	Igk-V1	Gained edge	> 0	0	DN	DN	2
Es22	Igk-V23	Gained edge	> 0	0	DN	DN	1
ENSMUSG00000076532	Igl-V1	Gained edge	> 0	0	DN	DN	1
LOC677445	LOC100046894	Gained edge	> 0	0	DN	DN	2
LOC56304	LOC100046894	Gained edge	> 0	0	DN	DN	1
Stat1	Nkg7	Gained edge	> 0	0	UP	UP	0
Slfn4	Oasl2	Gained edge	> 0	0	UP	UP	0
Ifi44	Oasl2	Gained edge	> 0	0	UP	UP	0
2810022L02Rik	Serpinb9	Gained edge	> 0	0	UP	UP	0
Irf7	Slfn4	Gained edge	> 0	0	UP	UP	0
Igk-C	Zfp467	Gained edge	> 0	0	DN	DN	1
Batf	Cyp2j12	Lost edge	0	< 0	UP	DN	0
Slc6a20a	Zfp295	Lost edge	0	< 0	DN	UP	0
Rps6ka5	Aldh1a1	Lost edge	0	< 0	DN	DN	0
Ube2l6	Casp3	Lost edge	0	< 0	UP	UP	0
Slc6a20a	Cdkn1c	Lost edge	0	< 0	DN	DN	0
Oas2	Cst7	Lost edge	0	< 0	UP	UP	0
Adh1	Cyp2j6	Lost edge	0	< 0	DN	DN	0
Cst7	D14Ertd668e	Lost edge	0	< 0	UP	UP	0
Tcra	D14Ertd668e	Lost edge	0	< 0	UP	UP	0
Cdadc1	EG433023	Lost edge	0	< 0	DN	DN	0
Slc6a20a	EG433023	Lost edge	0	< 0	DN	DN	0
Cxcl9	H2-K1	Lost edge	0	< 0	UP	UP	0
Igtp	H2-K1	Lost edge	0	< 0	UP	UP	0
Tcf7l2	H2-T22	Lost edge	0	< 0	UP	UP	0
1110032O16Rik	HV44_MOUSE	Lost edge	0	< 0	DN	DN	1
Phyhd1	HV44_MOUSE	Lost edge	0	< 0	DN	DN	1
Ube2l6	Pmvk	Lost edge	0	< 0	UP	UP	0
Ppm1j	Rps6ka5	Lost edge	0	< 0	DN	DN	0
Stat1	Sprr2b	Lost edge	0	< 0	UP	UP	0
Ddx60	Tcra	Lost edge	0	< 0	UP	UP	0
Gzma	Tcrb-J	Lost edge	0	< 0	UP	UP	0
Dtx1	Trafd1	Lost edge	0	< 0	UP	UP	0
Slc6a20a	Wdr45	Lost edge	0	< 0	DN	DN	0
Pmvk	Znfx1	Lost edge	0	< 0	UP	UP	0

Table A.2: *Differentially correlated pairs from BcKO study*

Gene symbol	Unique id	Clone	GB acc	UG cluster	Map Location	Regulation	DC gene
Gm1499	6330150	M400005733	0	Mm.305352	6 C1	0.17	1
HV44_MOUSE	6332411	M400008098	0	0	0	0.17	1
Igh	6334530	M400009955	111507	Mm.390473	12 F2|12 58.0 cM	0.17	1
Igh-1a	6334556	M400009956	0	0	0	0.17	0
Igh-6	6309934	M200003304	16019	Mm.342177	12 F1-2|12 58.0 cM	0.17	0
Ighv6-6	6327106	M400002980	238427	Mm.431322	12 F2	0.17	0
Igh-VJ558	6324882	M400000443	16061	0	12	0.17	1
Igj	6336624	M400012178	16069	Mm.1192	5	0.17	1
Igk-C	6334911	M400010428	0	0	0	0.17	1
Igk-V1	6323886	M300020618	16081	Mm.333124	6 30.0 cM	0.17	1
Igkv1-117	6333580	M400009425	16098	Mm.304143	6 C1|6 30.0 cM	0.0053	0
Igk-V23	6333584	M400009427	111735	Mm.423821	6 C1|6 30.0 cM	0.17	1
Igk-V38	6328662	M400004225	16120	Mm.288753	6 C1|6 30.0 cM	0.17	1
Igkv4-90	6330232	M400006035	434034	Mm.424510	6 C1	0.017	0
Igl-V1	6322228	M300013112	16142	Mm.326349	16 A3|16 13.0 cM	0.017	1
KV2G_MOUSE	6334909	M400010427	0	0	0	0.017	1
KV5A_MOUSE	6335026	M400010640	0	0	0	0.017	0
LOC100043977	6333077	M400008547	100043977	0	12	0.017	0
LOC100046546	6329615	M400005124	100046546	0	6	0.017	0
LOC100046894	6330789	M400006358	100046894	0	6	0.017	1
LOC636126	6333520	M400008850	636126	0	0	0.017	1
LOC636988	6330206	M400005749	636988	0	0	0.017	0
LOC675659	6332528	M400008311	675659	0	0	0.017	0
LOC676175	6328842	M400004552	676175	0	6	0.017	0
LOC676193	6328597	M400004394	676193	0	6	0.017	1
LOC677445	6330757	M400006354	677445	0	0	0.017	1

Table A.3: *Causal genes: BcKO*

Gene symbol	ID	GB_ACC	SPOT_ID	Gene Title	RefSeq Transcript ID	DC gene
NAT13	217745_s_at	NM_025146	NM_025146	NM_025146	80218	0
NAA50 217745_s_at	NM_025146	NM_025146	NM_025146	80218	0	
MCM2	202107_s_at	NM_004526	NM_004526	NM_004526	4171	0
TOPBP1	202633_at	NM_007027	NM_007027	NM_007027	11073	0
CEP70	1554488_at	BC016050	BC016050	BC016050	80321	1
GMPS	214431_at	NM_003875	NM_003875	NM_003875	8833	0
RFC4	204023_at	NM_002916	NM_002916	NM_002916	5984	0
LAMP3	205569_at	NM_014398	NM_014398	NM_014398	27074	0
CKS1B	201897_s_at	NM_001826	NM_001826	NM_001826	1163	0
DUSP12	218576_s_at	NM_007240	NM_007240	NM_007240	11266	0
NEK2	204641_at	NM_002497	NM_002497	NM_002497	4751	0
PSMB4	202243_s_at	NM_002796	NM_002796	NM_002796	5692	0
ADAR	201786_s_at	NM_001111	NM_001111	NM_001111	103	0
DTL	218585_s_at	NM_016448	NM_016448	NM_016448	51514	0
AIM2	206513_at	NM_004833	NM_004833	NM_004833	9447	0
EXO1	204603_at	NM_003686	NM_003686	NM_003686	9156	0
RFX5	202963_at	AW027312	AW027312	AW027312	5993	0
CDCA8	221520_s_at	BC001651	BC001651	BC001651	55143	0
CDC20	202870_s_at	NM_001255	NM_001255	NM_001255	991	0
S100PBP	218370_s_at	NM_022753	NM_022753	NM_022753	64766	0
IFI44L	204439_at	NM_006820	NM_006820	NM_006820	10964	0
RPA2	201756_at	NM_002946	NM_002946	NM_002946	6118	0
ITGB3BP	205176_s_at	NM_014288	NM_014288	NM_014288	23421	0
IFI44	214059_at	BE049439	BE049439	BE049439	10561	0
DEPDC1	220295_x_at	NM_017779	NM_017779	NM_017779	55635	0
HMGN2	208668_x_at	BC003689	BC003689	BC003689	3151	0
ISG15	205483_s_at	NM_005101	NM_005101	NM_005101	9636	0
NUP155	206550_s_at	NM_004298	NM_004298	NM_004298	9631	0
TPX2	210052_s_at	AF098158	AF098158	AF098158	22974	0
TYROBP	204122_at	NM_003332	NM_003332	NM_003332	7305	0
RAD54B	219494_at	NM_012415	NM_012415	NM_012415	25788	0
LAPTM4B	1554679_a_at	AF317417	AF317417	AF317417	55353	0
KPNA2	201088_at	NM_002266	NM_002266	NM_002266	3838	0
MMP9	203936_s_at	NM_004994	NM_004994	NM_004994	4318	0
BIRC5	202094_at	AA648913	AA648913	AA648913	332	0
RAD21	200607_s_at	BG289967	BG289967	BG289967	5885	0
MTFR1	203207_s_at	BF214329	BF214329	BF214329	9650	0

Table A.4: *Causal genes: Cervical Cancer*

	DEGs	DCPs	DEGs Correlation network
Missing allowed	30% max (BRB)	30% max	Inherited from DEGs
Mean p-value (BRB)	0.05		Inherited from DEGs
Regulation	same in all studies		Inherited from DEGs
Correlation p-value		if p-value > 0.2, marked as NOT significantly correlated	< 0.2
Correlation direction		same in all studies present for at least one state	Same in all studies, for each separate state
Difference of correlation p-value		< 0.1	
Minimum number of studies present	2 out of 2	2 out of 2	Inherited from DEGs
Sample size		> 2	
Difference of correlation direction		2 out of 2	
FDR on Fisher p-value	< 0.1	< 0.02	< 0.025
PUC			Applied after FDR filter
Procedure for duplicate Gene symbol		Select pair with lower difference of correlation fisher p-value. Remove daps that have the same gene symbol combination but different probe ids and have different change of correlation direction	If Different probe IDs have the same Gene symbol, they are going to be interpreted as the same gene in the network.

Table A.5: *All filters for all calculations on BcKO data*

	DCPs	DEGs Correlation network	DEGs Local Partial Correlation
Missing allowed	30% max	30% max in separate states	
Correlation p-value	if pv > 0.2, marked as NOT significantly correlated	< 0.1	Significant in DEGs Correlation network
Correlation direction	same in all studies present for at least one state	Same in all studies, for each separate state	
Local Partial correlation p-value			< 0.4
Local Partial correlation direction			Same in all studies, for each separate state
Difference of correlation p-value	< 0.1		
Minimum number of studies present	3 out of 5		
Sample size	> 2		
Difference of correlation direction	same in all studies present		
FDR on Fisher p-value	< 0.0025	$< 10 - 8$	< 0.05
PUC		Applied after FDR filter	
Procedure for duplicate Gene symbol	Select pair with lower difference of correlation fisher pv.	Select pair with lower correlation fisher pv. (Nothing done when they show different directions in correlation)	Select pair with higher correlation. (Nothing done when they show different directions in correlation)

Table A.6: *All filters for all calculations on cervical cancer data*

Reference	Accesion Number	Strain	# normal mice	# BcKO mice	Array Platform	Approximate number of transcripts
Shulzhenko et al, 2011	GSE23573	B10.A littermates	12	12	NIAID Mmca – Mouse	38K
Shulzhenko et al, 2011	GSE23573	BALB/c	10	10	NIAID Mmca – Mouse	38K

Table A.7: *Datasets included in the meta-analysis of gene expression microarray data for Bcell Knockout.*

Reference	Accesion Number	# normal tissue samples	# tumor tissue samples	Array Platform	Approximate number of transcripts
Mine et al, 2013	GSE26342	20	40	In house, NIAID, NIH	14K
Biewenga et al, 2008	GSE7410	5	35	Agilent-012391 G4112A	41K
Scotto et al, 2008	GSE9750	21	32	Affymetrix HG-U133A	39K
Pyeon et al, 2007	GSE6791	8	20	Affymetrix HG-U133_Plus_2	47K
Zhai et al, 2007	GSE7803	10	21	Affymetrix HG-U133A	39K

Table A.8: *Datasets included in the meta-analysis of gene expression microarray data for cervical cancer*

Appendix B

Algorithms and Scripts

B.1 Graphical structures

B.1.1 Edge Percent Random Tree

```
1   random.edge.percent.adj = function(p1, eta){
2   p = ceiling(eta*(p1*(p1-1))/2)
3   if(p<p1) stop("not enough edges. please choose a higher eta")
4   # assigning at least one edge for each node
5   A = matrix(data = 0, nrow = p1, ncol = p1)
6   i=1
7   while(i <= p1){
8     temp3 = which(apply(A,2,sum) > 0)
9     if(length(temp3) == 0)
10      temp2=sample(c(1:p1)[-i],1,replace = T) else{
11      if(length(temp)>1){
12        if(length(temp)>2)
13          temp2=sample(c(1:p1)[-c(i,temp3)],1,replace = T) else
14            if(length(temp)==2)
15              temp2=temp[-which(temp==i)]
16      }
17           else
18            if(length(temp)==1)
19              temp2=sample(c(1:p1)[-i],1,replace = T)
20      }
21
22    A[temp2,i]=1
23    A[i,temp2]=1
24    temp = which(apply(A,2,sum) == 0)
25    if(length(temp) > 0)
26      i = temp[1] else
27        i = p1+1
28  }
29
30  if(sum(A > 1)) stop("adjacency matrix with entries = 2. Correct this error!")
31
32  temp=A[upper.tri(A)]==0
33  tri.w = which(upper.tri(A))
34  tri.w=temp*tri.w
35  sig = ceiling(eta * length(tri.w))- (sum(A)/2)
36  tri.w=tri.w[tri.w!=0]
37
38  sig.node = sample(tri.w, sig)
39  A[lower.tri(A)]=0
40  A[sig.node] = 1
41  A = A + t(A)
42
43  return(A)
44 }
```

B.1.2 Galton Watson Tree

```
1
2 tree.gw=function(p, meanlog=0, sdlog=1){
3   A=data.frame(matrix(0, nrow = p, ncol = p), stringsAsFactors = F)
4   i=1
5   j=1
6   while (j<p) {
7     # nf=ceiling(rgeom(1,0.5))
8     nf= ceiling( rlnorm(1,meanlog = meanlog, sdlog = sdlog))
9     if(j+nf > p) nf=p-j
10    A[(j+1):(j+nf),i]=rep(1,nf)
11    j=j+nf
12    i=i+1
13  }
14  A=A+t(A)
15  return(A)}
```

B.1.3 Lattice with leaves

```
1
2 graphELeafs <- function(i,j,leafs=0) {
3   # verify input variables
4   if (i <= 0 || j <=0) {
5     print('error: i or j cannot be less than zero')
6     return (0)
7   }
8
9   # calculate the dimension of the adjacency matrix (max rows and colunns)
10  # that represents the number of max leafs
11  max_leafs = i*j
12
13  # validate the number of extra leafs
14  if(leafs <= 0) {
15    leafs = max_leafs
16  } else
17  if (leafs > max_leafs) {
18    leafs = max_leafs
19    cat('warning: leafs changed to ',max_leafs,'\n')
20  }
21
22  gLattice = make_lattice(dimvector = c(i,j))
23  mAdjacency = get.adjacency(gLattice)
24
25  # generate a new sparse matrix
26  mLeafs = Matrix(0, nrow=max_leafs, ncol=leafs, sparse=TRUE)
27
28  # generate a random sequence of numbers without repetition
29  # as we permutate the values from 1 to max_leafs (dimension)
30  new_values = sample(1:max_leafs)
31
32  for(i in 1:leafs) {
33    mLeafs[new_values[i],i] = 1
34  }
35  aMatrix = cbind2(mAdjacency, mLeafs)
36
37  # generate a sparse matrix to fill the remaining rows to complete a square
        matrix
38  mRows = Matrix(0, nrow=leafs, ncol=(max_leafs+leafs), sparse=TRUE)
39
40  # convert the matrix to a square matrix
41  final_aMatrix = rbind(aMatrix, mRows)
42
```

```
43   # generate the final graph
44   final_graph = graph_from_adjacency_matrix(final_aMatrix, mode=c("undirected"))
45
46   return (final_graph)}
```

B.2 Algorithms for Chapter 4

B.2.1 Algorithm for calculation partial correlations.

Algorithm for calculation partial correlations. Implementation of partial correlation is straightforward in R using the function cor2pcor from package corpcor. The input of this function is a correlation matrix, which should be a "positive definite", a mathematically required property. However, in the omics data it is common that the correlation matrix is not positive definite because thousands of variables are measured in tens or hundreds of samples. Thus, to apply the inverse method we should make the estimation of the covariance matrix positive definite. For this, we use shrinkage estimation for covariance matrix which is implemented in R in the same package corpcor.

```
1 #X is the data matrix where columns represent genes/variables and rows represent
      samples/individuals
2
3 C = cor(X)
4 C = cor2pcor(C)
```

B.2.2 Algorithm for Meta-analysis scheme

1. Select only genes (or pairs of DEGs in case of networks) with the same direction of difference of mean (or correlation in case of networks) throughout all data sets. Each gene or gene-gene correlation should pass a certain p-value threshold. This threshold controls for heterogeneity between datasets. 2. For each possible gene (or pair of DEGs in case of networks), calculate the Fisher p-values using the following function created in R:

```
1 # pvalue is a matrix where columns respresent pvalue vectors for each dataset
2
3 >pv.meta.analysis=function(pvalue){
4 # calculate fisher statistics
5 sum = log(pvalue[, 1])
6 for(j in 2 : ncol(pvalue)) sum = sum + log(pvalue[, j])
7 t=-2*(sum)
8
9 # calculate fisher pvalue
10 pv_fish=1-pchisq(t, 2*ncol(pvalue))
11 return(pv_fish)}
```

Compute FDR over the obtained vector of Fisher p-values using the R function

```
1 pv=pv.meta.analysis(pvalue)
2 pv.fdr=p.adjust(pv, method = "fdr")
```

3.Select pairs with FDR less than a threshold.

B.3 Algorithm for DCPs identification

1. Calculate a matrix of pairwise correlation $C1$ and $C2$ for two groups of genes expression matrices $X1$ and $X2$ (number of columns is number of samples and rows are genes)

```
1 C1 = cor(X1)
2 C2 = cor(X2)
```

2. We compute the p-values of difference of correlation using the function pv.dif.cor.pearson developed by the auther using function fisherz from the R package psych. Let n1 and n2 be the number of samples for corresponding groups (n1 is the number of rows in X1, n2 is the number of rows in X2).

```
1 pv.dif.cor.pearson = function(C1,C2,n1,n2){
2 if(!"psych" %in% installed.packages())
3 install.packages("psych")
4 library(psych)
```

```
5 # Convert correlations to z-scores
6 z1 = fisherz(C1)
7 z2 = fisherz(C2)
8 # Calculate vector of t-tests to compare
9 # correlations between classes
10 fisher = (z1 - z2) / sqrt((1/(n1 - 3)) + (1/(n2 - 3)))
11 # Calculate raw p-values
12 pv.dif.cor = 2*pt(-abs(fisher),Inf)
13 return(pv.dif.cor)}
```

3. Compute FDR over the obtained vector of Fisher p-values using the R function. Select pairs with FDR less than a threshold.

```
1 pv=pv. dif.cor.pearson(C1,C2,n1,n2)
2 pv.fdr=p.adjust(pv, method = "fdr")
```

Bibliography

[AGC+09] I. Amit, M. Garber, N. Chevrier, A. P. Leite, Y. Donner, T. Eisenhaure, M. Guttman, J. K Grenier, W. Li, O. Zuk et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 326(5950):257–263, 2009. 15

[AJS82] G. K. Ackers, A. D. Johnson e M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences*, 79(4):1129–1133, 1982. 9

[ASS13] D. Amar, H. Safer e R. Shamir. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol*, 9(3):e1002955, 2013. 23, 34

[BA99] Albert-László Barabási e Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. 6

[BAS+04] R. C. Borra, P. M. Andrade, I. D. C. G Silva, A. Morgun, L. L. M. Weckx, A. S. Smirnova e M. Franco. The th1/th2 immune-type response of the recurrent aphthous ulceration analyzed by cdna microarray. *Journal of oral pathology & medicine*, 33(3):140–146, 2004. 17

[BB13] B. Barzel e A.-L. Barabási. Network link prediction by global silencing of indirect correlations. *Nature biotechnology*, 31(8):720–725, 2013. 18

[BBI+06] S. Balaji, M.M. Babu, L.M. Iyer, N.M. Luscombe e L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of molecular biology*, 360(1):213–227, 2006. 9

[BBM+08] Petra Biewenga, Marrije R Buist, Perry D Moerland, Emiel Ver Loren van Themaat, Antoine HC van Kampen, Fiebo JW ten Kate e Frank Baas. Gene expression in early stage cervical cancer. *Gynecologic oncology*, 108(3):520–526, 2008. 24

[BEGd08] O. Banerjee, L. El Ghaoui e A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008. 41, 42

[BEL+05] E. Björck, S. Ek, O. Landgren, M. Jerkeman, M. Ehinger, M. Björkholm, C.A.K. Borrebaeck, A. Porwit-MacDonald e M. Nordenskjöld. High expression of cyclin b1 predicts a favorable outcome in patients with follicular lymphoma. *Blood*, 105(7):2908–2915, 2005. 29, 30, 34

[BH95] Y. Benjamini e Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, páginas 289–300, 1995. 24

[BHJ+04] J. A. Berger, S. Hautaniemi, A.-K. Järvinen, H. Edgren, S. K. Mitra e J. Astola. Optimized lowess normalization parameter selection for dna microarray data. *BMC bioinformatics*, 5(1):1, 2004. 17

[BPS+03] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara et al. Arrayexpress-a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, 31(1):68–71, 2003. 19

[BS09] Ed Bullmore e Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. 1

[BSS04] K. Baba, R. Shibata e M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, 46:657–664, 2004. 7, 11

[Buh93] S. L. Buhl. On the existence of maximum likelihood estimators for graphical gaussian models. *Scandinavian Journal of Statistics*, páginas 263–270, 1993. 11

[BW91] E.M. Berman e L.M. Werbel. The renewed potential for folate antagonists in contemporary cancer chemotherapy. *Journal of medicinal chemistry*, 34(2):479–485, 1991. 29, 30, 34

[BY97] Yaneer Bar-Yam. *Dynamics of complex systems*, volume 213. Addison-Wesley Reading, MA, 1997. 1

[CAT+14] J. C Chen, M. J. Alvarez, F. Talos, H. Dhruv, G. E. Rieckhof, A. Iyer, K. L. Diefes, K. Aldape, M. Berens, M. M. Shen et al. Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell*, 159(2):402–414, 2014. 15

[CCK15] N.G. Cost e M.F. Czyzyk-Krzeska. Regulation of autophagy by two products of one gene: Trpm3 and mir-204. *Molecular & Cellular Oncology*, 2(4):e1002712, 2015. 29, 30, 34

[CFF+08] R. Caspi, H. Foerster, C. A Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36(suppl 1):D623–D631, 2008. 19

[CHBA03] Reuven Cohen, Shlomo Havlin e Daniel Ben-Avraham. Structural properties of scale-free networks. *Handbook of graphs and networks*, 2003. 6

[CKK09] S. B. Cho, J. Kim e Ju H. Kim. Identifying set-wise differential co-expression in gene expression microarray data. *BMC bioinformatics*, 10(1):1, 2009. 23, 34

[CKP12] D.-Y. Cho, Y.-A. Kim e T. M. Przytycka. Network biology approach to complex diseases. *PLoS Comput Biol*, 8(12):e1002820, 2012. 23, 34

[CLZ+16] T Tony Cai, Weidong Liu, Harrison H Zhou et al. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016. 34

[CN06] G. Csardi e T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006. 34

[CS99] Paul Cilliers e David Spurrett. Complexity and post-modernism: Understanding complex systems. *South African Journal of Philosophy*, 18(2):258–274, 1999. 1

[CYYK05] J. K. Choi, U. Yu, O. J. Yoo e S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355, 2005. 23, 34

[DGP05] M. Dettling, Ed. Gabrielson e G. Parmigiani. Searching for differentially expressed gene combinations. *Genome biology*, 6(10):1, 2005. 23, 34

[dlF10] A. de la Fuente. From differential expression to differential networking - identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7):326–333, 2010. 2, 23, 34

[DLFBHM04] A. De La Fuente, N. Bing, I. Hoeschele e P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004. 18

[DRA+13] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013. 17

[DS02] M.H. DeGroot e M.J. Schervish. *Probability and Statistics*. Addison-Wesley, 3 edição, 2002. 13

[DSB03] S. Dudoit, J. P. Shaffer e J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, páginas 71–103, 2003. 17

[DSJ+10] D. L. Diamond, A. J. Syder, J. M. Jacobs, C. M Sorensen, K.-A. Walters, S. C. Proll, J. E. McDermott, M. A. Gritsenko, Q. Zhang, R. Zhao et al. Temporal proteome and lipidome profiles reveal hepatitis c virus-associated reprogramming of hepatocellular metabolism and bioenergetics. *PLoS Pathog*, 6(1):e1000719, 2010. 26

[DYCS02] S. Dudoit, Y. H. Yang, M. J. Callow e T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, páginas 111–139, 2002. 1, 23

[DYK12] J. A. Dawson, S. Ye e C. Kendziorski. R/ebcoexpress: an empirical bayesian framework for discovering differential co-expression. *Bioinformatics*, 28(14):1939–1940, 2012. 1, 23, 34

[DYR+15] X. Dong, A. Yambartsev, S. A. Ramsey, L. D. Thomas, N. Shulzhenko e A. Morgun. Reverse engeneering of regulatory networks from big data: A roadmap for biologists. *Bioinformatics and biology insights*, 9:61, 2015. 2, 23, 30

[EDL02] R. Edgar, M. Domrachev e A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002. 19

[Edw95] D. Edwards. *Introduction to Graphical Modelling*. Springer-Verlag New York, Inc., 1 edição, 1995. 7, 11, 13

[ER59] P ERDdS e A R&WI. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959. 6

[FHT08] J. Friedman, T. Hastie e R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. 1, 3, 18, 41, 42, 43

[Fis24] R. A. Fisher. The distribution of the partial correlation coefficien. *Metron*, 3:329–332, 1924. 13

[Fis25] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925. 24

[FMMK13] S. Feizi, D. Marbach, M. Médard e M. Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733, 2013. 18

[Fre78] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978. 26

[FSM+09] E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. H. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. O. Wakelam e E. A. Dennis. Update of the lipid maps comprehensive classification system for lipids. *Journal of lipid research*, 50(Supplement):S9–S14, 2009. 19, 29

[Fuk13] A. Fukushima. Diffcorr: an r package to analyze and visualize differential correlations in biological networks. *Gene*, 518(1):209–214, 2013. 1, 23, 34

[GCB+04] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1, 2004. 17

[GCJJPG+08] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado et al. Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic acids research*, 36(suppl 1):120–124, 2008. 9

[GLT+13] D. Ghosh, Z. Li, X. F. Tan, T. K. Lim, Y. Mao e Q. Lin. itraq based quantitative proteomics approach validated the role of calcyclin binding protein (cacybp) in promoting colorectal cancer metastasis. *Molecular & Cellular Proteomics*, 12(7):1865–1880, 2013. 29, 30, 34

[HCH+14] D.P. Hall, N.G. Cost, S. Hegde, E. Kellner, O. Mikhaylova, Y. Stratton, B. Ehmer, W.A. Abplanalp, R. Pandey, J. Biesiada et al. Trpm3 and mir-204 establish a regulatory circuit that controls oncogenic autophagy in clear cell renal cell carcinoma. *Cancer cell*, 26(5):738–753, 2014. 29, 30, 34

[HJG17] Yangbo He, Jinzhu Jia e Zhi Geng. Structural learning of causal networks. *Behaviormetrika*, páginas 1–19, 2017. 1

[HKJ+07] D.J. Hunter, P. Kraft, K.B. Jacobs, D.G. Cox, M. Yeager, S.E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson et al. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, 39(7):870–874, 2007. 29, 30, 34

[HLN+10] H.M. Horlings, C. Lai, D.S.A. Nuyten, H. Halfwerk, P. Kristel, E. van Beers, S.A. Joosse, C. Klijn, P.M. Nederlof, M.J.T. Reinders et al. Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clinical Cancer Research*, 16(2):651–663, 2010. 29, 30, 34

[HQG+09] R. Hu, X. Qiu, G. Glazko, L. Klebanov e A. Yakovlev. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. *BMC bioinformatics*, 10(1):1, 2009. 23, 34

[HRR+05] D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. De Atauri, J. D. Aitchison, L. Hood, A. F. Siegel et al. A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17296–17301, 2005. 20

[HS14] M.J. Ha e W. Sun. Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. *Biometrics*, 70(3):762–770, 2014. 1, 3, 18, 34, 41, 43

[HSA06] Cindy E Hmelo-Silver e Roger Azevedo. Understanding complex systems: Some core challenges. *The Journal of the learning sciences*, 15(1):53–61, 2006. 1

[HTF09] T. Hastie, R. Tibshirani e J. Friedman. 14.3. 12 hierarchical clustering the elements of statistical learning. edn, 2009. 23

[JBN+14] A. Jauhiainen, M. Basetti, M. Narita, M. Narita, J. Griffiths e S. Tavaré. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics*, página btu175, 2014. 17

[JMC13] I. S. Jang, A. Margolin e A. Califano. haracne: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface focus*, 3(4):20130011, 2013. 18

[JSP01] J.-H. Jang, K.-H. Shin e J.-G. Park. Mutations in fibroblast growth factor receptor 2 and fibroblast growth factor receptor 3 genes associated with human gastric and colorectal cancers. *Cancer Research*, 61(9):3541–3543, 2001. 29, 30, 34

[JTA+00] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai e A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000. 6

[JTB+03] N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell e J. K Lee. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, 19(15):1945–1951, 2003. 17

[Kat08] M. Katoh. Cancer genomics and genetics of fgfr2 (review). *International journal of oncology*, 33(2):233–237, 2008. 29, 30, 34

[KCHS15] Y. Kwak, H. Cho, W. Hur e T. Sim. Antitumor effects and mechanisms of azd4547 on fgfr2-deregulated endometrial cancer cells. *Molecular cancer therapeutics*, 14(10):2292–2302, 2015. 29, 30, 34

[KR09] L. Kaufman e P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 23

[KS04] D. Kostka e R. Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20(suppl 1):i194–i199, 2004. 20, 23, 34

[Lau96] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996. 7

[Li02] K.C. Li. Genome-wide coexpression dynamics: theory and application. Em *Proceedings of the National Academy of Sciences*, volume 99, páginas 16875–16880. National Acad Sciences, 2002. 1, 23, 34

[LJB+08] T. Landemaine, A. Jackson, A. Bellahcène, N. Rucci, S. Sin, B.M. Abad, A. Sierra, A. Boudinet, J.-M. Guinebretière, E. Ricevuto et al. A six-gene signature predicting breast cancer lung metastasis. *Cancer Research*, 68(15):6092–6099, 2008. 29, 30, 34

[LPO+07] H.-J. Lee, J.-O. Pyo, Y. Oh, H.-J. Kim, S.-h. Hong, Y.-J. Jeon, H. Kim, D.-H. Cho, H.-N. Woo, S. Song et al. Ak2 activates a novel apoptotic pathway through formation of a complex with fadd and caspase-10. *Nature cell biology*, 9(11):1303–1310, 2007. 29, 30, 34

[LPS+12] Jacob L., Neuvial P., Dudoit S. et al. Package degraph, 2012. 23, 34

[LWCZ04] Y. Lai, B. Wu, L. Chen e H. Zhao. A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics*, 20(17):3146–3155, 2004. 1, 23, 34

[LWLC07] W. K. Lim, K. Wang, C. Lefebvre e A. Califano. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, 23(13):i282–i288, 2007. 17

[MB06] N. Meinshausen e P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, páginas 1436–1462, 2006. 12, 41, 42

[MCK+12] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012. 18

[MDD+15] A. Morgun, A. Dzutsev, X. Dong, R. L. Greer, D. J. Sexton, J. Ravel, M. Schuster, W. Hsiao, P. Matzinger e N. Shulzhenko. Uncovering effects of antibiotics on the host and microbiota using transkingdom gene networks. *Gut*, páginas gutjnl–2014, 2015. 15, 23, 28

[MH14] P. J. McMurdie e S. Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531, 2014. 17, 26

[MHJ+05] L. Martens, H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove e R. Apweiler. Pride: the proteomics identifications database. *Proteomics*, 5(13):3537–3545, 2005. 19

[MKB79] K. V. Mardia, J. T. Kent e J. M. Bibby. *Multivariate Analysis*. Academic Press, London, UK, 10 edição, 1979. 7, 11

[MLW+08] K. M. Mani, C. Lefebvre, K. Wang, W. K. Lim, K. Basso, R. Dalla-Favera e A. Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. *Molecular systems biology*, 4(1):169, 2008. 23, 34

[MNB+06] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera e A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006. 18

[MPS+10] D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano e G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. Em *Proceedings of the National Academy of Sciences*, volume 107, páginas 6286–6291. National Acad Sciences, 2010. 9

[MRG+11] M. R. Morris, C. J. Ricketts, D. Gentle, F. McRonald, N. Carli, H. Khalili, M. Brown, T. Kishida, M. Yao, R. E. Banks et al. Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma. *Oncogene*, 30(12):1390–1401, 2011. 29, 30

[MSMF09] D. Marbach, T. Schaffter, C. Mattiussi e D. Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239, 2009. 9

[MSY+13] K. L. Mine, N. Shulzhenko, A. Yambartsev, M. Rochman, G. F. O. Sanson, M. Lando, S. Varma, J. Skinner, N. Volfovsky, T. Deng et al. Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. *Nature communications*, 4:1806, 2013. 2, 15, 17, 20, 23, 24, 28, 33, 34, 46

[MWM+08] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer e B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008. 17

[New04] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004. 26

[NHDQ11] M. Needham, R. Hu, S. Dwarkadas e X. Qiu. Hierarchical parallelization of gene differential association analysis. *BMC bioinformatics*, 12(1):374, 2011. 1, 23, 34

[NSH+07] X. Ning, S. Sun, L. Hong, J. Liang, L. Liu, S. Han, Z. Liu, Y. Shi, Y. Li, W. Gong et al. Calcyclin-binding protein inhibits proliferation, tumorigenicity, and invasion of gastric cancer. *Molecular Cancer Research*, 5(12):1254–1262, 2007. 29, 30, 34

[NYW+10] F. Nie, X.-L. Yu, X.-G. Wang, Y.-F. Tang, L.-L. Wang e L. Ma. Down-regulation of cacybp is associated with poor prognosis and the effects on cox-2 expression in breast cancer. *International journal of oncology*, 37(5):1261–1269, 2010. 29, 30, 34

[Pan02] Wei Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002. 17

[PDMS07] A. Perez-Diez, A. Morgun e N. Shulzhenko. Microarrays for cancer diagnosis and classification. Em *Microarray Technology and Cancer Gene Profiling*, páginas 74–85. Springer, 2007. 17

[Pea01] J. Pearl. Direct and indirect effects. Em *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 2001. 17, 18

[Pea10] J. Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2), 2010. 18

[PNL+07] Dohun Pyeon, Michael A Newton, Paul F Lambert, Johan A Den Boon, Srikumar Sengupta, Carmen J Marsit, Craig D Woodworth, Joseph P Connor, Thomas H Haugen, Elaine M Smith et al. Fundamental differences in cell cycle deregulation in human papillomavirus–positive and human papillomavirus–negative head/neck and cervical cancers. *Cancer research*, 67(10):4605–4619, 2007. 24

[PS99] C. A. B. Pereira e J. M. Stern. Evidence and credibility: Full bayesian significance test for precise hypotheses. *Entropy*, 1:99–110, 1999. 13

[PSS+13] T.E. Pronk, E.P. Someren, R.H. Stierum, J. Ezendam e J.L.A. Pennings. Unraveling toxicological mechanisms and predicting toxicity classes with gene dysregulation networks. *Journal of Applied Toxicology*, 33(12):1407–1415, 2013. 1, 23, 34

[PTVF07] W. H. Press, S. A. Teukolsky, W. T. Vetterling e B. P. Flannery. Section 16.4. hierarchical clustering by phylogenetic trees. *Numerical Recipes: The Art of Scientific Computing*, páginas 868–881, 2007. 23

[RO10] M. D Robinson e A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1, 2010. 17

[RYB03] A. Reiner, D. Yekutieli e Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003. 1, 19, 23

[RYS+04] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey e A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9309–9314, 2004. 20

[SBR+06] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz e M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006. 19

[SKV+11] J. Skinner, Y. Kotliarov, S. Varma, K.L. Mine, A. Yambartsev, R. Simon, Y. Huyen e A. Morgun. Construct and compare gene coexpression networks with dapfinder and dapview. *BMC bioinformatics*, 12(1):286, 2011. 1, 10, 17, 20, 23, 24, 33, 34

[SLL+07] R. Simon, A. Lam, M.-C. Li, M. Ngan, S. Menenzes e Y. Zhao. Analysis of gene expression data using brb-array tools. *Cancer informatics*, 3, 2007. 17

[SMH+11] N. Shulzhenko, A. Morgun, W. Hsiao, M. Battle, M. Yao, O. Gavrilova, M. Orandle, L. Mayer, A.J. Macpherson, K.D. McCoy et al. Crosstalk between b lymphocytes, microbiota and the intestinal epithelium governs immunity versus metabolism in the gut. *Nature medicine*, 17(12):1585–1593, 2011. 2, 17, 23, 26, 33, 34

[SMS05] G. K. Smyth, J. Michaud e H. S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005. 17

[SNL+07] S. Sun, X. Ning, J. Liu, L. Liu, Y. Chen, S. Han, Y. Zhang, J. Liang, K. Wu e D. Fan. Overexpressed cacybp/sip leads to the suppression of growth in renal cell carcinoma. *Biochemical and biophysical research communications*, 356(4):864–871, 2007. 29, 30, 34

[SNN+08] Luigi Scotto, Gopeshwar Narayan, Subhadra V Nandula, Hugo Arias-Pulido, Shivakumar Subramaniyam, Achim Schneider, Andreas M Kaufmann, Jason D Wright, Bhavana Pothuri, Mahesh Mansukhani et al. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes, Chromosomes and Cancer*, 47(9):755–765, 2008. 24

[SS+05a] J. Schäfer, K. Strimmer et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005. 20, 41

[SS05b] Juliane Schäfer e Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005. 41

[Str01] Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001. 1

[SYAP11] E. Shin, Y. Yoon, J. Ahn e S. Park. Tc-vgc: a tumor classification system using variations in genes' correlation. *Computer methods and programs in biomedicine*, 104(3):87–101, 2011. 1, 23, 34

[SYC+11] P. Sumazin, X. Yang, H.-S. Chiu, W.-J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva et al. An extensive microrna-mediated network of rna-rna interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–381, 2011. 15

[SYGP+05] N. Shulzhenko, A. Yambartsev, A. Goncalves-Primo, M. Gerbase-DeLima e A. Morgun. Selection of control genes for quantitative rt-pcr based on microarray data. *Biochemical and biophysical research communications*, 337(1):306–312, 2005. 17

[TFY12] L.D. Thomas, V. Fossaluza e A. Yambartsev. Building complex networks through classical and bayesian statistics - a comparison. Em *XI Brazilian Meeting on Bayesian Statistics*, volume 1490, páginas 323–331. AIP Conf. Proc., Março 2012. 18, 25, 29

[Tho12] L.D. Thomas. ConstruÃ§Ã£o de redes usando estatÃstica clÃ¡ssica e bayesiana- uma comparaÃ§Ã£o. Dissertação de Mestrado, Instituto de MatemÃ¡tica e EstatÃstica, Universidade de SÃ£o Paulo, Brasil, 2012. 1, 2, 8, 11, 12, 13, 41

[TVS+16] L. D. Thomas, D. Vyshenska, N. Shulzhenko, A. Yambartsev e A. Morgun. Differentially correlated genes in co-expression networks control phenotype transitions. *F1000Research [version 1; referees: 1 approved, 1 approved with reservations]*, 5:2740, 2016. 3, 26

[VAE+15] H. M. Vedeld, K. Andresen, I. A. Eilertsen, A. Nesbakken, R. Seruca, I. P. Gladhaug, E. Thiis-Evensen, T. O. Rognum, K. M. Boberg e G. E. Lind. The novel colorectal cancer biomarkers cdo1, zscan18 and znf331 are frequently methylated across gastrointestinal cancers. *International Journal of Cancer*, 136(4):844–853, 2015. 29, 30, 34

[Wat06] M. Watson. Coxpress: differential co-expression in gene expression data. *BMC bioinformatics*, 7(1):1, 2006. 1, 20, 23, 34

[Whi09] J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009. 18

[WLC+10] B.A. Walker, P.E. Leone, L. Chiecchio, N.J. Dickens, M.W. Jenner, K.D. Boyd, D.C. Johnson, D. Gonzalez, G.P. Dagrada, R.K.M. Protheroe et al. A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, 116(15):e56–e65, 2010. 29, 30, 34

[WS98] Duncan J Watts e Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998. 6

[WTK+07] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney et al. Hmdb: the human metabolome database. *Nucleic acids research*, 35(suppl 1):D521–D526, 2007. 19

[XFG+04] Y. Xiao, R. Frisina, A. Gordon, L. Klebanov e A. Yakovlev. Multivariate search for differentially expressed gene combinations. *BMC bioinformatics*, 5(1):1, 2004. 1, 23, 34

[YPK+16] A. Yambartsev, M. A. Perlin, Y. Kovchegov, N. Shulzhenko, K. L. Mine, X. Dong e A. Morgun. Unexpected links reflect the noise in networks. *Biology Direct*, 11(1):52, 2016. 19, 24

[YSH+13] D. Yang, Y. Sun, L. Hu, H. Zheng, P. Ji, C. V. Pecot, Y. Zhao, S. Reynolds, H. Cheng, R. Rupaimoole et al. Integrated analyses identify a master microrna regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer cell*, 23(2):186–199, 2013. 15

[ZKN+07] Yali Zhai, Rork Kuick, Bin Nan, Ichiro Ota, Stephen J Weiss, Cornelia L Trimble, Eric R Fearon e Kathleen R Cho. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies hoxc10 as a key mediator of invasion. *Cancer research*, 67(21):10163–10172, 2007. 24, 46

[ZS08] Y. Zhao e R. Simon. Brb-arraytools data archive for human cancer gene expression: a unique and efficient data sharing resource. *Cancer informatics*, 6, 2008. 17