Inference of gene networks from time series expression data and application to type 1 Diabetes

Miguel Lopes

Machine Learning Group (Faculté des Sciences, Départment d'Informatique)
ULB Center for Diabetes Research

Université Libre de Bruxelles

A thesis submitted for the degree of Doctor of Philosophy (PhD)

2015

This thesis has been written under the supervision of the Prof. Gianluca Bontempi and Décio Eizirik. The members of the jury are:

- Prof. Gianluca Bontempi (Université Libre de Bruxelles)
- Prof. Décio Eizirik (Université Libre de Bruxelles)
- Prof. Tom Lenaerts (Université Libre de Bruxelles)
- Prof. Maarten Jansen (Université Libre de Bruxelles)
- Prof. Kris Laukens (Universiteit Antwerpen))

This thesis has been written by the author, containing original work of his. Some described work is a product of a collaborative effort, whose contributors are acknowledged.

Abstract

The inference of gene regulatory networks (GRN) is of great importance to medical research, as causal mechanisms responsible for phenotypes are unravelled and potential therapeutical targets identified. In type 1 diabetes, insulin producing pancreatic beta-cells are the target of an auto-immune attack leading to apoptosis (cell suicide). Although key genes and regulations have been identified, a precise characterization of the process leading to beta-cell apoptosis has not been achieved yet. The inference of relevant molecular pathways in type 1 diabetes is then a crucial research topic.

GRN inference from gene expression data (obtained from microarrays and RNA-seq technology) is a causal inference problem which may be tackled with well-established statistical and machine learning concepts. In particular, the use of time series facilitates the identification of the causal direction in cause-effect gene pairs. However, inference from gene expression data is a very challenging problem due to the large number of existing genes (in human, over twenty thousand) and the typical low number of samples in gene expression datasets. In this context, it is important to correctly assess the accuracy of network inference methods.

The contributions of this thesis are on three distinct aspects. The first is on inference assessment using precision-recall curves, in particular using the area under the curve (AUPRC). The typical approach to assess AUPRC significance is using Monte Carlo, and a parametric alternative is proposed. It consists on deriving the mean and variance of the null AUPRC and then using these parameters to fit a beta distribution approximating the true distribution. The second contribution is an investigation on network inference from time series. Several state of the art strategies are experimentally assessed and novel heuristics are proposed. One is a fast approximation of first order Granger causality scores, suited for GRN inference in the large variable case. Another identifies co-regulated genes (ie. regulated by the same genes). Both are experimentally validated using microarray and simulated time series. The third contribution of this thesis is on the context of type 1 diabetes and is a study on beta cell gene expression after exposure to cytokines, emulating

the mechanisms leading to apoptosis. 8 datasets of beta cell gene expression were used to identify differentially expressed genes before and after 24h, which were functionally characterized using bioinformatics tools. The two most differentially expressed genes, previously unknown in the type 1 Diabetes literature (RIPK2 and ELF3) were found to modulate cytokine induced apoptosis. A regulatory network was then inferred using a dynamic adaptation of a state of the art network inference method. Three out of four predicted regulations (involving RIPK2 and ELF3) were experimentally confirmed, providing a proof of concept for the adopted approach.

Résumé

L'inférence des réseaux de régulation de gènes (GRN) est d'une grande importante pour la recherche médicale, car les mécanismes de causalité responsable des phénotypes sont détectés et les cibles thérapeutiques potentielles identifiées. Dans le diabète de type 1, les cellules beta du pancréas qui produisent l'insuline sont la cible d'une attaque auto-immune conduisant à l'apoptose (suicide cellulaire). Bien que des gènes et régulateurs clés ont été identifiés, une caractérisation précise du processus conduisant à l'apoptose des cellules beta n'a pas encore été accomplie. L'inférence des voies moléculaires importantes pour le diabète de type 1 est un sujet crucial de recherche.

L'inférence des GRN à partir des données d'expression de gènes (obtenues par microarray et par la technologie du RNA-seq) est un problème d'inférence causale qui pourrait tre résolue par des concepts bien établis de statistiques et de apprentissage automatique. En particulier, l'usage des séries temporelles facilite l'identification de la direction causal dans les paires de gènes cause-effet. Cependant, l'inférence à partir des donnés d'expression de gène est un problème très difficile du au large nombre de gènes existants (plus de 20000 chez l'homme) et le problème typique du peu d'échantillons disponible. Dans ce contexte, il est important d'évaluer correctement la prcision des méthodes d'inférence des réseaux.

La contribution de cette thèse porte sur trois aspects distincts. La première porte sur l'évaluation de l'inférence en utilisant les *precision-recall curves*, en particulier en utilisant l'aire sous la courbe (AUPRC). L'approche typique pour évaluer la significativité du AUPRC est d'utiliser Monte Carlo et une alternative paramétrique est proposée. Il consiste à dériver la moyennes et la variance de l'hypothèse nulle de l'AURPC et ensuite de les utiliser pour construire une distribution beta qui estime la vraie distribution. La deuxième contribution est une investigation sur les l'inférence des réseaux à partir des séries temporelles. Plusieurs stratégies actuelles sont expérimentalement évaluées et de nouveau heuristiques sont proposées. L'une est une approximation rapide des scores de causalités de Granger du premier ordre, appropriée pour l'inférence des GRN dans le cas de large variable. Une autre identifie les gènes co-régulés (i.e régulés par les mmes gènes). Les deux sont expérimentalement validées en utilisant les données de microarray et les séries temporelles simulées. La troisième contribution de cette thèse porte sur le diabètes de type 1 et est une étude sur l'expression des cellules beta exposées aux cytokines, qui simulent les mécanismes conduisant à l'apoptose. Huit jeux de données d'expression de gènes des cellules beta ont été utilisés pour identifier les gènes différentiellement exprimés avant et après 24 heures qui ont ensuite ont été caractérisés fonctionnellement en utilisant les outils bioinformatiques. Les 2 gènes les plus différentiellement exprimés, précédemment inconnus dans la littérature sur le diabètes de type 1 (RIPK2 et ELF3) ont été analysés et jouent un rle dans l'apoptose induite par les cytokines. Un réseau de régulation a été inferré en utilisant une adaptation dynamique des méthodes actuelles d'inférence de réseaux. Trois sur quatre de régulations prédites (impliquant RIPK2 et ELF3) ont été expérimentalement confirmées, donnant ainsi une *proof of concept* de l'approche que nous avons adoptée.

To my dear family, old friends and new.

Acknowledgements

This thesis was funded by the ARC project *Discovery of the molecular pathways regulating pancreatic beta cell dysfunction and apoptosis in diabetes using functional genomics and bioinformatics*, based on a collaboration between the Machine Learning Group (Computer Science Department, ULB) and the ULB Center for Diabetes Research. The objectives of the project are the identification of relevant genes and regulatory pathways in β-cell apoptosis in diabetes; characterization of their role on cell function and fate; and experimental validation of predicted hypothesis, through bioinformatics and molecular biology techniques. The work here presented was supervised by the Professors Gianluca Bontempi and Décio Eizirik.

I would like to thank my supervisors Gianluca and Décio for the invaluable guidance throughout these years as a PhD candidate; my friends and colleagues at the Machine Learning Group and the Center for Diabetes Research for all the enriching personal and professional interactions; Andrea, Elisa and Catharina for comments and suggestions regarding this thesis; and the members of the jury for kindly accepting to read and critically assess this work. Finally, I would also like to thank all my family and friends for all the support, availability, friendship and love.

Contents

List of Figures

List of Tables

Nomenclature

T1D	Type 1 Diabetes, Section 1.1
DNA, RNA	Nucleic acids, Section 1.2
Transcription and translation	Transcription refers to the generation of mRNA from DNA, translation refers to the generation of proteins from mRNA, Section 1.2
GRN	Gene regulatory network, Section 1.3
Microarrays , RNA-seq	Gene expression measurement technology, Section 1.4
Bias-variance trade off	Known phenomenon in statistics/machine learning, where the bias of an estimation decreases at the cost of a variance increase (and vice versa), Section 1.6
MSE	Mean squared error, Section 1.6
Filters, wrappers, embedded methods	Approaches to variable selection, Section 1.8
CPP	Common cause principle, Section 1.9.1
Pr-curve	Precision-recall curve, Section 1.10
AUPRC	Area under the precision-recall curve, Section 1.10
ROC curve	Receiving operator characteristic curve, Section 1.10
AUROC	Area under the receiving operator characteristic, Section 1.10
OLS	Ordinary least squares, method to estimate linear regression models, Section2.2.1
Lasso	L^1 norm regularization of linear models, Section 2.2.4
Lars	Least angle regression, a forward selection procedure closely linked with the Lasso, Section 2.2.4
SEM	Structural equation model, Section 2.3
Markov property and faithfulness	These properties state the association between separation in graphical models and dependences in probability distributions, section 2.4
BN	Bayesian network, Section 2.4
GGM	Gaussian graphical model, Section 2.4.3
MB	Markov blanket of a node in a graphical model, Section 2.4.5
RF	Random forests, Section 2.6
VAR	Vector auto-regressive model, Section 2.7.2
GC	Granger causality, Section 2.7.4
RSS	Residual sum of squares, Section 2.7.4
GC-TY	Modified GC test, Section 2.7.5
DBN	Dynamic Bayesian network, Section 2.7.6
GNW	GeneNetWeaver - simulator of gene expression data, Section 5.3
AIC, BIC	Akaike and Bayesian information criterion, appendix A.1
X	Random variables, or sets of random variables, are denoted by an upper case letter
X_t	Random variable or set of variables X at time point t
$X_{i,t}$	Random variable X_i at time point t
$\dot{X}_t$	Time derivative of X at time point t
X_i	Random variables may be distinguished by subscript/index - in this case denoting the i-th element on a set X
$X^{\setminus i}$	Sets of random variables may be denoted by superscript - in particular the symbol $\setminus$ denotes the set X excluding the elements after it - denoted by index

$X^{\setminus Z}$	The set X excluding the elements in Z
x_i	i-th sample (observation) of the variable X
$x_{i,j}$	j-th sample of the variable X_i
$\mathbf{x_i}$	Vector of samples of the variable X_i, the j-th element being $x_{i,j}$
$\mathbf{X}$	Matrix of samples of the set of variables X, the i-th column being $\mathbf{x_i}$
n and p	In the context of a sample matrix, p is the number of variables (columns) and n is the number of samples per variable (lines)
$\perp\!\!\!\perp$	Symbol for independence ($X \perp\!\!\!\perp Y$ means X and Y are independent)
$\hat{\theta}$	Estimation of a parameter θ
$\mathcal{P}(X = x)$	General symbol for probability or probability distribution, depending on the context
$\mathcal{P}(X = x \mid Y = y)$	Conditional probability that $X = x$, given that $Y = y$
$p(x)$	Probability mass function of a discrete variable X
$f(x)$	Probability density function of a continuous variable X
$p(x \mid y)$	Conditional probability function of a discrete variable x, given $Y = y$
S_x	Support of the random variable X
$\mathbb{E}(X)$	Expected value of a variable X
$H(X)$	Entropy of variable X
$H(X \mid Y)$	Conditional entropy of X, given Y
MI	Mutual information
$I(X, Y)$	Mutual information of X and Y, given Z
$I(X, Y \mid Z)$	Conditional mutual information of X and Y, given Z
σ_X^2	Variance of X
σ_X	Standard deviation of X
$\sigma_{X,Y}$	Covariance of X and Y
$\rho_{X,Y}$	Pearson (linear) correlation between X and Y
$\rho_{X,Y\mid Z}$	Conditional linear correlation between X and Y, given Z
Σ	Covariance matrix of a set of variables X. $\Sigma_{i,j}$ corresponds to σ_{X_i,X_j} and $\Sigma_{i,i}$ corresponds to $\sigma_{X_i}^2$
Ω	Inverse of the covariance matrix (known as the concentration matrix)
$\omega_{i,j}$	[i,j]-th element of the concentration matrix Ω
β	This symbol denotes linear coefficients. Depending on the context, it may denote a single value, or a vector. β_i may denote the $[i]$-th element of a vector of linear coefficients.
B	This symbol denotes a matrix of linear coefficients, used in Section 2.7
$\mathcal{P}_{sel}(k, n, N, P)$	Probability that the k-th positive selected instance is the n-th selected, when sampling without replacement N instances, out of which P are positive, used in Chapter 4
GC3	First order conditional Granger causality score, obtained with the method described in Section 5.2
G_{iy}	Outcome of a Granger causality test from $X_{i,t}$ to $X_{y,t}$, usually in the form of a standard score, Section 5.2.1
$G_{iy\mid z}$	Outcome of a conditional Granger causality test from $X_{i,t}$ to $X_{y,t}$, given $X_{z,t}$, Section 5.2.1

Part I

Introduction

1

Introduction

The contributions of this thesis focus on three aspects: 1. assessment of gene regulatory network (GNR) inference using precision recall curves; 2. GRN inference from time series; 3. knowledge inference in the context of pancreatic β-cell dysfunction and death in Type 1 Diabetes. In particular, we propose a novel approach to assess the significance (in the form of a p-value) of the AUPRC of GRN inference when a gold standard of regulations is available. Second, we propose novel algorithms for GRN inference from time series, designed to deal with the high variable to sample ratio, and setup an experimental session where several aspects of GRN inference from time series are assessed and the novel methods validated. Finally, we perform a meta-analysis using eight datasets of gene expression after β-cell exposure to pro-inflammatory cytokines, emulating the biological mechanisms leading to β-cell apoptosis in Type 1 Diabetes. A GRN is inferred, and previously unknown causal regulations experimentally validated.

The phenotypical characteristics of living beings are encoded in molecules called DNA (deoxyribonucleic acid), long chains of information of a double helix structure (deciphered by Watson and Crick (271)). DNA molecules are organized in structures called chromosomes, and are constituted by individual blocks (genes) which regulate the production of proteins inside cells (63). The complete set of DNA of an organism is called its genome. Gene DNA sequences are the basic units of inheritance in evolution, although properties other than DNA sequences may be temporarily transmitted trans-generationally (ie. epigenetic changes (80)). Variants of genes (alleles) are responsible for the phenotypical diversity within a species. Phenotypes are usually a result of complex networks between different interacting molecules. As an example, the Figure 1.1 describes the regulatory network leading to apoptosis, a mechanism of cell suicide following extra- or intra-cellular cues. It represents different types of molecules and regulations (eg. inhibitory and stimulatory).

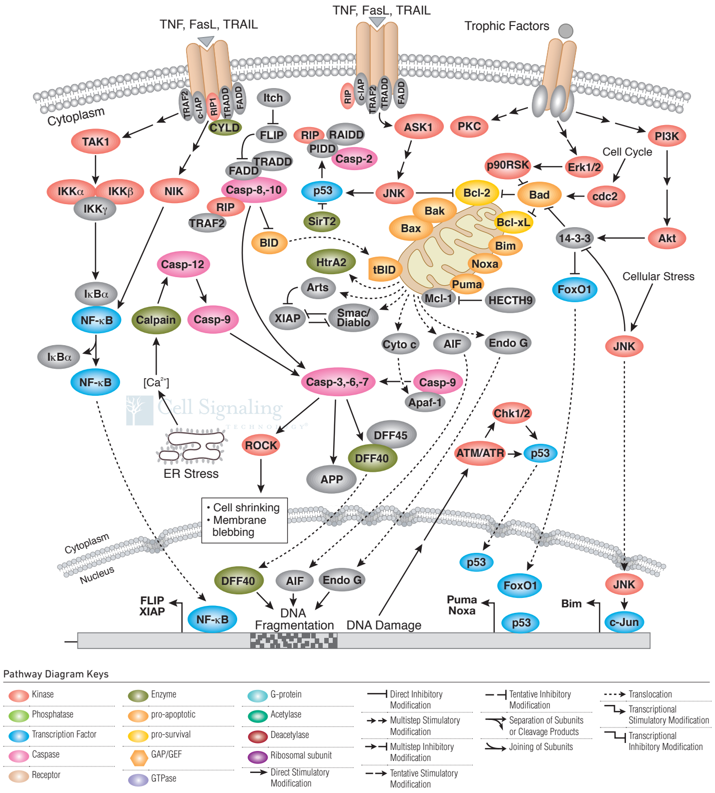


Figure 1.1: Regulation of Apoptosis Signaling Pathway - Image from Cell Signaling technologies, http://www.cellsignal.com/contents/science-pathway-research-apoptosis/regulation-of-apoptosis-signaling-pathway/pathways-apoptosis-regulation

Regulatory networks in which each element corresponds to a gene (or is a product of it) are known as Gene Regulatory Networks (GRNs, introduced in Section 1.3). GRN discovery is helpful to medical research, allowing for the identification of key genes in diseases or prediction of effects of gene targeting treatments. In type 1 diabetes, insulin producing pancreatic β-cells are the target of an auto-immune attack leading to local inflammation and β-cell apoptosis (cell suicide). This process has been shown to be regulated by the activity of key genes, but its precise characterization remains illusive.

Bioinformatics is a recent scientific field concerned with the analysis of biological data through computational and statistical methods. Topics in bioinformatics include biological data acquisition, organization, storage, accessibility and representation, and inference of knowledge from it, such as associations between genes and phenotypes. A bioinformatics landmark was the sequencing of the near-whole human genome in the Human Genome project (127). In humans, around 20000 protein-coding genes have been identified, constituting less than 2% of the human genome (280). In the last two decades, microarrays and RNA-seq technology (see Section 1.4) have made possible the measurement of the activity of thousands of genes in parallel.

Microarrays and RNA-seq have opened up the possibility of GRN inference from gene expression data. This is a problem of inferring causal relationships between variables (genes) from observational data. Using time series data facilitates causal inference as the realization of a cause induces an effect at a later time point. In practice, GRN inference is problematic due to the typical high variable to sample ratio (ie. thousands of genes to much fewer gene expression observations, see Section 1.7.1). This limitation is even more serious in time series where most datasets are composed of a few dozen samples at best. Currently, large scale gene expression measurement is still expensive and time consuming.

This thesis concerns GRN inference from gene expression data, in particular from time series, and its application to type 1 Diabetes. Its original contributions are threefold. The first contribution is on the assessment of inferred GRN when a gold standard is available (a set of known or putative regulations), in particular using the area under the precision recall curve (AUPRC) as the assessment criterion. The second contribution is on methodological aspects regarding GRN inference from (short size) time series. The third is on knowledge discovery (including GRN inference and experimental validation of predicted regulations) in the context of type 1 Diabetes. These are summarized in Section 1.11).

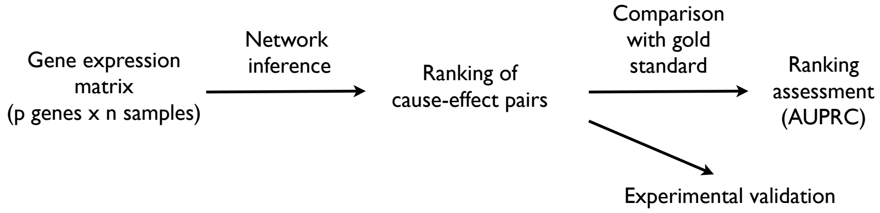

Gene expression matrix (p genes x n samples)
Network inference
Ranking of cause-effect pairs
Comparison with gold standard
Ranking assessment (AUPRC)
Experimental validation

Figure 1.2: Network inference and assessment - From a gene expression matrix, composed of n observations of p genes, pairwise regulations are ranked. If a gold standard of regulations is available, the ranking can be directly assessed (for instance, via the AUPRC). Alternatively, predicted regulations may be subject to experimental validation.

The inference and assessment aspects addressed in this work are illustrated in the Figure 1.2. This first chapter introduces gene networks, the acquisition of gene expression data, causal inference and inference assessment. Chapters 2 and 3 present preliminaries and state of the art of network inference.

Chapter outline The remainder of this chapter is as follows: Section 1.1 introduces type 1 Diabetes; Section 1.2 introduces basic notions of biology, including GRN and gene expression measurement; Section 1.7 describes the problem of GRN inference and its limitations; Sections 1.8 and 1.9 introduce relevant notions for GRN inference, in particular on variable selection and causal inference; Section 1.10 discusses GRN inference assessment; Section 1.11 summarizes the main contributions of this thesis.

1.1 Type 1 diabetes

Type 1 Diabetes (T1D) is an autoimmune disease characterized by a progressive loss of pancreatic β-cells. β-cells are located in the pancreas (in the islets of Langerhans) and are responsible for the production of insulin. Insulin regulates body fat - it removes excess glucose from the blood and converts it into glycogen, an energy storing molecule that is kept in the liver and in the muscles. When blood glucose levels fall, glycogen is transformed back into it. Due to β-cell loss, type 1 diabetic patients become dependent on insulin for life.

Pancreases from most patients affected by type 1 Diabetes have insulitis, a pancreatic islet inflammation characterized by infiltration of macrophages and T-cells (56, 78). These immune cells contribute to β-cell apoptosis by both cell-to-cell contact and release of pro-inflammatory cytokines[1] such as interleukin-1β (IL-1β), tumor necrosis factor-α (TNF-α) and interferon-γ (IFN-γ) (78). The prevalence of type 1 Diabetes is increasing at an alarming rate, and it was expected that new cases of type 1 Diabetes in young European children double between 2005 and 2020 (204).

Insulitis probably results from a 'dialog' between immune cells coming into the islets and the target β-cells. The latter contribute to this interaction by local production of cytokines and chemokines[2] and by delivering immunogenic signals as a consequence of cell death (78). The outcome of insulitis (ie. resolution or evolution to diabetes) is probably regulated by both the individuals genetic background and environmental elements. These include dietetic components, viral infections and other factors that remain to be determined, acting at both the immune system and β-cell levels (77, 78, 258). In particular, there is growing evidence that the triggering of type 1 Diabetes is associated with a viral infection (181). This emphasizes the importance of understanding the molecular mechanisms leading to β-cell death or survival.

The measurement of mRNA expression with microarray or RNA-seq technology following β-cell exposure to pro-inflammatory cytokines gives a snapshot picture of the changes in gene expression characterizing the path to β-cell dysfunction and death. It has been shown, in in vitro experiments, that exposure of beta-cells to IL-1β is sufficient to induce functional changes that

[1] Cytokines are proteins used in the cell signaling of the immune system.

[2] Chemokines are a particular group of cytokines that release chemical signals that direct the movement of the cell.

are similar to ones observed in pre-diabetic patients (53).[1] Unfortunately, most studies have focused up to now on islets from a single species (mostly human or rat) and used one or two time points only (20, 77, 180, 198, 286), precluding an accurate and dynamic understanding of the phenomena. A high number of mRNA expression measurements may arguably be informative for an accurate estimation of the networks mediating β-cell progressive destruction. Furthermore, comparisons between findings obtained in different species are helpful in defining important and recurrent pathways. In the work presented in the Chapter 6 multiple datasets of β-cell gene expression after cytokine exposure, of different species and at different time points, were used to identify and characterize relevant genes and to infer a GRN. The impact on apoptosis of selected genes was assessed, and predicted regulations were experimentally validated. The next sections present basic concepts for gene network inference.

1.2 Transcription and translation

The DNA (deoxyribonucleic acid) consists of two long complementary strands composed of four small molecules (nucleotides) (adenine, cytosine, guanine, thymine), connected in a double helix structure. In transcription, the enzyme RNA polymerase binds to certain regions of the DNA (called promoters), and synthesizes pre-mRNA from adjacent (to the promoter) regions. DNA transcription is ended in special terminating regions. Genes are the individual DNA regions coded into pre-mRNA, and are characterized by their location in the DNA sequence and constituting nucleotides (accounting for intra-species variation) (250).[2]

Pre-mRNA is composed of coding (exons) and non-coding regions (introns). After transcription, pre-mRNA is spliced and introns removed, resulting in the final mRNA.[3] The resulting mRNA is used to produce proteins by the ribosome (in a process designated by translation). In translation, sequences of three mRNA nucleotides (known as codons) are read sequentially, each producing an amino-acid. Translation ends with termination codons. The result of translation is an amino-acid sequence, the primary structure of a protein (250).

Information usually flows from DNA to RNA, and from RNA to proteins, and (usually) not in the reverse direction - this fact is the so-called central dogma of molecular biology (59).

[1] In order to be induced into apoptosis, beta cells must be exposed to a combination of IL-1β and IFN-γ, or of IL-1β and TNF-α (or of the three cytokines together).

[2] A same gene may encode different RNA molecules, and a protein may be the result of multiple RNA molecules. Also, genes may be overlapping, and a gene may be contained within another gene (210).

[3] Splicing variants may occur, resulting in different mRNA molecules (alternative splicing) and subsequent proteins. For instance, exons may be skipped, or some exons may be mutually exclusive (the presence of one in the final mRNA implies the absence of the other) (25). In humans almost all multi-exon genes (95 %) are subject to alternative splicing (200).

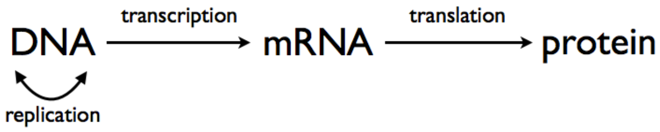


Figure 1.3: Central dogma of molecular biology - general information flow - General information flow, from DNA to DNA (replication), DNA to mRNA (transcription) and mRNA to proteins (translation).

Figure 1.3 illustrates the general cases of information transfer in biological systems. There are three general mechanisms of information transfer: transcription, translation and DNA replication (which is necessary for cell replication). There are other (special) cases of information transfer. Information transfer from RNA to DNA is known as reverse transcription (the obtained DNA is called complementary DNA, cDNA). In this case, RNA information is transformed into DNA and incorporated in the genome (examples are retrovirus, such as the HIV) (250). RNA replication and RNA editing (by RNA) is also known to happen in simpler organisms. Another common representation of biochemical regulations identifies three layers of regulations, at mRNA, protein and metabolite levels[1]) (32). It is illustrated in the Figure 1.4.

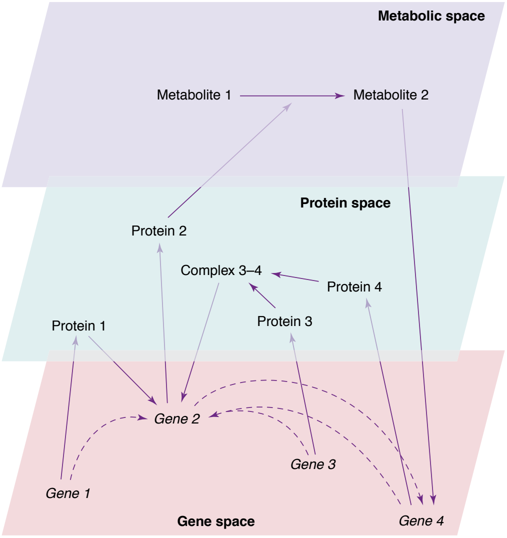


Figure 1.4: Three-layer biochemical network - Example of a biochemical network, organized in three layers corresponding to mRNA, protein and metabolite levels. Solid arrows indicate interactions and dashed arrows indicate the projection of interactions into the mRNA (gene) level. Figure from (32).

[1] Metabolites are small molecules, intermediate and final products of cell metabolism.

The ENCODE project started in 2003 with the goal of characterizing the activity of all genome (253). Although protein-coding genes constitute a small fraction of the genome (less than 2 % (280)), this project has found that near 80% of the genome takes part in biochemical activities, such as transcription of non protein-coding RNA (253).[1]

1.3 Gene regulatory networks

The proteins assembled by the mRNA have a broad range of functions, including the regulation of the transcription of other genes. Some proteins (called transcription factors) bind to non-coding DNA regions and regulate the rate of transcription of genes (not necessarily nearby), by promoting or repressing transcription by RNA polymerase. Other factors play a role in transcriptional regulation, described next.

DNA is compacted and protected by a complex of molecules called chromatin, in which histone proteins are responsible for the winding of DNA (which becomes organized in individual sections wrapped around 8 histones, called nucleosomes). One factor in transcriptional regulation is the modification of chromatin structure so that DNA regions becomes more or less accessible to RNA polymerase (chromatin remodeling). This process may take the form of modifications in histone proteins, such as by acetylation/deacetylation (histone acetylation weakens the association of DNA to histones, making the former more accessible to RNA polymerase and transcription), methylation, phosphorylation and ubiquitination (247). Another form of chromatin remodeling is the modification of the nucleosome structure, by energy carrying molecules (with an ATPase domain) (186). One other factor in transcriptional regulation are co-activators and co-repressors, proteins that do not bind to DNA, but to transcription factors, enhancing or reducing transcription (250).

Transcription regulating genes may regulate other transcription regulating genes, forming a cascade of direct and indirect regulations.[2] These may be represented as a network, composed of nodes and edges, where each node represents the expression of a gene (its mRNA level), and a directed edge represents the existence of a regulation, increasing or inhibiting the expression of the target (positive and negative regulations). An example of a simple GRN is represented in the Figure 1.5:

Such a (transcriptional) gene regulatory network (GRN) is a simplification as it does not consider regulatory factors occurring after transcription and regulating translation. An example of post-transcriptional regulation is mRNA degradation through its pairing with small RNA

[1] Controversy has ensued over the advertised importance of these findings, particularly on the usage of the term *function* to characterize these non-protein coding activities (97).

[2] This concept of gene regulation should not be confused with the one of gene interaction, or epistasis, which refers to the phenomenon of multiple genes controlling a single phenotypical characteristic (164).

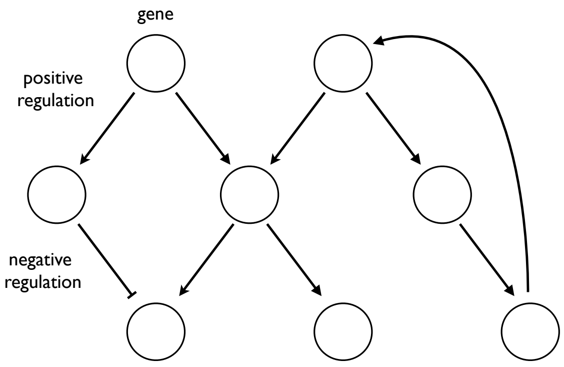


Figure 1.5: Example of a transcriptional gene regulatory network - Genes are represented by nodes, and regulations, positive or negative, are represented by edges.

fragments and subsequent cleavage. The most common forms of this type of regulatory RNAs are small interfering RNA (siRNA) and microRNA (miRNA) (2, 46).

Transcriptional GRN may be inferred from gene expression measurements, obtained by quantification of transcribed mRNA in microarrays and RNA-seq platforms (introduced in the next section).

1.4 Measuring gene expression

Single DNA or RNA strands bind to complementary sequences in a process termed hybridization. Double strands are usually heated to form single strands, which then hybridize with complementary strands after cooling. This process is used in common gene expression measurement techniques, such as blotting, polymerase chain reaction (PCR) and microarrays. Blotting consists in the separation of a target molecules by gel electrophoresis, transference to a carrier and quantification (190). PCR is the amplification of target sequences (by consecutive duplication of target strands) and subsequent quantification (43). Blotting and PCR techniques measure the quantity of a single or few sequences; microarrays and RNA-seq measure the expression of multiple genes at once and provide the data necessary for GRN inference. They are described next in more detail.

A microarray is a solid surface composed of spots of specific DNA sequences (called probes) (108). In a microarray experiment, mRNA of a biological sample is isolated and transformed into cDNA by reverse transcription. This cDNA is then labeled (eg. fluorescently) and hybridized with the microarray probes. The hybridization of each probe is then quantified, giving a measure of the amount of respective mRNA sequence in the sample (115). Several types of microarrays exist, which may be distinguished between manufacturing process[1] and single/dual channel types[2]. Microarray limitations are the following: they only measure RNA transcripts of designed probes and thus do not measure non considered sequences; there may be cross hybridization (hybridization of sequences not perfectly complementary); the range of RNA quantity that is measured is limited, both at the higher (signal saturation) and lower level (72). Microarrays should be subject to quality controls, identifying possible hybridization artifacts. Normalization techniques mitigates possible systematic errors in the arrays (eg. due to batch effects).[3]

RNA-seq technology measures mRNA using next generation (or second generation) sequencing technology (102). In RNA-seq, a population of RNA is fragmented into small segments (typically 30 to 400 base-pairs) and then converted to cDNA small-segments by reverse transcription. Alternatively, RNA may be transformed into cDNA and then segmented (270). These cDNA segments are then attached to a solid surface, which are read and assigned to a position in the genome corresponding to a coding exon. The number of cDNA sequences assigned to a given exon is a measure of its expression. Naturally, longer exons will tend to have more sequences mapped to them, therefore length-normalized gene expression measures are usually adopted.[4] The advantage of RNA-seq over microarray technology is that it is not limited to the identification of previously defined sequences, allowing for the identification of variations in the mRNA (ie. splice variants or mutations) (270).

1.5 Machine learning

Machine learning is a field of computer science concerned with the extraction and application of information from complex data. Main applications of machine learning are classification, regression, clustering, dimensionality reduction, or the inference of general models or structures, such as networks (24). Classification is the assignment of a class (a realization of a categorical

[1] Probes may be designed and spotted on the surface, allowing for customizable arrays. Alternatively, sequences may be synthesized directly (in situ) on the array surface (9, 13).

[2] Single channel microarrays measure the RNA quantities of a single sample. Dual channel measures the RNA quantities of a mixture of two samples, each labeled differently. It allows to compare samples of different conditions (eg. control vs treatment) (232). Dual channel microarrays are commonly associated with spotted microarrays, while single channel microarrays are to in situ microarrays (called oligonucleotide microarrays) (128).

[3] Normalization can be applied to single or multiple arrays. In the first case, each array is independently normalized (as in the MAS5 algorithm (122)). An example of multi-array normalization is the RMA algorithm, which consists of individual background correction; normalization of probe values so that all arrays have similar distributions; and log transformation (27). Other approaches can be found in the literature (experimental comparisons can be found in (109, 154)).

[4] Such as the RPKM - Reads Per Kilobase of transcript per Million mapped reads (182). For a comparison of different normalization methods see (68).

variable) to a sample, while regression usually refers to the assignment of a continuous numerical value to it. Clustering is the separation of samples into distinct groups, whose size and number may be unknown beforehand. Dimensionality reduction refers to the reduction of the number of variables in a problem, to ease its complexity and improve the accuracy of a model. It may take the form of the selection of a subset of the original variables, or of the selection of new variables obtained from the original ones. In the first case, variable selection may be used to quantify the importance of variables (107). In the latter case, a common approach is to transform the variable space into orthogonal components of decreasing explanatory content (a process known as principal component analysis).

Classification and regression are cases of *supervised* learning, whereas clustering, variable selection and network inference are cases of *unsupervised* learning (24). In supervised learning, models are estimated from samples associated with a class (classification), or a numeric value (regression). In unsupervised learning, this association is not given or is irrelevant. Machine learning models are often estimated on a training dataset and assessed on a test dataset, not part of the training part. Cross-validation consists in partitioning the data so that different partitions are used as the test dataset in different assessment rounds.

The fields of statistics and machine learning are overlapping, but the goals and methodology are perceived to differ (33). Statistics emphasizes formal theoretical results; machine learning emphasizes accuracy in solving real world problems. Statistical models assume well-defined processes and return valid conclusions given assumptions; machine learning models operate as black boxes in which the nature of the relationships between variables is often unclear. One example are random forests, described in the Section 2.6. Machine learning and statistics sometimes use different terminology. For instance, in machine learning, instances refer to samples, and features to variables. Popular machine learning models include nearest neighbors, neural networks, support vector machines or random forests (110)). The terms machine learning, pattern recognition, data mining or knowledge discovery in databases are often conflated and are difficult to clearly distinguish.

1.6 The bias-variance trade off

The estimation of models (a model is a set of assumptions regarding a data generating process) and parameters may be characterized in terms of bias and variance. The bias of an estimation is its tendency to make systematic errors, in 'one direction'. In the context of a single numeric random variable θ, the bias of an estimation $\hat{\theta}$ is defined as the difference between the expected value of the estimation, and the true value (110):

$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta \tag{1.1}$$

The variance of an estimation $\hat{\theta}$ is defined as:

$$\text{var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] \tag{1.2}$$

When assessing estimation accuracy over multiple samples, a measure of squared error is usually adopted (if the simple mean is used, underestimations and overestimations may cancel themselves out). The mean squared error (MSE) is the mean of the square of the errors of the estimated values.[1] The MSE of the estimator $\hat{\theta}$ is defined as (110):

$$\text{MSE} = \mathbb{E}[(\hat{\theta} - \theta)^2] \tag{1.3}$$

It can be rewritten as the sum of the variance and the square of the bias:

$$\text{MSE} = \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2 \tag{1.4}$$

It can happen that a biased estimator with small variance has a lower MSE than an unbiased estimator with high variance. This observation leads to the bias-variance trade off: reducing the variance of an estimator at the cost of a bias increase can be beneficial in terms of MSE (or other risk functions that are a function of bias and variance). This principle is the basis of Bayesian and shrinkage estimators.

A model with less parameters is less variant (more robust) at the cost of a possible higher bias. If a model has too many parameters it may over-fit the data. Over-fitting refers to the modeling of idiosyncrasies of particular observations, not representative of the variables true distribution. Models that over-fit are highly variant, as small differences in the data may cause large differences in the estimated model parameters. One strategy to improve the bias-variance trade-off is to combine multiple models to obtain a less variant and less biased one. Common strategies to do so (boosting, bagging) are introduced in the Section 2.6 in the context of decision trees. Random forests are a popular machine learning algorithm based on ensembles of decision trees.

1.7 GRN inference

GRN may be inferred through gene expression perturbations - by modifying the expression of genes, such as transcription factors, and identifying other genes whose transcription/translation is altered (are regulated by the former) (11, 126, 252, 291). A common perturbation is to temporarily reduce the expression of genes by degradation of the respective mRNA using siRNA

[1] The MSE is a risk function - the expected value of a loss function (which assigns a cost to an estimation, in this case the squared error).

(2). Due to the high number of genes in networks and necessary gene perturbations, it may be impractical (too costly and time consuming) to infer GRNs exclusively through this manner.

Another approach to GRN inference is through the association of transcription factors to genes, by matching transcription factor DNA binding motifs (short patterns of DNA (66)) to motifs in gene promoter regions (93, 112). However, this approach has the downside of returning an excessive number of false positives, as DNA binding motifs are often promiscuous and a large number of potential binding associations are thus predicted (112). Information on transcription factors binding profiles is available on databases such as JASPER (171) or TRANSFAC (172). Binding associations may be experimentally confirmed with ChIP-on-chip (5) and ChIP-seq (130) techniques (combining chromatin immunoprecipitation with microarray and RNA-seq technology respectively). They allow to identify, in vivo, which DNA regions bind to a particular transcription factor. As more data becomes available, ChIP data becomes a valuable resource for network inference.

GRN may be also inferred from (non-manipulated) gene expression data using statistical and machine learning models.These different approaches can be combined in integrative meta strategies (112, 132, 194, 215).

This thesis concerns exclusively with the case of network inference from gene expression. In this case, the problem of GRN inference amounts to the estimation of the parameters of a model characterizing a regulatory network (ie. existence and/or strength of a causal relationship between each pair of genes). This approach is limited by the fact that the number of variables (genes) is usually much higher than the number of observations. This limitation is described in the next section. Section 1.7.2 briefly refers the main strategies for GRN inference from expression data.

1.7.1 The high variable to sample ratio in GRN inference

Inference of GRN from gene expression data is currently conditioned by the fact that the number of genes (in the order of the thousands) is typically much higher than number of samples.[1] In this case, GRN inference is an ill-posed problem with an indeterminate linear solution - any n expression values of any gene can be represented as a linear function of any group of n genes plus a constant. Inference strategies that overcome inference indeterminacy are then adopted.

The high variable to sample ratio also means that complex network models may be highly variant - strongly modified by small variations in few samples. For this reason, simpler inference approaches (eg. bivariate) are known in the literature to outperform more sophisticated ones (12, 265). GRN models should strike a balance between robustness to sample variations, and

[1] This limitation is commonly known as the ”small n large p”.

sufficient complexity to accurately capture multivariate regulatory relations. This is an example of a bias-variance trade off, introduced previously in Section 1.6.

Even if the high variable to sample ratio makes it difficult to robustly infer GRN, this endeavor may be helpful to medical research - to rank putative gene regulations and prioritize confirmatory experiments.

1.7.2 Strategies for GRN inference

Several strategies for GRN network inference from expression data have been proposed in the literature. These can be based on assumptions regarding gene expression, using enzyme kinetics formulas such as the Michaelis-Menten or Hill equations (93, 133, 211). However, these assumptions require the estimation of several extra parameters, which may be deemed unfeasible due to the high variable to sample ratio in large networks.

A more practical alternative is to assume simple models (eg. linear), or to use concepts of statistical dependence or variable selection. This thesis deals exclusively with such strategies, which also have the advantage of being generalizable to any network inference problem. They include conditional independence tests between gene expression; Bayesian networks which use conditional independence tests to infer causal relations between genes; variable selection approaches that select regulators of each target gene using a variety of statistical/machine learning models (eg. information-theoretic, regularization-based, random forests, neural networks, support vector machines); Boolean networks, which involve a discretization of gene expression into two states (ie. active and inhibited); or dynamic approaches, designed to gene expression time series (such as lagged conditional independence tests and differential equations).

In time series, if the predictor and target genes are observed at separate time points, a causal aspect is implied in predictor-target associations. In static data, non-temporal associations can nevertheless be used to infer undirected networks. For reviews on network inference methods see (10, 93, 101, 131, 133, 170, 211, 265).

The next two sections present useful notions for GRN inference. Networks can be inferred using a variable selection strategy, consisting in the selection of predictors for each target gene in the network. It is described in more detail in the next section. Section 1.9 presents the link between statistical dependence and causality.

1.8 Variable selection

Variable (or feature) selection consists in the selection (or ranking) of predictor variables with respect to the modeling of a target (class) variable. It is commonly used in bioinformatics - eg. variable selection is used to identify genes which are good predictors of a given phenotype

(212, 224). It can be used to infer GRN, and in this case variable selection is applied to each gene at a time, considered the target variable. A network is reconstructed through the selection of predictors of each gene. Three main strategies for variable selection are commonly identified in the literature: filters, wrappers and embedded methods (106, 224).

Filters, wrappers, embedded methods Following the taxonomy presented in (224), filters are independent of statistical models (of the target variable), and only look at the characteristics (ie. dependences) present in the data. They can consist solely on a measure of dependence towards the target variable, such as the linear correlation or the mutual information (in this case they are bivariate). More sophisticated (multivariate) filters operate in a forward/backward selection manner, also considering the dependence (redundancy) towards previously selected predictors. The advantage of filters is their simplicity, computational speed, scalability and robustness to over-fitting. Due to these properties, filters have been used in GRN inference (179). One example is the mRMR filter (212) which sequentially selects predictors, the ones maximizing the difference between a measure of relevance towards the target, and the average of a measure of redundancy towards each previously selected predictor. In the context of type 1 Diabetes, in Chapter 6 mRMR is used in the inference of a GRN in β-cells following cytokine exposure. Several filter approaches are reviewed in the Chapter 3 and experimentally assessed in the Chapter 5.

Wrapper methods select variables to be included in a given statistical model, by validating that model on external data (ie. by cross validation). The search over the variable space may be done using a variety of methods (see (138, 224)), including forward and backward selection, simulated annealing, hill climbing, or genetic algorithms (224). The downside of wrappers is their higher computational complexity (due to the need of validating each considered variable subset) and risk of over-fitting. The use of wrapper variable selection in GRN inference is not common (to our knowledge), due to the high variable to sample ratio of gene expression datasets.

Embedded methods consist of statistical/machine learning models where variable selection or ranking is implicit (or from which it is easily obtained). These may include linear models where sparsity in the number of estimated non-zero coefficients is induced, or random forests returning an attribute of variable importance (36, 106). Variable selection in the context of linear models is described in the Section 2.2.4, and decision trees/random forests in Section 2.6. Adaptations for time series of these approaches are experimentally assessed in the Chapter 5.

1.9 Statistical dependence and causal inference

Two random variables X and Y are independent if $p(x|y) = p(x)$, or equivalently if $p(x, y) = p(x)p(y)$ (independence between X and Y is denoted by $X \perp\!\!\!\perp Y$). Analogously, X and Y are *conditionally* independent given a set of variables Z if $p(x|y, z) = p(x|z)$. A measure of the statistical dependence between two variables is the mutual information, a concept which originated in the field of information theory. The mutual information is a non-negative quantity, in which null mutual information implies no statistical dependence. It is based on the concept of information entropy, analogous to the concept of entropy used in statistical mechanics. The mutual information between two variables X and Y measures the similarity between the multivariate probability distribution $p(x, y)$ and the product of the marginal distributions $p(x)p(y)$ - in particular, it is the Kullback-Leibler divergence of $p(x)p(y)$ from $p(x, y)$ (57). The mutual information is introduced in more detail in Section 2.1. When variables follow multivariate elliptical distributions (such as the multivariate Gaussian), the mutual information and the linear correlation are a continuous bijective function of one another.

Graphical models of dependence Network inference approaches based on conditional independence tests often take the form of graphical models, characterized by nodes (ie. genes) and edges (ie. regulations) (as in the Figure 1.5). Edges represent statistical dependences between the nodes at its extremes. Undirected graphical models, in which an edge indicates a conditional statistical dependence are known as Markov networks (or Gaussian graphical models when variables are assumed to be Gaussian distributed and dependences linear). Directed and acylic graphical models are known as Bayesian networks, which use conditional independence relations to model causality. Graphical models are described in more detail in Section 2.4. The link between statistical dependence and causality is described in the next section.

1.9.1 On causality

Causality has been discussed in ancient Greek philosophy and then with renewed interest since the Renaissance and the advent of modern scientific thought (184). It has been viewed in the form of logical implications between single events (or binary variables, representing the occurrence or not of events), and also in the context of random numeric variables, in the form of asymmetrical functional relationships between them. Arguments in favor of this latter approach can be found in in (184). For references on the historical developments of causality see (118, 184, 206).

In the context of random variables, the existence of causality can be identified using interventions (ie. by setting the value of a variable, represented by the do operator). If there is an intervention on one variable such that the conditional (given the intervention) probability distribution of another variable is different than its marginal distribution, there is causality from the first to the second (184).[1]

Definition 1. *A variable X has a causal effect on a variable Y if:*

$$p(y|do(x)) \neq p(y) \tag{1.5}$$

If the causal network is known (ie. which variables cause which), the *back-door* and *front-door* adjustments, and the theory of do-calculus allows to estimate, when possible, the effects of hypothetical manipulations (counterfactuals) from (non-manipulated) observations, thus providing the link between non-manipulated and post-manipulation probability distributions (206, 208) (Section 2.4.6). In the remainder of this section we discuss the link between statistical dependence and causality.

When conditional dependence implies causality A rule of thumb (known as the common cause principle, CPP) is that if two variables X and Y are dependent, either one variable is a cause of the other; or the two variables have a common cause (118).[2] The cancelation or reversal of a dependence between two variables when a third (a common cause) is conditioned on is known as the Simpson's paradox (206).

Assuming the CPP, if all the possible common causes of X and Y are identified and conditioned on, the existence of a conditional dependence between X and Y implies that one is a cause of the other (assuming that none of the common effects of X and Y are also conditioned on - see next paragraph). Another result is that if X and Y are conditionally dependent given any set of conditioning variables, one is a *direct* cause of the other.

Full conditional dependence (given all other variables) does not imply causality because two independent variables may become conditionally dependent if conditioned on an effect of both of them. This phenomenon is known as the Berkson's paradox (21). In directed graphical models, these three-variable causal configurations are known as collisions (the effect variable is the collider, see Section 2.4).

Faithfulness Causality implies conditional dependence (given any set of conditioning variables) only under a condition of *faithfulness* (243).[3] A probability distribution $\mathcal{P}$ and a causal

[1] This definition concerns the identifiability of causality, and not on causality itself. It may be argued that using variable interventions to define causality is circular, as manipulations are causes themselves (184).

[2] Exceptions to this rule can be found in physics, in the form of instantaneous dependences between variables without an apparent cause-effect mechanism or a common cause (particle entanglement or electromagnetism). However, those variables may be seen as different manifestations of a same global variable. Equivalent examples are not found in the macroscopic realm.

[3] The term *stability* is also used (206).

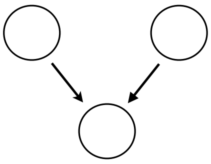


Figure 1.6: Collision in a directed graphical model - Two independent random variables may become conditionally dependent, given a common effect. This causal configuration is known as a collision, in directed graphical models.

model are faithful to each other if a conditional independence in $\mathcal{P}$ implies the absence of a direct causal relationship between the respective variables in the model. Faithfulness is described in the context of graphical models in Section 2.4.2. Figure 1.7 illustrates the relation between causality and conditional dependence.

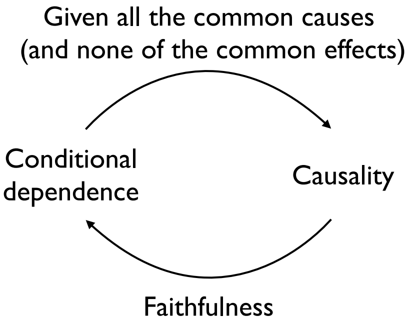


Figure 1.7: Conditional dependence and causality - When all the common causes and none of the common effects are conditioned on, conditional dependence implies causality. In the case of faithfulness, causality implies conditional dependence.

Causality and time The consideration of time-series (observations of variables over time) facilitates causal inference as effects are observed at later time points than causes. Representing variables in multiple time points allows to distinguish between causes (observed at earlier time points) and effects (observed at later time points). A definition of causality popularized by Suppes (249) is that there is causality from a cause to an effect if the past of the cause and the present of the effect are conditionally dependent, given the the past of all other possible causes. The dependence between two variables in separate time points is usually referred to as a lagged dependence. This form of causality is also known as Granger causality, following a test popularized by Granger (103, 159).

1.9.2 Causal inference

The enterprise of causal inference from data observations can be traced back to the end of the 19th century (eg. in the estimation of regression coefficients relating welfare and poverty (85, 290)). Advances on linear causal models led to the topic of structural equation modeling (SEM, also denoted by) (184), presented in Section 2.3. Alternatively, in econometrics linear models designed to time series were popularized, including vector auto-regressive models and Granger causality tests. The latter test lagged (conditional) dependences between causes and effects, and are described in Section 2.7).

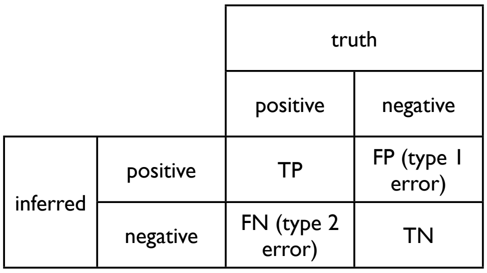
In the last decades, theoretical contributions on causality led to the field of Bayesian networks, directed and acyclic (or recursive) graphical models (BN) (206) (Section 2.4). BNs may be seen as non-parametric extensions of (recursive) SEMs (207). The acyclic condition prevents the modeling of feedback loops. This limitation is relevant to GRN inference as feedback loops are common in biological mechanisms, including gene regulation (131, 170). Feedback loop causality requires different models, as the causes and effects of variables are not distinguished (184). On non-recursive graphical models see (220, 244). Using time series allows to model feedback loops, as in this case variables may be represented in multiple time points.

In time series, auto-correlation is the statistical dependence of a variable at different time points. Non-stationarity occurs when the probability distribution of a variable is different for different time points. When estimating lagged dependences, these aspects require the consideration of extra parameters (lags) (159, 257). However, considering extra parameters increases the model variance, in a bias-variance trade off. This is relevant in GRN inference, typically of a high variable to sample ratio. Section 5 presents an empirical investigation on the optimal approach to model lags in Granger causality tests, on gene expression time series.

Recent methods aim to infer causality from the identification of peculiar characteristics (eg. non linearities) in the probability distributions of causes and effects (125, 129, 235) (these methods remain out of the scope of this thesis).

1.10 GRN inference assessment

GRN inference can be assessed through the experimental confirmation (ie. through siRNA experiments to inhibit the putative causal gene) of predicted gene regulations. A more practical alternative is to compare the predicted regulations with an available gold standard, constituted by ordered pairs of genes of known or putative regulations. Available gold standards are incomplete, as they do not contain all existing regulations (not all have been identified) and may contain errors. In simulated gene expression data, a complete gold standard is given. Assessing inference with a gold standard is a problem of binary classification, where known regulations are instances of a positive class (and the remaining possible regulations are of the negative class). A regulation which is correctly inferred, according to the gold standard, is a true positive (TP). False positives (FP), true negatives (TN) and false negatives (FN) are equivalently defined. Fig. 1.8 illustrates the difference between TP, FP, TN and FN.



		truth	
		positive	negative
inferred	positive	TP	FP (type 1 error)
	negative	FN (type 2 error)	TN

Figure 1.8: Confusion matrix - True positives, false positives, true negatives, false negatives.

The precision is the number of TP divided by the number of instances predicted as positives N, $\frac{TP}{N}$. The recall (or TP rate) is the number of TP divided by the number of positives P, $\frac{TP}{P}$. The FP rate is the number of false positives divided by the number of negatives. The specificity is 1 minus the FP rate. F-scores combine values of recall and precision into a single value. These measures require the classification of instances into positive and negative classes. When instances are ranked according to a score that they are of the positive class, ranking accuracy can be assessed with precision-recall (pr-curves) and receiving operator characteristic (ROC) curves.

ROC and pr-curves The use of curves to assess network inference requires that instances (ie. gene regulations) are scored and ranked (216). Instances are ranked according to a score that they are of the positive class and are incrementally selected. A pr-curve plots the precision (vertical axis) as a function of recall (horizontal axis), as instances are selected. ROC curves plot recall (vertical axis) as a function of the FP rate (horizontal axis). As instances are selected, both curves move along the horizontal axis (left to right). These curves may be characterized with a single value, the area under the curve (AUPRC or AUROC) - where a higher area means a better overall accuracy. Its maximum value of 1 corresponds to the optimal configuration where all the positive instances are ranked before all negative ones. Figure 1.9 presents precision-recall and ROC curves of the same ranking and gold standard. In particular, the number of total instances is 600 and the number of positives is 20.

Pr-curves have been considered to be a more informative indicator of performance than

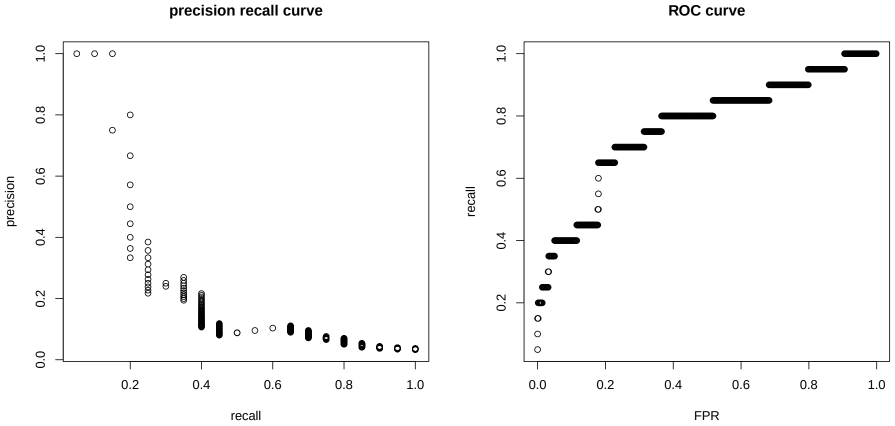


Figure 1.9: Precision-recall and ROC curves - Pr-curves plot precision as a function of recall; ROC curves plot recall as a function of the false positive rate (FPR).

ROC curves on class imbalanced problems (when the positive class is the minority). In this case, ROC analysis is less sensitive to variations in the number of false positives: a large change in the number of FPs (recall remaining constant, vertical axis) may lead to a small change in the false positive rate (horizontal axis) (62)). This phenomenon can be seen in the Figure 1.9, around recall 0.2. The drop in precision is much larger than the corresponding increase in the FPR.

One advantage of ROC analysis is that the expected ROC curve, in the null hypothesis of random selection, is invariant with the positive/negative class distribution and takes the form of a diagonal from 0 to 1 (with a respective area under the curve of value 0.5). Regarding pr-curves, on the contrary, the expected null curve depends on the class distribution and on the number of positive and negative instances.

Precision comparison and assessment Pr-curves depend on the characteristics of the inference problem, namely the number of positives and negatives. If the ratio of positive instances lowers, the expected precision lowers as well. For this reason, precision values obtained in different configurations (ie. number of positives and negatives) are not directly comparable. However, a measure of global precision (such as the area under the curve, AUPRC) may be transformed into a p-value (null hypothesis of random selection), which is comparable among different configurations. Such a p-value is also useful to assess statistical significance.

In the literature, the only approach to estimate AUPRC p-values is non-parametric, consisting in the estimation of the AUPRC probability distribution by Monte Carlo. The first contribution of this thesis, presented in the Chapter 4, is the analytical derivation of the mean and variance of the null AUPRC, which are used to estimate a continuous approximation of the AUPRC distribution.

1.11 Thesis contributions

1.11.1 On the null distribution of the precision-recall curve

The Chapter 4 presents the first contribution of this thesis, concerning the assessment of GRN inference through pr-curves. Different pr-curve interpolation strategies are discussed (Section 4.2), and the expected pr-curve of random selection is analytically derived (Section 4.3). From these results the expected value and variance of the null AUPRC are derived (Section 4.4). These parameters (together with the minimum and maximum AUPRC values) are used to estimate a continuous approximation of the AUPRC distribution based on the beta distribution (Section 4.4.3). This approach has the advantage (compared to Monte Carlo) of returning an exact solution, avoiding the high number of simulations required for an accurate approximation of the AUPRC distribution. Section 4.5 presents some experimental findings, including a comparison between the beta distribution-based AUPRC distribution and the one obtained by Monte Carlo, for different number of instances. It is shown that the beta distribution approximation is an accurate one, more so for larger number of instances. This result is expected: as the number of instances increases the discrete nature of the true pr-curve is smoothed out. Another experiment assesses the errors of the mean and variance of Monte Carlo approximations as a function of the number of simulations, and also number of instances. As expected, it is shown that as the number of instances grows larger, the number of simulations required to maintain a constant same error increases. We use this approach to compute AUPRC p-values in the experimental session of the Section 5.

1.11.2 GRN inference from time series

The Chapter 5 presents two algorithmic contributions to network inference from time series (Section 5.2), and an experimental investigation on its accuracy (Section 5.3).

The experimental session uses real and simulated gene expression time series. The former are 11 multivariate time series of gene expression from *Saccharomyces cerevisiae* (yeast), from two different datasets, of 25 and 18 time points (214, 238). A gold standard of interactions is obtained from the literature (1). The simulated time series are 100 multivariate time series (of 50 genes) generated by GeneNetWeaver, the software behind the popular DREAM challenges (166, 230), for different number of time points, from 20 up to 300. Using the simulated data we investigate how inference accuracy depends on the time series size. In the microarray time series we investigate inference accuracy in a real biological context.

A first experiment compares different approaches to model Granger causality (GC) between cause-effect pairs. A second experiment compares multiple state of the art approaches for GRN inference. In both the real and simulated data 100 networks of 50 genes were inferred. The inference of each network was assessed with a AUPRC p-value, as described in the previous section. The two experiments are bias-variance trade off investigations: the first on the modeling of causality between cause-effect pairs; the second on multivariate considerations in network inference. The two algorithmic contributions to network inference and the experimental results are briefly described next.

Conditional GC The first algorithmic contribution to network inference is a fast approximation of a first order conditional (ie. on a single third variable) GC test between two variables (described in Section 5.2.2). The resulting score is then used as a filter network inference method. This approach can be seen as a dynamic extension of the PC algorithm, restricted to first order conditional independence tests (see Section 2.5). In order to score and rank cause-effect pairs (direct regulations), useful for assessment using precision-recall or ROC curves, each pair is assigned the minimum of the respective first order conditional GC scores.

This strategy implies a search over the set of conditioning variables, which may be computationally intensive (each test involves two linear regressions of one and two independent variables). We propose an heuristic to prioritize the search, starting with the variables most likely to return low GC scores, and stopping when a criterion is met. The proposed method was experimentally validated in the Section 5.3, being the most precise inference approach in the simulated time series. The improvement in computational speed is critical in GRN inference from a very large number of genes (eg. the 20 thousand in human). The proposed approximation can be generalized to the static case.

The use of first order conditional dependences has been proposed as an attractive approach to GRN inference. Each regulation is scored as a function of three variables only, avoiding the high variable to sample ratio limitation (170). Examples in the context of static linear models (GGMs) can be found in (65, 163, 278, 279). Its adoption assumes that the correlation between a non cause-effect pair of genes is explained by a single third gene. Figure 1.10 illustrates an example of a case where this does not hold, of an indirect causal regulation via two genes. First order conditional independence tests are not sufficient to screen off such dependence. Figure 1.11 illustrates dependences mediated via a single gene. In directed graphical models, they correspond to the case where a single path (d-)connects two nodes (see Section 2.4).

Co-regulation identification The second algorithmic contribution to network inference is a simple method to identify co-regulated genes (regulated by the same genes) in time series,

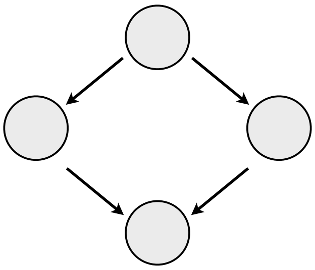


Figure 1.10: Illustration of an indirect regulation via two genes - The gene on the top regulates the gene on the bottom via two two intermediate genes. First order conditional independence tests do not screen off the dependence between the top and bottom genes.

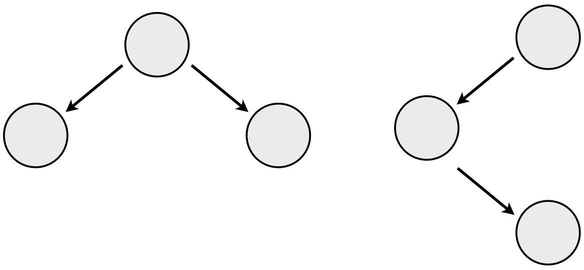


Figure 1.11: Illustration of indirect regulations via a single gene - In the left case, the genes on the bottom are regulated by the gene on the top. In the right case, the gene on the top regulates the gene on the bottom via a single intermediate gene. First order conditional independence tests are sufficient to screen off the dependence between the bottom genes (left case) and bottom and top genes (right case).

which can be combined with network inference methods. Co-regulated genes are determined by common causes, are statistically dependent and may be incorrectly inferred as a cause-effect pair. Under a linear assumption, we propose to identify as co-regulated the genes which exhibit a high linear correlation between them, and also a higher non-lagged correlation than a lagged correlation. The method is theoretically justified through the analysis of SEM path diagrams in Section 5.2.3.1, and experimentally validated in both real and in-silico gene expression data.

Experimental session In the experimental investigation of Section 5.3, different approaches to infer causality with respect to the number and estimation of lags and consideration of non-stationarity are assessed. We show that one-lag approaches are more accurate than multiple-lag approaches when the number of time points is low, but not when this number is higher than around one hundred points. In the network inference experiment, and in line with literature findings, when the number of samples is low, simpler methods outperform more sophisticated ones (which become more accurate as the number of samples increases). In the simulated time series, methods based on first order conditional dependences are the top performing, in particular the proposed conditional GC filter (see discussion in Section 5.4).

1.11.3 Knowledge inference in Type 1 Diabetes

The Chapter 6 presents a study on gene expression in the context of type 1 Diabetes, based on eight datasets of gene expression in β-cells after cytokine exposure. These datasets are from different species (human, mouse and rat), and of different time points after cytokine exposure. Two of the used datasets, time series of rat and human gene expression, were made publicly available in the context of the publication of the work of this chapter. A standard meta-analysis was performed on these datasets to identify sets of genes differentially expressed before and after 24 hours cytokine exposure. These genes were functionally characterized through functional enrichment tools, and compared with literature information in order to identify unknown genes, of potential relevance for β-cell dysfunction and apoptosis in the type 1 Diabetes context.

In parallel, a set of 84 genes, differentially expressed both before and after 24 hours, was selected to infer a GRN using the novel human gene expression time series dataset. The network was inferred using a temporal adaptation of the variable selection method mRMR (212) based on estimated lags (described in Section 6.3.5). From these lags the genes in the network were also ordered by time of regulation. Two genes (ELF3 and RIPK2), among the most up-regulated both before and after 24 hours (and the two most up-regulated among the unknown genes), were selected for an assessment of their impact in β-cell apoptosis. They were knocked down using siRNA and shown to have a protective role in β-cells as apoptosis went up after their knockdown. Four predicted regulations (present in the inferred network) involving the two selected genes as potential regulators were also subject to an experimental validation. Three causal links were confirmed. ELF3 was found to up-regulate CX3Cl1 and SP110. RIPK2 was found to up-regulate IRF7 (however, if the regulations are direct or indirect remains to be determined). These results provide a proof of concept of the network inference approach and identified novel genes of potential relevance to better understand β-cell loss in type 1 diabetes.

1.11.4 Publications used in this thesis

- Miguel Lopes, Gianluca Bontempi, ”Using Granger causality to infer gene regulatory networks from time series”, to be published. **Used in chapter 5.**

- Miguel Lopes, Gianluca Bontempi, ”On the null distribution of the precision and recall curve", ECML-PKDD, 2014. **Used in chapter 4.**

- Miguel Lopes, Burak Kutlu, Michela Miani, Claus H. Bang-Berthelsen, Joachim Strling, Flemming Pociot, Nathan Goodman, Lee Hood, Nils Welsh, Gianluca Bontempi, Decio Eizirik, "Temporal profiling of cytokine-induced genes in pancreatic β-cells by meta-analysis and network inference", Genomics, April 2014. **Used in chapter 6.**

The following publications are not used in this thesis, but consist of work which inspired or directly led to the presented contributions.

- Miguel Lopes, Gianluca Bontempi, "Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data", Frontiers in Genetics, December 2013.

- Miguel Lopes, Patrick Meyer and Gianluca Bontempi, "Estimation of temporal lags for the inference of gene regulatory networks from time series". Belgian-Dutch Conference on Machine Learning and Workshop on Predictive Modeling for the Life Sciences (Benelearn and PLMS), 2012.

The following publications are not directly related to the topics of the thesis, but have contributory work of my own and were developed during the writing of this thesis.

- Laura Marroqui, Miguel Lopes, Reinaldo S dos Santos, Fabio A Grieco, Merja Roivainen, Sarah J Richardson, Noel G Morgan, Decio L Eizirik, "Differential cell autonomous responses determine the outcome of coxsackievirus infections in murine pancreatic α and β cells", E-Life, 4, 2015.

- Flora Brozzi, Tarlliza R Nardelli, Miguel Lopes, Isabelle Millard, Jenny Barthson, Mariana Igoillo-Esteve, Fabio A Grieco, Olatz Villate, Joana M Oliveira, Marina Casimir, Marco Bugliani, Feyza Engin, Gkhan S Hotamisligil, Piero Marchetti, Decio L Eizirik, "Cytokines induce endoplasmic reticulum stress in human, rat and mouse beta cells via different mechanisms", Diabetologia, 2015.

- Baroj Abdulkarim, Marc Nicolino, Mariana Igoillo-Esteve, Mathilde Daures, Sophie Romero, Anne Philippi, Valrie Sene, Miguel Lopes, Daniel A Cunha, Heather P Harding, Cline Derbois, Nathalie Bendelac, Andrew T Hattersley, Dcio L Eizirik, David Ron, Miriam Cnop, Ccile Julier, "A missense mutation in PPP1R15B causes a syndrome including diabetes, short stature and microcephaly", Diabetes, 2015.

1.11.5 Data and software

The following data and software was made available and developed in the context of this thesis.

- A R package to estimate AUPRC significance, implementing the approach described in the Chapter 4. It takes as input a vector of scores and a gold standard.[1]
- A R package to estimate scores of Granger causality tests and other dynamic network inference methods. The novel methods are available, as well as all the methods assessed in the Section 5.3.[2]
- Datasets and R scripts to replicate the experiments of Section 5.3.[3]
- β-cell gene expression time series datasets, of human[4] and rat[5]. These are presented and used in the Chapter 6.

1.12 Structure of the thesis

This thesis is structured as follows: Chapter 2 consist of preliminaries for causal and network inference. Chapter 3 describes the state of the art of GRN inference. The three following chapters constitute the contributions part of this thesis. Chapter 4 is on the derivation of the mean and variance of the null AUPRC. Chapter 5 is on GRN inference from time series. Chapter 6 is on knowledge inference on β-cells in models of Type 1 Diabetes.

[1] https://github.com/miguelaglopes/pranker

[2] https://github.com/miguelaglopes/GCnetinf

[3] https://github.com/miguelaglopes/NetInfExps

[4] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53454

[5] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53453

Part II

Preliminaries and state of the art of network inference

2

Preliminaries

This chapter provides a foundation for the topic of causal inference and variable selection. It first presents relevant notions of statistical dependence in the general case (using information theory) and in the linear case. The theory behind causal inference (structural equation models, Bayesian networks) is then presented. Analysis and causal inference in time series is also discussed. Random forests are one example of a machine learning method that may be used to variable selection and network inference, and is also described in this chapter.

Chapter outline Basics of information theory are presented in the Section 2.1. Section 2.2 is on the estimation of linear dependencies. Section 2.3 introduces structural equation models and Section 2.4 the theory behind graphical causal models. The estimation of graphical models is addressed in 2.5. Decision trees and random forests are introduced in Section 2.6, and the analysis and inference in time series is addressed in the Section 2.7.

2.1 Basics of information theory

2.1.1 Introduction

Entropy was first introduced in the context of statistical thermodynamics, measuring the number of microscopic states of a system consistent with its observed macroscopic state, and playing a central role in the second law of thermodynamics (147, 153). Entropy was then extended by Shannon leading to the emergent field of information theory (233). The information entropy measures the amount of uncertainty, or informative content, associated with a random variable. Information-theoretic concepts have been applied in the last decades to a variety of biological problems (19), such as the analysis of neuron spike trains (188), cell signaling (219) or gene network inference (256).

Information theory allows to quantify the statistical dependence between two variables through the notion of mutual information. Mutual information is the reduction of the entropy of one variable when the other is conditioned on. Two variables are statistically independent if and only if their mutual information is zero (the mutual information is always non-negative).

If two variables are elliptically distributed (eg. multivariate Gaussian), its covariance is zero if and only if its mutual information is zero (same with partial correlation and conditional mutual information, regarding conditional dependence). In this case, statistical dependence is equivalent to linear dependence, and the mutual information and Pearson correlation are a bijective function of one another. Entropy and the mutual information are discussed in what follows (for a reference see (57)).

2.1.2 Entropy and the mutual information

Information entropy is often defined for discrete variables but can be extended for continuous variables, in this case it is called differential entropy (we will restrict to the discrete case).

Definition 2. *The entropy of a discrete random variable X with probability mass function $p(x)$ and support S_x is:*

$$H(X) = -\sum_{x \in S_x} p(x) \log p(x) \tag{2.1}$$

If the base of the logarithm is 2, the unit for entropy is the bit, if it is e, the unit is the nat. The conditional entropy is $H(X|Y) = H(X,Y) - H(Y)$. Conditioning does not increase entropy: $H(X|Y) \leq H(X)$. It follows that $H(X,Y) \leq H(X) + H(Y)$.

Definition 3. *The mutual information between two discrete random variables X and Y, with respective support S_x and S_y is:*

$$I(X;Y) = \sum_{x \in \ S_X} \sum_{y \in S_Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2.2}$$

Its relation with entropy is the following:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) \tag{2.3}$$

Figure 2.1 is a common representation of the mutual information and entropies of two variables. If two variables are independent, then $p(x,y) = p(x)p(y)$ and their mutual information is zero. A useful property is the following:

Theorem 1. *(Data processing inequality.) If $X_1 \perp\!\!\!\perp Y|X_2$:*

$$I(X_1;Y) \leq \min\{I(X_1;X_2), I(X_2;Y)\} \tag{2.4}$$

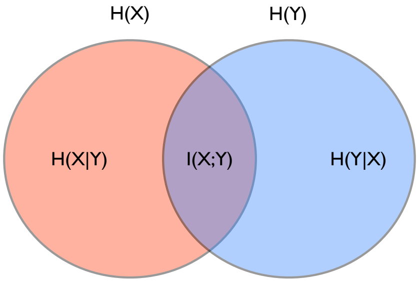


Figure 2.1: Mutual information - The mutual information between X and Y is $H(X) - H(X|Y)$.

The interaction information is the trivariate extension of the mutual information, defined next.

Definition 4. *The interaction information between three random variables X, Y and Z is: (175)*

$$I(X;Y;Z) = I(X;Y|Z) - I(X;Y) \quad (2.5)$$

$$= I(X;Z|Z) - I(X;Z) \quad (2.6)$$

$$= I(Y;Z|X) - I(Y;Z) \quad (2.7)$$

While conditioning always reduces entropy, conditioning may increase or decrease the mutual information. Thus, the interaction information may be positive or negative: tf the term $I(X;Y|Z)$ is lower than $I(X;Y)$ the interaction information is negative, and positive otherwise.

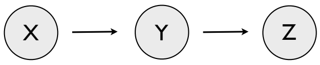


Figure 2.2: An example of negative interaction information - Conditioning on any variable reduces the mutual information between the other two variables: eg. $H(X;Z|Y) < H(X;Z)$.

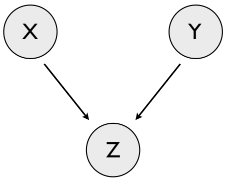


Figure 2.3: An example of positive interaction information - Conditioning on any variable increases the mutual information between the other two variables: eg. $H(X;Y|Z) > H(X;Y)$.

An example of negative interaction information is represented in the Figure 2.2. In this case, if we condition on Y, X and Z become independent, thus $I(X,Z) > I(X,Z|Y)$. An example of positive interaction information is represented in the Figure 2.3. Y and X are independent and $I(X;Y) = 0$, but if Z is conditioned they become dependent and $I(X;Y|Z) > 0$. This case is known as a collider in graphical models (Section 2.4). The mutual information can be generalized to a higher number of variables, although its interpretation is non-trivial - requiring reasoning with hypergraphs (16).

2.1.3 Estimation of entropy and mutual information

Mutual information can be estimated from entropy estimations. In the case of continuous variables, entropy is known as differential entropy and may be estimated through integration (analytical or numerical). It requires the estimation of a continuous probability density function, in a parametric or non-parametric way, eg. by kernel density estimation (202). Another approach is to discretize the observations (known as binning) and use the discrete entropy formula. Two common discretization strategies are equal-size binning (bins of the same size) and equal-frequency binning. In the latter, bins are created such that all bins are observed the same number of times, and in the resultant discretized probability function all bins have an equal area. For these and other approaches see (71).

In the discrete case, the maximum likelihood (ML) estimation and a bias corrected version of it are described in what follows. Consider a discrete random variable X, defined in a support $S_x = \{x_1, .., x_N\}$. Denote the number of times a value x_i is observed by n_i. The total number of observations by n. The ML estimation is based on the ML estimation of the density mass function.

$$\hat{H}(X) = -\sum_{i=1}^{N} \hat{p}(x_i) \log \hat{p}(x_i) \tag{2.8}$$

where $\hat{p}(x_i) = \frac{n_i}{n}$. The ML estimation of the density mass function is unbiased, however the entropy estimation is negatively biased.[1] The asymptotic bias of the ML entropy estimation has been shown to be $(m-1)/2n$, where m is the number of non-zero probability bins. The bias-corrected ML estimation is called the Miller-Madow estimator. It is:

$$\hat{H}(X) = \hat{H}(X) + \frac{m-1}{2n} \tag{2.9}$$

For improvements on the Miller-Madow estimation and a description of the ML entropy bias see (201). For a review on non-parametric methods see (15). Bayesian estimation approaches are described in (6, 188).

[1] This happens because an underestimation in $\hat{p}_X$ causes a greater error in the entropy estimation than an upwards overestimation of the same quantity.

In the case of variables following a multivariate Gaussian distribution, the mutual information is a bijective function of the Pearson correlation, presented next, in Section 2.2:

$$I(X;Y) = -\frac{1}{2}\ln(1-\rho^2_{X,Y}) \tag{2.10}$$

Conditional mutual information is analogously obtained from the partial correlation (introduced in Section 2.2.2):

$$I(X_1;X_2|X_3) = -\frac{1}{2}\ln\left(1-\rho^2_{X_1,X_2|X_3}\right) \tag{2.11}$$

2.2 Estimation of linear dependences

Two variables are linearly dependent if the realization of one above or below its mean increases the conditional probability of observing the other above or below its respective mean. The linear dependence of two variables may be quantified with their covariance. It measures how well one variable is predicted with a linear transformation of the other. The Pearson correlation is the covariance divided by the product of the variables' standard deviation and is bounded between -1 and 1 (in these extremes the variables are perfectly linearly correlated, negatively or positively). Analogously to the Pearson correlation, the partial correlation is a normalized measure of conditional linear dependence, measured in linear regression models. These concepts are presented in some detail in what follows.

2.2.1 Linear regression

In linear regression, a target variable Y is modeled as a linear function of predictor variables X (110).

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon_i \tag{2.12}$$

The vector β measures the influence of each X_i on Y. In particular, β_i is the change in Y after an unit increase in X_i, all things constant. A non-zero coefficient indicates a linear conditional dependence between the respective predictor variable and the target. The standard way to estimate β is through ordinary least squares (OLS), minimizing the empirical mean squared error ($\hat{\text{MSE}}$). Let $\mathbf{x_i}$ be a vector of realizations of the variable X_i and $\mathbf{X}$ a matrix of observations of the set of variables X. Assuming mean equal to zero and $\beta_0 = 0$, the $\hat{\text{MSE}}$ as a function of β, is: (110)

$$\hat{\text{MSE}}(\beta) = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{p} \beta_i x_{i,j} \right)^2 \tag{2.13}$$

or alternatively, in a matrix form:

$$\hat{\text{MSE}}(\beta) = \frac{1}{n} \left(\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}} (\mathbf{y} - \mathbf{X}\beta) \right) \tag{2.14}$$

By minimizing $\hat{\text{MSE}}$, we obtain the OLS coefficients $\hat{\beta}_{OLS}$:

$$\hat{\beta}_{OLS} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left((\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}} (\mathbf{y} - \mathbf{X}\beta) \right) \tag{2.15}$$

By taking the derivative relative to β, equaling it to zero and solving, we have:

$$\hat{\beta}_{OLS} = (\mathbf{X^T X})^{-\mathbf{1}} \mathbf{X^T y} \tag{2.16}$$

If the additive noise ϵ in equation (2.12) is uncorrelated and homoscedastic the OLS is the best linear unbiased estimator (BLUE) of the linear regression coefficients, MSE-wise (110).

2.2.2 Linear correlation

The Pearson (or linear) correlation measures the linear dependence between two variables.

Definition 5. *The Pearson correlation coefficient between two variables X and Y is:*

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{2.17}$$

Other measures of correlation exist, such as the Spearman correlation, which is the linear correlation of the ranks of the samples (the sample of lowest value has rank 1, the sample of highest value has rank n).

The partial correlation measures the conditional linear dependence between two variables. It is obtained from the inverse of the covariance matrix (the concentration matrix Ω) of the considered variables (146). In the following consider a set of variables $X = \{X_i\}, i \in \{1, .., p\}$.

Definition 6. *The partial correlation between X_i and X_j given $X^{\setminus\{i,j\}}$ is:*

$$\rho_{X_i,X_j|X^{\setminus\{i,j\}}} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}} \tag{2.18}$$

The cardinality of the set $X^{\setminus\{i,j\}}$ defines the order of the partial correlation. The partial correlation (changing notation, between X and Y conditioned on a set Z) may also be obtained recursively through the following formula: (146)

$$\rho_{X,Y|Z} = \frac{\rho_{X,Y|Z^{\setminus Z_0}} - \rho_{X,Z_0|Z^{\setminus Z_0}}\rho_{Y,Z_0|Z^{\setminus Z_0}}}{\sqrt{(1-\rho^2_{X,Z_0|Z^{\setminus Z_0}})(1-\rho^2_{Y,Z_0|Z^{\setminus Z_0}})}}, \quad Z_0 \in Z \tag{2.19}$$

A q-th order partial correlation is then computed from (q-1)-th order partial correlations. This makes it possible to compute any partial correlation recursively from a Pearson correlation (when $Z = Z_0$).[1]

2.2.3 Linear regression and partial correlation

The linear regression and partial correlation coefficients are closely related. If X_y is a linear function of $X^{\setminus y}$, the coefficient associated with X_i is given by: (146)[2]

$$\beta_{i,y} = \frac{-\omega_{i,y}}{\omega_{y,y}} \tag{2.20}$$

[1] If only full order partial correlations are to be kept, the covariance matrix inversion method is computationally preferable to the recursive method, as the latter computes a single partial correlation in approximately the same time $\mathcal{O}(p^3)$ it takes to invert the covariance matrix (and obtain all pairs of partial correlations at once) (139).

[2] When estimating the linear coefficient matrix, this approach is computationally preferable to regressing each variable at a time. It requires the estimation of a single $p \times p$ concentration matrix, whereas the latter case requires p $(p-1) \times (p-1)$ concentration matrices.

$\beta_{i,y}$ is related to the partial correlation in the following way:

$$\beta_{i,y} = \rho_{X_i,X_y|X^{\setminus i,y}} \sqrt{\frac{\omega_{i,i}}{\omega_{y,y}}} \tag{2.21}$$

In the bivariate case (one independent variable X and one dependent Y) the coefficient β is obtained from the Pearson correlation:[1]

$$\beta = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} \tag{2.22}$$

Full rank requirement The estimation of partial correlations or linear regression coefficients through OLS requires the covariance matrix to be invertible.[2] When that is not the case, the linear model does not have an unique optimal solution. Regularization tackles this limitation.

2.2.4 Regularization

Regularization is the consideration of parameter constraints in order to solve ill-posed problems and reduce model variance. Forms of regularization include bounds on the parameter space, or penalization of model complexity. In linear regression, regularization usually consists in the estimation of a full rank approximation of the covariance matrix (some strategies as referred to in the Section 2.5.1 on graphical models) or in the modification of the OLS error minimization function (equation (2.15)) to include a penalty term proportional to the L^p-norm of the β coefficients (23):

$$\hat{\beta}_{reg} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left((\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}} (\mathbf{y} - \mathbf{X}\beta) + \lambda |\beta|^p \right) \tag{2.23}$$

where $\lambda > 0$ and $|\beta|^p = \sum_i^n |\beta_i|^p$ (to avoid confusion, instead of the typical p, the number of variables is denoted by n). Usually, p is 1 or 2. The L^2 penalty regularization (called ridge regression, or Tikhonov regularization) adds a constant in the diagonal of the covariance matrix, making it invertible and enabling the estimation of β through OLS. The parameter λ can be selected with respect to various criteria, see (45, 100). The L^1 penalty regularization is called the lasso, and the combination of L^p-norm penalty regularization is called the elastic-net (297). Linear regression can be used for variable selection through the identification of non-null

[1] Both the concentration and covariance matrix are of dimension 2×2, and the division of the terms of the diagonal of a 2×2 matrix is equal to the inverse of the division of the diagonal terms of the inverse matrix.

[2] An invertible matrix is a square matrix that has full rank, meaning all its rows, or columns, are linearly independent in the linear algebra sense - no row/column can be obtained as a linear combination of the other rows/columns. The column rank and the row rank of a matrix are always equal. The maximum allowed rank of a matrix of dimension $n \times p$ is the minimum between n and p. The rank of a square matrix $X^T X$ of dimension $p \times p$ is the same as the rank of X, and $X^T X$ only has full rank if the rank of X is p, which is only possible if $n \geq p$.

coefficients. L^p-norm regularization is useful in this case as it induces sparsity and magnitude decrease in the estimated non-null coefficients.

L^p-norm regularization is discussed in more detail in what follows, as it is commonly used in network inference (see the Chapter 3).

The lasso The L^1-norm penalty regularization is known as the lasso (least absolute shrinkage and selection operator) (254). When $n > p$ and the predictors are uncorrelated (orthogonal) the lasso has a closed form solution. In the general case, several strategies exist in the literature to compute the lasso solution path (ie. the linear coefficients as λ decreases from infinity to zero). One is least angle regression (74) and makes use of the fact that the lasso path is piece-wise linear (see below). Another is coordinate descent, updating the lasso coefficients one at a time until convergence (89, 283). For other strategies see (199) or (203). The lasso has been shown to be consistent (selects the true model parameters as n tends to infinity) under certain conditions (293, 296), and unique (returning a single solution) in general conditions (predictors drawn from a continuous probability distribution) (255).

Least angle regression (lars) The lasso solution path can be obtained with least angle regression (74). Here, the linear regression model is initialized with the selection of the predictor most correlated with the target. The non-zero coefficients jump together in a direction equidistant (having equal angles) to the current predictors. The size of the jump corresponds to when the residual vector becomes as correlated with a new predictor as with the previously selected. This happens when the residual vector bisects the angle between the current predictors and the new one (74). This new predictor then enters the model. Lars is an improvement of a previous method called forward stagewise selection, where the coefficients are gradually incremented one a time. In lars, only a few steps (jumps, corresponding to when a predictor enters the model) are necessary to compute the stagewise selection path. If a regressor leaves the model when its coefficient hits zero, then the lars path is equivalent to the lasso path (74).

The elastic net Tikhonov regression tends to return few non-zero coefficients, difficulting the interpretation of the model in a context of variable selection (although linear coefficients may be transformed into partial correlations, enabling the ranking of coefficients and selection via a threshold). The lasso returns fewer non-zero coefficients (maximum n) and a more interpretable model. On the other hand, it tends to select only one of a group of correlated variables, and the constraint on the number of returned non-zero coefficients may be deemed too limiting (297). Tikhonov regression is preferable when there is a high number of non-zero coefficients of similar magnitude; the lasso is preferable when the model is characterized by a few large coefficients of uncorrelated variables. The elastic net (297) tackles the limitations of the lasso by combining Tikhonov regression and the lasso. It adds two penalty terms to the OLS MSE minimization equation, one proportional to the L^1 norm of β_i and the other proportional to the L^2 norm.

$$\hat{\beta}_{enet} = \underset{\beta \in \mathbb{R}^p}{\arg\min}((\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda_{\mathbf{1}}|\beta|^{\mathbf{1}} + \lambda_{\mathbf{2}}|\beta|^{\mathbf{2}}) \tag{2.24}$$

where $\lambda_1, \lambda_2 > 0$. Elastic net can be solved with adaptations of least angle regression (297) and coordinate descent (89).

Other variations When the regressors are grouped, L^2 regularization at the intra-group level may be combined with L^1 regularization at the inter-group level (289) (if the number of groups is 1 this is equivalent to ridge regression; if the number of groups is the number of regressors, to the lasso). This approach induces sparsity in the number of selected groups (but not at the intra-group level). It addresses the limitation of the lasso selecting only one variable from a group of correlated ones. Another extension to the lasso consists in weighting differently each coefficient (each multiplied by a penalty weight). This variation asymptotically identifies the true non-zero coefficients in all cases (whereas the lasso does not) (296).

2.3 Linear structural equation models

Linear structural equation models (SEMs) originated in the 20s, with Sewall Wright's path analysis (281), a graphical method to deduce variances and covariances in linear causal models. Linear SEMs are a formalization of path analysis, providing the association between the parameters (linear coefficients) of a linear causal model and the covariances of its variables. SEMs have been extensively used to model causality in econometrics and in the social sciences (184).[1]

SEMs may consider observed or latent (not observed) variables, exogenous or endogenous (the latter are a function of other variables and are determined within the system). Disturbances (eg. representing measurement errors) are latent exogenous variables uncorrelated with other exogenous variables, and influencing only one endogenous variables. Each endogenous variable is associated with a single disturbance.

In a SEM, variables are usually defined recursively, where each is a function of other previously defined variables. Non-recursive SEMs contain feedback loops and require particular treatment (condensing loops (184)).

Linear SEMs may be used for confirmatory purposes, by comparing observed covariances with predicted ones. SEMs may also be used in exploratory analysis, as model parameters (ie. its linear coefficients, in which non-zero values indicate a causal linear relation) may be estimated from data. For some approaches see (184).

The term SEM usually implies linear models, due to its origins and decades of linear SEM-based research and applications. SEMs have been generalized to the non-linear case and formalized in a context of causal models (206). Following a chronological order, this section discusses linear SEMs, while theory on causal models is introduced in Section 2.4.

An example of a (recursive) SEM, modeling endogenous variables as a function of exogenous, is:

$$\begin{cases} X_4 = \beta_{1,4}X_1 + \beta_{2,4}X_2 + \epsilon_4 \\ X_5 = \beta_{2,5}X_2 + \beta_{3,3}Y_3 + \epsilon_5 \\ X_6 = \beta_{5,6}X_5 + \beta_{8,6}X_8 + \epsilon_6 \\ X_7 = \beta_{3,7}X_3 + \beta_{8,7}X_8 + \epsilon_7 \end{cases} \tag{2.25}$$

X_1, X_2, X_3 and X_8 and the ϵ terms (the disturbances) are the exogenous variables. The variables X_4, X_5 and X_6 and X_7 are the endogenous variables. In SEMs, the variance/covariance of exogenous variables is usually given.

[1] There has been some debate over their adequacy to causal inference. One point in contention is the distinction between causal and statistical association, see (86) for negative perspective on SEM usage, see (26) for a rebuttal of common arguments against SEMs.

2.3.1 Path diagrams

The variances and covariances in a SEM can be obtained through the analysis of path diagrams. Figure 2.4 is a SEM path diagram representation of the SEM described in the system of equations (2.30).[1] Here, each variable manifest variable is represented with a square, and each latent variable with a circle. The model coefficients are represented by directed straight arrows (from the independent to dependent variable). Exogenous variables only have directed straight arrows pointed from them (not towards them). The covariance between exogenous variables (if non-zero) is represented with curved bi-directional arrows. The variance of exogenous variables is represented by bi-directional arrows (loops) from and to the variables.

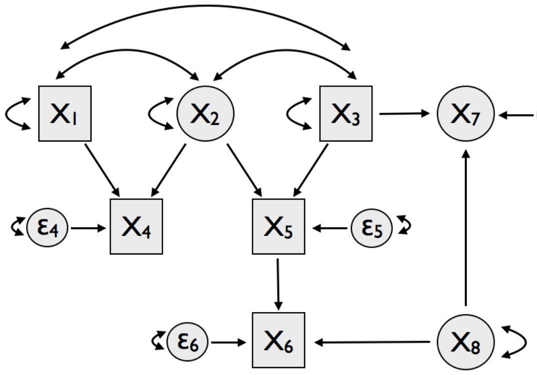


Figure 2.4: Path diagrams in structural equation models - Path diagram representation of a SEM. Exogenous variables (X_1, X_2, X_3, X_8 and disturbances) only have directed straight arrows pointed from them. Endogenous variables have arrows pointed towards them. Latent variables are represented in circles. Manifest variables are represented in squares. Each endogenous variable is influenced by a disturbance variable ϵ. The variance and covariance of exogenous variables are represented by curved bi-directional arrows.

2.3.1.1 Path tracing rules

Variances and covariances in a SEM path diagram can be obtained with a set of rules, described herein (based on (184)). A SEM path is a sequence of arrows (straight or curved) along variables, with the restriction that no two consecutive arrows in the path are pointed towards the same variable (ie. a SEM path does not include the sequence of arrows $X_2 \rightarrow X_5$ and $X_5 \leftarrow X_3$, in the Figure 2.4).

[1] In the system of equations (2.30), for simplicity, disturbances are assumed to have a unitary linear coefficient, and their contribution to the model is determined by their variance.

A SEM path may contain the same arrow twice, if separated by a variance loop (ie. a path may contain the sequence of arrows $X_4 \leftarrow X_2$ (variance loop) $\leftrightarrow X_2 \rightarrow X_4$, in the Figure 2.4). Finally, a SEM path must go through an exogenous variable, passing necessarily through either its variance loop or through a covariance bi-directional arrow.

Only one variance loop or covariance curved arrow may be part of a path (otherwise there is necessarily a sequence of two arrows pointed to the same variable). Two paths are distinct if they are not constituted by the same sequence of arrows. Straight arrows are associated with the respective coefficient of the SEM model, and curved arrows with a respective variance/covariance. In what follows, for simplicity, we may say that each arrow has an associated coefficient (linear coefficient in the case of straight arrows; variance/covariance in the case of curved arrows).

2.3.1.2 Variance

The variance of a variable is obtained by adding the product of the coefficients of each SEM path, for all distinct paths from and to that variable. These paths enter the variable with the arrow pointed towards it, as they must go through an exogenous variable. In the Figure 2.4, there are 5 distinct paths from and to the variable X_5: the first leaves X_5, goes to X_2 and around its variance loop, and then goes back to X_5. The second leaves X_5, goes to X_3 and around its variance loop, and then goes back to X_5. The third leaves X_5, goes to X_2, then to X_3 via the covariance arrow, and then to X_5. The fourth path does the same in the opposite direction. The fifth path goes around the disturbance variable. Note that there is no path through X_7, as all arrows are pointed towards it. The variance of X_5 is then:

$$\sigma_5^2 = \beta_{2,5}^2\sigma_2^2 + \beta_{3,5}^2\sigma_3^2 + 2(\beta_{2,5}\beta_{3,5}\sigma_{2,3}) + \epsilon_5^2 \tag{2.26}$$

2.3.1.3 Covariance

The covariance of two variables is obtained by adding the product of the coefficients of each path, for all distinct paths between the first variable and the second variable. Consider X_5 and X_6. The paths between them are the paths of the variance of the X_5 with the addition of the arrow between X_5 and X_6. Thus, the covariance between X_5 and X_6 is:

$$\sigma_{5,6} = \beta_{5,6}\left(\beta_{2,5}^2\sigma_2^2 + \beta_{3,5}^2\sigma_3^2 + 2(\beta_{2,5}\beta_{3,5}\sigma_{2,3}) + \epsilon_5^2\right) = \beta_{5,6}\sigma_5^2 \tag{2.27}$$

Consider now the case of variables X_4 and X_5. There are four paths from X_4 to X_5: via X_2, via X_1 and X_2, via X_1 and X_3, and via X_2 and X_3. The covariance is then:

$$\sigma_{4,5} = \beta_{2,4}\sigma_2^2\beta_{2,5} + \beta_{1,4}\sigma_{1,2}\beta_{2,5} + \beta_{1,4}\sigma_{1,3}\beta_{3,5} + \beta_{2,4}\sigma_{2,3}\beta_{3,5} \tag{2.28}$$

2.3.2 Covariance matrix estimation

The covariance matrix of all the manifest variables can be estimated through matrix operations, as a function of the model coefficients and covariances of the exogenous variables. Following (184), let η be a set of latent endogenous variables (of size m); Y a set of manifest endogenous variables (of size p); ξ a set of latent exogenous variables (of size n); X a vector of manifest exogenous variables (of size q); ϵ a set of disturbances (of size $m+p$, one for each endogenous variable). In a matrix form, the endogenous variables can be modeled as:

$$\begin{bmatrix} \eta \\ Y \end{bmatrix} = \mathbf{A} \begin{bmatrix} \eta \\ Y \end{bmatrix} + \begin{bmatrix} \mathbf{\Gamma}_{\xi} & \mathbf{\Gamma}_{\mathbf{X}} & \mathbf{\Gamma}_{\epsilon} \end{bmatrix} \begin{bmatrix} \xi \\ X \\ \epsilon \end{bmatrix} \tag{2.29}$$

$\mathbf{A}$ is a $(p+m) \times (p+m)$ matrix representing the contribution (ie. linear coefficients) of endogenous variables on the modeling of other endogenous variables. $\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{\xi} & \mathbf{\Gamma}_{\mathbf{X}} & \mathbf{\Gamma}_{\epsilon} \end{bmatrix}$ is a $(p+m) \times (n+q+p+q)$ matrix representing the contribution of exogenous variables (latent, manifest and disturbances) on the endogenous variables. The variance/covariance matrix of the manifest variables (endogenous and exogenous) is given by:

$$\Sigma = \mathbf{G}\mathbf{B}^{*^{-1}}\mathbf{\Gamma}^{*}\mathbf{\Phi}\mathbf{\Gamma}^{*^{\mathbf{T}}}\mathbf{B}^{*^{-1^{\mathbf{T}}}}\mathbf{G}^{\mathbf{T}} \tag{2.30}$$

This equation is known as the fundamental theorem of structural equation modeling (184). In the following we describe the meaning of these symbols, for the proof of equation (2.30) see (184). $\mathbf{G}$ = $\begin{bmatrix} \mathbf{G_y} & 0 \\ 0 & \mathbf{G_x} \end{bmatrix}$ where $\mathbf{G_y} = \begin{bmatrix} 0 & \mathbf{I} \end{bmatrix}$ with the null matrix of dimension $p \times m$ and the identity matrix is of dimension $p \times p$; and $\mathbf{G_x} = \begin{bmatrix} 0 & \mathbf{I} & 0 \end{bmatrix}$ with the first null matrix of dimension $q \times m$, the identity matrix of dimension $q \times q$ and the second null matrix of dimension $q \times (m+p)$. $\mathbf{B}^{*^{-1}}$ is a square matrix with number of rows/columns equal to the number of columns of $\mathbf{G}$ $(2m + 2p + n + q)$ and is equal to $\begin{bmatrix} \mathbf{B^{-1}} & 0 \\ 0 & \mathrm{I} \end{bmatrix}$, where $\mathbf{B} = \mathbf{I} - \mathbf{A}$. $\mathbf{\Gamma}^{*} = \begin{bmatrix} \mathbf{\Gamma} \\ \mathbf{I} \end{bmatrix}$, with dimension $(2m + 2p + n + q) \times (n + q + p + q)$. Finally:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_{\xi\xi} & \mathbf{\Phi}_{\xi\mathbf{X}} & 0 \\ \mathbf{\Phi}_{\mathbf{X}\xi} & \mathbf{\Phi}_{\mathbf{XX}} & 0 \\ 0 & 0 & \mathbf{\Phi}_{\epsilon\epsilon} \end{bmatrix} \tag{2.31}$$

$\mathbf{\Phi}_{\xi\xi}$ is the $n \times n$ covariance matrix of the latent exogenous variables (excluding disturbances); $\mathbf{\Phi}_{\mathbf{XX}}$ is the $q \times q$ covariance matrix of the manifest exogenous variables; $\mathbf{\Phi}_{\epsilon\epsilon}$ is the $(m+p) \times (m+p)$ covariance matrix of the disturbances (assumed to be diagonal); $\mathbf{\Phi}_{\mathbf{X}\xi}$ is the covariance matrix between the manifest exogenous and the latent exogenous. The covariances between disturbances and exogenous variables is assumed to be zero.

2.4 Graphical models

A graphical model is a representation of the conditional dependences of a multivariate probability distribution $\mathcal{P}$ in a graph characterized by nodes and edges (or vertices), $\mathcal{G} = \langle N, V \rangle$. Variables are represented in nodes, and conditional dependences between variables by edges.

The edges of graphical models may be directed or undirected. Direct graphical models usually have a causal interpretation (ie. edges are directed from causes to effects (206)). Nodes connected by an edge to another node are adjacent or neighbors of that node. In directed graphs, nodes are referred to in terms of parents and children (and ancestors/descendants): a node is a parent of another if there is a directed edge from the first to the second. Ancestors include parents, grandparents and so on. A path consists of a set of edges (directed or undirected) connecting two nodes along a set of intermediate nodes. A collider of a directed path is a node in which the edges of the path point towards it.

Causal models Following Pearl (206) (page 203), a causal model M is defined to be a composition of three elements: a set of background (or exogenous) variables U, determined by factors outside the model; a set of endogenous variables V, determined by variables within the model (exogenous and endogenous); and a set of functions uniquely determining the value of each endogenous variable, given the values of the other variables.

A probabilistic causal model (also referred to as a structural equation model (208), in which the linear case was discussed previously) is composed of a causal model M plus a probability distribution of U.

2.4.1 Separation and d-separation

We adopt the term separation for undirected graphs and d-separation for directed graphs, both denoted by the symbol $\perp\!\!\!\perp_G$. The symbol $\perp\!\!\!\perp_{\mathcal{P}}$ denotes independence in $\mathcal{P}$. The fact that Z separates/d-separates X and Y is represented with $X \perp\!\!\!\perp_{\mathcal{G}} Y|Z$.

Separation is defined as follows (266).

Definition 7. *(Separation in undirected graphs.) Let $\mathcal{G}$ be an undirected graph, and X, Y and Z three disjoint sets of nodes. Z separates X and Y if Z contains at least one node in each path between X and Y.*

D-separation is the analogous concept for directed acyclic graphs (266). A crucial characteristic is the role of colliders.

Definition 8. *(D-separation in directed acyclic graphs.) Let $\mathcal{G}$ be a directed graph, and X, Y and Z three disjoint sets of nodes. Z d-separates X and Y if, for all paths between X and Y,* *there is a node W satisfying one of the two conditions:*

- *W is a non-collider in the path and $W \in Z$.*
- *W is a collider in the path and does not - and neither do its descendants - belong in Z.*

If a path does not contain colliders, its edges follow the same direction and the path is *active*. If a path contains colliders, the path is *blocked*. A set of nodes present in all active paths between two nodes d-separates these two nodes. A collider *d-connects* nodes at opposite sides (relative to the collider) of a blocked path.

Markov property If all graphical separations/d-separations are associated with conditional independences in $\mathcal{P}$, then the graph has the Markov property with respect to $\mathcal{P}$ (206).

Definition 9. *(Markov property) A graph $\mathcal{G}$ has the Markov property with respect to $\mathcal{P}$ if and only if, for any pair of nodes X and Y and set of nodes Z:*

$$X \perp\!\!\!\perp_{\mathcal{G}} Y|Z \implies X \perp\!\!\!\perp_{\mathcal{P}} Y|Z \tag{2.32}$$

This property comes in three forms, which are equivalent if $\mathcal{P}$ is strictly positive (they are not distinguished in this section and are referred to as the Markov property) (146, 206). Note that a fully connected graph (where all nodes are adjacent to all other nodes) trivially respects the Markov property. The implication in the opposite direction is guaranteed by the property of faithfulness (see next section). In directed graphs, the Markov property is referred to as the causal Markov property. It implies that a node is independent of all its non-descendants (direct or indirect effects), given all its parents (direct causes) (206).

2.4.2 Faithfulness and minimality

Faithfulness The Markov property guarantees that each separation/d-separation relation in $\mathcal{G}$ is associated with a conditional independence in $\mathcal{P}$. This imposes a minimum number of edges in $\mathcal{G}$. However, if extra edges are added to the graph, the Markov property is still valid. Edge sparseness is guaranteed by the opposite implication (conditional independence implying graph separation). This implication, together with the Markov property, is known as faithfulness.[1]

Definition 10. *(Faithfulness.) A graph $\mathcal{G}$ is faithful to a probability distribution $\mathcal{P}$ if only if, for any pair of nodes X and Y and set of nodes Z:*

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z \iff X \perp\!\!\!\perp_{\mathcal{P}} Y \mid Z \tag{2.33}$$

[1] The term *stable* is also commonly used (206).

There are instances where a true causal structure (represented in a directed graph) is unfaithful to the probability distribution of its variables. For instance, if $X = \beta Y + \alpha Z + \epsilon_X$ and $Y = -\alpha Z + \epsilon_Y$, X is independent with Z (206). However, such a precise parameter (the linear coefficients) canceling has probability zero if the parameters follow a continuous probability distribution function and are independent (178, 206)). Non faithfulness may also occur when the multivariate probability distribution is not strictly positive and there are functional dependences (variables fully determined by other variables) (96, 243, 266). The requirement of a strictly positive probability distribution is formalized in the graphoid axioms (96, 209).

Minimality A faithful graph is the sparsest graph that respects the Markov property, as all conditional independences in $\mathcal{P}$ are represented in the form of absences of edges. However, as referred above, faithfulness may not be guaranteed. On the other hand, there is always a *minimal* graph with respect to a probability distribution (187). Together with faithfulness, the property of minimality ensures the adoption of simpler graphs, in the spirit of Occam's razor.

Definition 11. *(Minimality) A graph $\mathcal{G}$ is minimal with respect to a probability distribution $\mathcal{P}$ if the Markov property holds, but does not anymore if any edge is removed.*

2.4.3 Undirected and directed graphical models

Undirected graphical models in which the Markov property holds are known as Markov networks, or Markov random fields. If the variables are multivariate Gaussian distributed, dependence is measured with the linear (partial) correlation, and the Markov network is known as a graphical Gaussian model. Directed acyclic (or recursive) graphical models (DAGs) in which the causal Markov property holds are known as Bayesian networks (BN).

A difference between BNs and Markov networks is that Markov networks cannot represent colliders with independent parents. In this case, there is an edge between the two parents in the Markov network, as they are dependent given all other variables including the collider (such an edge does not exist in a faithful BN). Markov networks are then unfaithful if the underlying causal structure contains colliders with independent parents.

In BNs, the acyclic condition allows for a compact factorization of the multivariate probability distribution in terms of individual marginal distributions using the chain rule of probability (206). This makes use of the fact that any node is conditionally dependent of all its non-descendants, given its direct parents. Let X_i^{pa} be the set of parents of X_i:

$$\mathcal{P}(X_1, .., X_p) = \prod_{i=1}^{p} \mathcal{P}(X_i | X_i^{pa}) \tag{2.34}$$

The acyclic condition prevents the modeling of feedback loops. However, these may be inferred indirectly using time series data, if each variable is represented in multiple nodes corresponding to different time points. Such networks are known as dynamic Bayesian networks (DBNs) (99, 185).

In BNs, edge direction may be identified in open collider situations: if X and Y are independent but are conditionally dependent given Z, the latter d-connects X and Y and is a collider in a path involving them. This is the principle behind edge orientation in constraint-based BN inference algorithms (Section 2.5.2). Other edges may then be directed in order to respect the acyclic condition, but if there are equivalent networks (see below) it is not be possible to uniquely orient all edges (206). However, edge direction is trivially identified in DBNs, as edges are directed from nodes at past time points to nodes at future time points.

2.4.4 Equivalent networks

The Markov property, faithfulness and minimality aim to guarantee an optimal representation of the dependences encoded in $\mathcal{P}$. However, there may be multiple directed graphs respecting these properties. This the case of graphs with the same skeleton (edges disregarding direction) and the same v-structures (ie. colliders in an open triplet). Such networks are said to be equivalent (51, 267).

Consider a simple open triplet case, with three nodes and two directed edges, and in which we fix the unconnected pair of nodes. There are four possible ways to orient the edges (Figures 2.5 and 2.6). Conditional independence only allows for the unique identification of one case, the collider configuration (Figure 2.5). In this case, X and Y are independent but conditionally dependent given Z. However, if X and Y are dependent but conditionally independent given Z, any of the three structures of Figure 2.6 are possible.

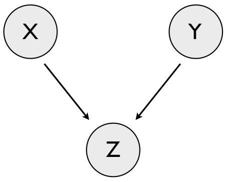


Figure 2.5: Collider configuration of an open triplet - X and Y are independent, but become conditionally dependent given Z.

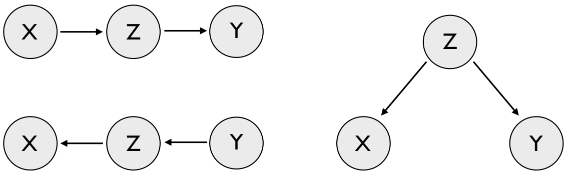


Figure 2.6: Non-collider configuration of an open triplet - X and Y are dependent, but become conditionally independent given Z. The three cases are d-separation equivalent.

2.4.5 Markov blankets

Definition 12. *The Markov blanket (MB) of a node X is the minimum set of nodes that (d-)separates X from all the other nodes in the network.*

Alternatively, it is the set of nodes that are not (d-)separated from X, given all other nodes in the network. In a Bayesian network, the MB of X is constituted by its parents, children and spouses (nodes that share with X a common child) (4, 205). Figure 2.7 illustrates the MB of a node X. Markov blankets are useful in the context of variable selection. Assuming the Markov property, given the MB of a variable X, no other variable is conditionally dependent with it. Thus, the MB of X is an optimal set of predictors of X: it is the minimum set of predictors for which there is no predictive gain in the consideration of extra predictors (4, 259). A Markov network (undirected model) is constituted by the edges connecting all nodes to the respective MB.

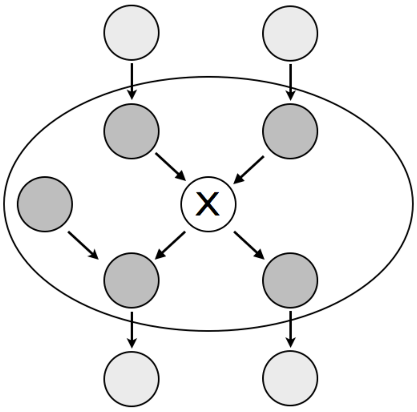


Figure 2.7: Markov blanket of node X in a DAG. - The Markov blanket of a node X is constituted by its parents, children and spouses (parents of a common child).

2.4.6 Computing the effects of interventions in Bayesian networks

In Bayesian networks, hypothetical effects of interventions (ie. setting a variable/node to a fixed value, represented with the do operator) may be obtained from observational data, if the causal effect is *identifiable* (206, 208). This condition means that the causal effect only depends on (and is uniquely identified from) the set of causal assumptions (encoded in the directed graph) and the probability distribution. Given them, the causal effect does not depend on possible variations of the functional parameters of the model (eg. linear coefficients). Any identifiable intervention effect may be estimated with a set of rules known as do-calculus (out of the scope of this thesis, see (121, 206, 237)).

Another approach to estimate the effects of do operations consists in covariate adjustments (meaning marginalizing out the covariates). Two known strategies to select these covariates are known as the *back-door* and *front-door* criteria (206). The back-door adjustment is briefly described next, providing the effect on Y of a $do(x)$ operation. It consists in adjusting for all the common causes of X and Y (and none of their effects).

Definition 13. *(Back-door criterion) A set of variables Z satisfies the back-door criterion with respect to two sets of variables X and Y if:*

- *Z blocks every path between X and Y with an arrow into X.*
- *No node in Z is a descendent of X.*

The adjustment for back-door covariates is:

Theorem 2. *(Back-door adjustment) If Z satisfies the back-door criterion with respect to X and Y, then:*

$$\mathcal{P}(y|do(x)) = \sum_{z \in S_z} \mathcal{P}(y|x,z)\mathcal{P}(z) \tag{2.35}$$

2.5 Estimation of graphical models

This section presents standard estimation approaches for undirected and directed graphical models, introduced in the previous section. Section 2.5.1 is on Gaussian graphical models, and Section 2.5.2 on Bayesian networks. Section 2.5.3 presents strategies to infer Markov blankets (constituting the Markov network), and sets of parents and children (constituting the skeleton of a Bayesian network).

2.5.1 Gaussian graphical models

As referred previously in Section 2.4, Gaussian graphical models (GGM) are undirected graphical models in which edges represent non-zero partial correlations between its nodes (146, 276).

When $n > p$ and the covariance matrix has full rank, the partial correlation matrix can be obtained by inverting the covariance matrix (Section 2.2.2). If the covariance matrix is not invertible, an invertible estimation may be obtained using variance-reducing techniques such as regularization or shrinkage (discussed next). Regularization in linear models was discussed in Section 2.2.4 and the estimation of the inverse of the covariance matrix with L^1-norm regularization is known as the graphical lasso (88). Alternatively, regularization may be applied to each target variable at once (returning the respective predictors linear coefficients / partial correlations). Another approach is the estimation of the nearest invertible correlation matrix with respect to a distance measure (116).

Full order partial correlations may also be approximated with lower order partial correlations. This strategy and shrinkage-based estimation are discussed next in more detail. It is also possible to infer GGMs by inferring the Markov blanket of each variable (under Gaussian assumptions). Markov blanket inference strategies are described in the Section 2.5.3.

2.5.1.1 Shrinkage estimation of the covariance matrix

If Σ^e is the empirical (possibly non-invertible) covariance matrix, the shrinkage estimation Σ^s, given a shrinkage target Σ^t is obtained by:

$$\Sigma^s = \lambda\Sigma^t + (1-\lambda)\lambda\Sigma^e \tag{2.36}$$

The shrinkage target matrix Σ^t may take various forms (see (229) for a review of commonly used shrinkage targets), from low to high complexity (ie. number of different elements). For instance, if Σ^t is diagonal, only the diagonal elements of the covariance matrix (the variances) are shrunk (towards the diagonal of Σ^t) while the non-diagonal elements are shrunk to zero.[1] Literature

[1] A shrinkage estimator moves (shrinks) an estimation towards a target point, estimated from the data. It has been shown to dominate, MSE-wise, maximum likelihood estimators in certain situations (73). Shrinkage is closely

approaches to estimate the parameter λ include empirical Bayes methods and cross-validation (229). The λ minimizing the MSE can also be derived analytically, a result due to Ledoit and Wolf (149, 229). This estimation has been shown to be biased and an unbiased estimate based on bootstrapping has been proposed in (141).

2.5.1.2 Lower order partial correlations

Lower order partial correlations can be used to approximate full order partial correlation graphs, assuming faithfulness (48). Let $\mathcal{G}^q$ denote a graph where the absence of an edge implies the absence of a non-null partial correlation up to order q between the respective nodes, and let $\mathcal{G}$ be the full partial correlations graph). When $q = 0$ edges correspond to a non-null Pearson correlation between the respective nodes. As q grows, under faithfulness[1], edges can only be removed (the ones whose nodes become conditionally dependent given any set of q variables). Thus, given $r < q < p$ (48):

$$\mathcal{G} \subseteq \mathcal{G}^q \subseteq \mathcal{G}^r \tag{2.37}$$

When estimating q-th order partial correlations graphs, only edges present in the $q-1$-th order graph have to be tested for conditional independence. One approach to estimate $\mathcal{G}$ is then to start with $\mathcal{G}^0$ and recursively remove edges as q increases (140). In the context of Bayesian networks this approach is known as the PC algorithm (see next section).

2.5.2 Bayesian networks

One application of Bayesian networks (when the graph is known) is the inference of conditional probability distributions of unobserved variables. Inference methods may be exact or approximate, both NP-hard (55, 60, 113). A related application is the inference of the most likely values of all unobserved variables (*most probable explanation*), or of only a subset (*maximum a posteriori*) (145).

Methods to infer the graph of a BN are commonly distinguished into approaches based on conditional independence (CI) tests; and based on a search over the network space (known as search and score methods) (145). Well-known algorithms based on CI tests are the PC algorithm, inferring the skeleton of the network (its non-directed edges), and the IC / SGS algorithms (similar and proposed independently), which return a directed network (206, 243). The skeleton of a BN may also be inferred by identifying the parents and children of each node. The skeleton

linked to empirical Bayes methods (221), and is a form of regularization (Section 2.2.4).

[1] On the assumption of faithfulness: a GGM (or a Markov network) is not faithful if there are colliders with independent parents in the underlying causal structure. When this is the case, there is an edge between the two parents in the full order GGM, while there is none in the linear correlation graph (of order 0).

can then be directed using IC / SGS. An approach to do so is by first identifying the Markov blanket of each node. The PC and the IC / SGS algorithms are described next. Algorithms for Markov blanket inference are described in Section 2.5.3. Section 2.5.2.2 refers some score and search strategies.

2.5.2.1 Inference based on CI-tests

Assuming faithfulness, two nodes connected by an edge are conditionally dependent, for any conditioning set of variables. If two nodes are found to be conditionally independent, there should be no edge connecting them. The PC algorithm (206, 243) (algorithm 2.1) considers conditioning sets of increasing cardinality (ie. number of elements) starting from 1. An edge between two nodes is discarded as soon as d-separation is found.

Algorithm 2.1: The PC algorithm.

```
input  : a multivariate probability distribution $\mathcal{P}$ of a set $X$ of $p$ variables
output : an undirected graph $\mathcal{G}$
initialize a fully connected (undirected) graph $\mathcal{G}$ ;
initialize $k = 1$ ;
while $k < p$ do
    for each $X_i \in X$ do
        for each $X_j \in X$ do
            (if there exists a set of conditioning variables which screens off the
            dependence between $X_i$ and $X_j$, of cardinality $k$)
            if $\exists X^z, X^z \subseteq X, |X^z| = k : I(X_i; X_j|X^z) = 0$ then
                remove the edge between $X_i$ and $X_j$ in $\mathcal{G}$
    $k = k + 1$ ;
```

The IC / SGS algorithm (206, 243) is a general scheme to infer the skeleton and infer edge direction and is described in the algorithm 2.2. The first step infers the skeleton and can be implemented with the PC algorithm.

Regarding the third step, rules have been identified as sufficient to identify all the common oriented edges in a class of equivalent networks, respecting the principle of no extra v-structures and no directed cycles (177, 206). They consist on four situations where an undirected edge between X and Y can be directed from X to Y, without creating new cycles and v-structures. These situations are illustrated in the Figure 2.8. Note that orienting the edge from Y to X necessarily creates v-structures or cycles. In the figure we see that:

Algorithm 2.2: The IC / SGS algorithm.

```
input : a multivariate probability distribution $\mathcal{P}$ of a set $X$ of $p$ variables
output : a partially directed graph $\mathcal{G}$
initialize a fully connected (undirected) graph $\mathcal{G}$ ;
(step1 - identification of conditional independences)
for each $X_i \in X$ do
    for each $X_j \in X$ do
        (if there exists a set of conditioning variables which screens off the dependence between $X_i$ and $X_j$)
        if $\exists X^z, X^z \subseteq X : I(X_i; X_j | X^z) = 0$ then
            remove the edge between $X_i$ and $X_j$ in $\mathcal{G}$
(step 2 - identification of collider structures)
for each triplet of variables $X_i$, $X_j$ and $X_z$, where $X_j$ and $X_i$ are non-adjacent, but are adjacent with $X_z$, in $\mathcal{G}$ do
    if $X_j \not\perp\!\!\!\perp X_i | X_z$ then
        orient the edges between $X_i$ and $X_z$, and between $X_j$ and $X_z$, to be directed towards $X_z$ ;
(step 3 - orientation of all other edges)
orient as many edges as possible in $\mathcal{G}$, on the condition that no directed cycles and new colliders are created ;
```

- In situation 1, orienting an edge from Y to X create a new v-structure (X is the collider).
- In situation 2, orienting an edge from Y to X creates a directed cycle.
- In situation 3, orienting an edge from Y to X implies that an orientation of the edges between X and Z_1 and between X and Z_2 creates either a cycle, or a new v-structure composed of X (the collider), Z_1 and Z_2.
- In situation 4, orienting an edge from Y to X implies that an orientation of the edge between X and Z_2 creates either a cycle, or a new v-structure composed of X (the collider), Z_2 and Y.

The difference between IC and SGS is that SGS specifies step 3, orienting edges based on the first two rules stated above.

It is possible to test whether a partially oriented graph can be fully extended into a fully oriented DAG. This tests consists in recursively removing any collider in open triplets without

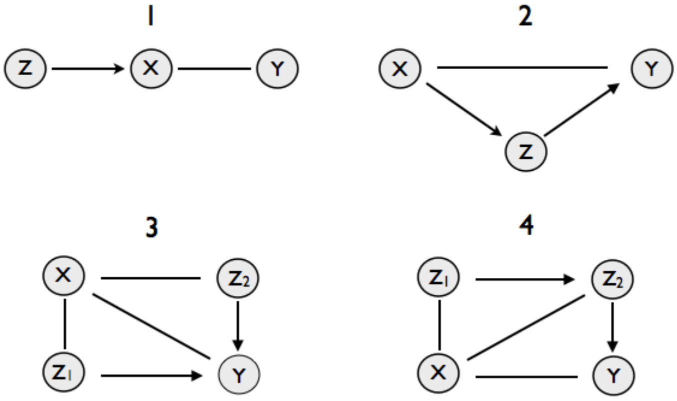


Figure 2.8: Rules for edge orientation in Bayesian networks - Situations where an edge can be directed from X to Y without creating cycles and new v-structures.

directed edges pointed from it, and where its adjacent nodes via undirected edges are adjacent to all nodes adjacent to the collider. If all colliders in open triplets can be removed, then the graph may be extended to a fully oriented DAG (70, 206). If this is the case, a way to direct undirected edges is to test hypothesis by checking if the extension remains valid.

2.5.2.2 Score and search approaches

Approaches based on CI tests may be computationally intensive due to the possibly large number of required conditional independence tests, particularly in large networks. Alternative methods infer BN by searching the network space, assessing each network with a scoring function. Search methods include hill climbing, simulated annealing (113), evolutionary algorithms (145) or particle swam optimization (269). Scoring functions take into account goodness of fit and network sparsity. Common scoring functions are the AIC, the BIC (see Section A.1), the Bayesian Dirichlet criterion, the minimum description length criterion (145), or the MIC (64). An empirical comparison between different scoring functions can be found in (47). These commonly assume that the data is discrete, requiring a discretization of continuous variables. Scoring functions for the continuous Gaussian case can be found for instance in (95, 114).

2.5.3 Markov blanket inference

The edges of a Markov network (undirected model, such as the Gaussian graphical model) connect each node to its Markov blanket (MB, Section 2.4.5). On Bayesian networks edges connect nodes to the respective set of parents and children, the MB except for spouses (parents of same children). This section describes some approaches to infer MBs and sets of parents and children (from these the skeleton of a BN is obtained, which may be directed using IC / SGS).

Inferring MBs One method to infer the MB of a variable Y is known as Grow-shrink (168) and involves two steps. In the first, variables are added to a MB candidate set if they are conditionally dependent with Y, given the previously selected variables. In the second step, variables are removed from the MB candidate set if they are conditionally independent with Y, given the remaining candidate set. A refinement of the selection part (known as IAMB (260)) selects at each step the variable that maximizes the conditional dependence with Y given the previously selected (eg. using the mutual information to quantify dependence).

Methods to infer the parents and children of Y work in a similar fashion. The difference is that while a variable is part of the MB of Y if it is conditionally dependent with Y given all other variables, it is a parent or child of Y if it is conditionally dependent with Y given any variable(s).

Inferring parents and children One method to infer the set of parents and children of Y is MMPC (261), which also proceeds in two steps. First, for each variable X, the subset of variables among the currently selected for which conditioning on minimizes the conditional dependence between X and Y is identified. MMPC then adds to a candidate set the variable X for which this (minimal) conditional dependence is maximum. If the dependence is null, the selection stops. In the second step, the algorithm removes from the candidate set any variable which is conditionally independent with Y, given any set of variables in the candidate set. Another method is the HITON algorithm (3). It adds variables to a candidate set of parents / children following an order based on their pairwise dependence with the target. After each variable is added, all previously selected variables which are conditionally independent with Y given any subset of the previously selected are discarded. The algorithm stops when all variables were considered.

From MB to parents and children The difference between the MB and the set of parents and children is that the MB also includes spouses. If the set of parents/children is known, for all variables, the spouses of Y may be identified as follows (3). Possible spouses of Y are be first identified as being the variables that are parents/children of each parent/children of Y, but are not parents/children of Y. Spouses of Y are then the ones which are conditionally dependent with Y given any of its parents/children. When the MB of a node Y is known, any variable in it that becomes conditionally independent with Y given any other variable in the MB must be a spouse of Y.

2.6 Variable selection with decision trees and random forests

Decision trees are machine learning models used in supervised classification and regression. Following the machine learning terminology, variables may be referred to as attributes, and samples as instances. A decision tree is composed by test nodes, branches and leaf nodes. An instance enters the tree at a root node, follows a path of branches, and ends up in a final leaf. Random forests are ensembles of decision trees, discussed in more detail below.

A test node represents a question on the nature of the instance (eg. if one of its attributes belongs to a given class or is higher than a threshold; or in which region of the attribute space it falls into). The outcome of test nodes are branches leading to other test nodes or leaves. Leaves are nodes without outgoing branches, representing the class of the instance to classify (or a numeric value in the case of regression trees).

Attributes appearing on top nodes should be more informative of the target/class variable. This observation can be used to perform variable selection, for instance by growing several trees and identify which attributes (variables) appear on top nodes (83, 298)).

Tree nodes are characterized in terms of impurity, measuring the class variance in the instances in the node (in classification). If all instances in a node are of the same class, the node is pure. Impurity is maximum when the classes have an equal frequency. Node impurity is usually measured with the entropy or the Gini index (248). In the case of regression trees, impurity may be quantified with the local residual sum of squares (considering only the node instances).

2.6.1 Learning decision trees

Several algorithms have been proposed to learn decision trees from data. A major group (ID3, C.45, CART) (34, 87, 217) [1] is characterized by a divide-and-conquer strategy. Instances enter the tree at an initial node and then each follows one of multiple downward branches, according to the values of a most informative attribute, according to some criterion. A branch follows into a subsequent node, where its instances are split again, according to the most informative attribute not previously used in node splitting of parent nodes. A leaf is an end node and is created when all the instances in its parent branch belong to the same class (the leaf is assigned the value of that class), or if there are no more attributes not used previously in splitting (in this case the leaf is assigned the majority class of the subset instances). It may happen that there are

[1] C.45 is an extension of ID3 to deal with continuous values, missing data, and which also implements a pruning step to mitigate over-fitting. In the case of CART, any test node only outputs two branches (this is not a restriction in ID3/C.45, for non-continuous data). Differences between CART and C.45 are relative to tree pruning, splitting criterion and handling of missing values.

no instances in a current branch, and in this case its subsequent node is a leaf whose class is the majority class of the instances in the parent node.

2.6.2 Boosting, bagging, random forests

Decision trees are commonly used as the base model in model ensembles. Common ensemble strategies (boosting, random subspace methods, bagging) are described next. Boosting is the process of iteratively combining weak models into a stronger one. It consists on assessing the classification error on the training set and then increasing the weight of the misclassified examples on the training set distribution. The classification is then repeated and so on. One popular example is the AdaBoost algorithm, using decision trees as the individual models (87).

Random subspace methods combine models obtained after randomly sampling the variable space (119, 136). Whereas boosting is an iterative procedure in which the final model is incrementally improved, in this approach weak models are learned in parallel and then combined. This approach has been shown to return robust models, as variant and susceptible to over-fitting as the weaker individual models, but more precise (less biased) than these (136). Bagging (bootstrap aggregating) (35) is the process of sampling with replacement the training dataset, training a model in each case and then combining the obtained individual models.

These methods combine model variations in the variable space (boosting, random subspace methods) and in the sample space (bagging). Random forests combine variations in both levels. They are defined as an ensemble of individual decision trees, each depending on a random vector sampled independently from the same probability distribution(36). A common implementation is a bagging-combination of decision trees, in which the attributes considered at node splitting are randomly selected (36).

Variable importance Random forests allow for variable selection through the consideration of the importance of each variable (attribute). One approach to estimate variable importance is through the reduction in the accuracy in out-of-bag samples (samples not used in training) when the values of the variable are randomly permuted (36).

Another approach to measure the importance of a variable is to estimate the average improvement of the tree after node splitting when that variable is considered in the split. This may be quantified by estimating the difference between the impurity of the node that is split, and the average (weighted on the number of instances) impurity of its children nodes (a difference which is necessarily positive (248)). Random forests have been applied in the context of GRN inference (124), see Section 3.2.5.

2.7 Analysis of time series

A time series is an ordered sequence of variable observations along time. Time series analysis is useful in a broad range of areas: in econometrics (in the temporal behavior of market values, unemployment or inflation), physical and environmental sciences (eg. evolution of global warming) or medicine (effect of treatments over time). Time series analysis may have a forecasting purpose (to predict variable realizations in future time points), or it may concern the inference of statistics of temporal variables.

2.7.1 Considerations using dynamic models

This section reviews approaches to model (lagged) statistical dependences between stochastic processes represented in multivariate time series. Due to the temporal direction of causality, these dependences may be used to infer causal relationships. We note two points of care in the modeling of causality using lagged dependences.

1. Causality may be missed if the sampling rate is too low. If the sampling period is higher than the time it takes for a change in X_i to influence X_y, the change in both X_i and X_y is observed at the same moment and the causal direction cannot be inferred from data alone.

2. Non-stationarity poses problems to standard estimation strategies, as the probability distribution of variables changes over time. Strategies to deal with non-stationarity should then be adopted.

The remainder of the section is organized as follows. Section 2.7.2 introduces basic notions of time series, including auto-correlation, stationarity and vector-autoregressive processes; Section 2.7.3 is on the estimation of auto-regressive models; 2.7.4 describes Granger causality; 2.7.5 describes strategies to deal with non-stationarity; 2.7.6 introduces dynamic Bayesian networks; 2.7.7 introduces analysis in the frequency domain.

2.7.2 Properties of time series

Time series are sequential observations of stochastic (ie. subject to uncertainty) processes called *data generation processes*. The values of the process at individual sampled points are modeled as distinct random variables, characterized by the same functional dependences (159). Some properties of stochastic processes are discussed in this section.

On notation The term time series usually denotes univariate time series, while the term multivariate time series refers to a set of univariate time series, corresponding to different variables. Regarding notation, the symbol X_t may denote both the variable(s) (data generation process) X at time t in the (multivariate) time series, or the process itself (which is also referred to as a variable). In the case of multiple variables, $X_{i,t}$ refers to the i-th variable in the set X_t. A variable X_{t-l} (at time $t-l$) is termed a lagged variable of X_t, of lag l. Lags may also be referred to using τ.

2.7.2.1 Auto and cross-covariance

The auto-covariance/auto-correlation of stochastic processes are the covariance/correlation between variables at different time points (159).

Definition 14. *The auto-covariance of a stochastic process X_t at two time points t_1 and t_2, with respective means μ_{t_1} and μ_{t_2}, is:*

$$\sigma_{t_1,t_2} = E[(X_{t_2} - \mu_{t_1})(X_{t_2} - \mu_{t_2})] \tag{2.38}$$

Note: in the context of a single variable X_t its expected value $E(X_t)$ is simply denoted by μ_t. The auto-correlation of X_t at t_1 and t_2 is then σ_{t_1,t_2} divided by $\sigma_{t_1}\sigma_{t_2}$. The cross-covariance is the covariance between two processes at different time points.

Definition 15. *The cross-covariance between two stochastic processes X_t and Y_t at time points t_1 and t_2 is:*

$$\sigma_{X_{t_1},Y_{t_2}} = E[(X_{t_1} - \mu_{X_{t_1}})(Y_{t_2} - \mu_{Y_{t_2}})] \tag{2.39}$$

The cross-correlation is defined analogously to the auto-correlation.

2.7.2.2 Stationarity

A stochastic process is stationary if its probability characteristics do not change over time. It is usually defined in a weak and strong (or strict) sense (159).

Definition 16. *(Weak sense stationarity) A stochastic process is stationary in the weak sense if and only if:*

$$E[X_t] = E[X_{t+\tau}], \quad \forall \tau \in \mathbb{Z} \tag{2.40}$$

and

$$\sigma_{t_1,t_1+\tau} = \sigma_{t_2,t_2+\tau}, \quad \forall t_1, t_2, \tau \in \mathbb{Z} \tag{2.41}$$

Weak-sense stationarity means that the expected value of the time series is the same for for all time points, and that its auto-covariance only depends on the temporal lag (τ) between the two considered points. It follows that the variance of a weak-sense stationary time series is constant for all values of t (corresponding to the auto-covariance when $\tau = 0$).

Definition 17. *(Strong sense stationarity) A stochastic process is stationary in the strong sense if and only if:*

$$\mathcal{P}(x_t) = \mathcal{P}(x_{t+\tau}), \quad \forall t, \tau \in \mathbb{Z} \tag{2.42}$$

Strong-sense stationarity is a stronger condition requiring that the probability distribution of X_t is the same for all values of t. A related definition is the one trend-stationarity. A process is trend-stationary if it contains a non-stationary component which is a function of t and whose removal renders the process stationary.

2.7.2.3 Vector auto-regressive models

Vector autoregressive (VAR) models are linear models in which set of variables X_t (at time t) are modeled as a linear function of themselves at previous time points, additive white noise and a constant (intercept) term. In the univariate case, the term auto-regressive (AR) model is used. VAR models are commonly used as an approximation of data generating processes, in various fields (159). VAR models are characterized by an order L indicating the number of considered lags. A VAR(L) is then:

$$X_t = c + \sum_{l=1}^{L} B_l X_{t-l} + \epsilon_i \tag{2.43}$$

In the general multivariate case, B_l is a $p \times p$ matrix of coefficients relating the effect of each $X_{i,t-l}$ into each $X_{i,t}$. The *reverse characteristic polynomial* of a VAR model is the determinant $\det(\mathbf{I} - \sum_{i=1}^{L} B_l z^i)$. A property of VAR models is the one of *stability*. It is defined as follows (159).

Definition 18. *A VAR process is stable if the determinant of its reverse characteristic polynomial does not have roots inside or on the unitary circle.*

Stability implies stationarity, however the converse is not true (there are cases of unstable stationary VAR processes, see below) (159).

2.7.2.4 Integration and cointegration

An univariate time series is said to be *integrated* of order n if it is non stationary and differentiation (ie. subtraction of the value at the previous time point) n times makes it stationary (79). In the AR process case, the time series is integrated of order n if its reverse characteristic polynomial has n *unit roots* (equal to 1). If an AR process has roots strictly inside the unitary circle, it exhibits an explosive behavior (the variance of the process increases with t at an exponential rate). Processes with one unit root are known as random walks, and their variance increases linearly with t. If additionally, the intercept term in equation (2.43) is different than zero, the expected value of the process also increases linearly with t. This process is known as a random walk with drift and has a deterministic trend (as opposed to a stochastic trend). Stationary and integrated processes are the most well studied cases in the time series literature (159).

In case of a VAR with n unit roots, its individual processes may be integrated of order $< n$ or even stationary (due to canceling terms). Also, as opposed to the univariate case, a constant term in equation (2.43) does not imply deterministic trends in the individual processes (159).

Differentiating n times the VAR necessarily makes its individual processes stationary. However, differentiating may result in a loss of information regarding relationships between variables (eg. a common trend). This limitation is addressed with the concept of *cointegration*. multivariate time series integrated of order n are cointegrated if there is a linear combination of them resulting in a stationary time series, or integrated of order $< n$. A slightly different definition is: multivariate time series whose maximum order of integration is n are cointegrated if there is a linear combination of them (with non-zero coefficients) resulting in a stationary time series, of integrated of order $< n$. A known framework in econometrics to deal with integrated variables is the vector error correction model (outside the scope of this work) (159).

2.7.3 VAR estimation

A direct approach to estimate the coefficients of a VAR model as in equation (2.43) is by OLS (Section 2.2.1). In the case of a VAR1 model, the OLS estimation of the B matrix is:

$$\hat{B}_{OLS} = (\mathbf{X_{t-1}^T X_{t-1}})^{-1} \mathbf{X_{t-1}^T X_t} \tag{2.44}$$

The solution can be modified to represent higher order VAR models. The solution for a VAR2 is obtained by replacing $\mathbf{X_{t-1}}$ with $\{\mathbf{X_{t-1}}, \mathbf{X_{t-2}}\}$. If the respective covariance matrix is invertible, a solution can be obtained with regularization (Section 2.2.4).

The OLS estimation is consistent and has the usual asymptotic properties if the VAR process is stationary and normally distributed (assuming white noise) (159). In the case of non-stationarity, differentiating until stationarity or using vector error correction models is then usually required. One exception is the the particular case of two cointegrated processes of order one, where the OLS estimation is consistent (79). Residual auto-correlation also appears if the lag length L in the estimated model is not as high as the order of the underlying process. An estimation of L may be obtained with standard model quality measures, such as the AIC (appendix A.1).

2.7.3.1 Tests of stationarity and existence of unit roots

A test whose null hypothesis is time series stationarity is the KPSS test (144). Tests whose null hypothesis is the existence of unit roots are the Dickey-Fuller tests (67), or the Phillips-Perron test (213). Testing the number of unit roots (order of integration) may be carried out by recursive stationarity/unit root testing and differentiation. One approach to estimate cointegration between two integrated time series of order one is the Engle-Granger method, based on the OLS consistency in this particular case (79)). It consists on estimating using OLS the residuals of a linear regression of one time series using the other. If the residuals are stationary, the two time series are cointegrated. For other tests see (37).

Figure 2.9 illustrates a simulated gene expression time series (generated with GNW) estimated to be non-stationary with the KPSS test (p-value cut-off of 0.05). Figure 2.10 illustrates two time series estimated to be cointegrated with the Engle-Granger method (p-value cut-off of 0.05).

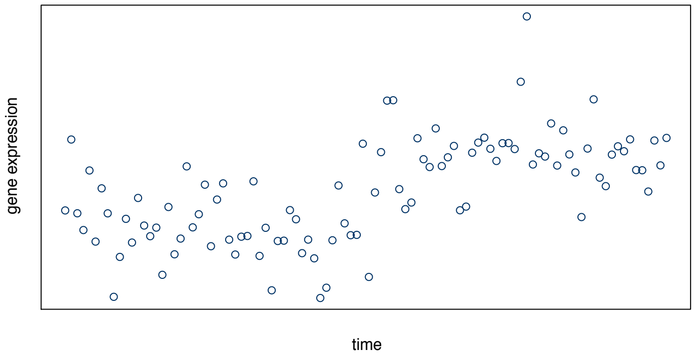


Figure 2.9: Example of a non-stationarity time series. - Gene expression time series of 100 points, generated by GNW, estimated to be non-stationary with the KPSS test (p-value cut-off of 0.05).

2.7.3.2 Reduced and recursive VAR models

The use of VAR models has been popularized in econometrics in the last decades, replacing a paradigm based on SEMs, with restrictions on causal links and distinctions between endogenous and exogenous variables (240, 245). On the contrary, VAR models do not use impose causal restrictions (each variable is modeled as a function of all other variables). In a retrospective review on the use of VARs in the context of econometrics, Stock and Watson point out their success in data description and forecasting (but not so much in causal inference and policy analysis) (245).

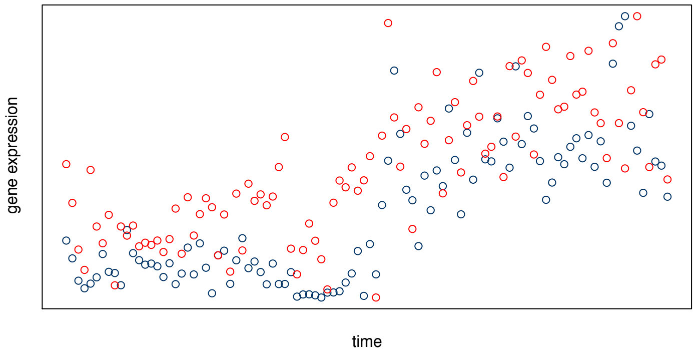


Figure 2.10: Example of two cointegrated time series. - Two gene expression time series of 100 points, generated by GNW, estimated to be cointegrated with the Engle-Granger method (p-value cut-off of 0.05).

VARs can be represented in two forms. The *reduced* form is the one presented in the equation (2.43), where each variable is modeled as a function of its past and the past of other variables only. If there is causality occurring faster than the sampling rate (as mentioned in Section 2.7.1), the error terms of different time series will be correlated. In this case *recursive* VAR models may be used (245). Here, variables are modeled as a function of the present of other variables also (and not only the past), such that the error terms of the individual time series are uncorrelated. Recursive VARs take the form of:

$$B_0^* X_t = c + \sum_{l=1}^{L} B_l^* X_{t-l} + u_i \tag{2.45}$$

B_l are $p \times p$ coefficient matrices, and the new matrix B_0^* represents the dependence between variables at time t. Recursive VARs model instantaneous causality and may be estimated from reduced form VARs. However, recursive VARs are not uniquely identified from reduced-form VARs, being dependent on the variable ordering (135, 245). This ordering should reflect an informed expert judgment - when this is the case, the recursive model is known as *structural* (245).

2.7.4 Granger causality

Granger causality (GC) is a test for causality on time series popularized by Granger in econometrics (103), and is often used as a synonym for the definition of causality of Suppes (159, 249). There is GC from a X (the cause) to Y (the effect) if there is a conditional dependence between the present of the Y and the past of X, conditioned on the past of other possible causes, including the effect itself.

An alternative definition of Granger causality is that there is GC from X to Y if future values of Y are better predicted using the present and past values of X, than otherwise (159). This definition concerns the predictive power of X_{t-l_1} on Y_{t+l_2}, where $l_1 \geq 0$ and $l_2 > 0$ (known as multi-step forecasting). When forecasting multiple future time points of a target (as opposed to just the first), and in the case of three or more variables, non (direct) causes may be useful (ie. if they are relevant in the prediction of the direct causes of Y_{t+l_2}, at intermediate time points $l_1 < l_2$) (159, 161). As the topic of forecasting is out of the scope of this thesis, we will restrict to the first characterization of GC, equivalent to the other in the one-step into the future case.

In the linear model, GC is inferred by testing the hypothesis that the linear coefficients of X (corresponding to different lags), in a VAR model of Y, are equal to zero. These coefficients may be tested individually or at once, addressing the problem of multiple testing. In the last case, two models are compared, one where the coefficients of X are included, the other where they are *restricted* to zero. If Y is better predicted in the first case, there is GC from the predictor to it. The bivariate case (no other variables are conditioned on) takes the following form (X_y is the target and X_i the predictor):

$$X_{y,t} = \alpha_0 + \left(\sum_{l=1}^{L} \alpha_l X_{y,t-l} \right) + \left(\sum_{l=1}^{L} \beta_l X_{i,t-l} \right) + \epsilon_{ut} \tag{2.46}$$

If $X_{y,t}$ does not depend on $X_{i,t-l}$ (Granger non-causality from X_i to X_y), $\beta_l = 0$ (these coefficients are restricted to 0), and equation (2.46) reduces to:

$$X_{y,t} = \alpha_0 + \left(\sum_{l=1}^{L} \alpha_l X_{y-l} \right) + \epsilon_{rt} \tag{2.47}$$

Under the assumption that the residuals are uncorrelated, homoscedastic and normally distributed, the F-test comparing the two models is as follows. Let RSS_U and RSS_R be the residual sum of squares of equations (2.46) and (2.47), and N the number of points in the time series. Define the statistic G_{iy} as (L being the number of lags):

$$G_{iy} = \frac{(RSS_R - RSS_U)/L}{RSS_U/(N - 2L - 1)} \tag{2.48}$$

Under the null hypothesis (Granger non-causality), G_{iy} follows a F distribution with degrees of freedom L and $N - 2L - 1$ (assuming a constant in the regression). The number of lags L in equations (2.46) and (2.47) can be estimated using a measure of model quality such as the AIC. We have assumed that L is the same for both variables, but that may not be the case. If they are different, equation 2.48 should be modified accordingly. Usually, the number of lags of each variable is estimated in its respective restricted model (equation 2.47 for $X_{y,t}$). The described method tests Granger non-causality (it is the null), however for simplicity we refer to it simply as a GC test.

In the one lag case, testing linear GC is equivalent to testing a lagged partial correlation between $X_{i,t-1}$ and $X_{y,t}$, conditioned on $X_{y,t-1}$. GC tests can be modified to incorporate conditioning variables, added to both the unrestricted and restricted models. The F-statistic of equation (2.48) should be modified - the denominator becomes $N - pL - 1$, where p is the number of variables in the unrestricted model. If $p > n$ the linear model is ill-posed due to over-fitting and the GC test is unfeasible.

GC has also been described in the frequency domain (98) and extended to the non-linear case with mutual information. This information-theoretic form is known as transfer entropy (231). In the bivariate case, the transfer entropy from X_i to X_y is

$$T_{X_i \to X_y} = I(X_{y,t}; X_{i,t-l}|X_{y,t-l}), l \in S_l \tag{2.49}$$

(S_l represents the set of considered lags.) The transfer entropy is the mutual information between X_i and X_y, conditioned on the past of X_i.

2.7.5 Granger causality in non-stationary time series

In the linear case, OLS-based inference requires residual properties which are not obtained in non-stationary time series (105). One way to overcome the non-stationarity limitation is to differentiate the time series until they are stationary and then apply OLS on the new stationary time series. Another strategy was proposed by Toda and Yamamoto (TY) (257) in the context of GC tests and is described next.

In the case of two cointegrated time series, one result is that there is necessarily GC between them (in one, or both directions) (104). In this case, standard GC tests maintain their asymptotic properties (160).

TY modification The TY-modified GC test consists in finding the maximum order of integration of all time series, adding an extra number of lags equal to this value in the restricted and unrestricted models (for all regressors), and then testing for GC. This approach has the advantage of avoiding pre-testing for cointegration. In the bivariate case described in the previous section, if the maximum order of integration of the multivariate time series is d, the TY approach compares the following restricted and unrestricted models:

$$X_{y,t} = \alpha_0 + \left(\sum_{l=1}^{L+d} \alpha_l X_{y,t-l} \right) + \left(\sum_{l=1}^{L+d} \beta_l X_{i,t-l} \right) + \epsilon_{rt} \tag{2.50}$$

$$X_{y,t} = \alpha_0 + \left(\sum_{l=1}^{L+d} \alpha_l X_{y,t} \right) + \left(\sum_{l=1}^{d} \beta_l X_{i,t-l} \right) + \epsilon_{ut} \tag{2.51}$$

The number of denominator degrees of freedom of the F-test in equation (2.48) changes to $N - 2(L + d) - 1$), and it becomes:

$$G_{iy} = \frac{(RSS_R - RSS_U)/L}{RSS_U/(N - 2(L + d) - 1)} \tag{2.52}$$

This approach has the advantage that the resulting statistic is well-behaved asymptotically, even in the case of integrated variables. However, the extra regressors added in the restricted and unrestricted models may result in a loss of power of the test, particularly in the small sample case (69).

2.7.6 Dynamic Bayesian networks

Dynamic Bayesian networks (DBN) are extensions of Bayesian networks in which each variable is represented in multiple nodes corresponding to consecutive time points (or slices). Edges connect nodes at different time points, in a temporal direction. DBN graphs are naturally acyclic, as edges are always directed from a node to another at a subsequent time point. It is possible to model feedback loops, as each node is represented at multiple time points (eg. $X_{1,t} \to X_{2,t+1} \to X_{1,t+2}$). Usually the number of considered time slices is only two. In this case, and under linear assumptions, DBN are similar to VAR1 models. See (185) for an extensive description of DBNs.

Some authors have extended graphical causal models (ie. Bayesian networks) to deal with time series (in this case each variable is represented in a single node, as opposed to in DBN). A ”global Granger causal Markov property” is described in (75), analogous to the Markov property (Section 2.4.1), with the difference that GC takes the role of statistical dependence. White and colleagues (273, 274) established an equivalence between GC and causality in the framework of Bayesian networks, under a condition of exogeneity of unobserved variables (ie. these are not determined by observed variables).

2.7.7 Differential equations

Ordinary differential equations model an n-th degree derivative as a function of lower degree derivatives. One example, of the first derivative of a variable X_y ($\dot{X}_y$) modeled as a function of itself, other variables and external perturbations $\theta_{y,t}$ is the following:

$$\dot{X}_{y,t} = f_i(X_{1,t}, .., X_{p,t}, \theta_{y,t}) \tag{2.53}$$

If f_i is a linear equation, $\dot{X}_y$ becomes:

$$\dot{X}_{y,t} = \sum_{i=1}^{p} \beta_{iy} X_{i,t} + \theta_{y,t} + \epsilon_{y,t} \tag{2.54}$$

In practice, when dealing with discrete time series, and in the absence of external perturbations, $\dot{X}_{y,t}$ may be approximated as $X_{y,t+1} - X_{y,t}$ (as in (11, 151)).

2.7.8 Analysis on the frequency domain

Time series can be analyzed on the frequency domain, using the Fourier transform (FT). A possible application of the FT in gene expression time series data is the identification of periodic biological processes, and noise removal. Examples can be found in (251), (239) or in (222). In the frequency domain, the ordering of cause-effect pairs may be identified as follows. If one time series causes the other, it precedes it in time, and in the frequency spectrum there is a phase difference between the time series (for each frequency). This phase difference becomes higher (in absolute value) as the frequency increases. It is then expected a positive (or negative) slope in the phase difference spectrum. The signal of the slope gives information about which time series precedes the other. This method is known as the phase-slope index (191).

2.8 Conclusion

This chapter presented the preliminaries for causal inference from variable observations. As discussed in the introduction (Section 1.9), causality is closely linked to statistical dependence. Statistical dependence may be quantified with the mutual information, a concept from information theory and introduced in Section 2.1. When variables are elliptically (eg. normal) distributed, linear dependence is equivalent to dependence. Linear models are extensively used and were introduced in Section 2.2. Conditional dependence between two variables implies causality between them when all their common causes are conditioned on, but two independent variables are conditionally dependent given a common effect. These insights are the basis of causal models, commonly represented in a graphical form - where nodes represent variables, and edges represent causal mechanisms. Due to the need of a clear distinction between the causes and effects of a variable, these models are usually acyclic (or recursive). Only this case was considered in this chapter. Linear causal models (known as structural equation models) were introduced in Section 2.3, and general graphical causal models (Bayesian networks) in Section 2.4. Undirected graphical models are often used to represent conditional dependences between variables, one example being the graphical Gaussian model, in which edges represent partial correlations between nodes (variables). Some strategies to infer causal models were presented in Section 2.5. Networks may also be inferred with the selection of predictors for each target variable. Variable selection may be implemented using a variety of statistical and machine learning methods, such as decision trees and random forests. These were introduced in Section 2.6. The final Section 2.7 addresses the topic of analysis and causal inference from time series. The use of time series allows to associate variable observations at different time points and to distinguish between possible causes and effects - as the effect of a cause is observed at a later time point than the cause itself. This is the basis of Granger causality tests and vector autoregressive models. The described concepts compose a foundation for network inference. Some literature applications of network inference in the context of gene regulatory networks are described in the next Chapter 3.

3

State of the art for network inference

This chapter gives a flavor of GRN inference approaches that can be found in the literature, and specific methods will be described in detail (some with available software implementations). As discussed in Section 1.7), we will deal exclusively with inference from observational gene expression data only and will not consider strategies incorporating prior information or experimental perturbations (see (11, 195)).

We distinguish between inference methods designed to non-temporal, static observations; and dynamic methods explicitly designed to time series. The former may be directed (Bayesian networks) or undirected, while the latter are directed. Due to the high variable to sample ratio in GRN inference, inference based on high order conditional independence tests is not feasible and alternative strategies are adopted. Reviews on GRN inference can be found in (10, 93, 101, 131, 170, 211, 265).

The number of GRN inference approaches in the literature is large and we do not aim to provide an extensive survey. In the particular case of dynamic inference, strategies that are left out of this chapter include Boolean networks, S-systems and state space models. Boolean networks model gene expression (of two states only) at a time point as a boolean function of gene expression at the previous point. They were among the first GRN inference approaches to be proposed in the literature (134, 152). S-systems have also been used to model gene expression dynamics (268). They are representations of non-linear functions using the Taylor series approximation (227). Under certain assumptions, the estimation of the parameters of the S-system is simplified and reduces to a linear problem (268). Finally, state space models have been used to map gene expression into a lower orthogonal space vector (117). The hidden state evolves over time, controlling the gene expression at the original space.

Chapter outline Section 3.1 describes experimental findings in the literature regarding network inference. Section 3.2 describes static GRN inference approaches, including based on filter variable selection (Section 3.2.1), information theory (Section 3.2.2), linear models (Section 3.2.3), Bayesian networks (Section 3.2.4), and random forests (Section3.2.5). Section 3.3 describes dynamic approaches, based on the estimation of optimal lags (Section 3.3.1); VAR models and dynamic Bayesian networks (Section 3.3.2); and differential equations (Section 3.3.3).

3.1 Network inference assessment in the literature

The DREAM challenges are a successful endeavor to assess and compare methods to solve problems in biology, including GRN inference. In this section we describe some relevant findings of these challenges. The DREAM5 challenge (166) consisted in the inference of three networks, two of around 5000 genes (of E.coli and S.cerevisai) and one of around 1600 genes of simulated gene expression data. In all cases a gold standard of regulations is available. Networks were inferred from between 500 and 900 gene expression samples of a variety of experiments, including perturbatory and time series. Figure 3.1 illustrates the results (from (166)). Some assessed methods are described in this section. The regression methods are composed of combinations of the lasso and data resampling techniques (Tigress is method 1, Section 3.2.3). Regarding the MI methods: method 1 is CLR, then MI and ARACNE (Section 3.2.2). Methods 4 and 5 are based on Markov blanket inference (Section 2.5.3). The correlation methods are the Pearson and Spearman and the first two Bayesian networks methods are search and score approaches (simulated annealing), the next two are based on Markov blanket inference. GENIE3 (Section 3.2.5) is the first of ”other” methods. Meta predictors combine multiple methods (described below). Although high precision was obtained in the in-silico network, the precision is much lower in the biological networks. In the S.cerevisai experiment the precision of all methods was close to random. The inference performance was measured with the AUPRC, which was assigned a p-value obtained with Monte Carlo.

Ensemble of predictions In the DREAM5 challenge, an ensemble of methods achieved the highest precision. It is based on a (non-weighted) combination of the methods’ ranking of regulations, using the Borda count method (each regulation is assigned a rank by each individual method, and these ranks are then averaged (264)). The higher performance of Borda count-based combinations (compared with individual methods) has also been observed in previous DREAM challenges (167, 246). It was also shown that weighting networks according to its individual precision results in a even higher precision.

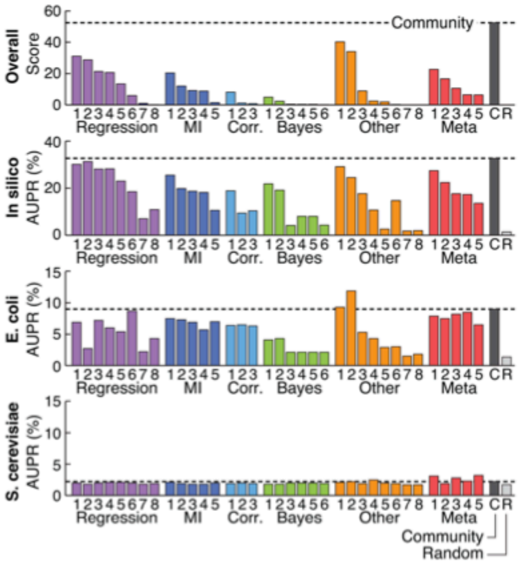


Figure 3.1: DREAM5 inference results - AUPRC results for three different network inference tasks (S. cerevisiae, E.coli and in silico) and combined score, for multiple methods. Reprinted by permission from Macmillan Publishers Ltd: Nature Methods (166), copyright 2012.

Simpler may be better It has been noted in previous DREAM challenges that simple GRN inference approaches often outperform more sophisticated ones (167, 265). The DREAM3 challenge consisted in the inference of networks of 10, 50 and 100 genes from data reflecting interventions (gene knock downs) and time series of 21 time points (216). The authors note that the highest inference precision could be obtained with a simple z-score using perturbation experiments only[1]. In this challenge, several assessed methods did not perform significantly better than random (11 out of 29). In previous challenges, a simple bivariate correlation was also shown to outperform more complex approaches (12)).

Gene expression time series Presently, the time series of gene expression that can be found in public repositories are composed of a few dozen of different time points at most. The high variable to sample ratio limitation is then more accentuated in inference strictly from time series. For this reason, GRN inference from time series is usually validated in medium sized and small networks.

One of the most informative gene expression time series currently available is of the human

[1] The expression of Y after X is perturbed is normalized into a z-score, which is used as a measure of the likelihood of a regulation from X to Y.

cancer cell cycle (Hela) (275). It consists on 3 multivariate time series of 12, 27 and 48 time points, and 1134 genes were identified as being periodically expressed along the cycle. Network inference on these genes and time series has been presented in (151, 158). In the literature, other used time series are of a shorter size (around 20 points and lower) (11, 223, 294, 295).

Due to lack of a complete gold standard, inference assessment is often done on smaller networks (an handful of genes to few dozens) (11, 223, 226, 294, 295). Due to the time series short size, multiple datasets may be combined: 95 multivariate time series of yeast, of 6 time points, were used with prior knowledge to infer a network of around 3500 genes (284); 44 multivariate time series of 10 time points were used to infer a network of 58 genes (196). A similar combination is assessed in 5, using yeast time series. Often, GRN inference validation is done on simulated data.

3.2 Static approaches to network inference

In this section we describe some state of the art network inference strategies for static data, based on filter variable selection, mutual information, Bayesian networks and random forests.

3.2.1 Filter forward selection variable selection

As discussed in Section 2.5.3 undirected graphical models (Markov networks) may be obtained by inferring the MB of each variable in the network. The described approaches in that section work by estimating high order conditional dependences, which may be unfeasible in the high variable to sample ratio case. In this case, an alternative is to resort to approximations based on lower order approximations. Some approaches are described next, working in a forward selection manner (as the methods of Section 2.5.3). These variable selection methods are considered to be filters (Section 1.8). From them, network scores may be obtained.

In what follows, dependences are formulated in terms of relevance towards a target and redundancy between predictors. We use MI as the dependence measure. The target variable is denoted by X_y and the set of predictor variables by $X^{\setminus y}$. Additionally, consider X_k^s to be the k-th selected predictor variable, and S^k the set of previously selected variables (before the k-th selection). $|S^k|$ denotes the number of elements in S^k.

FCBF FCBF (Fast Correlation Based Filter) (288) selects, at each selection step k, the predictor with the highest relevance towards X_y, among the predictors whose relevance to X_y exceeds the redundancy with any previously selected predictor in S^k.

mRMR mRMR (minimum-Redundancy-Maximum-Relevance) (212) selects, at each step, the predictor maximizing the difference between the relevance towards the X_y and the average redundancy towards each of the previously selected predictors. The k-th selected predictor is:

$$X_k^s = \underset{X_i \in X^{\setminus y} \setminus S^k}{\arg\max} \; (u_i - r_i) \tag{3.1}$$

where $u_i = I(X_i; Y)$ and

$$r_i = \frac{1}{|S^k|} \sum_{X_j \in S^k} I(X_i; X_j) \tag{3.2}$$

The term u_i represents the relevance of X_i towards X_y and the term r_i represents the redundancy of X_i with the previously selected predictors S^k. mRMR selects variables by taking into account redundancy and relevance, and by only considering pairwise dependences avoids the estimation of conditional dependences. If $k = 1$ the term r_i is set to zero.

MRNET MRNET (179) is an application of mRMR to network inference. To any pair of genes X_i and X_j a score is assigned which is equal to the maximum between two mRMR scores $u - r$: the mRMR score of X_i when X_j is used as the target gene, and the mRMR score of X_j when X_i is used as the target gene.

CMIM CMIM Conditional Mutual Information Maximization) (83) selects the predictor with the highest lowest conditional dependence with X_y, for all previously selected predictors used as the single conditioning variable. The k-th selected predictor is:

$$X_k^s = \underset{X_i \in X^{\setminus y} \setminus S^k}{\arg\max} \; (\min_{X_j \in S^k} (I(X_i; X_y | X_j))) \tag{3.3}$$

CMIM is based on conditional dependences of a single conditioning variable (of order 1), thus avoiding the over-fitting issue of high order conditional dependence estimation, when n is low.

mIMR mIMR (min-Interaction Max-Relevancy) (29) considers the interaction information (Section 2.1.2) to add a causal aspect to variable selection, as the interaction information of an open triplet with a collider is negative.[1]

At each step, mIMR selects the predictor maximizing the difference between its relevance towards X_y and the average of interaction terms including the predictor, X_y, and each of the previously selected predictors. The k-th selected predictor is:

[1] Due to inconsistent notation in the literature, the interaction information is sometimes defined as $I(X; Z|Z) - I(X; Z)$ or $I(X; Z) - I(X; Z|Z)$. In mIMR the last definition is used.

$$\begin{aligned}
X_k^s &= \operatorname*{arg\,max}_{X_i \in X^+ \setminus S^k} \left(I(X_i; X_y) - \frac{1}{|S^k|} \sum_{X_j \in S^k} I(X_i; X_j; X_y) \right) = \\
&= \operatorname*{arg\,max}_{X_i \in X^+ \setminus S^k} \left(I(X_i; X_y) - \frac{1}{|S^k|} \sum_{X_j \in S^k} (I(X_i; X_y) - I(X_i; X_y | X_j)) \right) = \\
&= \operatorname*{arg\,max}_{X_i \in X^+ \setminus S^k} (\sum_{X_j \in S^k} I(X_i; X_y | X_j)) \qquad (3.4)
\end{aligned}$$

X^+ is a subset of $X^{\setminus y}$, containing the variables that have a positive mutual information (non-zero) with X_y.

In the open tripet case, the interaction information is negative if there is a collider. Then, either X_y is the collider, or one of X_i and X_j is the collider. The situation of interest is when X_y is the collider (as X_i and X_j are then causes of X_y). If X_i is the collider, then X_j is independent with X_y and does not enter the initial set X^+. This way, mIMR tends to select predictors which are a cause of X_y.

3.2.2 Inference based on information theory

The mutual information is a measure of non-linear dependence and its use has been popularized in GRN inference. Under a Gaussian assumption the mutual information is a monotonous function of the linear correlation (Section 2.1), and the linear and the Spearman correlations have been shown to outperform other estimations of MI in GRN inference (193).

Relevance networks A simple approach to infer GRN is to score each possible regulation with the MI between the two respective genes. A network can be inferred by selecting the highest score regulations. This simple approach does not take into account multivariate dependences, and is liable to score highly pairs of genes which are only indirectly dependent. Other methods aim to improve this aspect through the consideration of multivariate aspects, and are described next.

ARACNE ARACNE (169) is based on the Data Processing Inequality (Section 2.1.2): if if two genes X and Y are indirectly dependent via a third one (Z), then the dependence between X and Y is weaker than the dependence between X and Z, and between Y and Z. In this case, ARACNE discards a regulation between X and Y to prevent the inference of false positives. ARACNE first computes the MI of each pair of genes. Then, it considers each combination of three genes and discards the weakest pairwise dependence, if the difference between the two weakest dependences (measured with the MI) is above a certain threshold. If the true network is a tree[1], ARACNE is able to reconstruct it, due to the DPI. This method enjoys some popularity (eg. successfully used in GRN reconstruction in B cells (14)).

CLR CLR (81) normalizes the pairwise MI so that all genes have the same MI mean and variance (between itself and all other genes). In CLR, the regulation between X_i and X_j is assigned a score equal to $w_{ij} = \sqrt{z_i^2 + z_j^2}$, where:

$$z_i = \max\left(0, \frac{I(X_i, X_j) - \mu_i}{\sigma_i}\right) \tag{3.5}$$

μ_i and σ_i are the mean and the standard deviation of the mutual information between X_i and all the other genes (z_j is analogously defined). CLR tends to select highest ranked regulations involving all genes.

3.2.3 Linear models

Due to their well-studied properties, linear models are extensively used in network inference. The high variable to sample ratio prevents the use of OLS and strategies to deal with inference indeterminacy are usually adopted (such as regularization, shrinkage, or use of lower order partial correlations, see Section 2.5.1).

GeneNet A extension to infer partially directed GGMs is described next (and available in a R package named GeneNet (197, 228)). A positive definitive estimation of the covariance matrix is obtained using shrinkage, which is then used to obtain a partial correlation matrix. An undirected GGM is obtained by selecting the highest partial correlations (edges between the respective nodes). In a second step GeneNet orients edges as follows. For each pair of connected nodes, the respective partial variances (diagonal entries in the partial correlation matrix) are obtained. The ratio between the partial variance and the variance is the proportion of variance that remains unexplained after a linear regression using all other variables as predictors (see Section 2.2.2). This is used as a measure of variable endogeneity, of how much a variable is explained from within the system (the lower it is, the higher the endogeneity). An edge between two nodes is directed from the least endogenous to the most endogenous.

Tigress Tigress (111) is based on Lars (Section 2.2.4), with an additional step of *stability selection*, described next. Tigress first perturbs the gene expression data, multiplying each

[1] A tree, in graph theory, is an undirected graph with only one path between two nodes - there are no fully connected triplets.

multivariate sample by a value between 0 and 1. Then it runs Lars on the new data, in a variable selection way (each variable is considered the target, one at a time), selecting the top k ranked predictor variables (of the first k Lars steps). This step is repeated a number of times, for different values of k. An area under the curve, relating the frequency of selection as a function of k, is used to score the predictors of the target gene. Software code (in Matlab) is available online.[1]

3.2.4 Bayesian networks

Several score and search approaches to infer BN (and DBN) from gene expression data are adopted and assessed using simulated time series data (designed by the authors) in (287). The search methods include a greedy search with multiple random restarts, simulated annealing and a genetic algorithm. The adopted scoring functions are the Bayesian Dirichlet equivalence and the BIC. The used methods are publicly available, an implementation named Banjo.[2]

An BN approach to deal with the high variable case is the Sparse Candidate algorithm (90, 91). First, for each node (gene) a candidate set of parents is selected. On this constraint, a BN maximizing a score is then obtained. Then, for each node, the candidate set of parents is modified, to include the current parents plus a number of genes maximizing a score, dependent on the current network. This number of genes is chosen so that each gene has a same number of candidate parents. Another network is obtained, and the procedure is repeated until convergence (ie. the returned network is not improved upon). Software implementing this method is publicly available online.[3]

An approach combining skeleton inference with MMPC (Section 2.5.3) and hill climbing search to orient edges has also been used in GRN inference (262).

3.2.5 Random forests

Random forests have been applied for GRN inference under the name GENIE3 (available as an R function) (124). For each target gene, a random forest is created using all the other genes as the predictor variables. These are scored by importance - considered to be the average decrease in node impurities resultant from splitting, in the cases when the predictor variable is considered in the decision. The impurity of a node is measured to be the residual sum of squares (of the local regression on the response) at that node. Random forests are created using the R library randomForest.

[1] http://cbio.ensmp.fr/ ahaury/svn/dream5/html/index.html

[2] http://www.cs.duke.edu/ amink/software/banjo/

[3] http://compbio.cs.huji.ac.il/LibB/programs.html

3.3 Dynamic approaches to network inference

This section describes dynamic approaches to GRN inference, based on the estimation of lags (Section 3.3.1), VAR models and dynamic Bayesian networks (Section 3.3.2), and differential equations (Section 3.3.3).

3.3.1 Lag estimation

Lagged dependences may be used to estimate the lag (τ^*) for which the dependence between two time series is maximum:

$$\tau^* = \arg\max_{\tau} I(X_{1,t}; X_{2,t+\tau}) \tag{3.6}$$

Under the Gaussian assumption, the mutual information may be estimated using the linear correlation. Lag estimation was first proposed in the context of biochemical pathways (7, 265) and has been used several times in GRN inference (50, 155, 157, 294). A particular implementation is described next.

Time-Delay Aracne The Time-Delay ARACNE (294) (available as an R package) extends ARACNE and is based on three steps. First, it estimates the first time point on which each gene is differentially expressed. Two thresholds then are chosen, one for up-regulation, and the other for down-regulation. A gene is differentially expressed if the ratio between the expression level at time t and the expression level at time $t = 1$ is higher or lower that the up-regulation or down-regulation thresholds, respectively. The set of possible regulations is restricted to the regulations where the target gene has a start-of-regulation time higher than the start-of-regulation time of the predictor gene (if two genes have the same start-of-regulation time, regulations in both directions are allowed).

The second step of the algorithm lags the temporal expression of each pair of genes, for a consecutive number of lags, and finds the lag which maximizes the mutual information between the genes. The lagged mutual information is estimated with a copula based approach. A score is assigned to the edge from X_i to X_y, being the maximum of the lagged mutual information $I(X_{i,t-l}; X_{y,t})$, for different lag values l. The directed edges whose score is higher than a defined threshold are kept in the graph. The threshold is calculated using a statistical test based on bootstrapping the data. The third and final step of the algorithm applies the DPI property to break up fully connected triplets.

3.3.2 VAR models and dynamic Bayesian networks

VAR models and DBNs have been extensively used in the context of gene regulatory network inference. Due to the high variable to sample ratio, regularization or low order conditional dependences are usually adopted (such as bivariate GC tests (183)). One lag models are typically the most popular ((92) or (234)) but models using multiple lags have also been proposed, such as an weighted lasso (Section 2.2.4) grouping predictors at different lags (158, 236). The estimation of different networks for different time points has also been proposed (292). In what follows, particular implementations are described in more detail.

Simone A weighted lasso approach to the inference of the VAR1 coefficients is described in (49), and implemented in the R package Simone. First, genes are grouped based on connectivity patterns. Two main groups are created: hubs, which are genes showing a high likelihood of regulating other genes, and leaves, mainly connected to hubs. Hubs are assumed to be transcription factors. Each gene is assigned into one of these two groups using prior information, or an estimation from the data. The latter case consists in estimating a matrix of coefficients using the standard lasso and then to group genes into hubs or leaves according to the L_1 norm of the respective coefficients (relative to the modeling of the other variables). After this step, a target gene $X_{y,t}$ is modeled as:

$$\hat{\beta}_{hubs} = \underset{\beta \in \mathbb{R}^p}{\arg\min}((X_{y,t} - X_{t-1}^{\setminus y}\beta)^T(X_{y,t} - X_{t-1}^{\setminus y}\beta) + \lambda \sum_{i=1}^{p} w_i \mid \beta_i \mid) \tag{3.7}$$

The weight w_i is a function of whether X_i is a hub or a leaf. Let $Z_{i,hub}$ and $Z_{i,leaf}$ be parameters taking values of either 0 or 1, indicating if $X_i^{\setminus y}$ is hub or a leaf. w_i is given by:

$$w_i = \rho\rho_{iy}(\rho_{hub}Z_{i,hub} + \rho_{leaf}Z_{i,leaf}) \tag{3.8}$$

where ρ_{hub} and ρ_{leaf} are parameters defining the weights of hubs and leaves; ρ is a general parameter (to be tuned); and ρ_{iy} is a parameter to incorporate prior knowledge (set to 1 if none is used). The output of Simone is a list of networks for various values of λ. Starting from a value returning an empty network, λ is progressively shrinked. Simone outputs all the different networks obtained along the shrinkage of λ.

Recursive elastic net A GRN inference proposal (called recursive elastic net (234)) combines the elastic net with a weighted lasso. The coefficients β, for a target variable Y and a respective vector of realizations y, in the recursive elastic net, at the iteration i, are estimated as:

$$\hat{\beta^i} = \underset{\beta \in \mathbb{R}^p}{\arg\min}(X_{y,t} - X_{t-1}^{\setminus y}\beta^i)^T(X_{y,t} - X_{t-1}^{\setminus y}\beta^i) + \frac{\lambda_2}{2} \parallel \beta^i \parallel_2 + \lambda_1 \sum_{k=1}^{p} w_k^i \mid \beta_k^i \mid) \tag{3.9}$$

The coefficient weights are computed recursively in a number of iterations. Weights are not considered in the first iteration, and the following are given by:

$$w_j^i = \frac{1}{\mid \beta_j^{i-1} \mid + \delta} \tag{3.10}$$

where β_j^{i-1} is the estimation of the coefficient β_j at the iteration $i-1$ and the parameter $\delta > 0$ is used so that zero coefficients are not assigned an infinite weight.

G1DBN G1DBN (available in an R package (148)) combines first-order and higher-order partial correlations. It is a two-step method: the first step consists in the computation of all the possible first order partial correlations (and significance testing), between each gene $X_{z,t-1}$ and each gene $X_{y,t}$, conditioned on each gene $X_{i,t}$. Each directed regulation, from $X_{z,t-1}$ to $X_{y,t}$, is assigned a score equal to the maximum of the p-values of the respective coefficients. Regulations whose p-values are lower than a cut-off are selected to create a network (in the description of the method, the authors use a cut-off of 0.7). The second step of the algorithm further prunes the network obtained in the first step. In it, each node $X_{y,t}$ has a set of parents (pa(X_y)), and is modeled as:

$$X_{y,t} = \sum_{i \in pa(X_y)} \beta_{i,y} X_{i,t-1} + \epsilon_i \tag{3.11}$$

The coefficients $\beta_{i,y}$ are estimated and assigned a p-value.

3.3.3 Differential equations

In an approach named inferelator (28), the first derivative of gene expression $\dot{X_{i,t}}$ is written as:

$$X_{i,t} + \tau \dot{X_{i,t}} = g(\sum_j^p \beta_j X_{j,t}) \tag{3.12}$$

τ is a time interval to be estimated. The function g implements a truncation of extreme values. The function is approximated as:

$$X_{i,t} + \tau \frac{X_{i,t+1} - X_{i,t}}{\Delta t} = g(\sum_{j=1}^p \beta_j X_{j,t}) \tag{3.13}$$

The inferelator starts with an initial estimate for τ and the coefficients β are estimated using the lasso. Then, assuming that estimate, τ is updated with the one minimizing the prediction error. This is repeated until convergence. The coefficients β are a measure of the likelihood of the respective gene regulations. A combination of the inferelator with a lagged version of CLR (one first lag) is proposed in (162). The inferelator coefficients and the normalized MI coefficients returned by CLR are combined in the form of the root square of the sum of their squares.

3.4 Conclusion

This chapter reviewed several approaches to GRN inference (static and dynamic) that can be found in the literature. These consist on the application of statistical and machine learning notions to deal with the high variable to sample ratio of GRN inference, presented in Chapter 2. Examples include regularized linear models, approaches based on low order conditional dependences or random forests. GRN inference strategies may be static or dynamic, designed to time series. In this case, common dynamic approaches are based on lagged dependences, between variables at different time points (separated by a lag), requiring a shift of one time series relative to the other. Although multi-lag strategies can be found in the literature, usually a single first lag is adopted.

Due to typical short size of gene expression time series, inference is very challenging and improving on random inference is not guaranteed, particularly on real datasets. In this context, inference assessment assumes a crucial role. This is the topic of the first contributions Chapter 4. In the Chapter 5, several methods here described are experimentally compared in GRN inference from time series.

Part III

Contributions

4

GRN inference assessment with precision-recall curves

4.1 Introduction

The accuracy of GRN inference may be assessed by comparing the inferred network with an available gold standard. This is an instance of assessing a binary classification task, where the edges present in the gold standard are of the positive class and the absent edges of the negative class. When network inference results in a ranking of edges (in which the top ranked are associated with higher edge scores), inference may be assessed for multiple threshold values of edge selection. In this case, an approach to assess inference accuracy is through precision and recall curves (pr-curves), associating precision (y-axis) with recall (x-axis), as edges are incrementally added to the network, from the highest to lowest ranked. An alternative to the pr-curve is the receiving operating characteristic (ROC) curve, associating recall (y-axis) to the false positive rate (x-axis), introduced in the Section 1.10. The average precision in pr-curves is typically measured with the area under the curve (AUPRC), where a higher area (maximum 1, when all the positive instances are ranked higher than all the negatives) means a higher average precision.

Pr-curves and the AUPRC are commonly used to assess GRN inference accuracy, for instance in the DREAM inference challenges (166, 246). The statistical significance of the AUPRC (relative to a null hypothesis of random selection) is typically obtained using Monte Carlo, which may be computationally intensive in the case of large networks.

In this chapter we propose an alternative to estimate AUPRC significance. We will consider the case when precision-recall values are computed for all elements in the ranking, and the pr-curve is interpolated between consecutive points of recall. We derive the expected null discrete pr-curve and the expected value and variance of the AUPRC. These (and the minimum and maximum of the AUPRC) are used to obtain a continuous beta distribution, which is used as an approximation of the true null AUPRC distribution. This approximation is used to assess network inference in the Chapter 5.

Documented R software implementing the described method is available online.[1]

Continuous and discrete curves Pr-curves are discrete, defined in a finite number of recall values. In order to obtain a continuous curve, an interpolation between consecutive points is then required. However, this may result in a curve with an accentuated discrete behavior (eg. saw-tooth shapes). One alternative is to estimate smooth pr-curves, by non-parametric (eg. boostrap-based (52)) or parametric means (eg. assuming an intrinsic continuous (eg. normal) probability distribution describing the class decision, for the two classes (of negatives and positives) (38)). An implicit assumption is that the set of considered instances is a realization of two class distributions.

We only consider the first case, where consecutive points in the pr-curve are interpolated. Also, we assume that precision-recall values are computed for each ranked element. The analysis of the AUPRC from this perspective is related to average precision measures, commonly used in information retrieval (241).

Statistical significance The significance analysis of pr-curves is less developed than the one of ROC curves, where the distribution of the curve and its area under is available (31). On pr-curves, one approach to test the difference between curves is through confidence intervals (30). The null distribution (corresponding to random selection) of the pr-curve and in particular of the AUPRC is useful to test the hypothesis of random precision. The fact that a p-value is obtained also makes it possible to compare different pr-curves (obtained in problems of different null distributions and not directly comparable).

To our knowledge the only used approach to estimate the null pr-curve/AUPRC distribution in the literature is nonparametric (by Monte Carlo). This approach can be computationally intensive, requiring a high number of simulations for a good approximation of the true distribution. The sampling error is also subject to uncertainty. In this chapter an alternative is proposed, based on the estimation of a continuous approximation of the true (discrete) AUPRC distribution.

Chapter outline Section 4.2 describes different strategies to interpolate pr-curves and compute the respective AUPRC. Section 4.3 contains the analytical derivation of the expected (maximum and minimum) precision for a given recall value, as they are sufficient for the expected null

[1] https://github.com/miguelaglopes/pranker

pr-curve. Section 4.4 uses the previous results to derive the mean and the variance of the AUPRC distribution, and proposes the beta distribution approximation to it. Section 4.5 describes how to use the proposed method to assess AUPRC significance, and presents an experimental comparison with the Monte Carlo approach. The impact of the number of Monte Carlo simulations in the respective expected error (in terms of mean and variance) is also experimentally investigated.

4.2 Interpolation of the discrete pr-curve

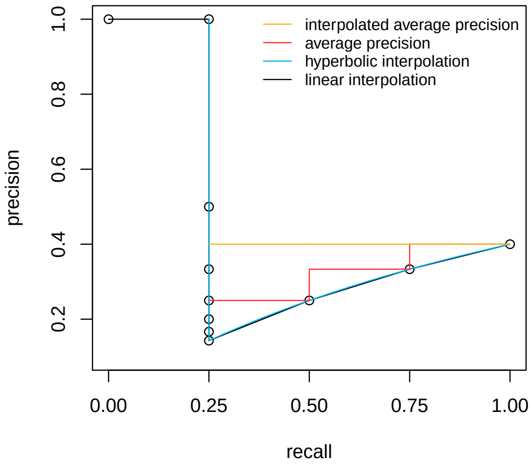


Figure 4.1: Different ways to interpolate an empirical pr-curve - Different ways to interpolate an empirical pr-curve, $P = 4$. The first selected instance is positive, the next six are negatives, and the last three are positives. Points represent the precision after the selection of each instance until all positives are selected.

Consider a ranking of N instances, of which P are positive, and that precision-recall values are obtained for each ranked element. The interpolation of the points in the curve is not straightforward since several values of precision can be associated to the same recall (i.e. each time the next instance in the ranking is not positive, precision falls and recall remains constant). A pr-curve may then characterized by a saw-tooth shape, which is the more accentuated the lower is the value of P.

Multiple interpolation strategies exist. A common approach (named *interpolated average precision*) assigns to each recall, the value of the maximum precision at that recall, or at higher recall values (165). An alternative consists in considering that the precision between two consecutive recall values is constant and equal to the maximal precision value associated to the higher recall (an approach named *average precision*). A more precise way to interpolate pr-curves is to connect the minimum precision at a given recall and the maximum precision at the subsequent recall (246). Consider a scenario with P positive instances. At a point A, after a selection of n instances, there are TP true positives and FP false positives. The precision at this point is $\frac{TP}{TP+FP}$ and the recall is $\frac{TP}{P}$. If the next selected instance is positive, the precision moves to $\frac{TP+1}{TP+FP}$ and the recall moves to $\frac{TP+1}{P}$. Let this be point B. Point A corresponds to the minimum precision at the recall $\frac{TP}{P}$, and point B corresponds to the maximum precision at the recall $\frac{TP+1}{P}$. In these two points, the precision p as a function of the recall r takes the following hyperbolic form: $p = \frac{rP}{rP+FP}$. The simpler linear interpolation between points A and B is a close approximation of the hyperbolic function, particularly between close recall points. It returns area values necessarily lower, making this approximation a close lower bound. The estimation of the AUPRC is straightforward in all interpolation strategies (except in the hyperbolic interpolation), reducing to a sum of trapezoid areas. Regarding the hyperbolic interpolation, a way to estimate the AUPRC is to incrementally increase it every time there is a recall increase. If we integrate p from point A to point B, we have the area below p, between these points. The integration of p for values of r between $\frac{TP}{P}$ and $\frac{TP+1}{P}$ (the area between these points) is easily derived[1] (246). Figure 4.1 illustrates the differences in the different interpolation approaches described, for (at least) 10 instances, 4 of which are positive. The first selected instance is positive, the next 6 are negatives, and the next 4 are the remaining positives.

4.3 The expected null pr-curve

This section derives analytically the expected maximum and minimum (and average) precision for a given recall in the case of random ranking. A value of recall can be associated with multiple values of precision, but only the maximum and minimum determine the pr-curve (as discussed in 4.2). This section also investigates the difference between the maximum and minimum null precision, as a function of P and N.

4.3.1 Expected maximum precision for a given recall

Let N be the total number of instances, and P the number of positive instances. Suppose we have a random ranking where n denotes the position of the k-th positive instance. This implies that for the recall $r = \frac{k}{P}$ we obtain the precision, $p = \frac{k}{n}$. This precision p is the maximum

[1] $\text{Area}_{AB} = \frac{1}{P}\left(1 - FP\ln(\frac{FP+TP+1}{FP+TP})\right)$

precision that can be obtained for the recall r, since further selections will either cause the precision at recall r to go down (false positive) or the recall to increase to $\frac{k+1}{P}$ (true positive). The probability that the k-th positive selected instance is the n-th selected can be obtained with the hypergeometric distribution (returning the probability of selecting k positive instances in n draws, without replacement, on a population of size N, with P positive instances (and $N-P$ negative instances). The probability that the k-th positive instance is the n-th selected instance is equal to the probability of selecting $k-1$ positive instances in $n-1$ draws (without replacement), multiplied by the probability of selecting a positive instance in the next draw (which is $\frac{P-(k-1)}{N-(n-1)}$). The first multiplicand is returned by the hypergeometric distribution:

$$\mathcal{P}_h(k-1,n-1,N,P) = \frac{\binom{N-n+1}{P-k+1}\binom{n-1}{k-1}}{\binom{N}{P}} \tag{4.1}$$

Let the probability that the k-th positive selected instance is the n-th selected instance be denoted by $\mathcal{P}_{sel}(k,n,N,P)$. This probability is also known as the negative (or inverse) hypergeometric probability (84). It is defined as:

$$\mathcal{P}_{sel}(k,n,N,P) = \frac{\binom{N-n+1}{P-k+1}\binom{n-1}{k-1}}{\binom{N}{P}}\left(\frac{P-k+1}{N-n+1}\right) = \frac{\binom{N-n}{P-k}\binom{n-1}{k-1}}{\binom{N}{P}} \tag{4.2}$$

Note that the probability that the first randomly ranked instance is a positive one (i.e. $n=1$ and $k=1$) is $\frac{P}{N}$ while the maximum precision at the recall level $\frac{k}{P}$ is $p_{max}(k)=\frac{k}{n}$. Therefore, the expected maximum precision for a recall $\frac{k}{P}$ of a random selection, is:

$$\langle p_{max}(k)\rangle = \sum_{n=k}^{n=N}\frac{k}{n}\mathcal{P}_{sel}(k,n,N,P) \tag{4.3}$$

4.3.2 Expected minimum precision for a given recall

The probability that the k-th positive instance has the n-th position in the ranking equals the probability that the minimum precision at the recall $\frac{k-1}{P}$ is $\frac{k-1}{n-1}$. Therefore, the expected minimum precision at recall $\frac{k-1}{P}$, of a random ranking, is:

$$\langle p_{min}(k-1)\rangle = \sum_{n=k}^{n=N}\begin{cases}\frac{k-1}{n-1}\mathcal{P}_{sel}(k,n,N,P), & n>1\\ \mathcal{P}_{sel}(k,n,N,P) & n=1\end{cases} \tag{4.4}$$

Note that if $n=k=1$ the first selected instance is a positive one and the value of precision is not defined when the recall is zero. In this case the value of such zero-recall precision is set to one (as in Figure 4.1). Otherwise, the precision at recall zero is also zero. Equation (4.4) can be simplified as it follows. If $k=n=1$, $\mathcal{P}_{sel}(k,n,N,P)$ is equal to $\frac{P}{N}$. If $k>1$ (i.e. the recall is $\frac{k-1}{P}$), equation (4.4) becomes:

$$\begin{aligned}\langle p_{min}(k-1)\rangle &= \sum_{n=k}^{n=N}\left(\frac{k-1}{n-1}\right)\mathcal{P}_{sel}(k,n,N,P) = \\ &= \sum_{n=k}^{n=N}\left(\frac{k-1}{n-1}\right)\frac{\binom{N-n}{P-k}\binom{n-1}{k-1}}{\binom{N}{P}} = \frac{1}{\binom{N}{P}}\sum_{n=k}^{n=N}\binom{n-2}{k-2}\binom{N-n}{P-k}\end{aligned} \tag{4.5}$$

Since according to the Chu-Vandermonde identity (8),

$$\sum_{a=0}^{c}\binom{a}{b}\binom{c-a}{d-b} = \binom{c+1}{d+1} \tag{4.6}$$

if $a = n-2$, $b = k-2$, $c = N-2$ and $d = P-2$, $\langle p_{min}(k-1)\rangle$ becomes:

$$\langle p_{min}(k-1)\rangle = \frac{1}{\binom{N}{P}}\binom{N-1}{P-1} = \frac{P}{N} \tag{4.7}$$

This implies that the expected minimum precision for any value of recall is constant and equal to $\frac{P}{N}$. An horizontal approximation of the null pr-curve of value $\frac{P}{N}$ necessarily underestimates the true null pr-curve (for any interpolation strategy that takes into account not only the minimum precision, but also the maximum precision). A discussion of this dissimilarity, and its dependence with P and N, is presented in 4.3.4.

4.3.3 Expected average precision for a given recall

For a comparative purpose, we also derive the expected average precision for a given recall $\frac{k}{P}$ of random selection. For a given recall $\frac{k}{P}$, the probability that the n-th selected instance is the k-th positive, and that the $(n+n^*+1)$-th selected instance is the $(k+1)$-th positive, is estimated. This probability is multiplied by the average of the precision $\frac{k}{n^{**}}$ for all values of n^{**} between n and $n+n^*$ (for k positive selected instances). The sum of this product, for all possible values of n and n^* gives the expected average precision for a given recall. Formally, it is as follows:

$$\langle p_{avg}(k)\rangle = \sum_{n=k}^{N}\left(\mathcal{P}_{sel}(n,k,N,P)\sum_{n^*=0}^{N-n}\mathcal{P}^*_{sel}p^*_{avg}\right) \tag{4.8}$$

where

$$\mathcal{P}^*_{sel} = \mathcal{P}_{sel}(1, n^*+1, N-n, P-k) \tag{4.9}$$

and

$$p^*_{avg}(n^*, n) = \begin{cases} \frac{k}{n} & n^* = 0 \\ \frac{1}{n^*+1}\left(\frac{k}{n} + \sum_{n^{**}=1}^{n^*}\frac{k}{n^{**}}\right) & n^* > 0 \end{cases} \tag{4.10}$$

4.3.4 Difference between expected maximum precision and expected minimum precision

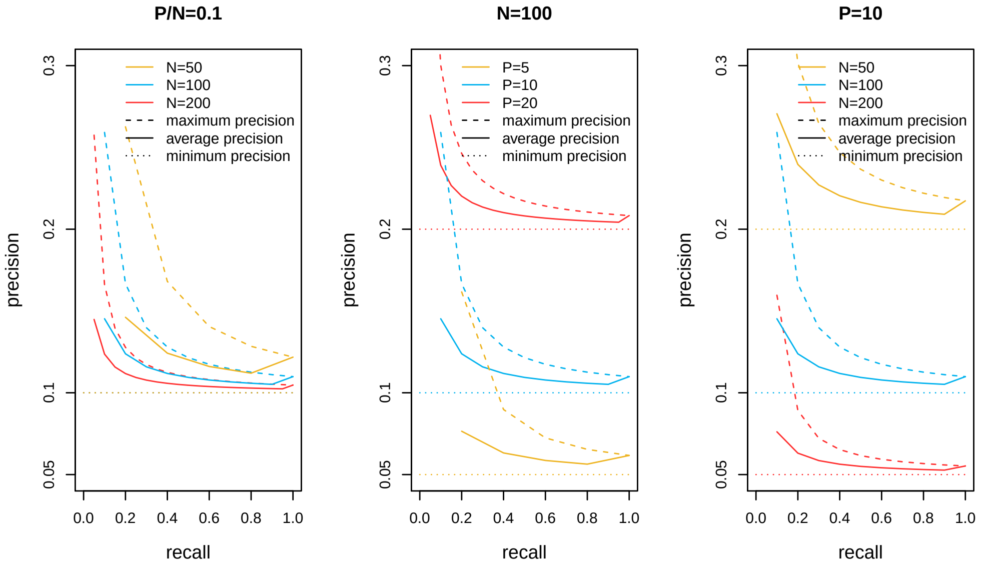


Figure 4.2: Expected maximum, average and minimum precision for different values of recall, and for different combinations of P and N - In the left plot $\frac{P}{N}$ is fixed (0.1), in the middle plot N is fixed (100), in the right plot P is fixed (10). Different colors represent different values of N and P.

On the basis of previous results it is possible to bound the gap between the expected maximum precision and expected minimum precision for a given recall $\frac{k}{P}$. The expected minimum precision, for any recall, is $\frac{P}{N}$. Since $\langle p_{min}(k)\rangle = \langle p_{min}(k-1)\rangle$, as $\langle p_{min}(k)\rangle$ is constant and does not depend on k, we obtain

$$\begin{aligned}
&\langle p_{max}(k)\rangle - \langle p_{min}(k)\rangle = \langle p_{max}(k)\rangle - \langle p_{min}(k-1)\rangle = \\
&= \sum_{n=k}^{n=N} \mathcal{P}_{sel}(k,n,N,P)\left(\frac{k}{n} - \frac{k-1}{n-1}\right) = \sum_{n=k}^{n=N} \frac{\binom{N-n}{P-k}\binom{n-1}{k-1}}{\binom{N}{P}}\left(\frac{n-k}{n(n-1)}\right) = \\
&= \sum_{n=k}^{n=N} \frac{\binom{N-n}{P-k}\binom{n-2}{k-1}}{\binom{N}{P}}\left(\frac{1}{n}\right) \qquad (4.11)
\end{aligned}$$

By replacing the term $\left(\frac{1}{n}\right)$ with $\left(\frac{1}{n-2}\right)$ we obtain an upper bound of the difference

$\langle p_{max}(k)\rangle - \langle p_{min}(k)\rangle$. This upper bound is:

$$\sum_{n=k}^{n=N} \frac{\binom{N-n}{P-k}\binom{n-2}{k-1}}{\binom{N}{P}} \left(\frac{1}{n-2}\right) = \sum_{n=k}^{n=N} \frac{\binom{N-n}{P-k}\binom{n-3}{k-2}}{\binom{N}{P}(k-1)} = \tag{4.12}$$

(and using again the Chu-Vandermonde identity, equation (4.6))

$$= \frac{\binom{N-2}{P-1}}{(k-1)\binom{N}{P}} = \frac{\binom{N}{P}\frac{(N-P)P}{(N-2)N}}{(k-1)\binom{N}{P}} = \frac{(N-P)P}{(k-1)(N-2)N} \tag{4.13}$$

Equation (4.13) represents the difference between the expected maximum precision at recall $\frac{k}{P}$ and the expected minimum precision $\frac{P}{N}$. Let us consider also the relative difference (divided by $\frac{P}{N}$). For a given recall, if P is fixed, this difference decreases with N (but the relative difference increases). If N is fixed, the difference (and relative) difference decreases with P. If $\frac{P}{N}$ is fixed, the difference increases with the number of instances.

This behavior is illustrated in the Figure 4.2, which illustrates the expected maximum, average and minimum precision as a function of recall, for different combinations of P and N. Note that there is an uptick in the average precision curve when it reaches the last recall value. This is due to the fact that when the last positive instance is selected, the curve is completed and there are no more selections - at recall 1, the maximum and average precision are the same.

The fact that the null pr-curve tends to $\frac{P}{N}$ in the asymptotic case (i.e. $P \to \infty$ and finite P/N) becomes clear. If a value of recall $\frac{k}{P}$ is finite and if $P \to \infty$, then k must also tend to infinite (and equation (4.13), representing an upper bound between the expected maximum and minimum null precision, tends to 0).

4.4 The null distribution of the AUPRC

Based on the previous results, we next derive the expected value and variance of the AUPRC. Three interpolation strategies presented in the Section 4.2 (average precision, hyperbolic interpolation and linear interpolation) are considered in what follows.

4.4.1 Expected value of the AUPRC of a random selection

Let AUPRC^{pmax} denote the AUPRC returned when using the average precision interpolation (average of the maximum precision for all recall values), by AUPRC^{hyp} the one returned by hyperbolic interpolation and by AUPRC^{lin} the one computed with linear interpolation. From equations (4.4-4.7) we can estimate the average maximum precision and the average minimum precision of a random ranking for all recall values. The average maximum precision for non-zero recall is:

$$\langle p_{max} \rangle = \frac{1}{P} \sum_{k=1}^{P} \langle p_{max}(k) \rangle \tag{4.14}$$

and the average minimum precision is:

$$\langle p_{min} \rangle = \frac{1}{P} \sum_{k=0}^{P-1} \frac{P}{N} = \frac{P}{N} \tag{4.15}$$

It follows that the expected AUPRC^{pmax} and AUPRC^{lin} of a random ranking of instances are:

$$\langle \text{AUPRC}^{pmax}_{random} \rangle = \langle p_{max} \rangle \tag{4.16}$$

$$\langle \text{AUPRC}^{lin}_{random} \rangle = \frac{\langle p_{max} \rangle + \langle p_{min} \rangle}{2} \tag{4.17}$$

The expected AUPRC$^{hyp}_{random}$ is estimated as follows: each time there is an increase in recall, there is an increase in the AUPRC. This increase is equal to:

$$\Delta A(k,n) = \begin{cases} \frac{1}{P}\left(1 - (n-k)\ln(\frac{n}{n-1}\right) & n > 1 \\ \frac{1}{P} & n = 1 \end{cases} \tag{4.18}$$

where $k = TP + 1$ and $n = TP + FP + 1$ (see Section 4.2). The value in (4.18) is the area that is added to the AUPRC when the k-th selected positive instance is the n-th selected. We can estimate the expected added area for any value of k:

$$\langle \Delta A(k) \rangle = \sum_{n=k}^{N} \Delta A(k,n) \mathcal{P}_{sel}(k,n,N,P) \tag{4.19}$$

The expected AUPRC of a random ranking, using the hyperbolic interpolation, is given by:

$$\langle \text{AUPRC}^{hyp}_{random} \rangle = \sum_{k=1}^{P} \langle \Delta A(k) \rangle \tag{4.20}$$

Finally, the expected AUPRC of a random ranking in the asymptotic case (ie. assuming that the number of instances is infinite, and $\frac{P}{N}$ is finite, see Section 4.3.4) is:

$$\langle \text{AUPRC}^{asymp}_{random} \rangle = \frac{P}{N} \tag{4.21}$$

Figure 4.3 shows the expected AUPRC for the different interpolation methods, and for the asymptotic continuous case, as a function of $\frac{P}{N}$. In the right plots, P is fixed (equal to 10), and in the left plots N is fixed (equal to 100). The plots on the top show the absolute expected AUPRC, whereas the bottom plots show the relative difference, the difference between the

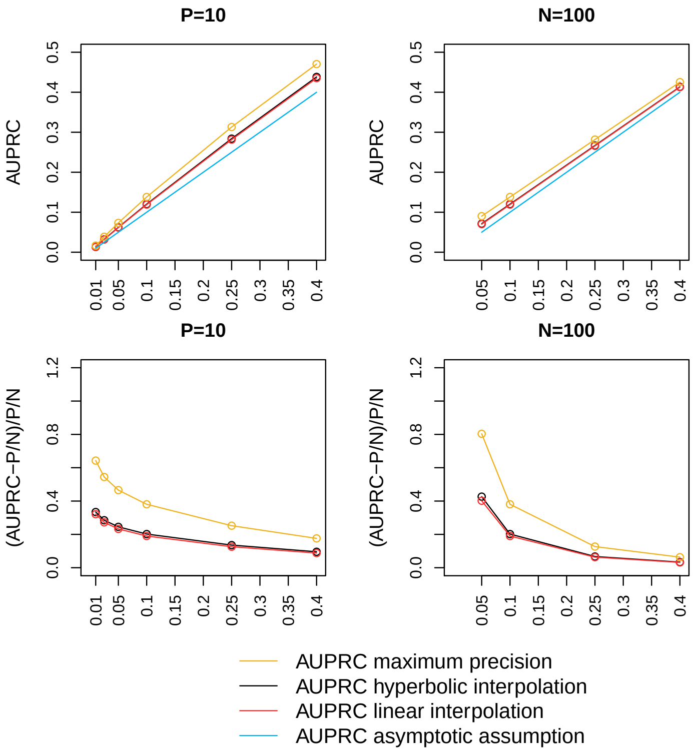


Figure 4.3: Estimated expected AUPRC of a random ranking as a function of $\frac{P}{N}$ when P=10, and N=100 - In the left plots P is fixed (10) and in the right plots N is fixed (100). The top plots represent the AUPRC. The bottom plots represent the relative difference to $\frac{P}{N}$.

expected AUPRC and the asymptotic AUPRC ($\frac{P}{N}$), divided by the latter. An increase in N leads to a decrease in the absolute difference between the AUPRC estimations (top-left plot), but leads to an increase in the relative difference (bottom-left plot). An increase in P leads to a decrease in both absolute and relative differences (in line with the results of Section 4.3.4).

4.4.2 Variance of the AUPRC of a random selection

For the sake of simplicity only the average precision approach is considered, though the obtained results can be easily extended to the linear interpolation approach, and with some more difficulty, to the hyperbolic interpolation. The expected AUPRC of a random selection, using the average precision interpolation, is given in (4.16): the AUPRC is the average of the expected maximum precision for the different recall values. This is equivalent to the sum of the expected maximum precision for the different values of recall, divided by P (equation (4.14)) The variance of a sum of random variables is equal to the sum of the values of their covariance matrix. In our case, we have:

$$\text{Var}(\text{AUPRC}^{pmax}_{random}) = \sum_{k=1}^{P}\sum_{j=k}^{P}\text{Cov}\left(\frac{1}{P}p_{max}(k), \frac{1}{P}p_{max}(j)\right) \tag{4.22}$$

For simplicity, let us define $X_k = \frac{1}{P}p_{max}(k)$ and $X_j = \frac{1}{P}p_{max}(j)$. The covariance between two random variables X_k and X_j is:

$$\text{Cov}(X_k, X_j) = \mathbb{E}(X_k X_j) - \mathbb{E}(X_k)\mathbb{E}(X_j) \tag{4.23}$$

The term $\mathbb{E}(X_k)\mathbb{E}(X_j)$ is straightforward to estimate, given that we already have $\mathbb{E}(X_k)$ and $\mathbb{E}(X_j)$ (they are given by $\frac{1}{P}\langle p_{max}(k)\rangle$ and $\frac{1}{P}\langle p_{max}(j)\rangle$, in the equation (4.3)). The term $\mathbb{E}(X_k X_j)$ is given by the sum of the product between all the possible values of X_k and X_j, multiplied by the probability of observing them:

$$\mathbb{E}(X_k X_j) = \sum_x \sum_y \mathcal{P}(X_k = x, X_j = y)xy \tag{4.24}$$

The probability $\mathcal{P}(X_k = x, X_j = y)$ can be stated as:

$$\mathcal{P}(X_k = x, X_j = y) = \mathcal{P}(X_k = x)\mathcal{P}(X_j = y|X_k = x) \tag{4.25}$$

$\mathcal{P}(X_k = x)$ is the probability that the maximum precision at the recall $\frac{k}{P}$ is xP. $\mathcal{P}(X_k = x)$ is therefore equal to $\mathcal{P}(p_{max}(k) = xP)$. The probability mass function of $p_{max}(k)$ is defined by $\mathcal{P}_{sel}(k, n_k, N, P)$ (equation (4.2)). This equation gives the probability that the maximum precision at recall $\frac{k}{P}$ is $\frac{k}{n_k}$, assuming N total instances and P positive instances. Concluding, $\mathcal{P}(X_k = x) = \mathcal{P}_{sel}(k, n_k, N, P)$, given that $x = \frac{k}{Pn_k}$ and $X_k = \frac{1}{P}p_{max}(k)$.

The conditional probability $\mathcal{P}(X_j = y|X_k = x)$ is the probability that the maximum precision at the recall $\frac{j}{P}$ is yP, given that the maximum precision at the recall $\frac{k}{P}$ is xP. If we define $yP = \frac{j}{n_j}$ and $xP = \frac{k}{n_k}$, and on the condition that $j > k$, $\mathcal{P}(X_j = y|X_k = x)$ is the probability that the $(j - k)$-th positive selected instance is the $(n_j - n_k)$-th selected instance, in a population of $N - n_k$ instances, and $P - k$ positive instances. This probability is given by $\mathcal{P}_{sel}(j - k, n_j - n_k, N - n_k, P - k)$ (4.2). Let's denote it simply by $\mathcal{P}^*_{sel}$. Equation (4.24) can be rewritten as:

$$\mathbb{E}(X_k X_j) = \sum_{n_k=1}^{N}\sum_{n_j=n_k+j-k}^{N}\mathcal{P}_{sel}(k, n_k, N, P)\mathcal{P}^*_{sel}\frac{k}{n_k P}\frac{j}{n_j P} \tag{4.26}$$

If $j < k$, $\mathbb{E}(X_k X_j) = \mathbb{E}(X_j X_k)$. If $j = k$, $\mathbb{E}(X_k X_k)$ is given by equation (4.3): $\mathbb{E}(X_k) = \frac{1}{P}\langle p_{max}(k)\rangle$, and $\mathcal{P}_{sel}(k, n, N, P)$ describes the probability mass function of $p_{max}(k)$. X_k

takes values in $\frac{k}{Pn}, n = k, k+1, ...N$. Therefore, we have:

$$\mathbb{E}(X_k X_k) = \sum_{n=k}^{n=N} \left(\frac{k}{Pn}^2\right) \mathcal{P}_{sel}(k, n, N, P) \tag{4.27}$$

Cov(X_k, X_j) can now be estimated as in (4.23), and the variance of the AUPRC of a random selection, if the pr-curve is interpolated using the average precision interpolation, is given in (4.22).

4.4.3 Distribution of the AUPRC

The knowledge of the mean and variance of the AUPRC, and the fact that it is contained in a finite interval (between a minimum and 1) suggests the adoption of the beta distribution as a parametric approximation to the AUPRC probability distribution. The beta is the most popular continuous distribution with finite support, fully described with two parameters characterizing the mean and variance, and two parameters defining the support interval. Being a continuous distribution, it can only be an approximation of the true (discrete) distribution. However, due to its ease of use, it is a practical alternative to Monte Carlo simulations.

The two shape parameters of the beta distribution, α and β, can be estimated through the methods of moments approach:

$$\alpha = x^* \left(\frac{x^*(1-x^*)}{v^*} - 1\right) \tag{4.28}$$

and

$$\beta = (1-x^*) \left(\frac{x^*(1-x^*)}{v^*} - 1\right) \tag{4.29}$$

where x^* and v^* are the normalized mean and variance (84):

$$x^* = \frac{\langle \text{AUPRC}_{random}^{pmax}\rangle - \min(\text{AUPRC}_{random}^{pmax})}{1 - \min(\text{AUPRC}_{random}^{pmax})} \tag{4.30}$$

and

$$v^* = \frac{\text{Var}(\text{AUPRC}_{random}^{pmax})}{(1 - \min(\text{AUPRC}_{random}^{pmax}))^2} \tag{4.31}$$

In the equations above, the value 1 in the denominator corresponds to the maximum AUPRC$_{random}^{pmax}$. The minimum is attained when the last P ranked instances are all positive. Since precision is non-zero only in the last k recall points the minimum is returned by:

$$\min(\text{AUPRC}_{random}^{pmax}) = \frac{1}{P} \sum_{k}^{P} \frac{k}{N - P + k} \tag{4.32}$$

The support of the resulting standard beta distribution should be re-transformed: multiplying by the range $(1 - \min(\text{AUPRC}_{random}^{pmax})$, and addition of $\min(\text{AUPRC}_{random}^{pmax})$.

4.5 Empirical assessment

This section assesses the beta distribution approximation against the true AUPRC distribution. The Figure 4.4 plots a Monte Carlo (empirical) obtained AUPRC distribution (10000 number of simulations) and the approximation based on the beta distribution. The ratio $\frac{P}{N}$ was fixed at 0.5, and the number of instances was set to 200, 800 and 3200. The AUPRC corresponding to a p-value of 0.1 in the beta distribution approximation is drawn, and the corresponding p-value in the Monte Carlo distribution is indicated. It is seen that the beta distribution approach reasonably approximates the true AUPRC distribution, and this approximation seems more precise as the number of instances increases. This is expected, as as the number of instances increases, the discrete saw-tooth shape of the null AUPRC distribution is smoothed out.

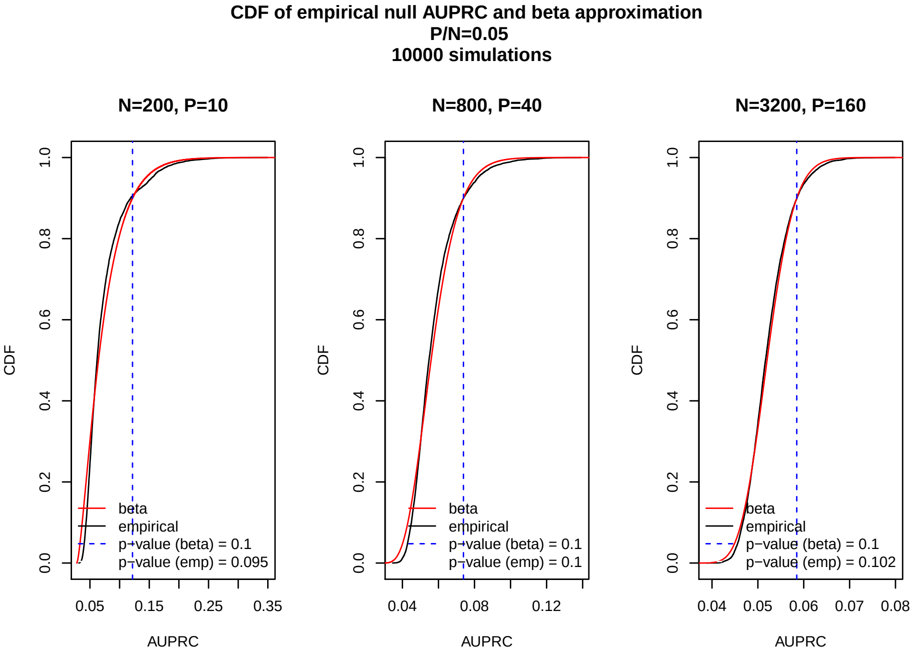


Figure 4.4: Cumulative distribution of the AUPRC (empirical and beta distribution) - The distribution of the AUPRC for three cases of increasing number of instances was estimated with 10000 simulations, or with the beta distribution approach. The ratio $\frac{P}{N}$ is fixed at 0.5.

The error of the approximation when the number of positive instances is low is illustrated in the Figure 4.5. For configurations of $N = 2500$ and P equal to 10, 50 and 100, 10000 AUPRC z-scores (standard scores of the normal distribution) of random rankings were obtained with the beta distribution approximation. These are shown in the figure, including a reference of values generated from a standard normal distribution. It can be seen that there is a positive bias, increasing as P decreases. Further investigation on this error should be carried out.

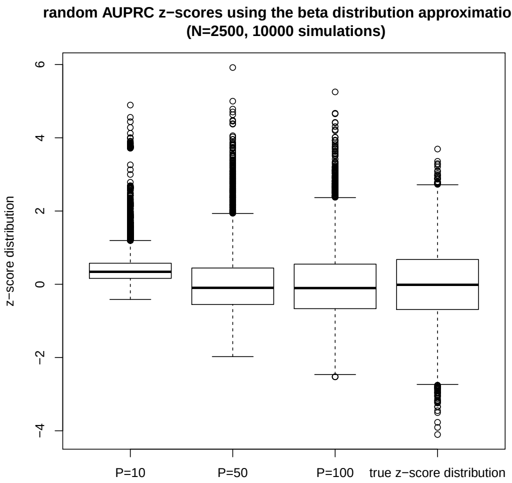


Figure 4.5: Boxplots of null AUPRC z-scores, for different values of P - Boxplots of 10000 null AUPRC z-scores, obtained with the beta distribution approximation, for $N = 2500$ and P equal to 10, 50 and 100. A reference of values generated from a standard normal distribution is also shown

A final experiment concerns the study of the impact of the number of simulations on the accuracy of the empirical AUPRC distribution. Figure 4.6 shows the relative difference between the empirical and expected AUPRC mean and variance (relative to the expected values), as a function of the number of simulated AUPRC. This experiment was performed for $N = 90, 180, 900$, keeping P fixed at 15. The values shown are the average of 1000 sets of simulations (the variance is also plotted for $N = 90$ and $N = 900$, for simplicity for a couple of simulation numbers only). The number of simulated AUPRC is shown in a logarithmic scale of base 2, and goes from 2 to 131072 (2^{17}). The relative difference is shown in a logarithmic scale of base 10. A threshold for the relative difference corresponding to 0.01 (10^{-2}) is drawn. As expected, the empirical mean and variance tend to the respective expected values, as the number of simulated AUPRC increases. When $N = 90$, the relative difference of both the mean and variance is below 0.01 when the number of simulations is higher than 32768 (2^{15}). When $N = 900$ this number increases to 131072 simulations (2^{17}). The number of needed simulations to approximate the true distribution increases with the number of instances. As described, the expected value and variance of the null AUPRC can be used to control the sampling error of

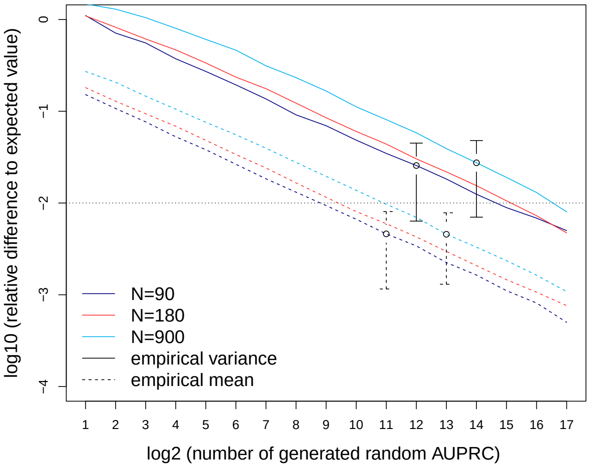


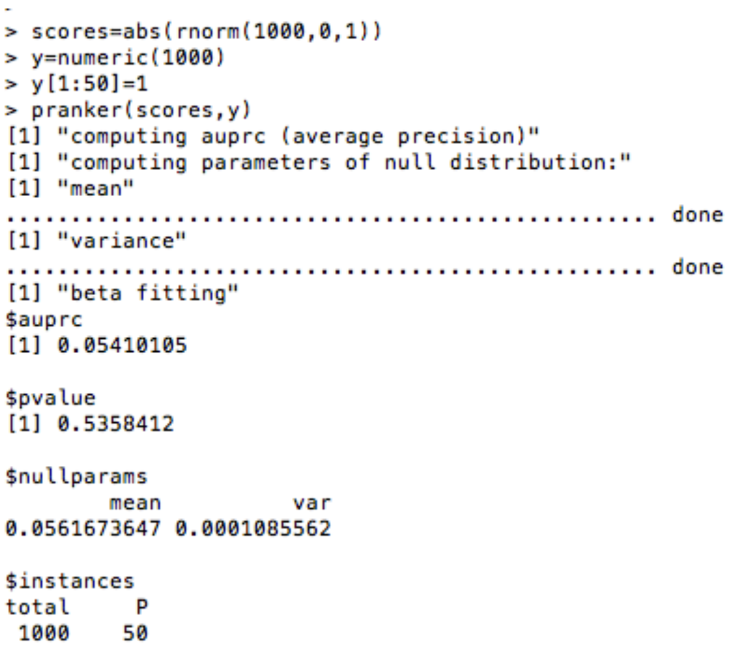
Figure 4.6: Comparison between empirical and the true parameters of the AUPRC distribution - Relative difference of the mean and variance of the empirical AUPRC distribution, relative to the true values, as a function of the number of random AUPRC generations.

empirical estimations.

4.6 Conclusion

Pr-curves (and the respective AUPRC) are commonly used to assess the accuracy of IR algorithms. The distribution of the null AUPRC (of random ranking) can be used to estimate the significance of single and multiple independent AUPRC values. A direct approach to obtain the AUPRC distribution is to compute the probability of each possible curve (by multiplying the probability at each point, conditioned on the previous one, for all points). An AUPRC is assigned to each curve and probability, and the AUPRC distribution may then be reconstructed. However, when the number of instances is high this approach is unfeasible due to the large amount of possible pr-curves (of order of magnitude close to N^P).

Monte Carlo methods are typically used to estimate the AUPRC probability distribution, however they can also be computationally intensive, requiring a high number of simulations to accurately approximate the true distribution. In this chapter we deduced analytically expressions for the expected null pr-curve, and for the null AUPRC mean and variance. These parameters can be used as performance baselines, and to assess the quality of Monte Carlo distributions. We also propose to use them to estimate a parametric continuous approximation of the null AUPRC distribution based on the beta distribution. This approximation is shown to become more precise as the number of instances increases, however a non-negligible bias was found for low values of P (in the particular case of $N = 2500$ and $P = 10$). It is used in the network inference assessment of chapter 5.



```
-
> scores=abs(rnorm(1000,0,1))
> y=numeric(1000)
> y[1:50]=1
> pranker(scores,y)
[1] "computing auprc (average precision)"
[1] "computing parameters of null distribution:"
[1] "mean"
.................................................. done
[1] "variance"
.................................................. done
[1] "beta fitting"
$auprc
[1] 0.05410105

$pvalue
[1] 0.5358412

$nullparams
        mean          var 
0.0561673647 0.0001085562 

$instances
total     P 
 1000    50 
```

Figure 4.7: Example of R code for AUPRC assessment. - Ranking assessment using a R implementation of the proposed beta-distribution approximation to the AUPRC, available in the author's page.

Implementation and computational limitation This beta distribution approximation of the AUPRC is proposed as an alternative to lengthy Monte Carlo simulations, however it is also computationally intensive when the number of instances is high. This is due to the computation of the AUPRC variance, obtained after considering all the probabilities in the support space (equations (4.22 - 4.24)). A software implementation in R (integrating C++ code) is available online[1], computing the expected value and variance more efficiently than the direct application of the described formulas. Another approach to speed up the variance computation, implemented in the available code, is to skip intermediate elements of the covariance matrix in the equation (4.22), and then interpolate them with splines. In a 2010 macbook, the AUPRC distribution is approximated in minutes when the number of instances is in the order of the tens of thousands. When it is in the order of hundreds of thousands and higher, the computation takes longer. However, once computed, the parameters may be kept for future use. A comparison between

[1] https://github.com/miguelaglopes/pranker

the computational performance of the Monte-Carlo approach and the proposed approach is left for future work. Alternatives to speed up the computation of the AUPRC variance should also be investigated. One alternative approach to overcome this limitation is to replace the AUPRC interpolation strategy with an approximation, described next. Nevertheless, even with the speed limitation in the large N case, the proposed method remains an attractive alternative to Monte Carlo when the number of instances is not too large (ie. hundreds of thousands) and not too low (ie. few dozens). The Figure 4.7 illustrates an application of the available R function. It takes as input a ranking (randomly generated in the example) and a gold standard (in the example, the first 50 instances are positive).

Alternatives and future research Instead of computing precision values for all values of recall to obtain the AUPRC, one alternative is to compute the precision for different points of n (number of selections) (eg. sampled at same intervals). A sensible AUPRC score would be an weighted average of these precision values, with the weights proportional to the difference between the associated recall and the one of the previous selected point (the sum of weights would be 1). The advantage of this AUPRC score is that its expected value and variance is less demanding to compute, even when the number of instances is high. In this case, for a given n, there is a maximum of P possible values of precision whose probability needs to be computed. On the contrary, in the considered interpolation strategies, which associate a precision value to each possible recall, there may be close to N possible values of precision for a single recall. Having the expected value and variance of this new AUPRC, its distribution can also be approximated with the beta distribution. The only non-trivial aspect in this procedure is the consideration of the weights, which depend on the associated and previous precision. Another alternative to AUPRC significance to both Monte Carlo and the proposed beta distribution approximation is to assign a p-value to the p-values of all the points in the obtained pr-curve. Note that the p-values are dependent and methods to combine them should be applied. However, current methods to combine dependent p-values often entail a continuous assumption on the variables distribution (40, 142). The alternatives here described constitute interesting points of future research.

5

GRN inference from gene expression time series

5.1 Introduction

This chapter presents an experimental investigation on GRN inference from gene expression time series. It also describes two novel techniques for GRN inference from time series: one filter network inference approach based on a fast measure of (first order) conditional GC (denoted by GC3), and one approach to identify co-regulated genes (ie. regulated by the same transcription factors), not a network inference approach by itself, but which can be incorporated in network inference methods. These were introduced in the Section 1.11 and are described in detail in the Section 5.2.

Experiments were performed on 11 multivariate time series (of 1731 considered genes) of *Saccharomyces cerevisiae* (yeast) microarray gene expression (of 18 and 25 time points), and on 100 simulated multivariate time series (of 50 genes) of increasing size, from 20 to 300 time points. A gold standard of known regulations was used in both cases. Due to the high number of considered genes and time series, the reported experiments and findings constitute a relevant contribution to network inference assessment and validation.

Three experiments were performed. The first is on the selection of lags in the estimation of temporal dependences, in particular of bivariate GC in cause-effect pairs of gene expression time series. We assess causal inference accuracy when using single or multiple lags (which may be the first, or estimated) and the Toda and Yamamoto (TY) modification of the GC linear test to take into account non-stationarity. As more lags (predictors) are used, the model becomes more variant. Using higher lags also means discarding a higher number number of used samples (at the extremes of the time series). This variance increase is likely to be critical when the time series are of short size, as in the gene expression case.

The second experiment assesses the proposed method to identify co-regulated genes by comparing it with the available gold standard and computing significance p-values. The third experiment assesses different dynamic strategies to GRN inference from time series, using the same lag selection strategy (the first lag). 100 networks of 50 genes (in both yeast and simulated time series) were inferred. The assessed strategies (described in the Chapter 3) are bivariate GC, dynamic adaptations of MI-based methods (MI, ARACNE, CLR), forward selection filter (mRMR, CMIM, mIMR) and embedded variable selection (lasso and random forests), methods from the literature (Simone, G1DBN, GeneNet), the proposed novel methods, and an ensemble of dynamic methods. GRN inference methods should be well-balanced in terms of bias and variance and this experiment aims to shed light on an optimal balance. Inference accuracy was assessed with the AUPRC, following the method described in Chapter 4. Both the lag selection and network inference experiments are bias-variance trade off investigations, in which simpler models are more biased but less variant.

Documented R software implementing the novel GC methods and the used dynamic adaptations of static methods is available online.[1] R data files and scripts used in the experimental session are also available.[2]

Results In the lag selection experiment, the most precise strategy in the yeast time series is to consider only the first lag. In the simulated time series, the selection of a higher number of lags (as well as the consideration of non-stationarity) results in a more precise inference as n becomes higher (> 40). The proposed co-regulation identification method is shown to identify co-regulated genes with a very significant precision.

In the network inference experiment, precision was very low in the yeast time series, with few methods performing better than random. The obtained precision was higher in the simulated time series, increasing with the time series size n. In this case, and when $n > 40$, the most precise approach was the proposed GC3 method.

Chapter outline Section 5.2 presents the two algorithmic contributions to GRN inference. Section 5.3 presents the experimental session and results. Section 5.4 is the discussion.

[1] https://github.com/miguelaglopes/GCnetinf

[2] https://github.com/miguelaglopes/NetInfExps

5.2 Contributions to GRN inference

In this section we present two novel extensions to GC based network inference. The first is a fast approximation of first order conditional GC scores which are used to infer networks. This method scores the predictors of each target with a measure of dependence, and may thus be considered a filter variable selection method (Section 1.8), adapted to time series. The second is a simple method to identify co-regulated genes, which can be incorporated in network inference methods.

5.2.1 Model justification

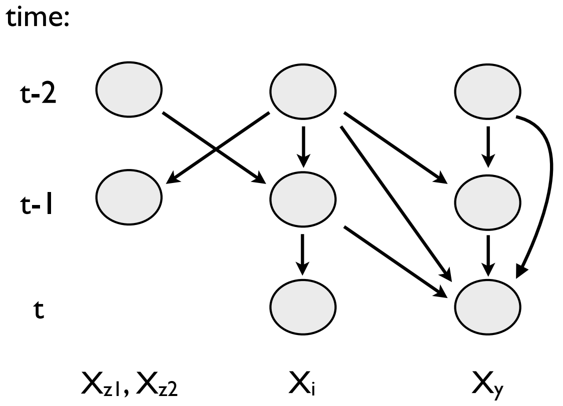


Figure 5.1: Simple causal network - Network with causal dependences up to 2 lags. There is causality from gene $X_{i,t}$ to $X_{y,t}$, but there is also a statistical dependence between $\{X_{z_1,t-l}, X_{z_2,t-l}\}$ and $X_{y,t}$ if $X_{i,t-l}$ is not conditioned on. $X_{z_1,t}$ is the node on top.

Consider two gene expression time series, $X_{i,t}$ and $X_{y,t}$, where $X_{i,t-1}$ is a cause of $X_{y,t}$ and where $X_{y,t}$ is a function of $X_{i,t-l}$ and $X_{y,t-l}$, for $l = 1, 2$. This case is represented in Figure 5.1. The figure also shows two other genes, $X_{z_1,t}$ and $X_{z_2,t}$, which have an indirect lagged dependence with $X_{y,t}$, via $X_{i,t}$. In network inference, $X_{i,t}$ should then be considered (conditioned on) when assessing the existence of direct causality from either $X_{z_1,t}$ and $X_{z_2,t}$ to $X_{y,t}$. In the high variable to sample ratio case, it is not feasible to condition on all possible variables. An alternative approach, sometimes used in the literature, is to condition on a single third gene (170). In dynamic GRN inference one example is G1DBN, described in Section 3.3.2. This strategy avoids the high variable to sample limitation, as regulation scores are only based on the expression of three variables (genes).

If all indirect regulations between two genes are mediated via a single third gene, an inference strategy based on first-order conditional dependences is sufficient to infer the true network, as discussed in the Section 1.11.2. In the context of graphical models, this condition means that there is a single path connecting each pair of connected nodes in the network. As mentioned, the limitation of this strategy is that only indirect dependences via a single gene are screened off.

If only the network structure is relevant (ie. indication of its edges) inference consists in applying conditional independence tests between each pair of nodes (genes), for all single conditioning nodes. An edge (regulation) is absent from the network if there is a null conditional independence between its nodes. This amounts to the PC algorithm restricted to first order conditional independence tests (Section 2.5.2). If possible edges are to be scored and ranked (useful for the assessment using precision-recall or ROC curves), an appropriate edge score is the minimum of conditional dependence scores between its nodes, for all single conditioning nodes. By conditional dependence score we mean a score (eg. the standard score of the normal distribution) obtained from the p-value resultant from a test on the null hypothesis of conditional independence (a lower p-value meaning a higher score). If the ranking of edges associated with a null conditional dependence is irrelevant, conditional independence tests may stop as soon as conditional independence is found.

We adopt this strategy for time series, considering linear conditional GC tests. The returned p-value is transformed into a z-score (standard score of the normal distribution) and used as the GC score. As mentioned in the Section 2.7.4, in the single lag case a GC test is equivalent to a test on the linear lagged partial correlation between the predictor and target, conditioned on the past of the target (ie. considering auto-correlation). The benefit of conditioning on the past of the target is seen in the Figure 5.1. $X_{i,t}$ (considered the target) and $X_{y,t-1}$ (considered the predictor) are dependent, due to a common cause in $X_{i,t-2}$. However, this dependence is screened off if $X_{i,t-1}$ or $X_{i,t-2}$ are conditioned on.

Scoring each possible edge with the minimum of first order conditional GC scores between its nodes requires a search over all conditioning nodes (each implying a conditional GC test). Scoring all possible edges this way may be computationally burdensome in large networks. Assume that the time complexity of a linear regression model with p predictors using OLS is $\mathcal{O}(p^2)$.[1] In a network with p genes, assuming a single lag, scoring each regulation requires $p-2$ (one for each conditioning variable) linear unrestricted models with three predictors (the lagged target, predictor and conditioning variable), each with complexity $\mathcal{O}(3^2)$, and $p-2$ restricted models with two predictors (the lagged target and conditioning variable), each with complexity $\mathcal{O}(2^2)$. This results in a complexity of $\mathcal{O}(13(p-2))$. The complexity of scoring all possible regulations (p^2-p, excluding auto-regulatory) is then approximately $\mathcal{O}(13p^3)$. Note

[1] Assuming $p < n$, it is $\mathcal{O}(np^2)$ asymptotically, but we drop the n for simplicity.

that we assume each edge is scored with the minimum of conditional GC scores, even for edges associated with null GC (if edge ranking is irrelevant for these cases, the search may stop as soon as null GC is found).

We propose an heuristic to speed up the search for the minimum conditional GC score of each possible edge, controlled by a stopping criterion. This method is designated by GC3 (as each score is a function of three variables) in the remainder of this chapter. In what follows, $G_{zy|i}$ means a conditional GC score from $X_{z,t}$ to $X_{y,t}$, conditioned on $X_{i,t}$ (and G_{zy} a non-conditional GC score).

5.2.2 Fast minimum first order conditional GC (GC3)

The linear GC score from $X_{z,t-l}$ to $X_{y,t}$, conditioned on $X_{i,t-l}$ ($G_{zy|i}$), is estimated with the following modification of the restricted/unrestricted models of equations (2.46) and (2.47):

$$X_{y,t} = \alpha_0 + \left(\sum_{l=1}^{L_y} \alpha_l X_{y,t-l} + \beta_l X^z_{t-l} + \gamma_l X_{i,t-l} \right) + \epsilon_{ut} \tag{5.1}$$

$$X_{y,t} = \alpha_0 + \left(\sum_{l=1}^{L_y} \alpha_l X_{y,t-l} + \gamma_l X_{i,t-l} \right) + \epsilon_{rt} \tag{5.2}$$

The GC score is obtained by comparing the two models using the F-test of equation (2.52), changing the number of denominator degrees of freedom to $N-3L-1$ (taking into account $X_{i,t}$). As mentioned above, when scoring network edges with a first order conditional dependence score (in this case, GC), an appropriate score is the score minimum, for all conditioning nodes. In this case, X_i is one such as $i = \arg\min_j(G_{zy|j})$. Scoring all possible edges this way may be computationally intensive in large networks and we propose a faster approximation, described next.

5.2.2.1 Searching for the conditioning gene

For each possible regulation from $X_{z,t}$ to $X_{y,t}$, the genes $X_i, i \in \{1,..p\}, i \neq z, y$ are ranked according to a parameter b_{ziy}, functioning as a proxy of the conditional GC score between $X_{z,t}$ and $X_{y,t}$ conditioned on $X_{i,t}$. The indices i are sorted by b_{ziy}, and $G_{zy|i}$ z-scores are estimated consecutively, following that order. The minimum GC z-score is kept along the process. If s consecutive $G_{zy|i}$ scores are higher than the current minimum, the process stops. As s increases, this approximation tends to the full search of all conditioning variables.

Two ranking approaches Consider the Figure 5.2. If the dependence between $X_{z,t-l_1}$ and $X_{y,t}$ is due to $X_{i,t-l_2}$ then $X_{i,t-l_2}$ is either an effect of $X_{z,t-l_1}$ and cause of $X_{y,t}$ or a common cause of $X_{z,t-l_1}$ and $X_{y,t}$.

In the first case, $G_{zi} > G_{iz}$, in the second $G_{zi} < G_{iz}$. In both cases, both G_{iy} and $\max(G_{zi}, G_{iz})$ should be higher than G_{ky} and $\max(G_{ki}, G_{kz})$ for a gene X_k which does not block the dependence between $X_{z,t-l_1}$ and $X_{y,t}$.

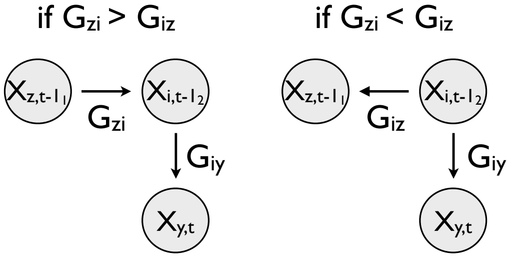


Figure 5.2: Spurious direct causality between $X_{z,t-l_1}$ and $X_{y,t}$ via $X_{i,t-l_2}$ - The two situations where $X_{z,t-l_1}$ and $X_{y,t}$ are dependent via a single gene $X_{i,t-l_2}$ are represented. Either $X_{i,t-l_2}$ is an intermediate gene in the causal chain (left side), or a common cause of $X_{z,t-l_1}$ and $X_{y,t}$ (right side). In the first case, $G_{zi} > G_{iz}$, in the second $G_{zi} < G_{iz}$ (G being GC scores).

We suggest two approaches to estimate b_{ziy}. The first takes into account the previous consideration and the two possible cases where $X_{i,t}$ blocks the dependence between $X_{z,t}$ and $X_{y,t}$:

$$b_{ziy} = G_{iy} + \max(G_{zi}, G_{iz}) \tag{5.3}$$

The second is a faster static approximation and approximates the GC score above with the linear correlation between the respective genes:

$$b_{ziy} = |\rho(X_{z,t}, X_{i,t})| + |\rho(X_{i,t}, X_{y,t})| \tag{5.4}$$

where ρ is the Pearson correlation. This score is simply the average of the non-lagged correlation between $X_{z,t}$ and $X_{i,t}$ and between $X_{i,t}$ and $X_{y,t}$.

This network inference approach (GC3) is described in the algorithm 5.1 (where the two strategies to compute b_{ziy} are designated by correlation and GC based). This approach, for different values of s (1, 3 and 5), is assessed in the experimental session. The advantage of the correlation based strategy is that it is faster, requiring a bivariate correlation matrix. The GC strategy is better motivated but requires a matrix of bivariate GC scores. As s increases, the outcomes of these strategies converge.

Algorithm 5.1: A (fast) measure of first-order conditional GC.

```
input : a $n \times p$ gene expression time series matrix; parameters $l_{max}$ and $L_{max}$; bivariate
        GC z-scores $G_{ij}, i, j \in \{1, .., p\}$; a stopping criteria $s$; ranking method
        (correlation or GC)
output: GC3 z-scores $G^3_{ij}, i, j \in \{1, .., p\}$
for each target $X_{y,t}, y \in \{1, .., p\}$ do
    for each predictor $X_{z,t}, z \in \{1, .., p\}, z \neq i$ do
        for each conditioning variable $X_{i,t}, i \in \{1, .., p\}, i \neq y, z$ do
            if ranking method = GC-based then
                $b_{ziy} = G_{iy} + \max(G_{zi}, G_{iz})$
            if ranking method = correlation-based then
                $b_{ziy} = |\rho(X_{z,t}, X_{i,t}| + |\rho(X_{i,t}, X_{y,t})|$
        $i^*$ = indices $i$ ordered by $b_{ziy}$ value;
        initialize $G^3_{zy}$ with a high value;
        for each $j \in \{i^*_1, .., i^*_{p-2}\}$ do
            estimate the conditional GC, $G_{zy|j}$;
            if $G_{zy|j} < G^3_{zy}$ then
                $c$=0;
                $G^3_{zy} = G_{zy|j}$;
            if $G_{zy|j} > G^3_{zy}$ then
                $c$=$c$+1;
            if $c > s - 1$ then
                break;
```

Computational time The computational time of GC3 is difficult to pin down exactly (as it depends on the quality of the proxy variable), but is controlled by s. In the best case scenario (for each regulation, the stopping criteria is met after $1 + s$ GC estimations), GC3 involves the estimation of $(1 + s)p^2$ linear regression models with three and two predictors (not considering the ranking computation). The complexity of the best-case search is then $\mathcal{O}((1 + s)(13p^2))$ (assuming the complexity of linear regression as in Section 5.2.1 above).

It is known that the average number of coin flips to get straight n heads (with p_h being the probability of observing heads) is $\frac{{p_h}^{-n}-1}{1-p_h}$.[1]

[1] http://www.quora.com/Whats-the-expected-number-of-coin-flips-until-you-get-two-heads-in-a-row

Assuming that the ranking of the proxy variable is a better approximation of the ranking of conditional GC scores than random ranking, the expected computational time is lower than $\mathcal{O}((1+\frac{0.5^{-t}-1}{0.5})13p^2)$. This gives us $\mathcal{O}(7*13p^2)$ for $s=2$, $\mathcal{O}(15*13p^2)$ for $s=3$ and so on. If, for instance, the average probability that the next conditional GC score in the ranking is lower than the previous one is 0.8 (if the proxy ranking reasonably approximates the true ranking), we have $\mathcal{O}(6*13p^2)$ for $s=3$. This improves the $\mathcal{O}(13p^3)$ time of the full search by a magnitude equal to $p/6$. Note that this does not consider the computation of the ranking, which if estimated with GC scores is approximately $\mathcal{O}(5p^2)$ (p^2-p linear regressions with two and one predictors).

5.2.3 Co-regulation identification

GRN inference using conditional independence tests is vulnerable to type 2 errors (to miss true regulations) if either the predictor or target genes are co-regulated (and highly correlated) with another gene. In this case, conditioning on this gene reduces the mutual information between the predictor and target genes, and the significance of conditional independence tests.

Co-regulation may also lead to the inference of spurious regulations (type 1 errors) if a bivariate measure is used to infer GRN (eg. inferring a regulation between the genes X_{z_2} and X_y in the Figure 5.1). Identifying co-regulation may be then be useful to GRN inference: by discarding regulations between co-regulated genes prior to, or after inference. In the following section a simple test to identify co-regulation is proposed, theoretically validated in the linear case under a parameter constraint.

5.2.3.1 A co-regulation identification algorithm

The proposed test to identify co-regulation is as follows: two genes are co-regulated if their expression time series exhibit a high correlation, and a stronger non-lagged correlation than otherwise (in any temporal direction). The first condition is justified as we wish to identify the pairs of correlated co-regulated genes that may be incorrectly identified as a cause-effect pair (or screen off the dependence involving co-regulated genes). The second condition is true in the linear case under a condition described in the next section, through the analysis of the covariance structure of linear causal models (Section 2.3).

Algorithm 5.2 describes the adopted co-regulation identification test (for simplicity, only the first lag is considered). The correlation should be transformed into p-values (in the experimental session we adopted the Fisher's transformation, appendix A), to take into account differences in the number of samples due to lagging. A potential problem of this approach is the incorrect identification of genes involved in a regulation as being co-regulated, preventing the inference of true regulations (type 2 errors). This aspect, and the precision of the method, are investigated in the experimental session (in Section 5.3.3).

In the Section 5.3.4 we assess the inclusion of this method in network inference using bivariate GC and the GC3 filter introduced above. In both cases, regulations between co-regulated genes are discarded (with the aim of preventing type 1 errors). Additionally, in the conditional GC case, genes co-regulated with the predictor or with the target variables are not used as the conditioning variables (with the aim of preventing the type 2 errors).

Algorithm 5.2: Identification of co-regulation.

```
input  : a $n \times p$ gene expression time series matrix; a correlation threshold $k$
output : a co-regulation matrix $C_{i,j}, i, j \in \{1, .., p\}$ (1 meaning co-regulation)
for each $X_i, i \in \{1, .., p\}$ do
    for each $X_j, j \in \{1, .., p\}, j \neq i$ do
        (estimate lagged correlation):
        $\rho_{t-1} = \max(|\rho(X_{i,t-1}, X_{j,t})|, |\rho(X_{j,t-1}, X_{i,t})|)$ ;
        (estimate (non-lagged) correlation):
        $\rho_t = |\rho(X_{i,t}, X_{j,t})|$ ;
        if $\rho_t > \rho_{t-l}$ then
            if $\rho_t > k$ then
                $C_{i,j} = 1$ ;
                ($X_i$ and $X_j$ are identified as co-regulated)
```

5.2.3.2 Method justification

Assume three variables X_t, $Y_{1,t}$ and $Y_{2,t}$. X_t is a cause of the two other variables; and all variables are dependent on its past (at a single previous time point) and error terms, representing non-modeled sources of variation (which are assumed to be independent). In the linear case, we have:

$$X_t = \beta_X X_{t-1} + \epsilon_X \tag{5.5}$$

$$Y_{1,t} = \beta_{Y_1} Y_{1,t-1} + \beta_{X,Y_1} X_{t-1} + \epsilon_{Y_1} \tag{5.6}$$

$$Y_{2,t} = \beta_{Y_2} Y_{2,t-1} + \beta_{X,Y_2} X_{t-1} + \epsilon_{Y_2} \tag{5.7}$$

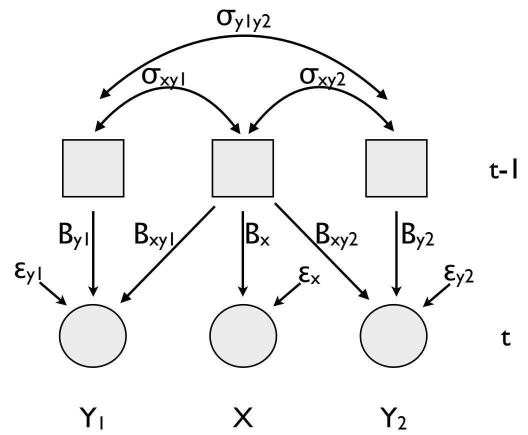


Figure 5.3: Co-regulation case involving three genes - X_t regulates both $Y_{1,t}$ and $Y_{2,t}$. All genes are also dependent of its past (the single previous time point) and subject to error terms.

Figure 5.3 is a SEM representation (Section 2.3) of the causal network for two consecutive time points (t and t-1), with each edge characterized by the respective linear coefficient. The covariances between the variables at t-1 (they are exogenous variables) are represented by bi-directed arcs. These covariances are equal to the ones at the subsequent time point, which can be inferred by path tracing rules.[1]

The covariance between X_t and $Y_{1,t}$ is obtained by multiplying the variance of X_t by the sum of the products of the coefficients of each distinct SEM path between X_t and $Y_{1,t}$ (Section 2.3). There are two SEM paths betweenX_t and $Y_{1,t}$: through X_{t-1}, and through X_{t-1} and Y_{t-1}. Note that bi-directed arcs can only be passed once in each path. Assume that all variables have unit variance (without loss of generality). The covariance between X_t and $Y_{1,t}$ is the same as the covariance between X_{t-1} and $Y_{1,t-1}$ (denoted by σ_{X,Y_1}) and is given by:

$$\sigma_{X,Y_1} = \beta_{X,Y_1}\beta_X + \beta_{Y_1}\sigma_{X,Y_1}\beta_X \tag{5.8}$$

The above can be reformulated:

$$\sigma_{X,Y_1} = \frac{\beta_{X,Y_1}\beta_X}{1 - \beta_{Y_1}\beta_X} \tag{5.9}$$

There are four open paths between $Y_{1,t}$ and $Y_{2,t}$. The covariance σ_{Y_1,Y_2} is given by:

$$\sigma_{Y_1,Y_2} = \beta_{X,Y_1}\beta_{X,Y_2} + \beta_{X,Y_1}\sigma_{X,Y_2}\beta_{Y_2} + \beta_{Y_1}\sigma_{Y_1,Y_2}\beta_{Y_2} + \beta_{Y_1}\sigma_{Y_1,X}\beta_{X,Y_2} \tag{5.10}$$

[1] In the drawn graph, the exogenous (at t-1) variables are distinguished by symbol (squares) - following the SEM representation of Section 2.3.

and we obtain:

$$\sigma_{Y_1,Y_2} = \frac{\beta_{X,Y_1}(\beta_{X,Y_2} + \sigma_{X,Y_2}\beta_{Y_2}) + \beta_{Y_1}\sigma_{X,Y_1}\beta_{X,Y_2}}{1 - \beta_{Y_1}\beta_{Y_2}} \tag{5.11}$$

There are two open paths between $Y_{1,t}$ and $Y_{2,t-1}$. From them we obtain the covariance between those two variables:

$$\sigma_{Y_{1,t},Y_{2,t-1}} = \beta_{Y_1}\sigma_{Y_1,Y_2} + \beta_{X,Y_1}\sigma_{X,Y_2} \tag{5.12}$$

We wish to prove that $|\sigma_{Y_1,Y_2}| > |\sigma_{Y_{1,t},Y_{2,t-1}}|$ - that the non-lagged covariance is higher than the lagged one (in absolute terms). The condition is then:

$$|\sigma_{Y_1,Y_2}| > |\beta_{Y_1}\sigma_{Y_1,Y_2} + \beta_{X,Y_1}\sigma_{X,Y_2}| \tag{5.13}$$

Which is true if:

$$|\sigma_{Y_1,Y_2}| > |\beta_{Y_1}\sigma_{Y_1,Y_2}| + |\beta_{X,Y_1}\sigma_{X,Y_2}| \tag{5.14}$$

The inequality above is implied by the following inequality (they are equivalent if β_{Y_1} is positive):

$$|\sigma_{Y_1,Y_2}(1 - \beta_{Y_1})| > |\beta_{X,Y_1}\sigma_{X,Y_2}| \tag{5.15}$$

Substituting σ_{Y_1,Y_2} with equation (5.11) and σ_{X,Y_2} with equation (5.9) (σ_{X,Y_2} is defined analogously as σ_{X,Y_1}), we obtain:

$$\left| \frac{\beta_{X,Y_1}(\beta_{X,Y_2} + \sigma_{X,Y_2}\beta_{Y_2}) + \beta_{Y_1}\sigma_{X,Y_1}\beta_{X,Y_2}}{(1 - \beta_{Y_1}\beta_{Y_2})}(1 - \beta_{Y_1}) \right| > \left| \frac{\beta_{X,Y_1}\beta_{X,Y_2}\beta_X}{1 - \beta_{Y_2}\beta_X} \right| \tag{5.16}$$

Substituting σ_{X,Y_1} and σ_{X,Y_2} with equation (5.9) we obtain:

$$\left| \beta_{X,Y_1}\left(\beta_{X,Y_2} + \frac{\beta_{X,Y_2}\beta_X\beta_{Y_2}}{1 - \beta_{Y_2}\beta_X}\right) + \frac{\beta_{Y_1}\beta_{X,Y_1}\beta_X\beta_{X,Y_2}}{1 - \beta_{Y_1}\beta_X} \right| > \left| \frac{\beta_{X,Y_1}\beta_{X,Y_2}\beta_X(1 - \beta_{Y_1}\beta_{Y_2})}{(1 - \beta_{Y_2}\beta_X)(1 - \beta_{Y_1})} \right| \tag{5.17}$$

Simplifying:

$$\left| \left(1 + \frac{\beta_X\beta_{Y_2}}{1 - \beta_{Y_2}\beta_X} + \frac{\beta_X\beta_{Y_1}}{1 - \beta_{Y_1}\beta_X}\right) \right| > \left| \frac{\beta_X(1 - \beta_{Y_1}\beta_{Y_2})}{(1 - \beta_{Y_2}\beta_X)(1 - \beta_{Y_1})} \right| \tag{5.18}$$

After equaling the denominators on the left hand side, we obtain:

$$\left| \frac{1 - \beta_X^2\beta_{Y_1}\beta_{Y_2}}{1 - \beta_{Y_1}\beta_X} \right| > \left| \frac{\beta_X(1 - \beta_{Y_1}\beta_{Y_2})}{(1 - \beta_{Y_1})} \right| \tag{5.19}$$

The inequality only depends on the coefficients β_{Y_1}, β_{Y_2} and β_X, representing the autocorrelation of the variables in consecutive time points. The subspace of β_{Y_1} and β_{Y_2} for which the inequality is true, for $\beta_X = \{0.3, 0.5, 0.7\}$ is represented in the Figure 5.4. It can be seen

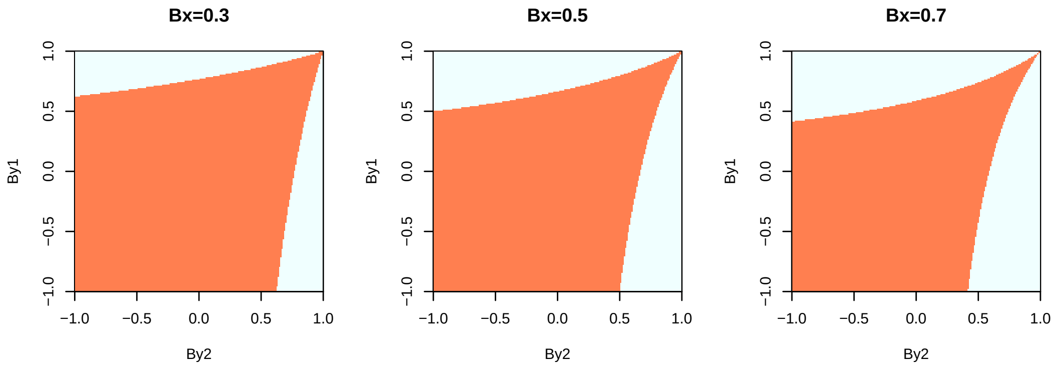


Figure 5.4: Coefficient subspace for which the co-regulation criterion is valid - If $Y_{1,t}$ and $Y_{2,t}$ have a common cause (X_t), and follow a linear model as described in the Figure 5.3, their non-lagged correlation is higher than their lagged correlation in a subspace of β_{Y_1}, β_{Y_2} and β_X, as described in the equation (5.19). The figure illustrates this subspace, between -1 and 1, for $\beta_X = \{0.3, 0.5, 0.7\}$ in orange.

that the inequality is true for a large subspace of β_{Y_1} and β_{Y_2}. The coefficients were restricted to the interval between -1 and 1. This results from the assumption of unit variable variance. The variance of X_t is obtained as the sum of the squares of the coefficients of its causes (X_{t-1} and ϵ_X) as they are independent:

$$\sigma^2_{X_t} = \beta^2_X + \epsilon^2_X \Rightarrow 1 = \beta^2_X + \epsilon^2_X \Rightarrow 1 > |\beta_X| \tag{5.20}$$

Trivially, the result remains valid if other causes are considered (as in the case of $\sigma^2_{Y_{1,t}}$ and $\sigma^2_{Y_{2,t}}$).

5.3 Experimental session

The experimental section consists of a set of experiments using real microarray and simulated gene expression time series. The simulated time series were generated with GeneNetWeaver (GNW (230)), in which a gold standard of gene regulations is provided. Multiple multivariate time series of different sizes were generated. The GNW time series were considered as they are used in the well-known DREAM network inference challenges.

The microarray gene expression data are two publicly available sets of multivariate yeast gene expression time series (described below). Gene regulations reported in the literature were used as the gold standard. Three experiments were performed, described next.

- Section 5.3.2 assesses different approaches to model GC, regarding number of lags and consideration of non-stationarity, in the causal inference of the direction of known gene regulations, as a function of n (time series size).
- Section 5.3.3 assesses the precision of the co-regulation identification approach of Section 5.2.3.
- Section 5.3.4 compares a variety of state of the art network inference approaches, ensemble of methods, and the proposed GC3 filter and co-regulation identification method. Inference is assessed with the AUPRC, following the approach proposed in the Chapter 4).

5.3.1 Methods and materials

5.3.1.1 Data and gold standard

Yeast microarray data The microarray time series datasets are from the same species and are composed of multiple multivariate time series:

- 2 multivariate time series of gene expression of *Saccharomyces cerevisiae* (yeast) (214), along two cell cycles. The time series are of 25 points, from 0 to 120 minutes, sampled at a regular interval of 5 minutes. The data is available in the NCBI GEO database, accession number GSE4987.
- 9 multivariate time series of gene expression of different strands of *Saccharomyces cerevisiae* (238), composed of time series of 18 time points, sampled at irregular times from from 0 to 340 min. These were linearly interpolated resulting in time series of 35 points, of 10 min interval. The NCBI GEO accession number is GSE24771.

A gold standard of 2960 regulatory interactions involving the genes present in both yeast datasets was obtained from YeastMine, at the Saccharomyces Genome Database [1]. These regulations were originally obtained from YEASTRACT, a repository of yeast gene regulations (1). In YeastMine, the YEASTRACT regulations were filtered in order to remove the ones curated from two or fewer publications. This way, only high confidence regulations are considered. YEASTRACT collects gene regulations documented in the literature, inferred from changes in gene expression due to deletions or mutations in transcription factors, or from binding associations, for instance obtained with ChIP technology. In the first case, a limitation of this gold standard is that direct and indirect regulations are not distinguished. Only the genes involved in a gold standard regulation were considered, 1734 in total.

GNW data Time series simulating E.Coli gene expression (1565 genes, 3758 regulations) were generated with GeneNetWeaver (v3.1 Beta) (230). 10 networks of 50 genes were initialized, and for each of them 10 multivariate time series of 300 time points were generated (resulting in 100 multivariate time series).[2] To assess the impact of the number of time points in the inference quality, only the n central points were considered, $n = \{20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 300\}$ (in the last case all points are considered).

5.3.1.2 Network inference methods

We assessed several state of the art network inference methods, both static and dynamic, including the proposed GC3 filter (Section 5.2) and a bivariate GC test (designated henceforth by GC2). GC tests were carried out using the F-test as described in Section 2.7.4, returning a p-value/z-score. The co-regulation test (Section 5.2) was included in both GC2 and GC3, as described in Section 5.2.3.1. The adopted correlation cut-off was 0.7 (the precision and recall for this value, in the considered time series, is presented in Section 5.3.3).

We also assessed ready-to-use network inference methods (available in R implementations) and dynamic adaptations of static state of the art network inference methods. All were presented in the Chapter 3. We also combined all the dynamic methods with the Borda count method, as done in the DREAM challenges (Section 3.1)).

In most of the approaches, edge scores were measured with (conditional) linear correlations, or with (regularized) linear regression coefficients, entailing a Gaussian assumption. For instance, in the MI-based methods, MI was estimated as a function of the linear correlation (see Section

[1] http://yeastmine.yeastgenome.org/yeastmine/ data from September 2013.

[2] Regarding the parameters of the GNW time series generation: the duration of time series is 14950, number of measured points is 300, the other being the default options.

2.1.3). The only exception is random forests. All the dynamic approaches use one lag (the first). The assessed approaches are described next.

Literature methods We implemented three ready-to-use methods to infer gene regulatory networks (Simone, G1DBN, GeneNet), available in R packages and described in the Chapter 3. GeneNet is a static approach, although it predicts directed networks (Section 3.2.3). Simone and G1DBN are dynamic approaches, described in Section 3.3.2. We adopted default parameters on all methods. In particular G1DBN uses a p-value cut-off of 0.7 (as described in (148)). In Simone, we score each regulation with the sum of its scores, for all returned networks (for different values of λ).The proposed clustering into two groups is also adopted.

RF and Lasso We assessed two network inference methods based on embedded variable selection, using random forests (RF, Section 2.6 and the Lasso, Section 2.2.4). Each gene, one at a time, is considered the target and the remaining genes considered as predictors. We considered dynamic adaptations of these methods, by lagging the predictors (one lag) relative to the target. In the case of random forests, a static (non-lagged) version was also assessed.

We implemented RF using the R package *randomForest*, with the parameters as in the Genie3 method (Section 3.2). Regarding the Lasso, we used the appropriate modification to Lars (see Section 2.2.4) available in the R package *lars* to compute the Lasso coefficients along the path of the penalty term, from maximum to zero (the least squares solution), sampled in a regular intervals of the ratio between the L1 norm of the coefficient vector and the norm at the least squares solution. Each regressor is assigned a score which is the average of its lasso coefficients along the lasso path.

MI A simple bivariate filter approach consists in the estimation of the MI of the expression of each pair of genes (eg. $I(X_i, X_j)$). We score each (directed) regulation with an estimation of the lagged MI (1 lag) of the respective genes (eg. $I(X_{i,t-1}, X_{j,t})$).

ARACNE ARACNE (Section 3.2.2) removes an edge of a closed triplet if the corresponding (bivariate) MI is the lowest of the three edges MI, and if the difference between that MI and the second lowest MI is above a threshold. We implemented ARACNE with a temporal modification: a regulation from X_i to X_y is removed if there is a gene X_j for which the following condition holds. Let $\min(I(X_{i,t-1}, X_{j,t}), I(X_{j,t-1}, X_{y,t}))$ be denoted as $\min I_{ijy}$:

$$|I(X_{i,t-1}, X_{j,t}) - I(X_{j,t-1}, X_{y,t})| < \min(I_{ijy} - I(X_{i,t-1}, X_{y,t})) \tag{5.21}$$

CLR The CLR (Section 3.2.2) consists of a normalization of the pairwise MI. In the adopted dynamic modification, the normalized MI between $X_{i,t-1}$ and $X_{y,t}$ is:

$$I^*(X_{i,t-1}, X_{y,t}) = \sqrt{I_i(X_{i,t-1}, X_{y,t})^2 + I_y(X_{i,t-1}, X_{y,t})^2} \tag{5.22}$$

where

$$I_i(X_{i,t-1}, X_{y,t}) = \left(\frac{I(X_{i,t-1}, X_{y,t}) - \mu_i}{\sigma_i}\right) \tag{5.23}$$

and μ_i and σ_i are the mean and standard deviation of $I(X_{i,t-1}, X_k), k \in \{1, .., p\}$. $I_y(X_{i,t-1}, X_{y,t})$ is defined equivalently, where μ_y and σ_y are the mean and standard deviation of $I(X_{k,t-1}, X_y), k \in \{1, .., p\}$.

MRMR, CMIM, MIMR MRMR, CMIM, and mIMR (Section 3.2.1) consist in a forward selection of predictor variables, the selected variable at each step being the one maximizing a criterion considering the pairwise and/or conditional dependences with the target and the previously selected variables. Following the approach implemented in MRNET using mRMR, we apply these methods to GRN inference by scoring each regulation with the criterion value (at the time of selection) of the respective gene, for a given target gene.

The adopted dynamic modification consists in using lagged correlations (1 lag) between target and predictor variables. Dependences between predictors are estimated with non-lagged correlations. In both mIMR and CMIM implementations, the selection stops as soon as the dependence (linear correlation) between the last selected predictor and the target is below a cut-off (pv< 0.05, using Fisher's transformation, appendix A). This was done to avoid intensive computations for lowly scored regulations.

5.3.2 Experiments - GC lag selection

The first experiment assesses GC lag selection in GRN inference. In the yeast experiment, a GC z-score was assigned to each direction (correct and incorrect) of a directed cause-effect pair (ie. directed regulation) present in the gold standard, for each of the multivariate time series. The scores of all multivariate time series were then averaged (as with the Stouffer's method, appendix A.4). The inferred direction was the one of the highest value. Significance was assessed with the binomial distribution (probability of a number of successes each with probability 0.5, in a sequence of independent experiments). A similar experiment was done on the GNW time series, the difference being that each directed regulation in each multivariate time series was assessed individually.

Causal direction inference strategies Three methods to infer the causal direction were assessed. The first is a lagged correlation approach (LC), returning the maximum lagged correlation from a predictor to target gene, for a given number of lags ($l \in \{1, .., l_{max}\}$). The second uses the GC F-test. The third uses the TY modified GC F-test (considers non-stationarity, Section 2.7.5). In the GC tests (both standard and TY-modified) the predictor lags are either the first or estimated in a forward selection manner (by RSS minimization). In all cases, the lags of the target variable are the first, and the number of target and predictor lags are the same.

Regarding LC, the lag was defined as the first or it was estimated ($l_{max} = 3$). The first case is designated as LC-F (first), the second as LC-E (estimated). In the GC approaches, the lag selection was the following:

1. The number of lags L is 1, and the lag is the first (GC-F1).
2. The number of lags L is 1, and the lag of the predictor is estimated (GC-E1).
3. The number of lags L is estimated (may be multiple), and the lags are the first ones (GC-FM).
4. The number of lags L is estimated, and the lags of the predictor are estimated (GC-EM).

The parameter L in the multi-lag settings is defined as the value that optimizes the AICc (appendix A.1) in the restricted model (equation (2.47)).

Regarding the TY test for GC, it was implemented by first selecting the lags as above, then by estimating the maximum order of integration d of the two time series (Section 2.7.2.4). The KPSS test was used to test the null of stationarity (Section 2.7.3.1, p-value cut-off of 0.1). The TY modified test was carried out as described in Section 2.7.5. This approach is referred to as TY.

The causal inference precision is presented for the top 10% and the top 10-50% ranked regulations (ordered by z-score average). The precision of the highest ranked regulations is of particular importance, as GRN inference consists in a selection of the most highly ranked (according to some criteria) network edges.

Results - yeast The results for the yeast microarray data are in the Figure 5.5. It presents the proportion of correctly inferred causal directions, for the top 10% and the top 10-50% ranked regulations. In this experiment, the optimal GC modeling strategy is GC-F1 (described above). The consideration of estimated and multiple lags causes a performance decrease, as well as the adoption of the TY test. LC has a high precision considering the top 10% regulations, but this precision is lower than all other approaches when considering the lower ranked interactions.

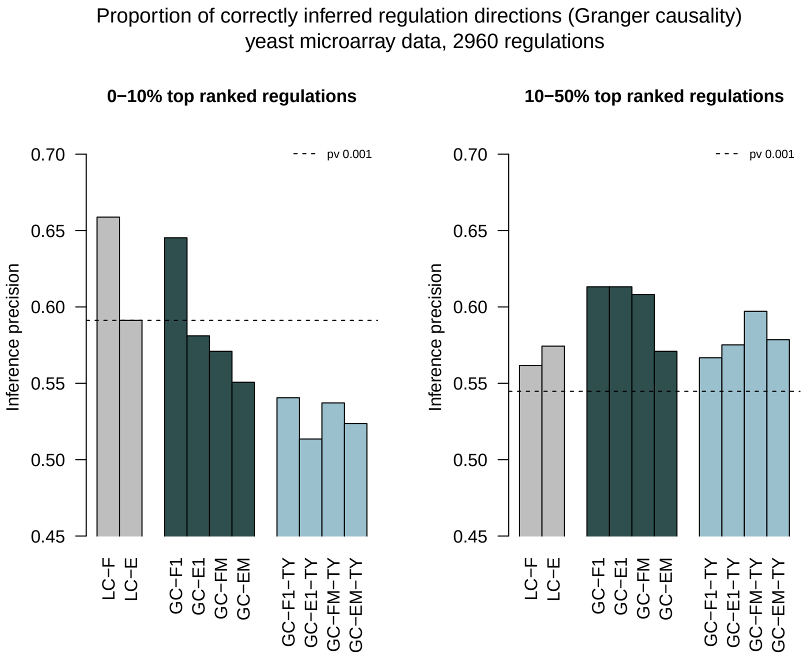


Figure 5.5: GC modeling assessment (yeast) - Proportion of correctly inferred regulation directions, for the yeast microarray time series. Direction was inferred as being the one returning the highest average GC z-score (for 11 times series sets). Results are for the different approaches to infer GC. The top 10% and the top 10-50% ranked regulations (of the respective method) are considered. The number of regulations is 2960. The dashed line corresponds to a p-value equal to 0.001, obtained using the binomial distribution.

The adoption of the TY modification has a negative impact, more visible in the 10% ranked interactions. The best achieved precision on the top 10% regulations (LC-F and GC-F1), is around 0.65 (considering 296 regulations, its p-value is 1^{-7}).

Results - GNW The results for the GNW data are in the Figure 5.6, showing the proportion of correctly inferred causal directions, for the top 10% and the top 10-50% ranked regulations. Results are presented as a function of the time series size n. As expected, inference precision increases with n. It is also higher when only the top regulations are considered (see also the difference in scale in the figure plots). Considering the top 10% regulations, when n is high, the highest precision is obtained considering multiple lags.

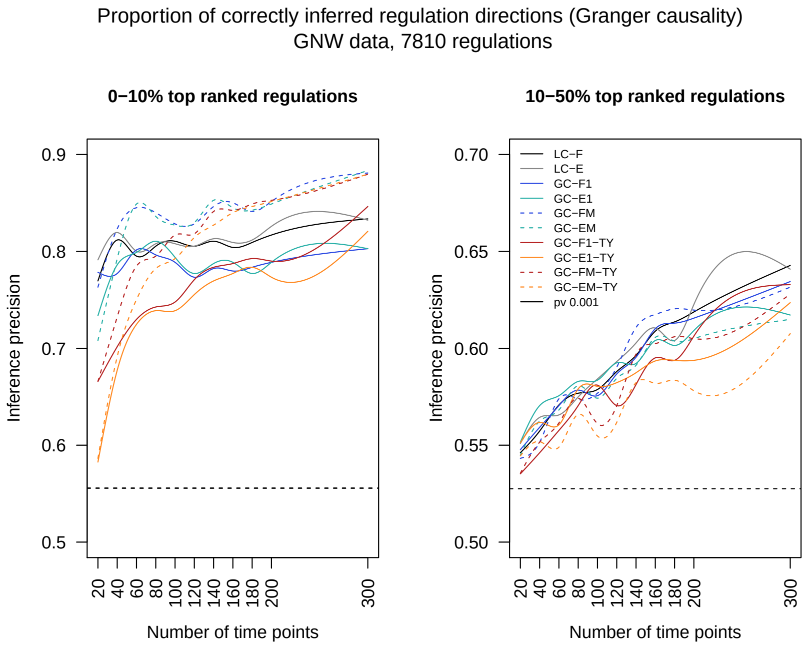


Figure 5.6: GC modeling assessment (GNW) - Proportion of correctly inferred regulation directions, for each pair of GNW time series (of the same set) whose genes are involved in a causal regulation. Direction was inferred as being the one returning the most significant GC. Results are for the different approaches to infer GC and are presented as a function of the time series size (interpolated with splines). The top 10% and the top 10-50% ranked regulations (of the respective method) are considered. The number of regulations is 7810. The dashed line corresponds to a p-value equal to 0.001, obtained using the binomial distribution. The legend is only shown in the right plot, to avoid redundancy.

5.3.3 Experiments - identification of co-regulation

The second experiment consists in the assessment of the co-regulation identification method proposed in Section 5.2.3. Each pair of time series in each time series set, in the yeast and GNW datasets, was tested for co-regulation following the algorithm 5.2, for $k = \{0.9, 0.7, 0.5\}$. A gold standard of co-regulation is directly obtained from the gold standard of regulation. The tables 5.1, 5.3 and 5.5 represent the accuracy of the method for the yeast and for the GNW datasets, when $n = \{20, 300\}$. The tables present the number of pairs of genes estimated as co-regulated ($\#\hat{cr}$), the number of true co-regulations ($\#cr$), and the number of the correctly estimated as co-regulated ($\#(\hat{cr} \cap cr)$). Precision, recall and a p-value (hypergeometric distribution) are also presented. Regarding the gold standard of co-regulation, only co-regulated genes not involved in a causal regulation, were considered as co-regulated.

Table 5.1: Precision of co-regulation inference (yeast) - $\#\widehat{cr}$ is the number of estimated co-regulated pairs of genes; $\#cr$ is the number of true co-regulated; $\#(\widehat{cr} \cap cr)$ is the number of pairs of genes correctly identified as co-regulated; precision and recall are relative to the identification of true co-regulated pairs, among the estimated as co-regulated; N is the number of considered pairs of genes. The p-value is given by the hypergeometric distribution.

	$\#\widehat{cr}$	$\#(\widehat{cr} \cap cr)$	Precision	Recall	$\#cr$	N	p-value
k=0.9	15089	4376	0.290	0.00252	1733039	16527621	0 (e^{-1977})
k=0.7	454799	73746	0.162	0.0425	1733039	16527621	0 (e^{-7155})
k=0.5	1883050	242827	0.129	0.140	1733039	16527621	0 (e^{-6232})

Table 5.2: Incorrect identification of regulation as co-regulation (yeast) - $\#\widehat{cr}$ is the number of estimated co-regulated pairs of genes; $\#r$ is the number of regulations; $\#(\widehat{cr} \cap r)$ is the number of regulations whose genes were incorrectly identified as co-regulated; precision and recall are relative to the identification of regulations, among the pairs of genes estimated as co-regulated; N is the number of considered pairs of genes. The p-value is given by the hypergeometric distribution.

	$\#\widehat{cr}$	$\#(\widehat{cr} \cap r)$	precision	recall	$\#r$	N	p-value
k=0.9	15089	18	0.00119	0.000553	32527	1652762	0.985
k=0.7	454799	908	0.00200	0.0279	32527	16527621	0.385
k=0.5	1883050	3983	0.00212	0.122	32527	16527621	7.92^{-7}

Table 5.3: Precision of co-regulation inference (GNW, n=20) - legend as table 5.1.

	$\#\widehat{cr}$	$\#(\widehat{cr} \cap cr)$	precision	recall	$\#cr$	N	p-value
k=0.9	2628	1784	0.678	0.0275	64810	122500	2.04^{-56}
k=0.7	7318	4584	0.626	0.0707	64810	122500	1.38^{-67}
k=0.5	12737	7607	0.597	0.117	64810	122500	2^{-60}

Table 5.4: Incorrect identification of regulation as co-regulation (GNW, *n*=20) - legend as table 5.2.

	$\#\widehat{cr}$	$\#(\widehat{cr} \cap r)$	precision	recall	$\#r$	N	p-value
k=0.9	2628	157	0.0597	0.0203	7760	122500	0.764
k=0.7	7318	419	0.0572	0.0539	7760	122500	0.986
k=0.5	12737	765	0.0601	0.0985	7760	122500	0.944

A potential problem of the co-regulation filter is the incorrect identification of regulation as co-regulation, preventing the inference of true regulations (type 2 errors). To assess the degree

Table 5.5: Precision of co-regulation inference (GNW, n=300) - legend as table 5.1.

	$\#\hat{cr}$	$\#(\hat{cr} \cap cr)$	precision	recall	$\#cr$	N	p-value
k=0.9	3528	2713	0.768	0.0418	64810	122500	1.33^{-196}
k=0.7	7445	5038	0.676	0.0777	64810	122500	3.82^{-157}
k=0.5	9933	6529	0.657	0.101	64810	122500	1.02^{-160}

Table 5.6: Incorrect identification of regulation as co-regulation (GNW, *n*=300) - legend as table 5.2.

	$\#\hat{cr}$	$\#(\hat{cr} \cap r)$	precision	recall	$\#r$	N	p-value
k=0.9	3528	148	0.0419	0.0191	7760	122500	$1 - 1.33^{-8}$
k=0.7	7445	336	0.0451	0.0433	7760	122500	$1 - 1.90^{-12}$
k=0.5	9933	441	0.0444	0.0568	7760	122500	$1 - 1.39^{-17}$

of this (possible) limitation, direct regulations were considered as the positive instances in the described test (results in tables 5.2, 5.4 and 5.6). In this case, $\#r$ is the number of regulations, $\#(\hat{cr} \cap r)$ is the number of regulations whose genes were (incorrectly) identified as co-regulated, and precision, recall and p-value are relative to the identification of regulation.

In both experiments (yeast and GNW), the proposed co-regulation identification method achieves a (relatively) very high precision (illustrated by the extremely low p-values, particularly on the yeast experiment - tables 5.1, 5.3 and 5.5). On the contrary, the number of regulations whose genes were identified as co-regulated (type 2 errors) is in the range of random selection in the yeast experiment (except when $k = 0.5$) and GNW experiment when $n = 20$. In the GNW case, this number is significantly low when $n = 300$. This is likely due to the fact that in the GNW time series direct regulations are clearly characterized by lagged dependences - ie. they tend to be stronger than non-lagged dependences in cause-effect pairs. In this case, cause-effect pairs are not selected by the co-regulation filter.

5.3.4 Experiments - network inference

The network inference experiment consists in the inference of 100 networks of 50 genes, in both the yeast microarray and GNW time series, using the inference methods previously described in Section 5.3.1.2. For each network and inference method a matrix of regulation scores was obtained, which was assessed with the AUPRC (average precision approach, Section 4.2).

In the yeast experiment, 1000 sets of 50 genes were randomly selected. The 100 gene sets with the highest number of regulations (present in the gold standard) were then selected as the network genes. Each 50-gene network was then reconstructed by each method, each time using a multivariate time series. The edge scores obtained with the different multivariate time series were then combined (Borda count method) to obtain a single score for each edge, on each network. Alternatively, each network obtained in each multivariate time series was assessed individually (results in the appendix B, not shown in this section). In the GNW experiment, each network was inferred from an unique multivariate time series.

The yeast networks have a number of edges between 7 and 24, with the average number being 10. The total number of possible edges in a 50-gene network is $50^2 - 50 = 2450$ (excluding auto-regulations). For this range of total and positive instances ($N = 2500$ and $P \sim 10$) there is a relevant bias in the AUPRC beta distribution approximation of Chapter 4, as illustrated in the Figure 4.5. For this reason, we assessed AUPRC significance with Monte Carlo simulations (100000 simulations for each different network configuration and AUPRC distribution).

The GNW networks have a number of edges between 51 and 118, and the average number of regulations is 84.6. The number of edges (positive instances) in these networks is higher than in the yeast experiment, and the bias of the AUPRC beta distribution is lower. It was considered acceptable (see Figure 4.5), and AUPRC z-scores were obtained following the proposed beta distribution approximation.

In the following results, Granger causality methods (GC2 and GC3), possibly incorporating the co-regulation filter, are designated by:

- GC2: bivariate GC z-scores, using one lag (the first)
- GC2+CF: bivariate GC z-scores, using the first lag, with the co-regulation filter (with parameter $k = 0.7$) as described in Section 5.2.3.
- GC3-SR1: The proposed GC3 filter, as described in Section 5.2.2, using the first lag. The ranking method (Section 5.2.2.1) is static, and the stop parameter is $s = 1$.
- GC3-DR1: GC3 filter with dynamic ranking and $s = 1$.
- GC3-DR3: GC3 filter with dynamic ranking and $s = 3$.
- GC3-DR5: GC3 filter with dynamic ranking and $s = 5$.
- GC3-DR3+CF: GC3 filter with dynamic ranking, $s = 3$, and with the co-regulation filter (with parameter $k = 0.7$) as described in Section 5.2.3.

The assessed literature and state of the art methods were described previously in Section 5.3.1.2. Only two static methods were implemented (GeneNet and RF). The Borda-count combination of dynamic methods is designated by META.

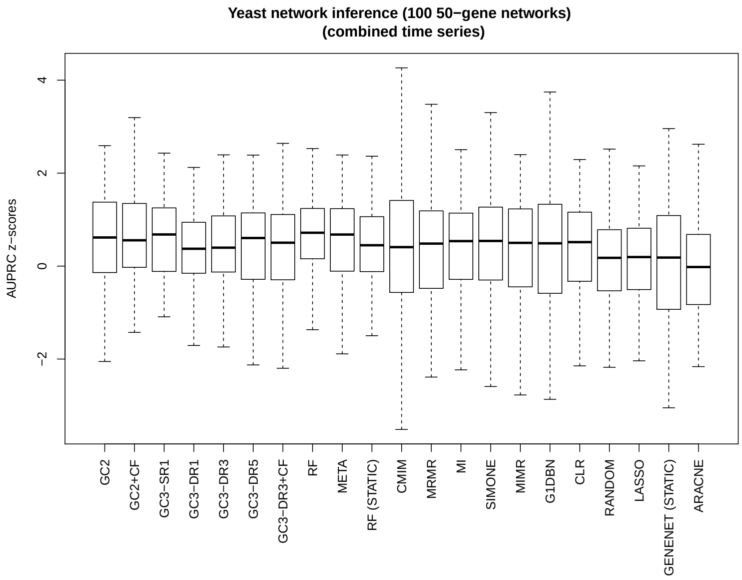


Figure 5.7: GRN inference performance (yeast time series) - Box plots of the AUPRC z-scores of the assessed methods. 100 GRN, of 50 genes, were inferred from the 11 yeast multivariate time series. Each network was inferred individually in all multivariate time series, and the obtained predictions combined with the Borda count method. The AUPRC of each inferred network was obtained and transformed into a z-score, obtained with Monte Carlo, 100000 simulations.

The results are presented in a series of figures. Figure 5.7 presents boxplots of the AUPRC z-scores in the yeast experiment, and Figure 5.8 presents the boxplots of the scores in the GNW experiment, for time series size of 20 and 300 time points. The results of the yeast experiment when assessing network inference on each multivariate time series individually is present on the appendix B. The appendix also shows the GNW experiment results for the intermediate number of considered time points (between 40 and 200).

The difference between different methods was assessed using the Wilcoxon signed rank test (see appendix A). The results are presented in the heatmap Figures 5.9 (yeast experiment), 5.10 and 5.11 (GNW experiment, for $n = 20$ and $n = 300$). In these matrix figures, different colors represent different obtained p-values.

In the boxplots, the first method (on the right) is bivariate GC (GC2), followed by the novel GC-based methods. The next presented methods are the state of the art approaches, ranked as in the significance heatmap figures (best performing methods on top).

Yeast results On the yeast experiment, inference precision is very low, and the only methods achieving a precision significantly higher than random (pv < 0.01) are GC2 (including or

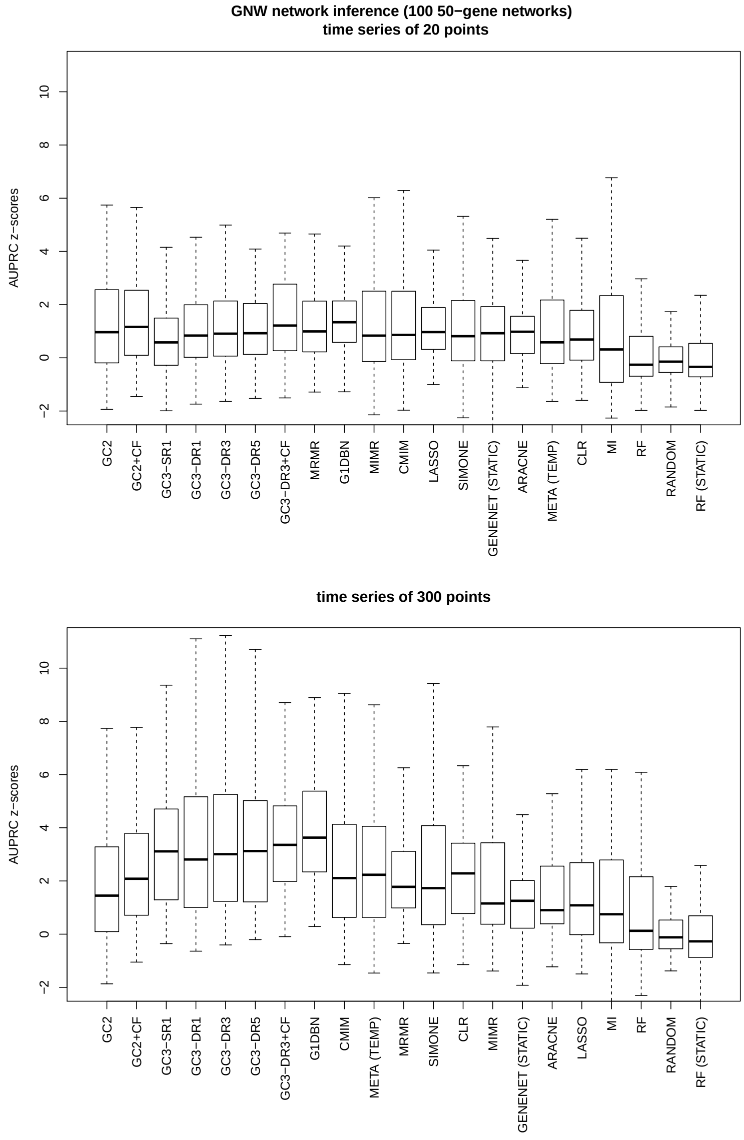


Figure 5.8: GRN inference performance (GNW time series) - Box plots of the AUPRC z-scores of the assessed methods. 100 GRN, of 50 genes, were inferred from GNW multivariate time series (one network corresponding to one multivariate time series). The AUPRC of each inferred network was obtained and transformed into a z-score, following the beta-distribution approximation proposed in Chapter 4. Results are shown using the 300 points of the time series, and the middle 20.

not the co-regulation filter), GC3-SR1 (with static ranking), and random forests (RF, dynamic implementation) and the ensemble of dynamic methods (META). The top ranked approach is the dynamic implementation of RF. The adoption of the co-regulation filter results in a small

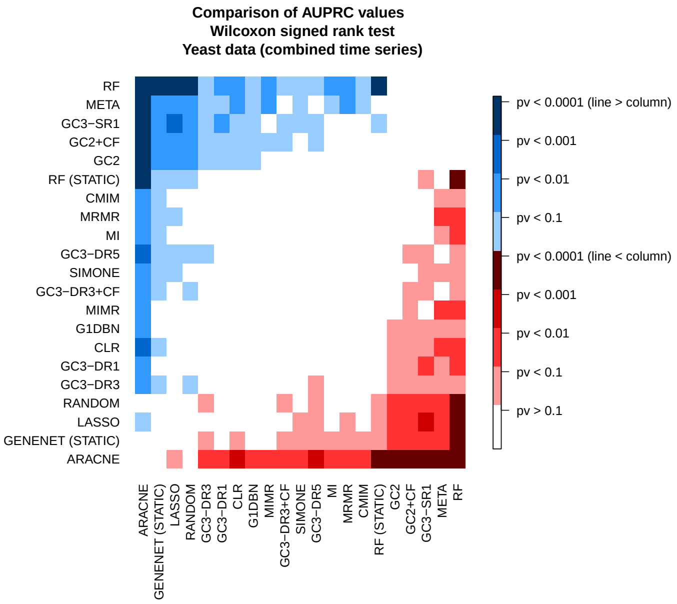


Figure 5.9: GRN inference statistical comparison (yeast experiment) - Comparison between the different GRN inference methods AUPRC z-scores (100 in total). The Wilcoxon signed rank test was used and the obtained (two-tailed) p-values are represented in the matrix. If the element $[i, j]$ is blue, then method i performs better than method j. Methods on top are the best performing. Results for the yeast experiment.

precision increase, although the difference is not significant. When networks are inferred from a single multivariate time series, no method is significantly more precise than random (appendix B).

GNW results On the GNW experiment, the precision of almost all methods increases with the size of the time series (the exception being the static implementation of RF, which is not significantly better than random even when $n = 300$). When the number of samples is low, several methods perform similarly. When $n = 20$ the top 8 ranked methods on Figure 5.10 do not exhibit a significant difference among them. However, even in this case almost all methods perform significantly better than random (exception being RF, static and dynamic).

When $n = 300$, the most precise methods are GC3-DR5, G1DBN and GC3-DR3 (with and without the co-regulation filter), not significantly different between them. The next most precise approaches are the remaining variations of GC3. These methods are significantly more precise than the remaining (pv < 0.0001).

The co-regulation filter has a beneficial impact in both GC2 and GC3-DR3, however only significant on the first case. GC3-DR5 is more precise than GC3-DR3, which in turn is more

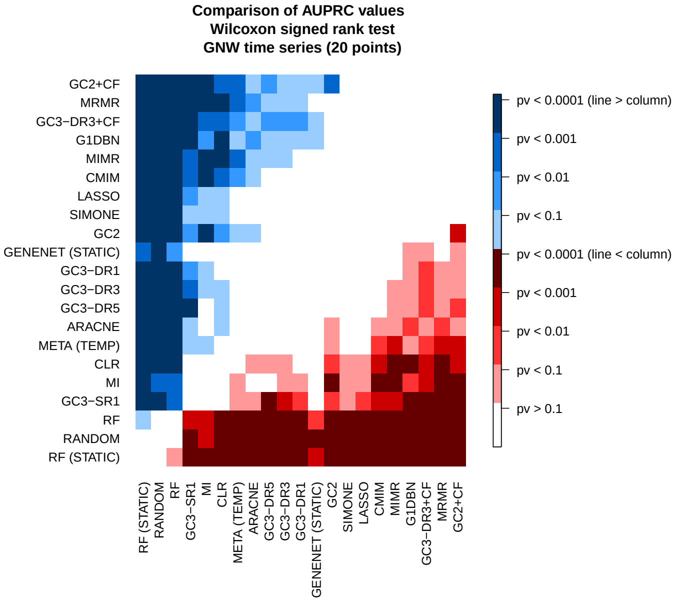


Figure 5.10: GRN inference statistical comparison (GNW, n=20) - Legend as of Figure 5.9. Results for the GNW time series, considering the middle 20 points of the time series.

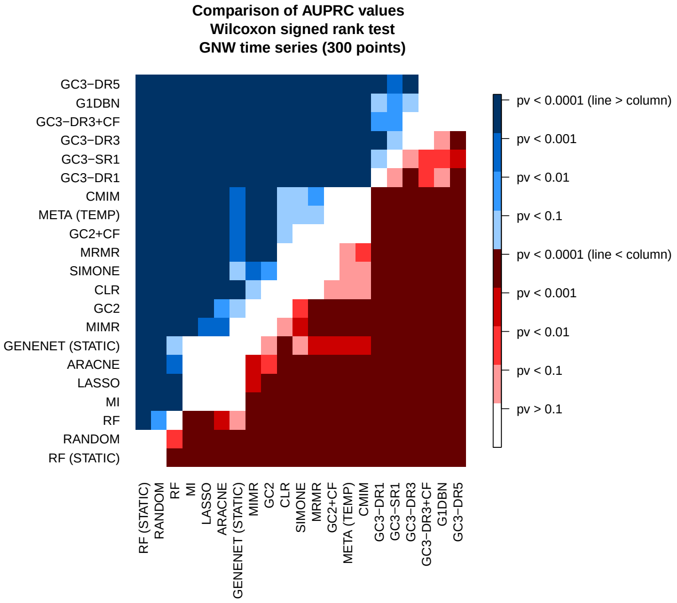


Figure 5.11: GRN inference statistical comparison (GNW, n=300) - Legend as of Figure 5.9. Results for the GNW time series, considering the 300 points of the time series.

precise than GC3-DR1. The effect size is small (Figure 5.8), but significant (pv < 0.01, Figure 5.11). GC3-SR1 performs similarly as GC3-DR1.

5.4 Discussion

5.4.1 GC lag selection

In the experiment of Section 5.3.2, the optimal approach to model GC on the yeast time series is using the first lag (Figure 5.5). The consideration of multiple and/or estimated lags results, and the adoption of the TY test, results in a performance decrease. This is explained by an increase in inference variance, more critical in the small sample case (the yeast time series are 25 and 35 points long, after interpolation).

In the GNW time series, the overall inference precision increases with n. When n becomes larger than around 60, the most accurate approaches to model GC are the ones considering multiple and/or estimated lags (considering only the top 10% ranked regulations). When n is lower, an higher precision is obtained with a single lag (lagged correlation, or GC with one lag). These is expected, as when n increases the extra variance of considering extra lags is mitigated. Regarding the GC-TY modification, its adoption only marginally improved the inference precision at the highest values of n. A more beneficial effect may be observed at values of n higher than the considered.

5.4.2 Co-regulation identification

The proposed method to identify co-regulation is based on the simple assumption that if two gene expression time series are co-regulated, respective changes in gene expression occur simultaneously. As shown in the Section 5.3.3, the proposed approach correctly identifies co-regulation with a very high precision, on both microarray and simulated time series. On the contrary, the spurious identification of regulation as co-regulation is under the range of random selection. These results validate the proposed co-regulation identification method. It is justified theoretically in the linear case, under the constraint identified in Section 5.2.3.2.

5.4.3 Network inference

In the yeast experiment, the obtained overall precision was very low, with few methods performing better than random. The results are even worse when considering inference on the individual 11 multivariate time series, as opposed to the adopted combination (these results are shown in the appendix B), where no method performed better than random. Apart from the short size of the time series, the low precision may be due to low quality of the microarray data and of the adopted gold standard, which is likely incomplete. Another potential problem emerges from the fact that inference was applied on relatively small networks of 50 genes, which do not contain all the common causes (regulators) of the respective genes. Due to this, indirect dependences via common causes may be a cause of type 1 errors. Finally, the expression of some genes may be relatively stable along the time series, resulting in low entropy behavior and small regulatory effect sizes.

On the contrary, the precision is higher on the GNW experiment, where almost all methods achieve a precision significantly better than random, even when $n = 20$. As opposed to the yeast networks, the GNW simulated networks contain all the regulators, and the gold standard is complete. As such, the inference of the gold standard is easier. The time series are also likely to be less noisy and more informative (the GNW time series reflect responses to gene expression perturbations). In the GNW experiment when $n = 20$ several methods achieve a similar precision. When $n > 40$ (appendix B) the most precise methods are the proposed first order conditional GC filter (GC3) and G1DBN, based on first order conditional dependence scores. These depend on three variables only and are protected against over-fitting, overcoming the high variable to sample ratio limitation. Our results suggest that in this case, first order conditional independence tests are an appropriate inference strategy. The limitation of this approach is that indirect regulations via multiple genes may be incorrectly inferred as direct regulations, as discussed in Section 5.2.1.

The adopted co-regulation filter has a beneficial impact in both GC2 and GC3 in the yeast and GNW experiments (significant in the GC2 and GNW case). The fact that the filter benefit is small is likely due to the the low number of pairs of genes estimated to be co-regulated (as seen in Section 5.3.3).

All assessed methods, except random forests, are based on the estimation of linear dependences. These may be (more or less trivially) extended to the non-linear case using mutual information. The adequacy of the linear assumption and how it compares to a non-parametric estimation of MI is out of the scope of this work.

GC3 GC3 consists in a heuristic to conduct first order conditional independence tests in the estimation of graphical models. It reduces the number of necessary tests to obtain a graph representing first order conditional independences, and was applied in a dynamic context using Granger causality. Its application to a static context is trivial. This heuristic consists on the ranking of individual conditioning variables, based on non-conditional dependence scores. A similar heuristic may be applied to higher order conditional independence tests, by ranking sets of conditioning variables of cardinality k according to $k - 1$ order conditional dependence scores. This strategy becomes then a fast approximation of the PC algorithm, useful in large networks.

Two strategies were suggested to compute the ranking of conditioning variables, for each cause-effect pair (static and dynamic, Section 5.2.2.1). The precision of the two approaches is similar in the GNW experiment, while the static is more precise in the yeast experiment. The search over the ranked conditioning variables depends on a parameter s. As it increases (maximum being $p - 2$), the number of conditional independence tests required to obtain the minimum conditional dependence score tends to $p - 2$ (the full search). As expected, the precision increases with s (GNW experiment). However, the difference is small (Figure 5.8), and the high precision obtained even when $s = 1$ suggest that both rankings are accurate and that the proposed heuristic is a good approximation of the search over all conditioning variables.

As discussed in Section 5.2.1, the complexity of inferring a network with GC3 using the full search (ie. setting $s = p - 2$) is $\mathcal{O}(13p^3)$ (assuming the complexity of linear regression as in Section 5.2.1), which can be considered too slow in the large p case. The proposed approximation decreases the complexity time to close to $\mathcal{O}(p^2)$. Under the 0.8 probability assumption of Section 5.1 (not considering the previous computation of the ranking of conditioning variables) we have $\mathcal{O}(2 * 13p^2)$ for $s = 1$ and $\mathcal{O}(6 * 13p^2)$ for $s = 3$. Under this assumption, using $s = 1$ reduces the computational time of GC3 by a factor of $p/2$. This is a substantial improvement in large networks (thousands of genes).

State of the art methods A curious result is the precision of RF in both experiments (it is the most precise method in the yeast experiment, and the least precise in the GNW experiment). Due to combining bagging with multiple random variable subspaces, RF are known to be protected against over-fitting and model variance (Section 2.6). On the other hand, regulation scores are a complicated non-linear function of all genes in the network. The different characteristics of the microarray and simulated time series may explain the difference in RF performance, but this point requires further investigation.

CMIM, MRNET, mIMR, and the Lasso (also implemented by Simone) are forward selection procedures, selecting predictors of a target variable incrementally, at each step considering the dependence towards the target and the previously selected predictors. While this favors the selection of non-redundant predictors, the selection and scoring of predictors is dependent on previous selections. For this reason, a forward selection strategies adequate for predictive purposes may not be optimal for causal inference.

The combination of methods resulted in a high precision, although lower than the most precise individual methods. In this second aspect we failed to replicate observations previously reported in the literature (Section 3.1).

Precision of GRN inference The AUPRC z-score used to assess inference precision is relative to its null distribution, which depends on the number of possible regulations and true regulations. The relation between the AUPRC and the z-score, as a function on the number of genes, and number of true regulations, is illustrated in the table 5.7, where the AUPRC corresponding to a z-score $= \{2, 4, 6, 8\}$ is presented, for number of genes and number of regulations both $= \{50, 100\}$. When estimating the AUPRC the number of positive instances P is the number of directed regulations, and the number of instances N is given by $N^2 - N$ (all possible directed regulations, except auto-regulatory). Although the mean of the distribution increases with $\frac{P}{N}$, its variance decreases with both P and N. This explains why a z-score of 8, when N (given by the number of genes) is fixed, is associated with a lower AUPRC when P=100 than when P=50.

Table 5.7: Relation between the AUPRC and z-score, for different network characteristics - the table presents the AUPRC corresponding to a z-score $= \{2, 4, 6, 8\}$, for the possible combinations of number of genes (50 or 100) and number of regulations (50 or 100). The μ and σ (standard deviation) of the null distribution are derived, and the z-score estimated, as in Section 4.

network characteristics				**AUPRC of z-score**			
# genes	**# regulations**	**null μ**	**null σ**	**z=8**	**z=6**	**z=4**	**z=2**
50	50	0.0234	0.00537	0.117	0.0831	0.0562	0.0361
50	100	0.0437	0.00544	0.115	0.0917	0.0720	0.0558
100	50	0.00593	0.00213	0.0561	0.0365	0.0216	0.0114
100	100	0.0110	0.00185	0.0394	0.0295	0.0214	0.0152

The number of edges in the GNW networks is between 51 and 118. An average z-score approaching 4 is obtained with GC3 when n=300. This corresponds to an AUPRC between 0.05 - 0.07 (for a number of edges between 50 and 100). These are low values of precision; however, the AUPRC measures the average precision over all values of recall. In network inference, precision at the lowest values of recall is what matters in practice (confirmatory experiments check the highest ranked predictions). The Figures 5.12 and 5.13 illustrate the obtained precision of GC3 (in particular, the setting GC3-DR3+CF as described in Section 5.3.4) in the GNW experiment when $n = 20$ and $n = 300$. The precision values in the 100 networks are represented in boxplots for different values of recall (until 0.5). The precision is considerably higher at the lowest levels of recall. When $n = 300$, an average precision close to 0.2 is obtained at the lowest recall.

5.5 Conclusion

We summarize the major conclusions:

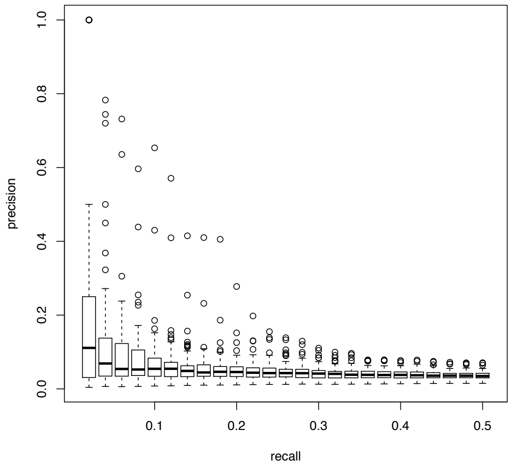


Figure 5.12: Boxplots of precision as a function of recall (GNW data, n=20) - Precision values of the 100 networks (GC3 with the co-regulation filter, in particular GC3-DR3+CF), as a function of recall.

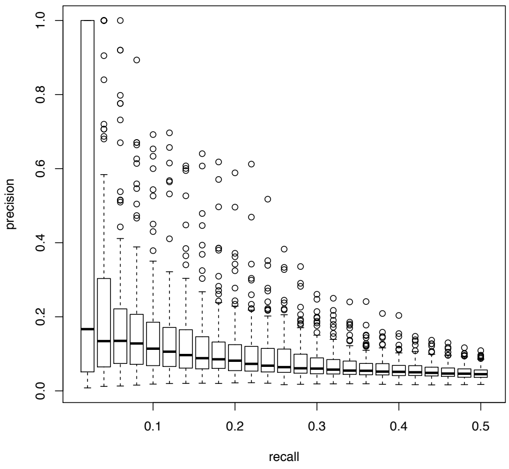


Figure 5.13: Boxplots of precision as a function of recall (GNW data, n=300) - Legend as Figure 5.12

- When modeling temporal dependences, it is important that the considered lags capture the lagged regulatory dependences. However, particularly when n is small (in the microarray experiment, and in the GNW experiment when $n < 40$), the consideration of high and multiple lags results in an increase in model variance. In this case, considering a single (first) lag is the most adequate strategy to infer gene regulation causality. As n increases, strategies to model GC that consider multiple, estimated lags, and non-stationarity, become more precise (as shown in the Figure 5.6).

- On the microarray experiment, network inference precision is low, and not guaranteed to be better than random. Apart from the high variable to sample ratio, another reason may be low data quality and incompleteness of the gold standard. In this low precision case, a precision significantly higher than random inference is not guaranteed. The approach to assess AUPRC significance (null of random selection) of Chapter 4 is relevant in such a context. In order to improve inference precision, longer time series should be used. Additionally, multivariate time series may be combined, together with data obtained from different experiments.

- An heuristic to obtain the minimum first order conditional GC dependence is proposed (GC3) as a faster alternative to the full search over all conditioning variables. An approach to identify co-regulation is also proposed, which is theoretically justified under certain conditions.

- In the GWN network inference experiment, the most precise inference method is GC3 and G1DBN, also based on first order partial correlations. This suggests that using first order conditional dependences are indeed a good strategy for GRN inference. The co-regulation filter was shown to identify co-regulation with a very high precision, and its incorporation in network inference methods increased their inference precision.

Data files and R scripts to replicate the described experiments are available online, as well as R software implementing the used dynamic inference methods (including GC3 and the co-regulation filter).[1][2]

[1] https://github.com/miguelaglopes/GCnetinf

[2] https://github.com/miguelaglopes/NetInfExps

6

Temporal profiling of cytokine-induced genes in pancreatic β-cells by meta-analysis and network inference

6.1 Introduction

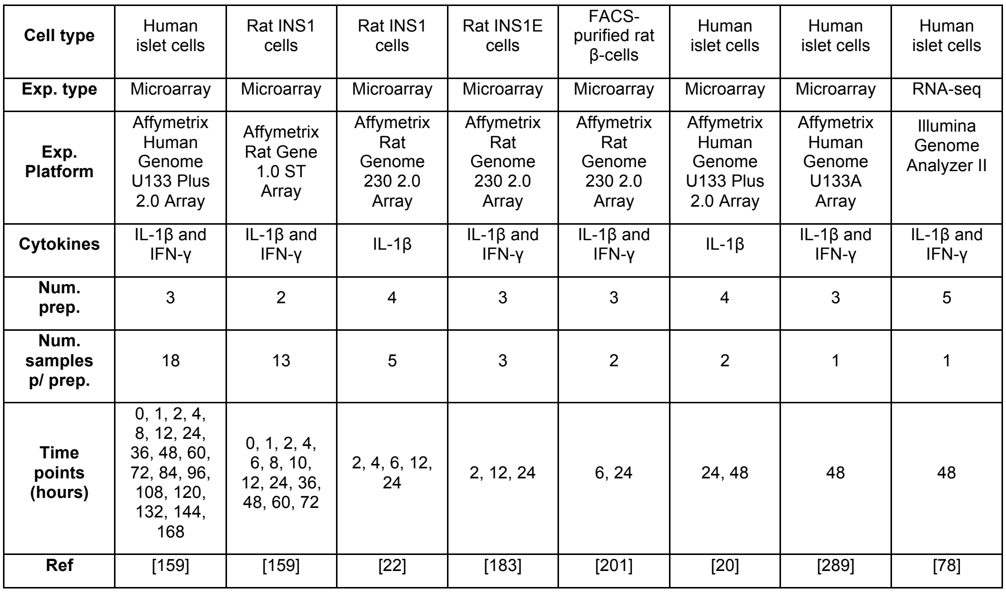
This chapter presents contributions regarding the inference of relevant genes and regulations in type 1 diabetes. Type 1 diabetes (T1D) is an autoimmune disease where local release of cytokines such as IL-1β and IFN-γ contribute to β-cell apoptosis. From 8 datasets of β-cell gene expression after exposure to IL-1β and IFN-γ, we identify a set of relevant genes to T1D. Two of these datasets were made available as a result of this work, and contain time-series expressions in human islet cells and rat INS-1E cells. Genes were ranked according to their differential expression within and after 24 hours from exposure, and characterized by function and prior knowledge in the literature. A regulatory network was then inferred from the human time expression datasets, using a temporal network inference method. The two most differentially expressed genes previously unknown in T1D literature (RIPK2 and ELF3) were found to modulate cytokine-induced apoptosis, and three predicted causal regulations were experimentally confirmed. The inferred regulatory network is thus supported by the experimental validation, providing a proof-of-concept for the proposed statistical inference approach.

Chapter outline Section 1.1 is an introduction to T1D; Section 6.2 describes the adopted methodology; Section 6.3 describes the methods and materials; Sections 6.4, 6.5 and 6.6 presents the results, discussion and conclusion, respectively.

6.2 Inference methodology

The meta-analysis consists of eight gene expression datasets from microarray and RNA-seq experiments measuring gene expression levels of human pancreatic islets, rat FACS-purified β-cells and rat INS1E cells (a cell line emulating many of the characteristics of pancreatic β-cells (242)) exposed to IL-1β or IL-1β and IFN-γ (20, 22, 77, 180, 198, 198) (see details in table 6.1).

Table 6.1: Datasets used in the meta-analysis study - Characterization of the different datasets used in the present study. Columns represent the gene expression datasets. Rows represent the different dataset characteristics. These are: cell type; type of experiment (microarray or RNA-seq); experiment platform; combination of cytokines used; the number of independent preparations used in the dataset (e.g., 3 human islets, indicating islets from three individuals); the total number of samples (time points) examined per individual preparation; the time points; the dataset reference.



Cell type	Human islet cells	Rat INS1 cells	Rat INS1 cells	Rat INS1E cells	FACS-purified rat β-cells	Human islet cells	Human islet cells	Human islet cells
Exp. type	Microarray	Microarray	Microarray	Microarray	Microarray	Microarray	Microarray	RNA-seq
Exp. Platform	Affymetrix Human Genome U133 Plus 2.0 Array	Affymetrix Rat Gene 1.0 ST Array	Affymetrix Rat Genome 230 2.0 Array	Affymetrix Rat Genome 230 2.0 Array	Affymetrix Rat Genome 230 2.0 Array	Affymetrix Human Genome U133 Plus 2.0 Array	Affymetrix Human Genome U133A Array	Illumina Genome Analyzer II
Cytokines	IL-1β and IFN-γ	IL-1β and IFN-γ	IL-1β	IL-1β and IFN-γ	IL-1β and IFN-γ	IL-1β	IL-1β and IFN-γ	IL-1β and IFN-γ
Num. prep.	3	2	4	3	3	4	3	5
Num. samples p/ prep.	18	13	5	3	2	2	1	1
Time points (hours)	0, 1, 2, 4, 8, 12, 24, 36, 48, 60, 72, 84, 96, 108, 120, 132, 144, 168	0, 1, 2, 4, 6, 8, 10, 12, 24, 36, 48, 60, 72	2, 4, 6, 12, 24	2, 12, 24	6, 24	24, 48	48	48
Ref	[159]	[159]	[22]	[183]	[201]	[20]	[289]	[78]

Two of these datasets, based on human pancreatic islets and rat insulin-producing INS-1E cells, are described in (156), and are available at GEO (reference numbers GSE53454 and GSE53453). The human islet dataset consists of a detailed time course analysis, ranging from 1 to 168 hours (7 days; this very late time point was included to model the long-term and protracted immune assault to which beta cells are exposed in diabetes) and is composed of 18 time points; to our knowledge, this is the most detailed time course array analysis performed in cytokine-treated human islets, up to this date.

The meta-analysis is performed to identify a group of genes whose expression levels are consistently and strongly modified by cytokines on both human and rat datasets, before or after 24 hours. This provides a unique indication of the dynamics of gene expression leading to apoptosis. An enrichment tool is used to assign biological terms to a majority of identified genes. A set of novel genes is identified, which were not previously reported in T1D literature. A network of regulations between a set of highly differentially expressed genes is then inferred from the human-islet dataset by using a temporal network inference algorithm. The predicted network is also used to order genes by time of regulation.

As a confirmatory procedure, two genes, ELF3 and RIPK2, found to be among the 10 most up-regulated both before and after 24 h, and the two most up-regulated of the novel genes, had their expression levels validated by RT-PCR analysis in INS1E cells exposed for 24 h to IL-1β and IFN-γ. Furthermore, these genes were knocked down for an assessment of their role (as predicted by the inferred network) on cytokine-induced β-cell apoptosis and expression of downstream genes. Both ELF3 and RIPK2 are found to have an impact on apoptosis and to influence the expression of genes predicted to be downstream. The experimental approach followed in the present study is summarized in the Figure 6.1.

6.3 Materials and methods

6.3.1 T1 Diabetes datasets

The datasets used in the meta-analysis (6.1) consist of cytokine-treated β-cells or clonal insulin-producing cell lines. These datasets result from experiments where the cells were treated with both IL-1β and IFN-γ, or with IL-1β alone (therefore the cytokine IL-1β is the common cytokine to all the datasets). Note that human islets are usually exposed for longer time points to cytokines than rat β-cells due to their increased resistance to cytokine-induced apoptosis (76).

6.3.2 Pre-processing of microarray data

The output of microarray experiments were stored in DAT files, processed by Affymetrix software to perform background subtraction and saved as .CEL files, containing intensity values for the probes in the chip. The pre-processing of the .CEL files of all the datasets was done using Custom Chip Definition (CDF) files, based on ENTREZ gene IDs, following the approach described in (61). These files map probe values into gene values for different microarray platforms (in our case, hgu133plus2 and hgu133a for human cell lines, and Ragene10stv1 and Rat2302 for rat/INS1 cell lines). The R package ”SimpleAffy”, taking as input a dataset CDF

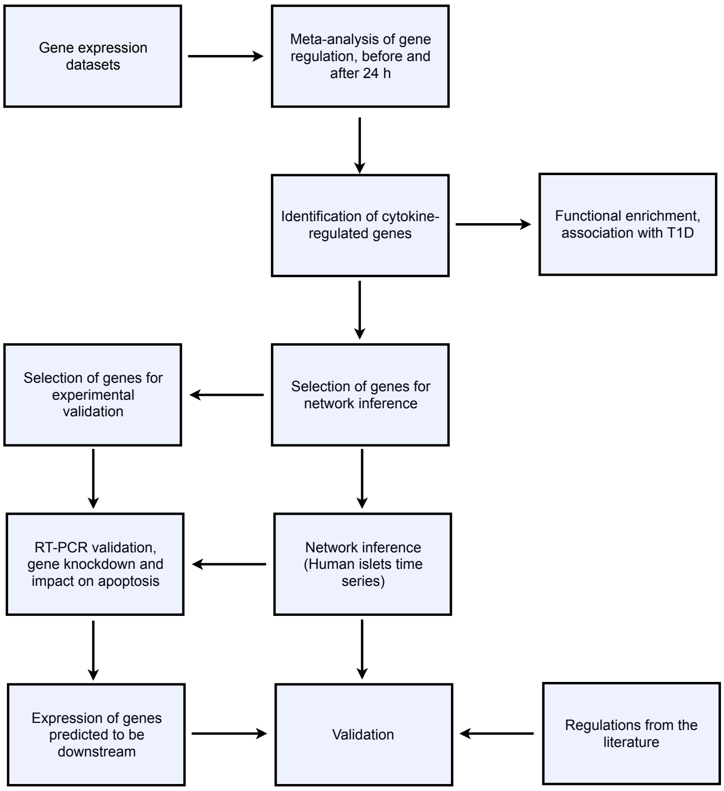

Gene expression datasets
Meta-analysis of gene regulation, before and after 24 h
Identification of cytokine-regulated genes
Functional enrichment, association with T1D
Selection of genes for experimental validation
Selection of genes for network inference
RT-PCR validation, gene knockdown and impact on apoptosis
Network inference (Human islets time series)
Expression of genes predicted to be downstream
Validation
Regulations from the literature

Figure 6.1: Overview of the adopted methodology - The methodology starts at the top of the diagram and ends at the bottom. Several gene expression datasets were used in a meta-analysis of cytokine-induced gene expression modulation, before and after 24h. Cytokine-modulated genes were identified and function allocated based on the datasets. A gene subset was selected to infer a regulatory network, using a newly presented time series dataset. Selected genes were subject to validation experiments based on gene knockdown and measurement of impact in apoptosis. The impact on genes predicted to be downstream, by the network inference, was also measured.

file, was used to normalize the respective CEL files and to compute the RMA expression levels. The quality of the individual arrays, after RMA normalization, was assessed using MA plots (R package arrayQualityMetrics). No array was found to be problematic, using the Hoeffding's independence test D statistic (default cut-off value of 0.15).

6.3.3 Selection of common genes

As the gene expression datasets relate to different species and come from different microarray platforms, preliminary correspondence mapping and filtering was implemented. Only the genes in human and rat experiments sharing common homolog were selected (two genes from different species are considered homologs if they descend from the same gene, on the last common ancestor of the species). The list of homolog genes (of rat and human) was retrieved in the NCBI Homologene database (94), enabling the mapping of genes from one species to another and to select a list of common genes to rat and human species. The final number of genes, common to all species and platforms, was reduced to 8148.

6.3.4 Meta-analysis of correlation coefficients

The adopted measure of differential expression is the Pearson sample correlation between gene expression and a dichotomous (binary) vector denoting cytokine exposure or control. This correlation is known as the point-biserial correlation coefficient. It was chosen as the measure of cytokine regulation effect size, as it is commonly used in meta-analysis, along with Cohens d (123, 176). For each gene, point-biserial correlations were estimated for all datasets. Meta-analysis on individual dataset correlation values was done by weighted average, and we defined the weights to be proportional to the number of samples in the respective dataset. This is a simple form of the random effects meta-analysis method proposed by Hunter and Schmidt (123), which is described and compared favorably with other methods in (82). The effect size, on a gene, is measured as:

$$r = \frac{\sum_i r_i}{\sum_i n_i} \tag{6.1}$$

where r_i is the correlation score of the gene in the dataset i, and n_i is the number of samples of the dataset. A returned score close to 1 indicates that the respective gene is strongly up-regulated with cytokines, where a score close to -1 indicates that the gene is strongly down-regulated. The experimental samples were partitioned into two groups of samples taken before or at 24 hours and of samples taken after or at 24 hours. For each gene, a meta-correlation score was calculated for the two temporal intervals. The threshold of 0.5 was used to identify large effects (54). Statistical significance was estimated by controlling the expected false discovery rate at 0.05, to account for multiple testing (following the Benjamini-Hochberg procedure (17) (see appendix A). Individual p-values were estimated as described in (82) (see Hunter and Schmidt method).

6.3.5 Network inference

Before network inference, the human islet dataset was preprocessed by i) averaging the three time series repetitions and ii) interpolating the resulting 18 time points vector to create a vector of 169 points, one for each hour from 0h to 168h. Network inference was performed using a lagged adaptation of the variable selection mRMR (Section 3.2.1). The adaptation consists in the use of a symmetric lagged mutual information matrix, in which each pairwise lag is the one that maximizes the lagged mutual information between the two time series (considering both directions). For each target gene Y, only the genes X for which $I(X_t; Y_{t-l}) < I(X_{t-l}; Y_t)$ are considered as predictors, l being the estimated lag:

$$l = \arg\max_i I(X_{t-l}; Y_t), l \in \{-l_{max}, .., -1, 0, 1, .., l_{max}\} \tag{6.2}$$

The mutual information was estimated with the linear correlation and transformed into MI using the Gaussian assumption (Section 2.1.3). The lags were estimated with a maximum lag of 5 time points (5 hours was assumed to be a reasonable upper bound of the lag of transcriptional regulations). In order to improve the decrease the variance of the inference, a re-sampling estimation based on leave-one-out was adopted (inspired by the jack-knife, appendix **??**). It consists in repeating the network inference algorithm n times (the number of experimental samples), each time leaving one sample (time point) aside. The resulting $n + 1$ (n plus the original, where no time point is removed) networks are combined by first selecting the $N = 800$ strongest edges in each network, and then by selecting the common edges on all resultant networks.

6.3.6 Temporal mapping of genes

An outcome of the adopted GRN inference method (described in Section 5) is the lag associated to each inferred regulation. Using these lags, the network genes were mapped into a temporal reference, where their position relative to the remaining genes indicates if they are regulated before or after them. The approach is the following: for each pair of genes in the predicted network, all the possible shortest undirected paths (in the predicted network) connecting them are calculated. Each of these paths is assigned a value, which is the sum of the lags of the respective edges (predicted interactions). Note that the lags can be positive or negative, depending on the causal direction. These values are averaged and the resulting value is assigned to the considered pair of genes. This value is an estimation of the time between the regulations of the two genes. By repeating the procedure for all different pairs of genes a square matrix of temporal distances between genes is obtained. In order to rank all the genes in time, each gene is assigned the average of the temporal distances between it and all the remaining genes. Note that the genes that are not part of any predicted interaction can not be included in this analysis.

6.3.7 Gene confirmation by RT-PCR and functional analysis by knockdown using specific siRNAs

Small interfering RNAs (siRNAs) were used to specifically silence ELF3 and RIPK2. Regarding the statistical analysis, data are expressed as mean SEM. Comparisons were performed by

ANOVA followed by Students t test with Bonferroni correction or two-tailed paired Students t test, as indicated. A significance p-value cutoff of 0.05 was adopted.

6.4 Results

To confirm the hypothesis that the receptors for the two cytokines (IL-1β and IFN-γ) are expressed in human pancreatic islets, the RNAseq dataset was used to measure the expression of IL-1R (19.6 RPKM), and of IFN-γR1 and IFN-γR2 (respectively 18.1 and 20.3 RPKM). For comparison, expression of the GLP-1R, a crucial receptor for pancreatic β-cell function was 9.8 RPKM (77).

6.4.1 Selection of differentially expressed genes before and after 24 hours

Table 6.2: Ranked down-regulated genes - Ranked down-regulated genes, differentially expressed and with a meta-score lower than -0.5, before 24 hours, after 24 hours, and both. For each gene, and for before and after 24 hours, a meta-score measuring cytokine modulation was calculated. This table shows the genes with a meta-score lower than -0.5 (meaning they are down-regulated), in the two temporal intervals , and ranked by absolute value (from most to least differentially expressed).

before 24 hours		before and after 24 hours			after 24 hours			
ISL1	-0.693	*EIF4EBP2*	-0.523	-0.512	*SLC3A1*	-0.756	*HPGD*	-0.563
ZNF395	-0.59				*HADH*	-0.683	*CTRB2*	-0.557
NR0B2	-0.572				*SPP1*	-0.66	*MAOB*	-0.555
NKX6-1	-0.572				*CPA2*	-0.65	*C11orf95*	-0.555
RUNX1T1	-0.568				*PDLIM1*	-0.65	*TIMP2*	-0.554
USP31	-0.539				*PCBD1*	-0.636	*EPB41L1*	-0.553
NCALD	-0.535				*NPTXR*	-0.627	*ACSM3*	-0.547
APPL2	-0.533				*GJB1*	-0.626	*POLE2*	-0.546
APEX1	-0.532				*ADRA2A*	-0.619	*PDGFC*	-0.536
ING2	-0.531				*SLC15A2*	-0.616	*ANKH*	-0.535
FAM171A1	-0.515				*RBP2*	-0.615	*CTGF*	-0.535
SSTR1	-0.505				*PEBP1*	-0.611	*VAV3*	-0.533
BLCAP	-0.502				*RAMP1*	-0.601	*CIRBP*	-0.533
					PFKM	-0.6	*COL1A2*	-0.53
					RUNDC3B	-0.592	*FGFR3*	-0.529
					ASB9	-0.585	*SPC25*	-0.525
					ALDH7A1	-0.583	*IGFBP2*	-0.523
					RAB40B	-0.58	*BBS1*	-0.522
					UPK1B	-0.572	*RIN2*	-0.514
					ENTPD3	-0.57	*CAV2*	-0.512
					CTNNBIP1	-0.57	*PCSK5*	-0.507
					CELSR2	-0.566	*SLC25A3*	-0.503

In order to identify genes that are most differentially expressed at early and late stages after cytokine exposure, the set of experimental samples (control and cytokine-exposed) was split into two groups: the first one contained samples measured within (and including) 24 hours from cytokine exposure while the second one contained samples measured after (and including) 24 hours. For each gene (8148 in total) and each group a meta-analysis correlation score, measuring

Table 6.3: Ranked up-regulated genes - Ranked up-regulated genes, differentially expressed and with a meta-score higher than -0.5, before 24 hours, after 24 hours, and both. For each gene, and for before and after 24 hours, a meta-score measuring cytokine modulation was calculated. This table shows the genes with a meta-score higher than 0.5 (meaning they are up-regulated), in the two temporal intervals , and ranked by absolute value (from most to least differentially expressed).

before 24 hours				before and after 24 hours						after 24 hours	
CCL20	0.797	*JUN*	0.572	*CXCL2*	0.892	0.82	*BCL2L14*	0.632	0.545	*WARS*	0.835
SDC4	0.789	*TJP2*	0.571	*IRF1*	0.879	0.87	*IFI35*	0.631	0.758	*BST2*	0.781
STX11	0.752	*RAPGEF5*	0.569	*NFKBIA*	0.864	0.82	*TAPBP*	0.63	0.788	*CD74*	0.78
CXCL1	0.751	*OGFR*	0.562	*ZC3H12A*	0.851	0.885	*STAT3*	0.628	0.557	*IFI30*	0.772
BMP2	0.736	*IL7*	0.561	*CXCL3*	0.843	0.859	*TAP2*	0.626	0.801	*C1S*	0.77
IER3	0.734	*FBXO7*	0.561	*ICAM1*	0.832	0.824	*ACSL5*	0.618	0.76	*IFITM1*	0.759
RND1	0.718	*LYN*	0.56	*RIPK2*	0.806	0.818	*EPHA2*	0.618	0.512	*FEZ1*	0.756
IL15	0.711	*NPPC*	0.559	*CXCL11*	0.803	0.803	*OPTN*	0.615	0.621	*ASS1*	0.745
IFNGR2	0.7	*CLIC1*	0.556	*ELF3*	0.801	0.568	*PTPN2*	0.614	0.519	*CXCL9*	0.74
SLC37A1	0.69	*RBM47*	0.553	*NFKB2*	0.799	0.678	*IRF7*	0.606	0.692	*HLA-DMB*	0.722
BIRC2	0.689	*BCL10*	0.552	*CSF1*	0.796	0.603	*LGMN*	0.605	0.738	*HLA-DRA*	0.72
GADD45B	0.668	*SPINT1*	0.55	*CXCL10*	0.789	0.805	*PARP12*	0.6	0.742	*C2*	0.71
ZFP36	0.664	*LRRC16A*	0.55	*TAP1*	0.778	0.856	*IFI44*	0.596	0.572	*ST5*	0.7
PPP1R15A	0.66	*EXT1*	0.549	*RHBDF2*	0.765	0.816	*C19orf66*	0.59	0.715	*PDZK1IP1*	0.696
RND3	0.658	*TP53*	0.548	*BIRC3*	0.765	0.757	*RELA*	0.586	0.596	*HLA-DRB1*	0.693
NFKBIB	0.651	*CD40*	0.546	*TNF*	0.742	0.681	*MAFF*	0.583	0.715	*BID*	0.685
SEMA4A	0.646	*TPBG*	0.546	*HLA-G*	0.74	0.741	*IRAK3*	0.583	0.649	*HLA-DQA1*	0.676
RFX5	0.639	*CCDC109B*	0.543	*TRAFD1*	0.738	0.763	*ISG20*	0.574	0.799	*PPAP2B*	0.669
CDKN1A	0.634	*LZTS1*	0.543	*SOD2*	0.737	0.803	*FAM82A2*	0.574	0.644	*IL1B*	0.659
RNF114	0.627	*CIDEC*	0.543	*MAP3K8*	0.737	0.666	*SERPINB9*	0.573	0.717	*RARRES1*	0.654
ABTB2	0.627	*PION*	0.538	*NFKB1*	0.73	0.646	*SAMD9*	0.573	0.509	*IFI27*	0.652
JUNB	0.625	*APAF1*	0.532	*TNIP1*	0.725	0.764	*CASP4*	0.569	0.777	*CFLAR*	0.648
F11R	0.624	*LSR*	0.53	*CD83*	0.721	0.529	*ISG15*	0.563	0.731	*MX1*	0.648
CLIC2	0.624	*ARID5A*	0.529	*UBD*	0.718	0.783	*OAS1*	0.56	0.743	*SLC11A2*	0.643
LGALS3BP	0.621	*SEC14L3*	0.524	*GBP2*	0.705	0.835	*TANK*	0.56	0.503	*IFIH1*	0.632
PARP8	0.618	*GJD2*	0.524	*CD69*	0.699	0.512	*LMO2*	0.559	0.549	*BCL2A1*	0.631
PSMA4	0.617	*IL1A*	0.523	*LTB*	0.697	0.718	*B2M*	0.548	0.654	*LAMB3*	0.613
IRF9	0.613	*ZC3H7A*	0.522	*DENND2D*	0.691	0.792	*DHX58*	0.544	0.64	*TNFAIP2*	0.606
SEC14L2	0.609	*CNP*	0.522	*RTP4*	0.688	0.711	*KARS*	0.535	0.523	*WTAP*	0.601
MAPK6	0.605	*NBN*	0.521	*PSMB9*	0.686	0.849	*ARAP1*	0.535	0.665	*OASL*	0.598
SLC25A28	0.6	*TBC1D22B*	0.518	*PSMB8*	0.682	0.822	*CD82*	0.525	0.539	*IFIT3*	0.596
TRIM26	0.59	*FRMD8*	0.517	*TMEM140*	0.681	0.763	*TRAF3*	0.524	0.594	*FGD6*	0.586
TRIM5	0.589	*FNDC3A*	0.516	*STAT2*	0.671	0.75	*HOPX*	0.524	0.622	*CSF2RB*	0.58
EHD1	0.588	*RASGRP3*	0.513	*TRIM25*	0.668	0.672	*BPGM*	0.516	0.566	*SP140L*	0.577
SGK1	0.588	*PRKCSH*	0.51	*UBE2L6*	0.661	0.828	*NMI*	0.512	0.636	*PPPDE2*	0.577
AGRN	0.587	*MAFK*	0.51	*PPP2R5B*	0.661	0.592	*C1R*	0.512	0.847	*KIT*	0.573
FASN	0.584	*EZR*	0.508	*HLA-E*	0.658	0.838	*ADM*	0.51	0.634	*GSDMD*	0.556
PVRL2	0.583	*BCL3*	0.508	*NAMPT*	0.651	0.722	*ERAP1*	0.509	0.593	*TNFRSF1A*	0.549
MVP	0.582	*PFKP*	0.507	*CX3CL1*	0.647	0.732				*ANXA1*	0.545
PDE9A	0.579	*PPP1R11*	0.506	*STAT1*	0.646	0.752				*SLCO5A1*	0.537
MYCBP2	0.579	*USP25*	0.505	*SP110*	0.642	0.752				*HPX*	0.527
TBK1	0.578	*LGALS13*	0.503	*TRAF3IP2*	0.641	0.73				*STX4*	0.521
IL10RB	0.578	*PAWR*	0.502	*JAK2*	0.64	0.642				*ADAR*	0.514
IFNGR1	0.574	*NELL1*	0.5	*FAS*	0.638	0.594				*DCN*	0.511
ATF3	0.573			*PSMB10*	0.633	0.839					

its differential expression across all datasets, and a significance measure, was calculated. This meta-analysis correlation was obtained from individual dataset correlation scores (see Section 6.3.4). 274 of the most differentially expressed genes, before or after 24 hours, were selected for further analysis. The selected genes satisfied the following three criteria: i) they had a meta-correlation score higher than 0.5 in absolute values; ii) the sign of the individual dataset correlations was the same (meaning that they are similarly regulated on all datasets) and iii) the meta-correlations were considered significant, admitting a false discovery rate of 0.05. These criteria were defined to make sure that the selected genes are significantly and strongly regulated by cytokines, and exhibit a common behavior on all datasets. The threshold of 0.5 was chosen as it is appropriate for the identification of large effect sizes (54). Tables 6.2 and 6.3

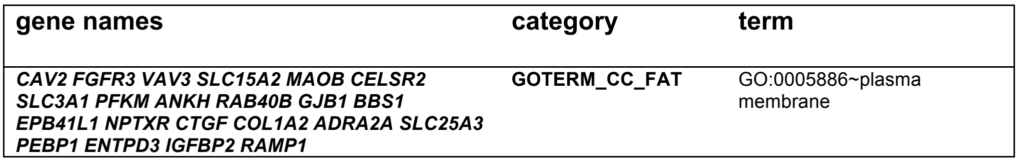
show the ranked lists, as well as the meta-correlation, of the 274 selected genes. Note that an up(down)-regulated gene corresponds to a positive (negative) correlation.

6.4.2 Gene characterization using enrichment analysis

In order to identify biologically meaningful terms associated with the selected genes, these were submitted to an enrichment analysis using the online software DAVID (120). Four lists of genes (down-regulated before 24h, down-regulated after 24h, up-regulated before 24h, up-regulated after 24h) were submitted to DAVID using the default options. Only terms returning a p-value lower than 0.001 were kept. The returned terms belong to different categories: biological processes present in the Gene Ontology (GO) annotation (GOTERM BP FAT); cellular processes, also from the GO annotation (GOTERM CP FAT); protein domains from the InterPro consortium (INTERPRO); molecular pathways from the KEGG pathway database (KEGG PATHWAY); and functional categories of proteins, present in the Swiss Prot and Protein Information Resource databases (SP PIR KEYWORD).

We associated each gene with a term, and in the case of multiple terms assigned to a gene the selected term was the one of the lowest p-value. Tables 6.4, 6.5 and 6.6 show the results for the groups of genes down-regulated after 24 hours or up-regulated before and after 24 hours. There were no terms enriched for the list of down-regulated genes before 24h. About two thirds of the differentially expressed genes (173 out of 274) were found to be enriched for terms using DAVID, before and after 24h, and are likely to play a role in the functions associated to them. As expected from previous studies (reviewed in (78)), there is a clear predominance of terms related to apoptosis and defense mechanisms.

Table 6.4: Enriched terms for genes down-regulated after 24 hours - Attributed terms to genes down-regulated after 24 hours. Enrichment analysis was performed with DAVID, with default parameters and p-value threshold < 0.001. Only genes associated with enriched terms are shown. Only one term is associated to each gene (the one with the lowest p-value).

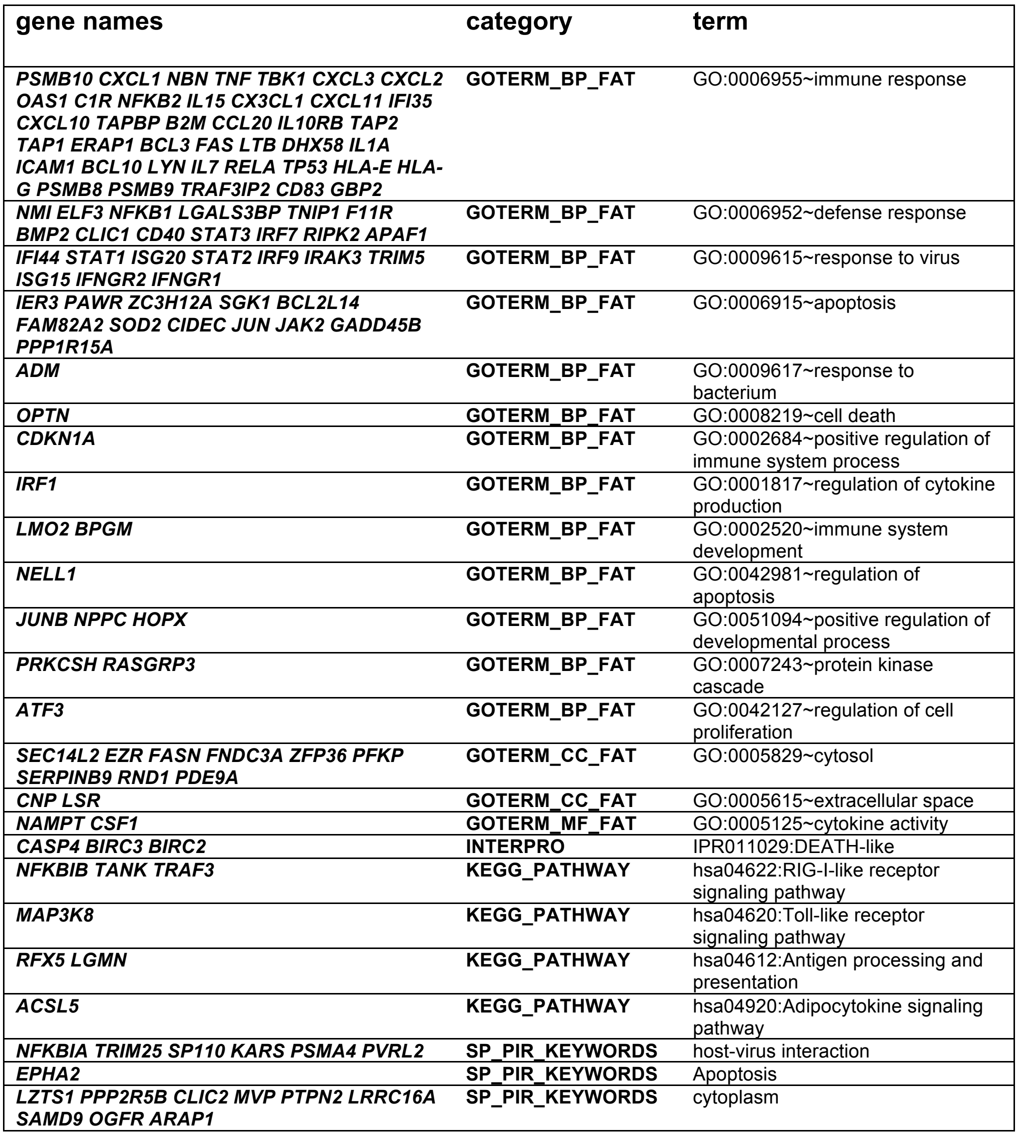


gene names	category	term
CAV2 FGFR3 VAV3 SLC15A2 MAOB CELSR2 SLC3A1 PFKM ANKH RAB40B GJB1 BBS1 EPB41L1 NPTXR CTGF COL1A2 ADRA2A SLC25A3 PEBP1 ENTPD3 IGFBP2 RAMP1	GOTERM_CC_FAT	GO:0005886~plasma membrane

6.4.3 Gene characterization using T1D literature

The 101 non-enriched genes (before or after 24h) were analyzied by text mining the T1D literature. The following sources were taken into consideration: NCBI GENE (44); GeneCards

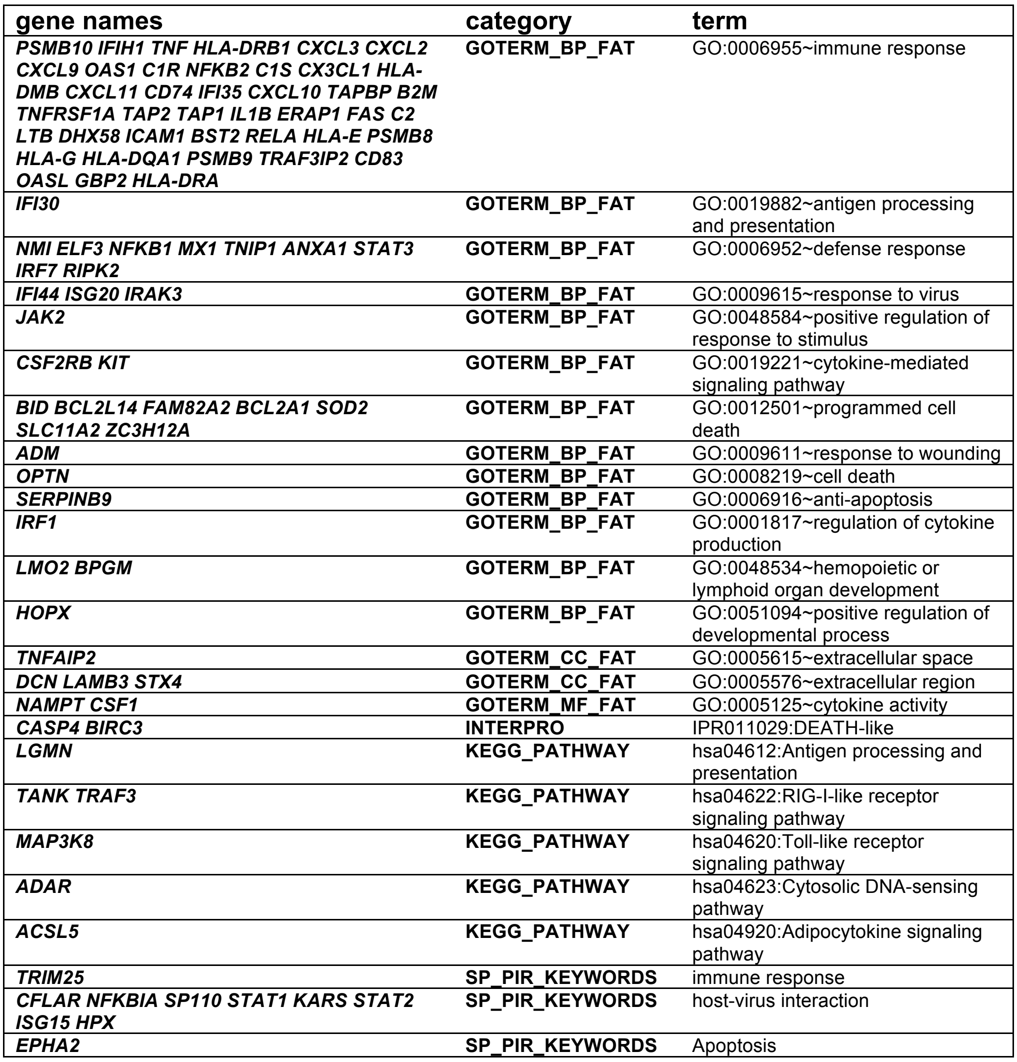
Table 6.5: Enriched terms for genes up-regulated before 24 hours - Attributed terms to genes up-regulated before 24 hours. Enrichment analysis was performed with DAVID, with default parameters and p-value threshold <0.001. Only genes associated with enriched terms are shown. Only one term is associated to each gene (the one with the lowest p-value).



gene names	category	term
PSMB10 CXCL1 NBN TNF TBK1 CXCL3 CXCL2 OAS1 C1R NFKB2 IL15 CX3CL1 CXCL11 IFI35 CXCL10 TAPBP B2M CCL20 IL10RB TAP2 TAP1 ERAP1 BCL3 FAS LTB DHX58 IL1A ICAM1 BCL10 LYN IL7 RELA TP53 HLA-E HLA-G PSMB8 PSMB9 TRAF3IP2 CD83 GBP2	**GOTERM_BP_FAT**	GO:0006955~immune response
NMI ELF3 NFKB1 LGALS3BP TNIP1 F11R BMP2 CLIC1 CD40 STAT3 IRF7 RIPK2 APAF1	**GOTERM_BP_FAT**	GO:0006952~defense response
IFI44 STAT1 ISG20 STAT2 IRF9 IRAK3 TRIM5 ISG15 IFNGR2 IFNGR1	**GOTERM_BP_FAT**	GO:0009615~response to virus
IER3 PAWR ZC3H12A SGK1 BCL2L14 FAM82A2 SOD2 CIDEC JUN JAK2 GADD45B PPP1R15A	**GOTERM_BP_FAT**	GO:0006915~apoptosis
ADM	**GOTERM_BP_FAT**	GO:0009617~response to bacterium
OPTN	**GOTERM_BP_FAT**	GO:0008219~cell death
CDKN1A	**GOTERM_BP_FAT**	GO:0002684~positive regulation of immune system process
IRF1	**GOTERM_BP_FAT**	GO:0001817~regulation of cytokine production
LMO2 BPGM	**GOTERM_BP_FAT**	GO:0002520~immune system development
NELL1	**GOTERM_BP_FAT**	GO:0042981~regulation of apoptosis
JUNB NPPC HOPX	**GOTERM_BP_FAT**	GO:0051094~positive regulation of developmental process
PRKCSH RASGRP3	**GOTERM_BP_FAT**	GO:0007243~protein kinase cascade
ATF3	**GOTERM_BP_FAT**	GO:0042127~regulation of cell proliferation
SEC14L2 EZR FASN FNDC3A ZFP36 PFKP SERPINB9 RND1 PDE9A	**GOTERM_CC_FAT**	GO:0005829~cytosol
CNP LSR	**GOTERM_CC_FAT**	GO:0005615~extracellular space
NAMPT CSF1	**GOTERM_MF_FAT**	GO:0005125~cytokine activity
CASP4 BIRC3 BIRC2	**INTERPRO**	IPR011029:DEATH-like
NFKBIB TANK TRAF3	**KEGG_PATHWAY**	hsa04622:RIG-I-like receptor signaling pathway
MAP3K8	**KEGG_PATHWAY**	hsa04620:Toll-like receptor signaling pathway
RFX5 LGMN	**KEGG_PATHWAY**	hsa04612:Antigen processing and presentation
ACSL5	**KEGG_PATHWAY**	hsa04920:Adipocytokine signaling pathway
NFKBIA TRIM25 SP110 KARS PSMA4 PVRL2	**SP_PIR_KEYWORDS**	host-virus interaction
EPHA2	**SP_PIR_KEYWORDS**	Apoptosis
LZTS1 PPP2R5B CLIC2 MVP PTPN2 LRRC16A SAMD9 OGFR ARAP1	**SP_PIR_KEYWORDS**	cytoplasm

(225) ; eGIFT ((263), using the keyword diabetes); T1Dbase ((42) searching for genes associated with T1D relevant publications); and the Beta Cell Gene Bank (42) (any referred gene). 45 (out of 101) genes were found to be associated with T1D terms (see table 6.7). We considered the remaining genes as unknown yet potentially relevant for the pathways leading to T1D β-cell

Table 6.6: Enriched terms for genes up-regulated after 24 hours - Attributed terms to genes up-regulated after 24 hours. Enrichment analysis was performed with DAVID, with default parameters and p-value threshold <0.001. Only genes associated with enriched terms are shown. Only one term is associated to each gene (the one with the lowest p-value).



gene names	category	term
PSMB10 IFIH1 TNF HLA-DRB1 CXCL3 CXCL2 CXCL9 OAS1 C1R NFKB2 C1S CX3CL1 HLA-DMB CXCL11 CD74 IFI35 CXCL10 TAPBP B2M TNFRSF1A TAP2 TAP1 IL1B ERAP1 FAS C2 LTB DHX58 ICAM1 BST2 RELA HLA-E PSMB8 HLA-G HLA-DQA1 PSMB9 TRAF3IP2 CD83 OASL GBP2 HLA-DRA	GOTERM_BP_FAT	GO:0006955~immune response
IFI30	GOTERM_BP_FAT	GO:0019882~antigen processing and presentation
NMI ELF3 NFKB1 MX1 TNIP1 ANXA1 STAT3 IRF7 RIPK2	GOTERM_BP_FAT	GO:0006952~defense response
IFI44 ISG20 IRAK3	GOTERM_BP_FAT	GO:0009615~response to virus
JAK2	GOTERM_BP_FAT	GO:0048584~positive regulation of response to stimulus
CSF2RB KIT	GOTERM_BP_FAT	GO:0019221~cytokine-mediated signaling pathway
BID BCL2L14 FAM82A2 BCL2A1 SOD2 SLC11A2 ZC3H12A	GOTERM_BP_FAT	GO:0012501~programmed cell death
ADM	GOTERM_BP_FAT	GO:0009611~response to wounding
OPTN	GOTERM_BP_FAT	GO:0008219~cell death
SERPINB9	GOTERM_BP_FAT	GO:0006916~anti-apoptosis
IRF1	GOTERM_BP_FAT	GO:0001817~regulation of cytokine production
LMO2 BPGM	GOTERM_BP_FAT	GO:0048534~hemopoietic or lymphoid organ development
HOPX	GOTERM_BP_FAT	GO:0051094~positive regulation of developmental process
TNFAIP2	GOTERM_CC_FAT	GO:0005615~extracellular space
DCN LAMB3 STX4	GOTERM_CC_FAT	GO:0005576~extracellular region
NAMPT CSF1	GOTERM_MF_FAT	GO:0005125~cytokine activity
CASP4 BIRC3	INTERPRO	IPR011029:DEATH-like
LGMN	KEGG_PATHWAY	hsa04612:Antigen processing and presentation
TANK TRAF3	KEGG_PATHWAY	hsa04622:RIG-I-like receptor signaling pathway
MAP3K8	KEGG_PATHWAY	hsa04620:Toll-like receptor signaling pathway
ADAR	KEGG_PATHWAY	hsa04623:Cytosolic DNA-sensing pathway
ACSL5	KEGG_PATHWAY	hsa04920:Adipocytokine signaling pathway
TRIM25	SP_PIR_KEYWORDS	immune response
CFLAR NFKBIA SP110 STAT1 KARS STAT2 ISG15 HPX	SP_PIR_KEYWORDS	host-virus interaction
EPHA2	SP_PIR_KEYWORDS	Apoptosis

apoptosis.

6.4.4 Network inference using the Human Islets time series

The next step of the analysis is the inference - from the human-islet time expression dataset - of a network of regulatory interactions between the 84 most differentially expressed genes. These selected genes correspond to a subset of the previously selected 274 genes, namely the ones that

Table 6.7: Genes not associated with a biological term, associated or not with T1D in the literature - Differentially expressed genes, before 24 hours, after 24 hours, or both, not associated with an enriched term. Enrichment analysis was performed with DAVID, with default parameters and p-value threshold < 0.001. These genes were grouped according to being found, or not, associated with T1 Diabetes in various available sources (NCBI GENE, GeneCards, eGIFT, T1DBase, The Beta Cell Gene Bank).

	Genes associated with T1D	**Genes not associated with T1D**
down regulated, before 24h	*APEX1 ISL1 NKX6-1 EIF4EBP2 NCALD NR0B2 SSTR1*	*APPL2 BLCAP FAM171A1 ING2 RUNX1T1 USP31 ZNF395*
down regulated, after 24h	*PDLIM1 RUNDC3B TIMP2 ACSM3 ALDH7A1 CPA2 CTNNBIP1 CTRB2 EIF4EBP2 HADH HPGD PDGFC PCSK5 PCBD1 SPP1*	*RIN2 SPC25 ASB9 CIRBP C11ORF95 POLE2 RBP2 UPK1B*
up regulated, before 24h	*CD69 RAPGEF5 TRAFD1 AGRN GJD2 MAPK6 PARP12 PARP8 PPP1R11 RHBDF2 SDC4 TJP2 TRIM26 UBD MAFK*	*ARID5A CD82 DENND2D EHD1 FBXO7 FRMD8 MYCBP2 RBM47 RND3 SEC14L3 TBC1D22B ABTB2 C19ORF66 CCDC109B EXT1 LGALS13 PION RTP4 RNF114 SEMA4A SPINT1 SLC25A28 SLC37A1 STX11 TMEM140 TPBG USP25 UBE2L6 MAFF ZC3H7A*
up regulated, after 24h	*ARAP1 CD69 PDZK1IP1 PPPDE2 TRAFD1 WTAP ASS1 IFITM1 PPAP2B PARP12 PTPN2 RHBDF2 WARS UBD*	*CD82 DENND2D FGD6 SP140L C19ORF66 FEZ1 GSDMD IFI27 IFIT3 PPP2R5B RTP4 RARRES1 SLCO5A1 SAMD9 ST5 TMEM140 UBE2L6 MAFF*

were selected at both before and after 24 hours, corresponding to the central columns of tables 6.2 and 6.3. The used algorithm is a temporal adaptation of the mRMR variable selection method (as described in Section 6.3.5). The final outcome of the algorithm is a set of 213 inferred directed regulations illustrated in Figure 6.2, where nodes represent genes and arrows represent regulations (only genes that were predicted to take part in an regulation are represented). All the represented genes are up-regulated genes. An in silico validation of the inferred network was done using the TRANSFAC (172) and the GeneCards databases (225). From these databases a list of 85 presumptive gene regulations was gathered (between the 84 genes of the inferred network). Nine of these annotated regulations are present in the predicted network (represented in red in the Figure 6.2). Taking the TRANSFAC/ GeneCards regulations as reference, the precision of the inferred network is 0.0422 (9 true positives out of 213 positives), a low yet significant value. The expected precision of a random selection of 213 regulations is 0.0122 and the probability of randomly attaining a precision as high as 0.0422 amounts to 0.001 (p-value returned by the hypergeometric distribution).

In order to visualize the gene dynamics captured by the network inference algorithm, the gene expression levels of gene pairs taking part in inferred regulations (Figure 6.3) was plotted. Four putative regulator-target pairs implying unknown genes were considered. These four interactions were subject to a posterior experimental validation (see next section). The

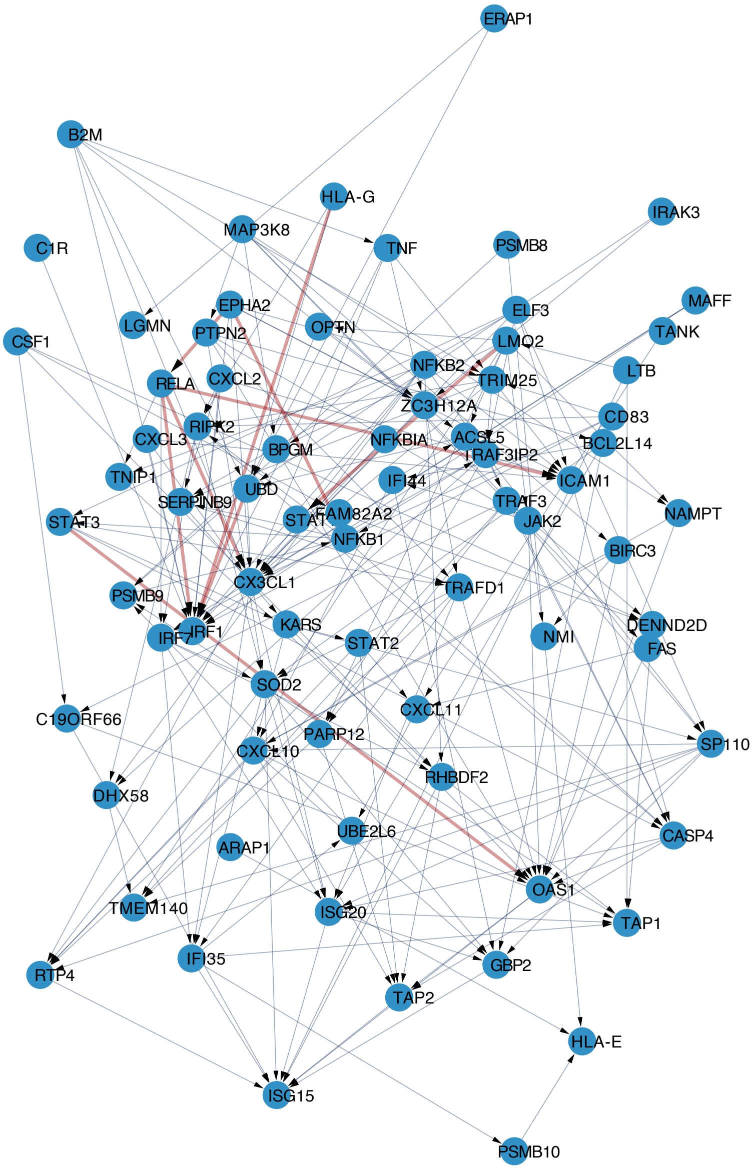


Figure 6.2: Predicted regulations with the adopted network inference algorithm - The nodes represent genes and the directed edges represent gene regulations (the arrow direction indicates the causal direction). The genes present in the represented network were consistently modulated by cytokines, before and after 24 hours. Edges in red represent regulations that were reported in the literature.

expression time series of the regulator was plotted in red while the one of the target gene in blue. Figures a and b show the behavior of ELF3 and two predicted downstream genes: CX3CL1 and SP110, while Figures c and d show the behavior of RIPK2 and two predicted downstream

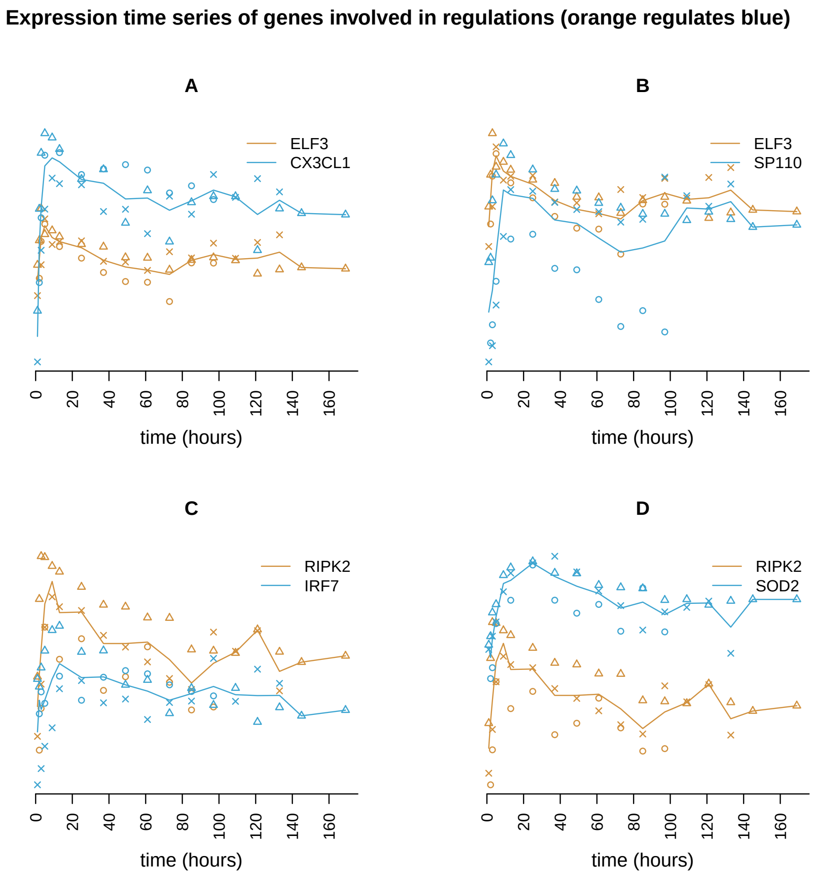


Figure 6.3: Expression time series of genes involved in regulations - The temporal characterization of genes taking part in interactions present in the inferred regulatory network is shown. Gene expression is plotted for different time points along a range of 168 hours. A linearly interpolated average is also shown. The human islet time series dataset is composed of three sequences of gene expression observations, of different temporal ranges, and therefore some time points are represented by less than 3 observations. Y-axis represents RMA expression levels. Figures a) and b) show ELF3 and the downstream-predicted CX3CL1 and SP110. Figures c) and d) show RIPK2 and the downstream-predicted IRF7 and SOD2. ELF3 and RIPK2 are shown in red. In each figure, the average expression of the genes shown in orange and blue were optimally correlated with a lag (the gene in blue following the gene in blue).

genes: IRF7 and SOD2. Of note, the regulation patterns of the gene pairs in Figures a and b have similar shapes. The gene pairs in Figures c and d have a somewhat less correlated behavior, which can be due to the existence of multiple regulators of IRF7 and SOD2.

The inferred network (and inferred regulation lags) was used to order genes by regulation

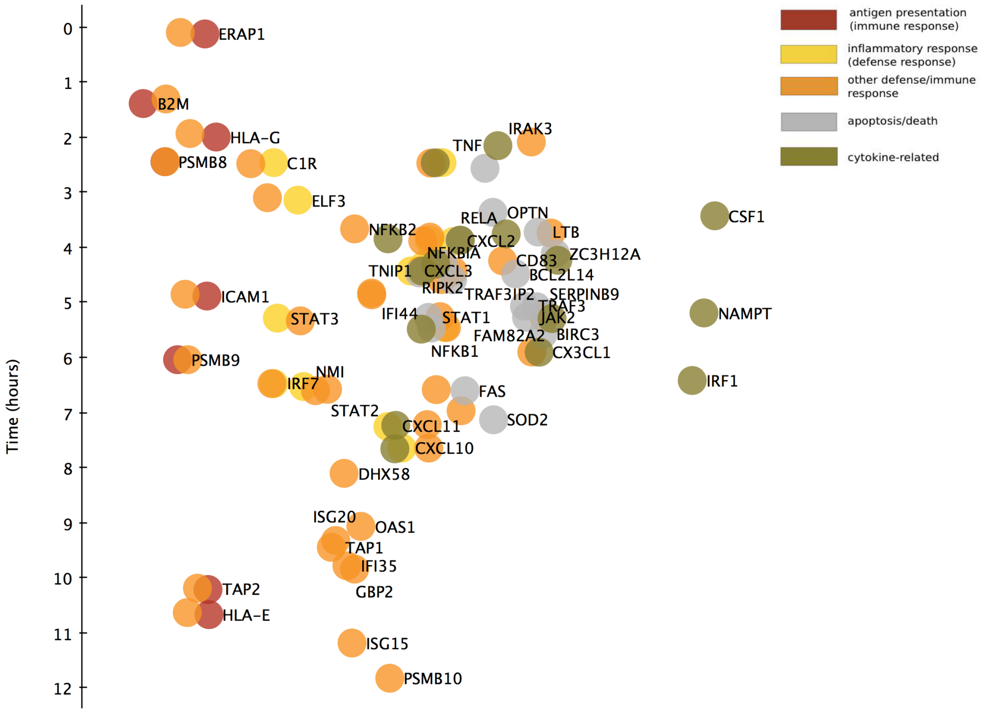


Figure 6.4: Temporal mapping of genes - Genes were mapped into a common temporal reference and associated with functional terms. The network inference algorithm returned a lag characterizing each gene interaction. From these lags a time of regulation of each gene was estimated. The figure shows the genes in the network ordered by time of regulation (from top to bottom). The figure also indicates the functional role that was attributed to each gene (circle colors). A gene is represented by more than one circle if it is associated with multiple roles.

time. The genes of the network were mapped into a temporal reference using the approach of Section 6.3.6. Figures 6.2 and 6.4 show the resulting mapping where the time arrow goes top down (i.e. lower genes are regulated later than upper genes). In order to map functions to such temporal framework (Figure 6.4), the GO terms returned by DAVID were used. Clusters of similar terms, associated with the majority of genes, were identified: antigen presentation, inflammatory response, other immune/defense response terms, apoptosis/death, cytokine-related. Note that there is some overlap in these terms, and that a majority of genes are associated with multiple terms. The first three clusters are rather similar: the only difference is that the first group is associated with the adaptive immune response, while the second group is associated with the innate immune response. The data shown in Figure 6.4 suggests a biologically meaningful sequence of molecular mechanisms. Thus, the first group of regulated genes is associated with defense/immune response; this is followed by modulation of genes related to cytokine activity and apoptosis, and then another group of defense/immune response genes is activated.

6.4.5 Biological validation of the network inference results

The genes RIPK2 and ELF3 were selected for biological validation since they were among the 10 most differentially expressed mRNAs, before and after 24 hours of cytokine treatment (see table 6.3), and the 2 most differentially expressed with an unknown role in basal and cytokine-induced apoptosis. In both INS-1E cells and FACS-purified primary rat β-cells cytokines up-regulated the expression of ELF3 (Figure 6.5) and RIPK2 (Figure 6.6). Small interfering RNA (siRNA) was used to knock down (KD) ELF3 and RIPK2, and to observe their role in cytokine-induced apoptosis and regulation of gene expression. ELF3 KD increased β-cell apoptosis both under basal condition and following cytokine exposure (Figure 6.5 D), while KD of RIPK2 increased only cytokine-induced apoptosis, Figure 6.6 D).

In the inferred network (Figure 6.2) ELF3 was predicted to regulate CX3CL1, SP110, TRAF3IP2, ICAM1, SERPINB9 and STAT3, while RIPK2 was expected to modulate IRF7, SOD2, DHX58 and CASP4. Two predicted targets for both ELF3 and RIPK2 were selected for an experimental validation of the inferred regulations. KD of ELF3 decreased expression of two predicted downstream targets, namely the chemokine CX3Cl1 and the nuclear body protein SP110 (Figure 6.5(E and F). Similarly, KD of RIPK2 decreased expression of the predicted target IRF7 (a key regulator of chemokine expression in pancreatic β-cells (198)), and induced a trend for lower expression of the free radical scavenger enzyme SOD2 (Figure 6.6 (E and F).

6.5 Discussion

This section described a meta-analysis, based on 8 different datasets, of gene expression of pancreatic β-cells (rat and human) after exposure to the pro-inflammatory cytokines IL-1β and IFN-γ. The comparison between the two different species, and the use of a large number of independent experiments (table 6.1 and references herein), based on both array analysis and RNAseq, provided a depth of information that is unique in the field. Two of the used datasets were made available in the course of this work (156). They are gene expression time series of human islets and rat clonal β-cells after cytokine exposure, and represent a valuable resource due to their comparatively high number of different time points studied.

Genes were ranked by magnitude of cytokine regulation at time points before and after 24 hours. 24 hours was selected as the dividing time point because it is around this time point that a progressive increase in cytokine-induced apoptosis is observed (204, 286). A list of genes that are strongly regulated by cytokines, on both human and rat experiments, and before or after 24

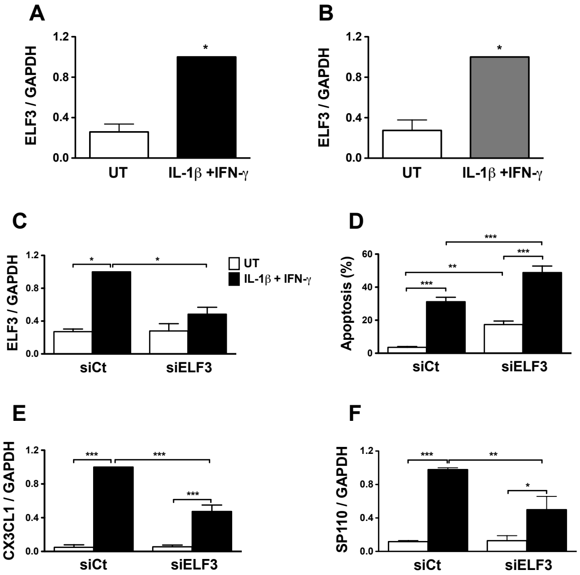


Figure 6.5: Confirmation of predicted changes in the expression of ELF3 and downstream targets - INS-1E cells (A) and FACS-purified primary rat β-cells (B) were treated for 24 h with IL-1β+IFN-γ (respectively 10 + 100 U/ml for INS-1E and 50 + 500 U/ml for primary β-cells) or left untreated (UT). Cells were collected for evaluation of ELF3 mRNA expression by qRT-PCR. The values obtained were corrected by the housekeeping gene GAPDH. Results are the mean ± SEM of 3 independent experiments. * p ¡ 0.05, paired t-test. (C-F) INS-1E cells were transfected with a siRNA against ELF3 or with negative control siRNA (siCt). 48 h after transfection cells were treated with IL-1β+IFN-γ (respectively 10 and 100 U/ml) or left untreated (UT) for 24 h. (C) Cells were collected and mRNA was evaluated for ELF3 expression. The values obtained were corrected by the housekeeping gene GAPDH. (D) Cell death was evaluated by Hoechst 33342/propidium iodide staining. Cells were harvested, the mRNA was collected and assayed for CX3CL1 (E) and SP110 (F) by qRT-PCR. The values obtained were corrected by the housekeeping gene GAPDH. Results are the mean ± SEM of 5-8 independent experiments. * p ¡ 0.05, ** p ¡ 0.01 and *** p ¡ 0.001 as indicated, ANOVA.

hours, is presented. A majority of these genes are characterized with a main functional term, related to apoptosis and cell inflammation, but the role of a relevant fraction remains unclear. Further investigations on these genes without known function in β-cells could provide valuable novel knowledge regarding the mechanisms of cytokine-induced apoptosis in T1D.

A group of key genes was selected as the nodes of a regulatory network. This network was inferred using a dynamic adaptation of a state of the art filter variable selection method (mRMR), using the human islet gene expression time series. The inference of a network from a limited amount of observed samples poses a difficult challenge and its validation is

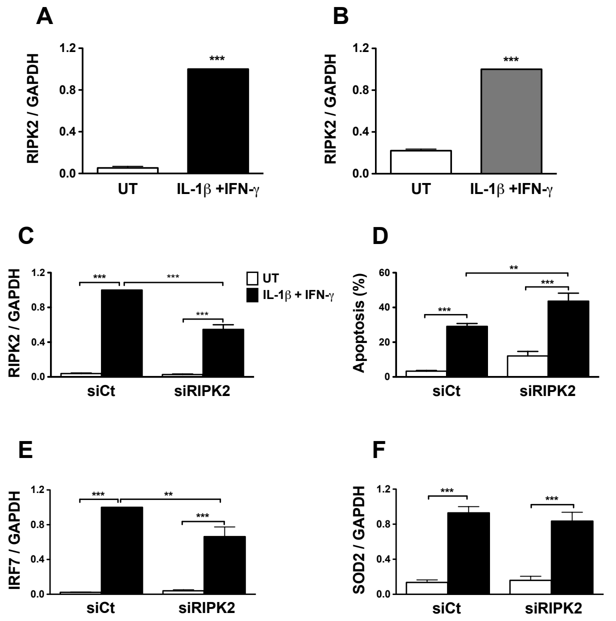


Figure 6.6: Confirmation of predicted changes in the expression of RIPK2 and downstream targets - INS-1E cells (A) and FACS-purified rat primary β-cells (B) were treated for 24 h with IL-1β+IFN-γ (respectively 10 + 100 U/ml for INS-1E and 50 + 500 U/ml for primary β-cells) or left untreated (UT). Cells were collected for evaluation of RIPK2 mRNA expression by qRT-PCR. The values obtained were corrected by the housekeeping gene GAPDH. Results are the mean ± SEM of 3 independent experiments. *** p ¡ 0.001, paired t-test. (C-F) INS-1E cells were transfected with siRNA against RIPK2 or with negative control siRNA (siCt). 48 h after transfection cells were treated with IL-1β+IFN-γ (respectively 10 or 100 U/ml) or left untreated (UT) for 24 h. (C) Cells were collected and mRNA was evaluated for RIPK2 expression. The values obtained were corrected by the housekeeping gene GAPDH. (D) Cell death was evaluated by Hoechst 33342/propidium iodide staining. Cells were harvested, the mRNA was collected and assayed for IRF7 (E) and SOD2 (F) by qRT-PCR. The values obtained were corrected by the housekeeping gene GAPDH. Results are the mean ± SEM of 4-6 independent experiments. ** p ¡ 0.01 and *** p ¡ 0.001 as indicated, ANOVA.

often inaccurate because of the large number of false negatives in annotated databases. This limitation in validation can only be overcome with more comprehensive knowledge regarding gene interactions. The accuracy of network inference algorithms tends to be low, due to typical low number of samples of gene expression datasets (as observed in the Chapter 5). In spite of this limitation, the inferred network achieved a precision significantly higher than random selection.

Two of the most differentially expressed genes before and after 24 hours, ELF3 and RIPK2, were subject to an experimental validation. Their knockdown led to an increase in β-cell apoptosis. Predicted interactions involving ELF3 and RIPK and 4 downstream genes were also submitted to experimental validation.

The first gene analyzed, ELF3, is a transcription factor that is mainly expressed in epithelial cells (192). It has been previously shown that ELF3 is implicated in the regulation of the morphogenesis and differentiation of mature endothelial cells in the small intestine (189) and in the inflammation of airway epithelia (143). In human bronchial airway epithelial cell lines, IL-1β + TNF-α treatment up-regulates ELF3 expression via NF-kB transcriptional activation (282). In line with these findings we observed an up-regulation of ELF3 in both INS-1E cells and in rat primary β-cells after IL-1β + IFN-γ treatment. Using a specific siRNA against ELF3 we found that ELF3 KD significantly increases β-cell apoptosis. The analysis of two putative downstream target genes (CX3CL1 and SP110) confirmed that their expression is inhibited after ELF3 KD. CX3CL1 is a chemokine that promotes leukocyte migration, cytotoxicity of NK cells and cytotoxic T lymphocytes and is induced after cytokine treatment in β-cells (61, 204), while SP110 is a nuclear body-associated protein that is induced by IFN-γ treatment in other cell types (218).

The second gene analyzed is RIPK2, a serine / threonine kinase that activates the NF-kB pathway (173) and it is up-regulated by NF-kB after cytokine treatment in endothelial cells (285). Recent findings suggest that RIPK2 is implicated in the regulation if the alternative NF-kB pathway and its KD increases cell apoptosis (44). It was confirmed in β-cells the up-regulation of RIPK2 after IL-1β + IFN-γ treatment and observed an increase in cytokine-induced β-cell apoptosis after the KD of the gene. RIPK2 KD had a significant effect on one of the two predicted downstream genes, namely IRF-7. IRF7 is a transcription factor that plays a key role in cytokine-induced chemokine expression by pancreatic β-cells (198).

ELF3, RIPK2 and the validated downstream genes, were attributed a main functional term related to defense mechanisms (defense response, immune response, host virus interaction, in the tables 6.5 and 6.6).This suggests that these genes may play a protective role for β-cells facing, for instance, a viral infection. In line with this hypothesis, ELF3 has been associated with airway inflammation (39, 143) while RIPK2 mediates signals in both the innate and adaptive immune systems (137).

It can be concluded that ELF3 is indeed likely to be upstream CX3CL1 and SP110, and RIPK2 upstream IRF7, but the question of whether these causal effects are direct or indirect (via intermediate gene(s)) remains open. These experimental results are consistent with the inferred network, and show that the statistical inference of gene regulatory networks, from time series, can be a helpful tool in the prediction of novel gene interactions.

The presented novel approach to map genes that are part of an inferred network, taking into account a temporal reference of regulation time, can be helpful in the temporal ordering of molecular events once genes are assigned functional terms. Previous approaches to deal with this problem included the simple clustering of gene expression time series (e.g. based on the Euclidean distance as in (272). In (150), similarly to the proposed approach, a relative time of regulation of genes was inferred from the pairwise temporal shift between genes. This approach is based on the ordinary least squares method, and the estimation of the pairwise lags is more complex than our approach.

The results presented suggest that molecular mechanisms related to defense/immune response are active throughout the process that will culminate in β-cell death. Furthermore, key mechanisms related to the triggering of apoptosis seem to be activated early following cytokine exposure, indicating that novel approaches to protect β-cells in type 1 diabetes may need to be implemented in the early stages of insulitis.

6.6 Conclusion

A meta-analysis study is presented, using a collection of gene expression datasets of pancreatic β-cells conditioned by an environment similar to the one observed in T1D induced-apoptosis (i.e. exposure to pro-inflammatory cytokines) to identify a set of relevant and differentially expressed genes. These genes were characterized by function and prior knowledge in the literature, and used to infer temporal regulatory networks. Biological validation experiments showed that inhibition of two of the most relevant genes, previously unknown in T1D literature, have an impact in apoptosis. Predicted regulatory interactions involving those genes were also consistent with experimental results. The inferred regulatory network is thus supported by the experimental validation of predicted causal effects involving the validated genes, providing a proof-of-concept for the proposed inference approach.

7

Conclusions

7.1 GRN inference

The inference of gene regulatory networks (GRN) is helpful to medical research, through the identification of relevant genes and causal mechanisms associated with phenotypes. One application is a prioritization of gene regulations subject to posterior experimental validation. An example is found in this thesis, with the inference of a regulatory network of differentially expressed genes in β-cell genes after cytokine exposure, emulating the molecular mechanisms leading to β-cell apoptosis in type 1 Diabetes. Top predicted regulations were then experimentally validated.

GRN inference from expression data is a causal inference problem which may be tackled with concepts and techniques developed in statistics and machine learning. The use of time series facilitates the problem of causal inference, as the direction of causality may be identified with lagged statistical dependences. However, GRN inference is currently very challenging due to the typical very high number of variables (genes) to observations. This limitation prevents the use of standard inference techniques and calls for the adoption of alternative approaches suited to the high variable to sample ratio. Such strategies include measures of pairwise or low order conditional dependence, regularization, or filter variable selection. The preliminaries for causal inference were presented in the Chapter 2, and the state of the art of gene network inference in Chapter 3.

When a gold standard of regulations is available, it may be used to assess inferred networks. If regulations are ranked, one measure of assess inference accuracy is the area under the precision-recall curve (AUPRC), commonly adopted in the field (166). A null (relative to random selection) AUPRC distribution is required to assess statistical significance, and is helpful to compare AUPRC scores in different inference tasks, as the AUPRC distribution depends on the number of total and positive instances. The common approach to obtain AUPRC distribution is via Monte-Carlo, which may be computationally intensive when the number of instances is high.

The contributions in this thesis are summarized and discussed in the next section. They are three-fold and consist on: a parametric approximation to the AUPRC distribution; novel inference algorithms and an experimental investigation of GRN inference from time series; a meta-analysis and network inference from gene expression in β-cells after cytokine exposure, in the context of type 1 Diabetes.

7.2 Summary and discussion of contributions

7.2.1 Performance assessment with the AUPRC

The first contribution of this thesis is the analytical derivation of the mean and variance of the null distribution of the AUPRC, presented in Chapter 4. These parameters are used to approximate the AUPRC null distribution, using the beta distribution. The beta distribution is a continuous distribution, defined by its support (minimum and maximum) and two parameters - the mean and variance. This approach stands as an alternative to a distribution estimation based on Monte-Carlo, which may be computationally intensive. For instance, as seen in Section 4.5, for 900 possible regulations (a situation of a small network of 30 genes), around 130000 simulations are needed so that the expected relative variance error is below 0.01. The proposed approach is used to assess network inference in the experimental session described in the Chapter 5. The characteristics of the expected null precision-recall curve are also extensively discussed.

Limitations and future work The beta distribution is a continuous distribution and how well it approximates the true (discrete) AUPRC distribution was only briefly investigated in Section 4.5. When the number of positive instances becomes lower, the discrete nature of the AUPRC curve becomes more accentuated and naturally diverges from a continuous approximation. In the case of a very low number total or positive instances, the beta distribution approximation exhibits a relevant bias and should not be adopted (Monte Carlo should be used instead, as in the yeast network inference experiment of Chapter 5).

The degree of accuracy of the proposed approximation should be investigated in more detail for different configurations of total and positive instances. As discussed in the Section 4.6, the proposed approximation is also computationally intensive (in particular, the computation of the variance) when the number of instances is high. A software implementation made available tackles this issue with an approximation of the true variance, based on a spline interpolation of the covariance elements of the covariance matrix of the pr-curve. Alternative approaches for AUPRC significance were also discussed in the Section 4.6.

7.2.2 GRN inference from time series

The Chapter 5 consists of an experimental investigation on GRN inference from time series. Two methodological contributions were presented: a fast approximation of the minimum of first order GC scores, which is used as a filter network inference method; and a method to identify gene co-regulation. The proposed inference approach (GC3) consists in attributing to each directed regulation the minimum of first order conditional GC scores. As a full search on all conditioning genes is computationally expensive in large networks, an heuristic to conduct the search is proposed whose speed and approximation accuracy is controlled by a user-given parameter. This method is experimentally validated in GRN inference using simulated gene expression time series, where it achieves the highest precision among state of the art approaches. Its accuracy is due to the fact that each possible regulation is scored as a function of only three genes, thus minimizing issues of over-fitting and high variance. The method can be easily adapted to static inference, through the use of first order partial correlations instead of GC scores. A similar heuristic may also be applied to higher order conditional independence tests, as discussed in Section 5.4.

The proposed strategy to identify co-regulation is also experimentally validated, through the precise identification of a very high number of co-regulation occurrences in both real and simulated data.

The described experimental session is composed of real microarray data (of around 20 time points) and short to medium sized simulated time series (20 to 300 time points). First, we assess different approaches to model linear Granger causality in cause-effect pairs of time series. Simple one-lag bivariate Granger causality models outperform more complex ones, considering multiple lags and non-stationarity, when the number of samples (n) is low. When n becomes higher than 100 time points, inference using more sophisticated approaches becomes more precise.

Secondly, we conduct a comparison between state of the art network inference approaches. Using the microarray and simulated time series, we infer 100 networks of 50 genes. Each network ranking is assessed with the AUPRC which is subsequently transformation into a z-score. Methods are then compared with the respective AUPRC z-scores, including a significance analysis. Regarding this aspect we point out a flaw in own previously published work, where AUPRC values coming from different distributions (networks) were simply averaged (155). This aspect was already observed and referenced at the end of that paper, and led to the investigation resulting in the work of Chapter 4.

Limitations and future work One limitation of the proposed methods for network inference is that they are based on linear models. However, Granger causality may be extended to the non-linear case using transfer entropy. Its application to GRN inference should be investigated in future work. A similar observation is applied to the co-regulation filter, which may be extended to the non-linear case by considering the mutual information (instead of the linear correlation). A point left for future work is a theoretical investigation of this general case (as on Section 5.2.3.2 for the linear case).

A limitation of the presented experimental session is the limited variety of data used. Only two microarray datasets were used, and the simulated time series were generated by a single software (GNW). It is unclear how well the GNW time series are representative of real gene expression time series. It is possible that there are particular characteristics in the GNW time series which stand responsible for the observed results. Also, the GNW time series are only up to 300 time points and a similar analysis in longer time series is left for future work. In particular, it would be interesting to investigate at which range of time points inference methods based on high order conditional dependences become more precise than based on first order conditional dependences. Another limitation of the experimental session is the fact that only a few network inference methods were assessed (mostly linear, except for random forests). The mutual information in information-theoretic methods was estimated following a Gaussian assumption, and as a function of the linear correlation. An investigation on the inference precision of non-parametric MI estimations, as a function of n, is an interesting point for future work.

7.2.3 Knowledge inference in type 1 Diabetes

In Type 1 diabetes (T1D), insulin-producing β-cells undergo apoptosis, due to local release of cytokines such as IL-1β and IFN-γ. In the work described in Chapter 6, 8 time-series datasets of β-cell gene expression after exposure to IL-1β and IFN-γ (two of them made available in the context of this work) were used to identify genes differentially expressed after cytokine exposure. Differentially expressed genes were identified, before and after 24h, functionally characterized and compared with available literature information. The two most differentially expressed genes previously unknown in T1D literature (RIPK2 and ELF3) were found to modulate cytokine-induced apoptosis. The knockdown of these genes caused an increase in β-cell apoptosis.

A regulatory network was inferred using a temporal adaptation of a known filter network inference approach (based on the estimation of lags) and three out of four predicted regulations (involving RIPK2 and ELF3) were experimentally confirmed. ELF3 knockdown inhibited the expression of two genes predicted to be downstream, CX3CL1 and SP110. RIPK2 knockdown inhibited the expression of IRF7, also predicted to be down-regulated by it. Functional enrichment analysis suggests that the genes involved in these regulations may play a role in defense mechanisms in $beta$-cells. Finally, a strategy to map genes to a time of regulation was presented, based on the inferred network and estimated regulation lags, which was used in the ordering of biologically meaningful events after cytokine exposure.

Limitations and future work One limitation of the presented study is that although three causal regulations were experimentally confirmed, they cannot be concluded to be direct, as the observed effect may be due to an indirect regulation. Regarding the network inference, one possible point of contention may be the processing of the time series: the three time series were averaged prior to network inference, and interpolated to obtain a constant time interval. The question on whether in these situations, the time series should be analyzed individually, or concatenated to form longer time series is left for future investigation. The inference method is a forward selection-based filter variable selection approach (mRMR), but other approaches might be more appropriate, as discussed in the Chapter 5. It is based on an estimation of the lags regulating gene expression, and an upper bound of 5 hours was adopted. This lag may be considered too high, and a future point of research is an investigation on more appropriate lower and upper bounds for the lags of gene regulations. The proposed method to order genes by time of regulation should also be subject to a more critical analysis and validation.

7.2.4 Concluding remarks

The inference of gene regulatory networks is a useful application of bioinformatics, valuable for medical research. This thesis focuses on the problem of GRN inference from (mRNA) expression data only. A proof-of-concept is presented in the form of a bioinformatics analysis and GRN inference in the context of type 1 Diabetes. It lead to the identification of novel genes playing a role in β-cell apoptosis and experimentally confirmed findings, in particular of causal links between predicted cause-effect gene pairs.

GRN inference only from mRNA expression data entails a biological simplification as the regulatory mechanisms occurring post transcription are ignored. It is an extremely challenging problem due to the typical low number of available gene expression measurements. This difficulty is observed in the low values of precision in the experiments reported in this work. A sensible approach to overcome this limitation is to combine multiple gene expression datasets and alternative relevant information, such as information on binding sites, perturbation experiments, available information in the literature. The development and scrutiny of integrative approaches is then an essential next step for current research. Still, each source of relevant information should be well studied, and causal inference from time series stands as valuable tool in the endeavor of GRN inference. The work presented in this thesis is a contribution to this particular topic of knowledge.

The adequate assessment of inference methods is an important issue and a contribution to this topic is also presented, in the form of an approximation of the statistical significance of precision-recall curves.

Appendix A

Model assessment and hypothesis testing

This appendix presents some tools used in this thesis, notably on the assessment of statistical models and testing of hypothesis.

A.1 Assessing statistical models

Common measures to assess statistical models take into account goodness of fit while penalizing model complexity. These include the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), extensively studied and with interesting properties (41). In these cases the lower is the measure, the better is the model. Assume a model with likelihood $\mathcal{L}$ and k parameters, estimated from n observations. The AIC measure:

$$\text{AIC} = 2k - 2\ln(\mathcal{L}) \tag{A.1}$$

The AICc is the AIC with a correction for finite sample sizes. It is given by:

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{A.2}$$

In OLS linear regression the model likelihood $\mathcal{L}$ is given by:

$$\ln(\mathcal{L}) = -\frac{n}{2}\ln\frac{RSS}{2} \tag{A.3}$$

Resulting in a AIC of:

$$\text{AIC} = \frac{n}{2}\ln\frac{RSS}{2} + 2k \tag{A.4}$$

This form of the AIC (in particular the AICc version) is used in the experiments of Section 5.3.

A.2 Linear correlation coefficient

The p-value of a sample estimate of the linear correlation (Pearson and partial) relative to the null hypothesis of zero correlation may be be approximated with the Fisher's transformation (58). The sample partial correlation between X and Y (conditioned on a set of variables Z, possibly empty) is transformed into a statistic named z.

$$z = \frac{1}{2} \ln \left(\frac{1 + r_{X,Y|Z}}{1 - r_{X,Y|Z}} \right) \tag{A.5}$$

z follows approximately a normal distribution with mean $\frac{1+\rho_{X,Y|Z}}{1-\rho_{X,Y|}}$ ($\rho_{X,Y|Z}$ is the true partial correlation) and standard deviation $\frac{1}{\sqrt{N-|Z|-3}}$, where $|Z|$ is the cardinality of the set Z. This test is used in network inference methods implemented in the Section 5.3, assessing the null hypothesis of zero correlation.

A.3 Testing sample differences

A common and useful statistical test is on whether groups of samples follow the same distribution. In the two variable case, these tests are paired or unpaired - in the first case, the samples of the two groups are observed simultaneously (paired tests are more powerful). One well known paired test is the Wilcoxon signed-rank test (277). This non-parametric test was used in the network inference experiment of Section 5.3 and is as follows. First the differences of the paired samples are obtained; then the values are ordered by magnitude and assigned a rank (lowest value is 1); the ranks are multiplied by the sign of the respective difference and are summed up, resulting in W. If the median difference is zero, W follows a normal distribution with mean 0 and standard deviation $\sqrt{N(N+1)(2N+1)}6$.

A.4 Meta-analysis of statistical tests

The field of meta-analysis tackles the problem of combining the results of multiple experiments. Two common approaches to test multiple p-values are the methods of Fisher and Stouffer. For instance, the latter transforms the p-values into z-scores, sums them up and divides by $\sqrt{n}$. The resulting statistic (if all the null hypothesis are true) follows approximately a standard normal distribution. One application of meta-analysis is the estimation of a global effect size in multiple experiments. In this case it is common practice to weight each experiment differently, on the basis of its quality (eg. number of used samples, variance). The effect of the experimental condition can be assumed to be a fixed value (same for all experiments), or following a random distribution (considering differences in the experiments). These two models are known as fixed-effects and random-effects models. For a reference see (123). One simple meta-analysis was performed in the Chapter 6 (Section 6.3.4), weighting different experiments on the number of samples.

A.5 The multiple testing problem

If we consider n independent p-values, the probability that at least one of them is lower than p is $1-(1-p)^n$. As we increase the number of tests, the probability increases that a test will return a p-value below some arbitrary level (leading to an erroneous rejection of the null). This is commonly referred to as the multiple testing problem (174).

The Bonferroni correction Assume n experiments, where the null is true, and that we set $\frac{\alpha}{n}$ to be a cut-off below which p-values are considered statistical significant. By Boole's inequality [1], the probability that at least one of the p-values will fall below the cutoff is lower than $\sum_n \frac{\alpha}{n} = \alpha$. Setting the significance cut-off at $\frac{\alpha}{n}$ assures that the probability that one null hypothesis is incorrectly rejected (a type 1 error) is lower than α. Trivially, this result remains valid if we also consider p-values of experiments where the null hypothesis is false (ie. if we do not know in which experiments the null is true). The probability of incorrectly rejecting the null decreases as we add experiments where the null should indeed be rejected. This approach controls the probability of at least one type 1 error (known as the familywise error rate), and is known as the Bonferroni correction.

False discovery rate The Bonferroni correction may be too conservative when the number or hypothesis is high (for instance in gene expression experiments). In these cases, applying the Bonferroni correction results in a very low $\frac{\alpha}{n}$ and a very stringent significance level. Another approach is to control the rate of type 1 errors (the number of errors divided by n, known as the false discovery rate). A known procedure to do so was proposed by Benjamini and Hochberg (17) and is as follows. First, the p-values are ordered from the lowest to highest (the index is k). Then, a false discovery rate α is defined used to obtain the largest k such that $p_k < \frac{k}{m}\alpha$. Finally, the null hypothesis is rejected in the first k p-values. The expected false discovery rate can be shown to be below α. Further refinements of this procedure have been proposed (18). The Benjamini and Hochberg was used in the Chapter 6, Section 6.3.4).

[1] Boole's inequality states that the probability of at least one event (out of multiple events) happening is equal or lower than the sum of the individual probabilities of all events.

Appendix B

GRN inference from gene expression time series - supplementary results

This section presents supplementary results for the network inference experiment of Section 5.3.

Figure B.1 presents the AUPRC z-scores obtained in the yeast network inference task. Contrary to the results presented in the Figure 5.7 where the results obtained in each multivariate time series were combined, here networks were inferred from each multivariate time series individually (resulting in the inference of 100 × 11 networks). Figure B.2 presents the results of the Wilcoxon signed rank test for this experiment. Note that there are multiple significant differences, not perceived in the boxplot representation. This is due to the higher number of considered networks (1100 compared to 100 when the inference from the individual time series was combined, as shown in Section 5.3). Note also that random inference is one of the top performing methods.

Figures B.3, B.4 and B.5 present the AUPRC z-scores obtained in the GNW network inference task, when the considered time points are the middle 40, 60, 80 (Figure B.3), 100, 120, 140 (Figure B.4), 160, 180 and 200 (Figure B.5) points. The Figures B.6 and B.7 present the results of the Wilcoxon signed rank test for this experiment. To avoid redundancy the results are presented only for the middle 40 and 80 time points (the results for higher number of time points follow a similar pattern).

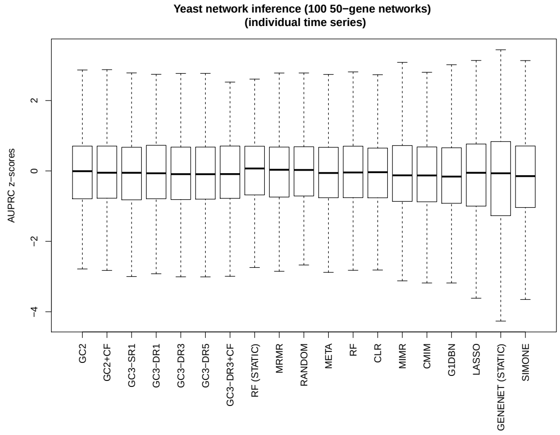


Figure B.1: GRN inference performance (yeast time series) - inference on individual time series - Box plots of the AUPRC z-scores of the assessed methods. 100 GRN, of 50 genes, were inferred from the 11 yeast multivariate time series. Each network was inferred individually in all multivariate time series (resulting in 100 × 11 inferred networks). The AUPRC of each inferred network was obtained and transformed into a z-score, obtained with Monte Carlo, 100000 simulations. The difference to Figure 5.7 is that in that case the inference on the multivariate time series was combined.

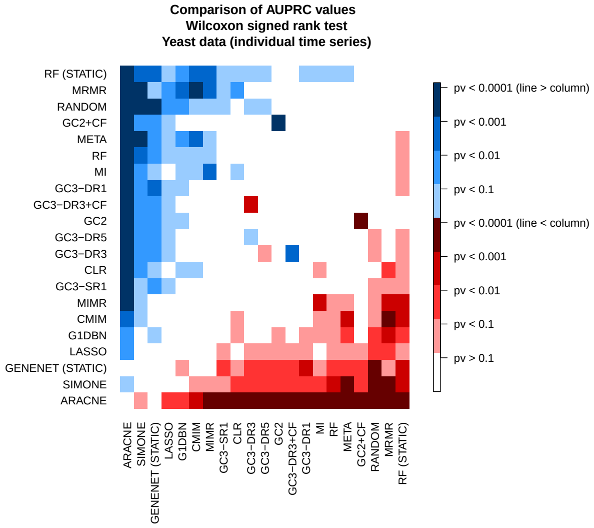


Figure B.2: GRN inference statistical comparison (yeast experiment) - inference on individual time series - Comparison between the different GRN inference methods AUPRC z-scores (1100 in total). The Wilcoxon signed rank test was used and the obtained (two-tailed) p-values are represented in the matrix. If the element $[i, j]$ is blue, then method i performs better than method j. Methods on top are the best performing. Results for the yeast experiment, when inferring networks from each multivariate time series individually.

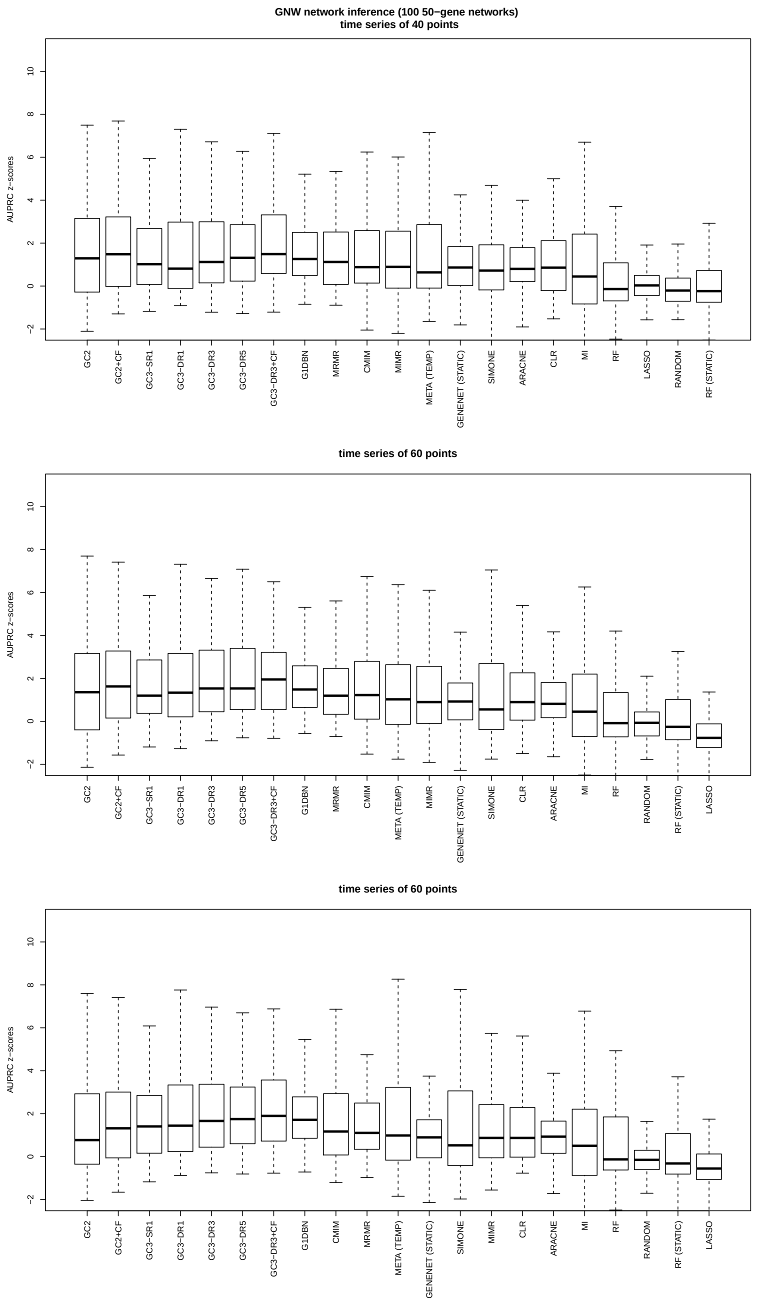


Figure B.3: GRN inference performance (GNW time series), using 40, 60 and 80 time points. - Box plots of the AUPRC z-scores of the assessed methods. 100 GRN, of 50 genes, were inferred from GNW multivariate time series (one network corresponding to one multivariate time series). The AUPRC of each inferred network was obtained and transformed into a z-score, following the beta-distribution approximation proposed in Chapter 4. Results are shown using the middle 40, 60 and 80 points of an original 300 point multivariate time series.

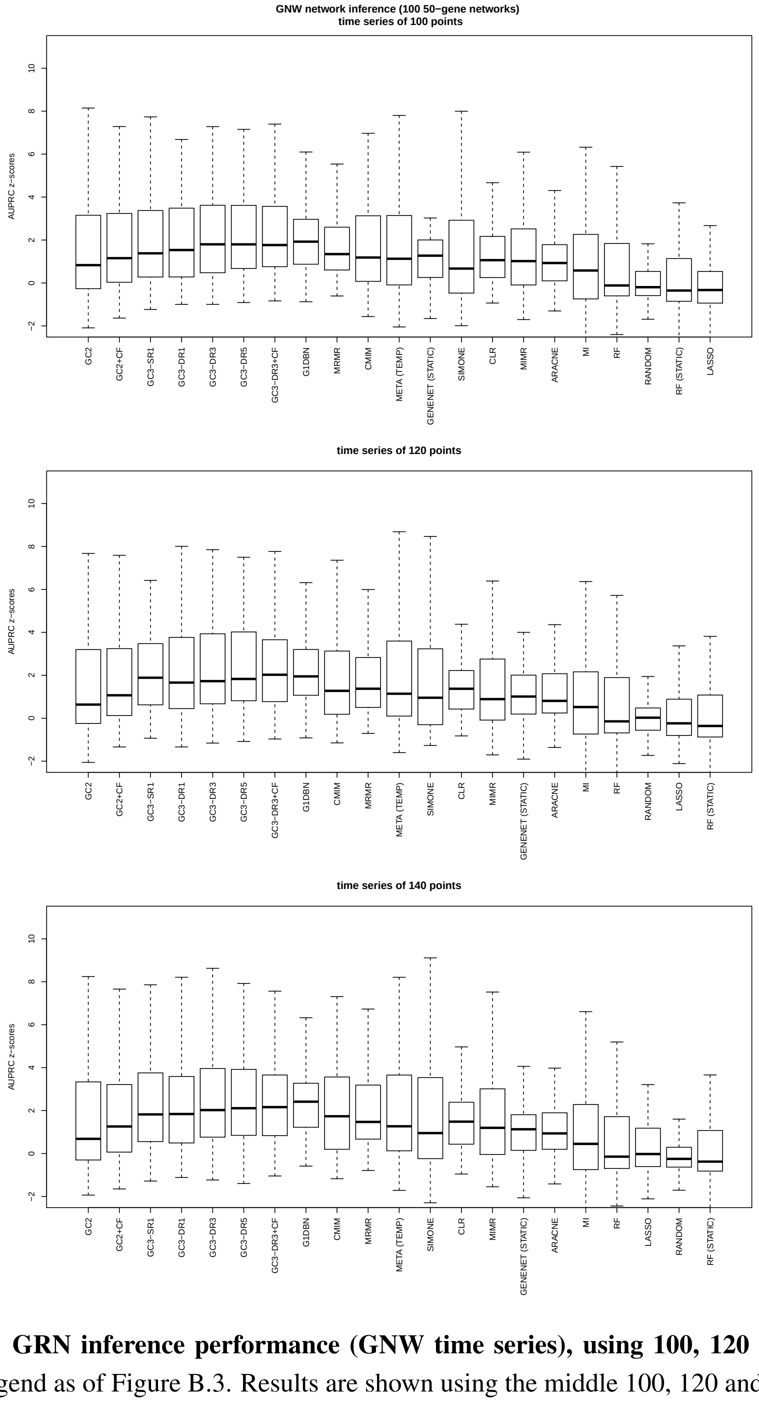


Figure B.4: GRN inference performance (GNW time series), using 100, 120 and 140 time points. - Legend as of Figure B.3. Results are shown using the middle 100, 120 and 140 points of an original 300 point multivariate time series.

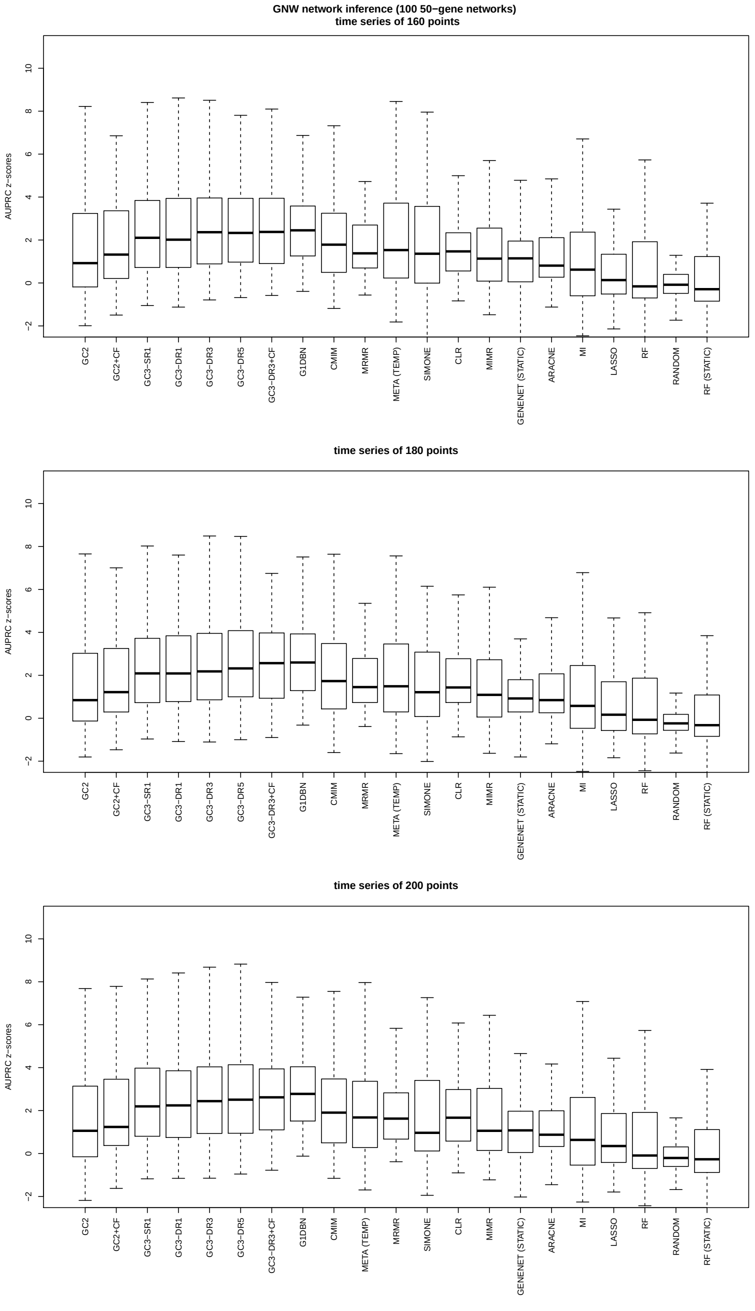


Figure B.5: GRN inference performance (GNW time series), using 160, 180 and 300 time points. - Legend as of Figure B.3. Results are shown using the middle 160, 180 and 200 points of an original 300 point multivariate time series.

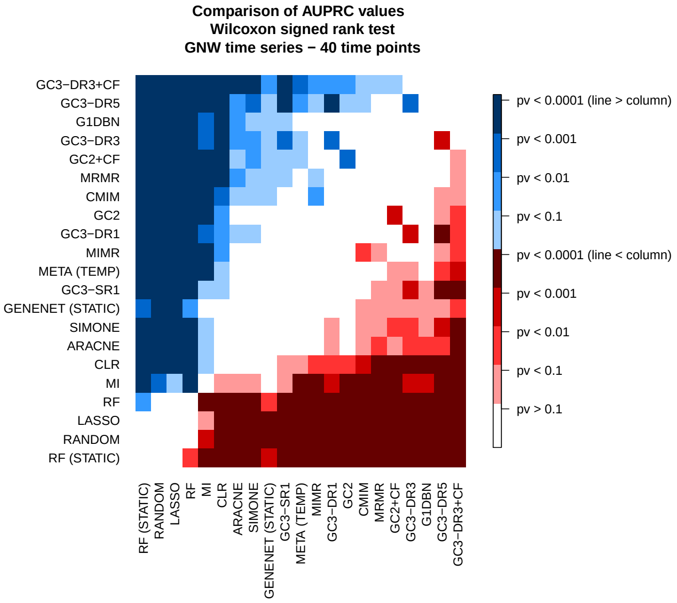


Figure B.6: GRN inference statistical comparison (GNW, n=40) - Legend as of Figure B.2. Results for the GNW time series, considering the middle 40 points of the time series.

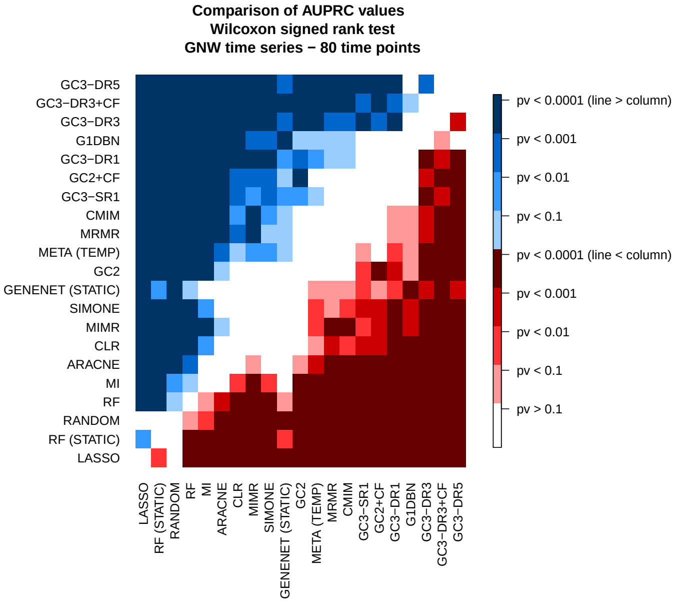


Figure B.7: GRN inference statistical comparison (GNW, n=80) - Legend as of Figure B.2. Results for the GNW time series, considering the middle 80 points of the time series.

References

[1] Dario Abdulrehman, Pedro Tiago Monteiro, Miguel Cacho Teixeira, Nuno Pereira Mira, Artur Bastos Loureno, Sandra Costa dos Santos, Tnia Rodrigues Cabrito, Alexandre Paulo Francisco, Sara Cordeiro Madeira, Ricardo Santos Aires, Arlindo Limede Oliveira, Isabel S-Correia, and Ana Teresa Freitas. **YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface**. *Nucleic Acids Research*, **39**(suppl 1):D136–D140, 2011. 23, 116

[2] Neema Agrawal, P. V. N. Dasaradhi, Asif Mohmmed, Pawan Malhotra, Raj K. Bhatnagar, and Sunil K. Mukherjee. **RNA Interference: Biology, Mechanism, and Applications**. *Microbiology and Molecular Biology Reviews*, **67**(4):657–685, December 2003. 10, 14

[3] Constantin Aliferis, Ioannis Tsamardinos, Alexander Statnikov, C. F. Aliferis M. D, Ph. D, I. Tsamardinos Ph. D, and Er Statnikov M. S. **HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection**, 2003. 56

[4] Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos. **Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation**. *J. Mach. Learn. Res.*, **11**:171–234, March 2010. 49

[5] O. Aparicio, J. V. Geisberg, E. Sekinger, A. Yang, Z. Moqtaderi, and K. Struhl. **Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo.** *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, **Chapter 21**, February 2005. 14

[6] Evan Archer, Il Memming Park, and Jonathan W. Pillow. **Bayesian Entropy Estimation for Countable Discrete Distributions**. *J. Mach. Learn. Res.*, **15**(1):2833–2868, January 2014. 34

[7] Adam Arkin, Peidong Shen, and John Ross. **A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements**. *Science*, **277**(5330):1275–1279, August 1997. 79

[8] Richard Askey. **Orthogonal polynomials and special functions**, 1975. Lectures given at the National Science Foundation regional conference held at Virginia Polytechnic Institute in June 1974. 90

[9] Richard P. Auburn, David P. Kreil, Lisa A. Meadows, Bettina Fischer, Santiago Sevillano Matilla, and Steven Russell. **Robotic spotting of cDNA and oligonucleotide microarrays**. *Trends in Biotechnology*, **23**:374–379, 2005. 11

[10] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. **How to infer gene networks from expression profiles**. *Molecular Systems Biology*, **3**(1):n/a–n/a, 2007. 15, 71

[11] Mukesh Bansal, Giusy Della Gatta, and Diego Di Bernardo. **Inference of gene regulatory networks and compound mode of action from time course gene expression profiles**. *Bioinformatics*, **22**(7):815–822, April 2006. 13, 68, 71, 74

[12] Angela Baralla, Wieslawa I. Mentzen, and Alberto de la Fuente. **Inferring gene networks: dream or nightmare?** *Annals of the New York Academy of Sciences*, **1158**(The Challenges of Systems Biology Community Efforts to Harness Biological Complexity):246–256, March 2009. 14, 73

[13] J. Carl Barrett and Ernest S Kawasaki. **Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression**. *Drug Discovery Today*, **8**:134–141, 2003. 11

[14] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. **Reverse engineering of regulatory networks in human B cells**. *Nature Genetics*, **37**:382–390, 2005. 77

[15] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Meulen. **Nonparametric Entropy Estimation: An Overview**. *International Journal of the Mathematical Statistics Sciences*, **6**:17–39, 1997. 34

[16] Anthony J. Bell. **The co-information lattice**. In *in Proc. 4th Int. Symp. Independent Component Analysis and Blind Source Separation*, pages 921–926, 2003. 34

[17] Yoav Benjamini and Yosef Hochberg. **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1):289–300, 1995. 139, 163

[18] Yoav Benjamini and Yosef Hochberg. **On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics**. *Journal of Educational and Behavioral Statistics*, **25**(1):60–83, 2000. 163

[19] Toby Berger. **Living Information Theory**. *IEEE Information Theory Society Newsletter*, **53**(1):1+, March 2003. 31

[20] R Bergholdt, C Brorsson, A Palleja, L A Berchtold, T Flø yel, C H Bang-Berthelsen, K S Frederiksen, L J Jensen, J Stø rling, and F Pociot. **Identification of novel type 1 diabetes candidate genes by integrating genome-wide association data, protein-protein interactions, and human pancreatic islet gene expression.** *Diabetes*, **61**(4):954–962, April 2012. 7, 136

[21] J. Berkson. **Limitations of the application of fourfold table analysis to hospital data.** *Biometrics*, **2**(3):47–53, June 1946. 18

[22] Claus Berthelsen, Lykke Pedersen, Tina Floyel, Peter Hagedorn, Titus Gylvin, and Flemming Pociot. **Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes**. *BMC Genomics*, **12**(1):97, February 2011. 136

[23] Peter Bickel, Bo Li, Alexandre Tsybakov, Sara Geer, Bin Yu, Tefilo Valds, Carlos Rivero, Jianqing Fan, and Aad Vaart. **Regularization in statistics**. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **15**(2):271–344, 2006. 38

[24] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 11, 12

[25] Douglas L. Black. **Mechanisms of Alternative Pre-Messenger RNA Splicing**. *Annual Review of Biochemistry*, **72**(1):291–336, 2003. 7

[26] Kenneth A. Bollen and Judea Pearl. **Eight Myths About Causality and Structural Equation Models**. In Stephen L. Morgan, editor, *Handbook of Causal Analysis for Social Research*, Handbooks of Sociology and Social Research, pages 301–328. Springer Netherlands, 2013. 41

[27] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics*, **19**(2):185–193, January 2003. 11

[28] Richard Bonneau, David Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin Baliga, and Vesteinn Thorsson. **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo**. *Genome Biology*, **7**(5):R36+, 2006. 81

[29] Gianluca Bontempi and Patrick Emmanuel Meyer. **Causal filter selection in microarray data**. In Johannes Furnkranz and Thorsten Joachims, editors, *ICML*, pages 95–102. Omnipress, 2010. 75

[30] Kendrick Boyd, Kevin H. Eng, and C. David Page Jr. **Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals.** In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezn, editors, *ECML/PKDD (3)*, **8190** of *Lecture Notes in Computer Science*, pages 451–466. Springer, 2013. 86

[31] Andrew P. Bradley. **The use of the area under the ROC curve in the evaluation of machine learning algorithms**. *Pattern Recognition*, **30**:1145–1159, 1997. 86

[32] P. Brazhnik, A. Delafuente, and P. Mendes. **Gene networks: how to put the function in genomics**. *Trends in Biotechnology*, **20**(11):467–472, November 2002. 8

[33] L. Breiman. **Statistical modeling: The two cultures**. *Statistical Science*, **16**(3):199–215, 2001. 12

[34] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984. 57

[35] Leo Breiman. **Bagging Predictors**. *Mach. Learn.*, **24**(2):123–140, August 1996. 58

[36] Leo Breiman. **Random Forests**. *Mach. Learn.*, **45**(1):5–32, October 2001. 16, 58

[37] Joerg Breitung and M. Hashem Pesaran. **Unit Roots and Cointegration in Panels**. CESifo Working Paper Series 1565, CESifo Group Munich, 2005. 63

[38] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. **The Binormal Assumption on Precision-Recall Curves**. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 4263–4266, Washington, DC, USA, 2010. IEEE Computer Society. 86

[39] C Brown, J Gaspar, A Pettit, R Lee, X Gu, H Wang, C Manning, C Voland, S R Goldring, M B Goldring, T A Libermann, E M Gravallese, and P Oettgen. **ESE-1 is a novel transcriptional mediator of angiopoietin-1 expression in the setting of inflammation.** *The Journal of biological chemistry*, **279**(13):12794–12803, March 2004. 153

[40] Morton B. Brown. **400: A Method for Combining Non-Independent, One-Sided Tests of Significance**. *Biometrics*, **31**(4):987+, December 1975. 101

[41] Kenneth P. Burnham and David R. Anderson. **Multimodel Inference: Understanding AIC and BIC in Model Selection**. *Sociological Methods & Research*, **33**(2):261–304, 2004. 161

[42] Oliver S. Burren, Ellen C. Adlem, Premanand Achuthan, Mikkel Christensen, Richard M. R. Coulson, and John A. Todd. **T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research.** *Nucleic Acids Research*, **39**(Database-Issue):997–1001, 2011. 144

[43] S. A. Bustin. **Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays**. *Journal of Molecular Endocrinology*, **25**(2):169–193, October 2000. 10

[44] X Cai, M Wang, H Kong, J Liu, Y Liu, W Xia, M Zou, J Wang, H Su, and D Xu. **Prokaryotic expression, purification and functional characterization of recombinant human RIP2.** *Molecular biology reports*, **40**(1):59–65, January 2013. 143, 153

[45] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari. **Tikhonov Regularization and the L-curve for Large Discrete Ill-posed Problems**. *J. Comput. Appl. Math.*, **123**(1-2):423–446, November 2000. 38

[46] Richard W. Carthew and Erik J. Sontheimer. **Origins and Mechanisms of miRNAs and siRNAs.** *Cell*, **136**(4):642–655, February 2009. 10

[47] Alexandra M. Carvalho. **Scoring functions for learning Bayesian networks**. Technical report, Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento, 2009. 55

[48] Robert Castelo, Alberto Roverato, and Max Chickering. **A robust procedure for gaussian graphical model search from microarray data with p larger than n**. *Journal of Machine Learning Research*, **7**:2006, 2006. 52

[49] Camille Charbonnier, Julien Chiquet, and Christophe Ambroise. **Weighted-LASSO for structured network inference from time course data.** *Statistical applications in genetics and molecular biology*, **9**(1), 2010. 80

[50] Ting Chen, Vladimir Filkov, and Steven S. Skiena. **Identifying gene regulatory networks from experimental data**. In *3rd Annual International Conference on Computational Molecular Biology (RECOMB'99)*, pages 94–103. ACM-SIGACT, 1999. 79

[51] David Maxwell Chickering. **Learning Equivalence Classes of Bayesian-network Structures**. *J. Mach. Learn. Res.*, **2**:445–498, March 2002. 48

[52] Stéphan Clémençon and Nicolas Vayatis. **Nonparametric Estimation of the Precision-recall Curve**. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 185–192, New York, NY, USA, 2009. ACM. 86

[53] Miriam Cnop, Nils Welsh, Jean-Christophe Jonas, Anne Jorns, Sigurd Lenzen, and Decio L. Eizirik. **Mechanisms of pancreatic beta-cell death in type 1 and type 2 diabetes: many differences, few similarities**. *Diabetes*, Dec 2005. 7

REFERENCES

[54] Jacob Cohen. **A power primer**. *Psychological Bulletin*, **112**(1):155–159, July 1992. 139, 142

[55] Gregory F. Cooper. **The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks (Research Note)**. *Artif. Intell.*, **42**(2-3):393–405, March 1990. 52

[56] K T Coppieters, F Dotta, N Amirian, P D Campbell, T W Kay, M A Atkinson, B O Roep, and M G von Herrath. **Demonstration of islet-autoreactive CD8 T cells in insulitic lesions from recent onset and long-term type 1 diabetes patients.** *The Journal of experimental medicine*, **209**(1):51–60, January 2012. 6

[57] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. 17, 32

[58] N. J. Cox. **Speaking Stata: Correlation with confidence, or Fisher's z revisited**. *Stata Journal*, **8**(3):413–439(27), 2008. 162

[59] Francis Crick. **Central Dogma of Molecular Biology**. *Nature*, **227**(5258):561–563, August 1970. 7

[60] Paul Dagum and Michael Luby. **Approximating probabilistic inference in Bayesian belief networks is NP-hard**. *Artificial Intelligence*, **60**(1):141 – 153, 1993. 52

[61] Manhong Dai, Pinglang Wang, Andrew Boyd, Georgi Kostov, Brian Athey, Edward Jones, William Bunney, Richard Myers, Terry Speed, Huda Akil, Stanley Watson, and Fan Meng. **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucl. Acids Res.*, **33**(20):e175–e175, January 2005. 137, 153

[62] Jesse Davis and Mark Goadrich. **The relationship between Precision-Recall and ROC curves**. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM. 22

[63] R Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, UK, 1976. 3

[64] Luis M. de Campos. **A Scoring Function for Learning Bayesian Networks Based on Mutual Information and Conditional Independence Tests**. *J. Mach. Learn. Res.*, **7**:2149–2187, December 2006. 55

[65] Alberto de la Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. **Discovery of meaningful associations in genomic data using partial correlation coefficients**. *Bioinformatics*, **20**(18):3565–3574, 2004. 24

[66] Patrik D'haeseleer. **What are DNA sequence motifs?** *Nature Biotechnology*, **24**(4):423–425, April 2006. 14

[67] David A. Dickey and Wayne A. Fuller. **Distribution of the Estimators for Autoregressive Time Series With a Unit Root**. *Journal of the American Statistical Association*, **74**(366):427–431, June 1979. 63

[68] Marie-Agns Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Cline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Lalo, Caroline Le Gall, Brigitte Schaffer, Stphane Le Crom, Mickal Guedj, and Florence Jaffrzic. **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Briefings in Bioinformatics*, 2012. 11

[69] Juan J. Dolado and Helmut Lutkepohl. **Making Wald Tests Work for Cointegrated Var Systems**. Working papers, Centro de Estudios Monetarios Y Financieros-, 1994. 67

[70] Dorit Dor and Michael Tarsi. **A simple algorithm to construct a consistent extension of a partially oriented graph**. Technical Report R-185, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA, USA, October 1992. 55

[71] J. Dougherty, R. Kohavi, and M Sahami. **Supervised and unsupervised discretization of continuous features**. In *ICML-95*, 1995. 34

[72] Sorin Draghici, Purvesh Khatri, Aron C. Eklund, and Zoltan Szallasi. **Reliability and reproducibility issues in DNA microarray measurements.** *Trends in genetics : TIG*, **22**(2):101–109, February 2006. 11

[73] B. Efron and C. Morris. **Stein's Paradox in Statistics**. *Scientific American*, **236**:119–127, May 1977. 51

[74] BRADLEY EFRON, TREVOR HASTIE, IAIN JOHNSTONE, AND ROBERT TIBSHIRANI. **Least angle regression**. *Annals of Statistics*, **32**:407–499, 2004. 39

[75] MICHAEL EICHLER. **Granger causality and path diagrams for multivariate time series**. *Journal of Econometrics*, **137**(2):334–353, April 2007. 67

[76] D L EIZIRIK, D G PIPELEERS, Z LING, N WELSH, C HELLERSTRÖM, AND A ANDERSSON. **Major species differences between humans and rodents in the susceptibility to pancreatic beta-cell injury**. *Proceedings of the National Academy of Sciences*, **91**(20):9253–9256, September 1994. 137

[77] D L EIZIRIK, M SAMMETH, T BOUCKENOOGHE, G BOTTU, G SISINO, M IGOILLO-ESTEVE, F ORTIS, I SANTIN, M L COLLI, J BARTHSON, L BOUWENS, L HUGHES, L GREGORY, G LUNTER, L MARSELLI, P MARCHETTI, M I MCCARTHY, AND M CNOP. **The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines.** *PLoS genetics*, **8**(3):e1002552, 2012. 6, 7, 136, 141

[78] DÉCIO L EIZIRIK, MAIKEL L COLLI, AND FERNANDA ORTIS. **The role of inflammation in insulitis and beta-cell loss in type 1 diabetes.** *Nature reviews. Endocrinology*, **5**(4):219–26, April 2009. 6, 143

[79] ROBERT F ENGLE AND CLIVE W J GRANGER. **Co-integration and Error Correction: Representation, Estimation, and Testing**. *Econometrica*, **55**(2):251–76, March 1987. 61, 62, 63

[80] BY EVA JABLONKA AND GAL RAZ. **Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution**. *The Quarterly Review of Biology*, **84**(2):pp. 131–176, 2009. 3

[81] JEREMIAH J FAITH, BORIS HAYETE, JOSHUA T THADEN, ILARIA MOGNO, JAMEY WIERZBOWSKI, GUILLAUME COTTAREL, SIMON KASIF, JAMES J COLLINS, AND TIMOTHY S GARDNER. **Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles**. *PLoS Biol*, **5**(1):e8, 01 2007. 77

[82] A P FIELD. **Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods.** *Psychological methods*, **6**(2):161–180, June 2001. 139

[83] FRANÇOIS FLEURET. **Fast Binary Feature Selection with Conditional Mutual Information**. *J. Mach. Learn. Res.*, **5**:1531–1555, December 2004. 57, 75

[84] CATHERINE FORBES, MERRAN EVANS, NICHOLAS HASTINGS, AND BRIAN PEACOCK. *Statistical Distributions*. John Wiley and Sons, Inc., 2010. 89, 96

[85] D. A. FREEDMAN. *Linear Statistical Models for Causation: A Critical Review*. John Wiley and Sons, Ltd, 2005. 20

[86] DAVID FREEDMAN. **From Association to Causation via Regression**. *Advances in Applied Mathematics*, **18**(1):59 – 110, 1997. 41

[87] YOAV FREUND AND ROBERT E. SCHAPIRE. **Experiments with a New Boosting Algorithm**. In *International Conference on Machine Learning*, pages 148–156, 1996. 57, 58

[88] JEROME FRIEDMAN, TREVOR HASTIE, AND ROBERT TIBSHIRANI. **Sparse inverse covariance estimation with the graphical lasso**. *Biostatistics*, **9**(3):432–441, July 2008. 51

[89] JEROME H. FRIEDMAN, TREVOR HASTIE, AND ROB TIBSHIRANI. **Regularization Paths for Generalized Linear Models via Coordinate Descent**. *Journal of Statistical Software*, **33**(1):1–22, 2 2010. 39, 40

[90] NIR FRIEDMAN, MICHAL LINIAL, IFTACH NACHMAN, AND DANA PE'ER. **Using Bayesian Networks to Analyze Expression Data**. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, RECOMB '00, pages 127–135, New York, NY, USA, 2000. ACM. 78

[91] NIR FRIEDMAN, IFTACH NACHMAN, AND DANA PEÉR. **Learning Bayesian Network Structure from Massive Datasets: The Sparse Candidate Algorithm**. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 206–215, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. 78

[92] ANDRÉ FUJITA, JOÃO R. SATO, HUMBERTO M. GARAY-MALPARTIDA, RUI YAMAGUCHI, SATORU MIYANO, MARI C. SOGAYAR, AND CARLOS E. FERREIRA. **Modeling gene expression regulatory networks with the sparse vector autoregressive model**. *BMC systems biology*, **1**:39, xx 2007. 80

REFERENCES

[93] Timothy S. Gardner and Jeremiah J. Faith. **Reverse-engineering transcription control networks**. *Physics of Life Reviews*, **2**(1):65 – 88, 2005. 14, 15, 71

[94] Lewis Geer, Aron Marchler-Bauer, Renata Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen Bryant. **The NCBI BioSystems database.** *Nucleic acids research*, **38**, January 2010. 139

[95] Dan Geiger and David Heckerman. **Learning Gaussian Networks**. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pages 235–243, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. 55

[96] Dan Geiger, Thomas Verma, and Judea Pearl. **Identifying independence in Bayesian Networks**. *Networks*, **20**:507–534, 1990. 47

[97] Pierre-Luc Germain, Emanuele Ratti, and Federico Boem. **Junk or functional DNA? ENCODE and the function controversy**. *Biology and Philosophy*, **29**(6):807–831, 2014. 9

[98] John Geweke. **Measurement of Linear Dependence and Feedback Between Multiple Time Series**. *Journal of the American Statistical Association*, **77**(378):304–313, 1982. 66

[99] Zoubin Ghahramani. **Learning dynamic Bayesian networks**. In *Adaptive Processing of Sequences and Data Structures*, pages 168–197. Springer-Verlag, 1998. 48

[100] Gene H. Golub and Urs von Matt. **Tikhonov Regularization for Large-Scale Problems**. In Gene H. Golub, S. H. Lui, F. T. Luk, and R. J. Plemmons, editors, *Workshop on Scientific Computing*, pages 3–26. Springer, 1997. 38

[101] J. Goutsias and N. H. Lee. **Computational and experimental approaches for modeling gene regulatory networks**. *Curr. Pharm. Design*, page 2007. 15, 71

[102] Ayman Grada and Kate Weinbrecht. **Next-Generation Sequencing: Methodology and Application**. *Journal of Investigative Dermatology*, **133**(8), August 2013. 11

[103] C. W. J. Granger. **Investigating Causal Relations by Econometric Models and Cross-spectral Methods**. *Econometrica*, **37**(3):424–438, August 1969. 19, 64

[104] C. W. J. Granger. **Some recent development in a concept of causality**. *Journal of Econometrics*, **39**(1-2):199–211, 1988. 66

[105] C. W. J. Granger and P. Newbold. **Spurious regressions in econometrics**. *Journal of Econometrics*, **2**(2):111–120, July 1974. 66

[106] Isabelle Guyon and André Elisseeff. **An Introduction to Variable and Feature Selection**. *J. Mach. Learn. Res.*, **3**:1157–1182, March 2003. 16

[107] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 12

[108] Hinrich W. H. Ghlmann and Willem Talloen. *Gene Expression Studies Using Affymetrix Microarrays.* Chapman and Hall / CRC mathematical and computational biology series. CRC Press, 2009. 10

[109] Bettina Harr and Christian Schlötterer. **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons**. *Nucleic Acids Research*, **34**(2), 2006. 11

[110] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. 12, 13, 36

[111] Anne-Claire C. Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe P. Vert. **TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.** *BMC systems biology*, **6**(1), 2012. 77

[112] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. **Gene regulatory network inference: data integration in dynamic models-a review.** *Bio Systems*, **96**(1):86–103, April 2009. 14

[113] D. Heckerman, D. Geiger, and D. M. Chickering. **Learning Bayesian Networks: The Combination of Knowledge and Statistical Data**. *Machine*

Learning, **20**(3):197–243, September 1995. Available as Technical Report MSR-TR-94-09. 52, 55

[114] David Heckerman and Dan Geiger. **Learning Bayesian Networks: A Unification for Discrete and Gaussian Domains**. *CoRR*, **abs/1302.4957**, 2013. 55

[115] M. J. Heller. **DNA microarray technology: devices, systems, and applications**. *Annu Rev Biomed Eng*, **4**:129–153, 2002. 10

[116] Nicholas J. Higham. **Computing the nearest correlation matrixa problem from finance**. *IMA Journal of Numerical Analysis*, **22**(3):329–343, 2002. 51

[117] Osamu Hirose, Ryo Yoshida, Seiya Imoto, Rui Yamaguchi, Tomoyuki Higuchi, Stephen D. Charnock-Jones, Cristin G. Print, and Satoru Miyano. **Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models.** *Bioinformatics*, **24**(7):932–942, 2008. 71

[118] Christopher Hitchcock. **Probabilistic Causation**. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012. 17, 18

[119] Tin Kam Ho. **The Random Subspace Method for Constructing Decision Forests.** *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**(8):832–844, 1998. 58

[120] Da Wei Huang, Brad Sherman, and Richard Lempicki. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic acids research*, **37**(1):1–13, January 2009. 143

[121] Yimin Huang and Marco Valtorta. **Pearl's Calculus of Intervention Is Complete.** In *UAI*. AUAI Press, 2006. 50

[122] Earl Hubbell, Wei-Min Liu, and Rui Mei. **Robust estimators for expression analysis**. *Bioinformatics*, **18**(12):1585–1592, 2002. 11

[123] J E Hunter and F L Schmidt. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. SAGE Publications, 2004. 139, 163

[124] Vân A. Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. **Inferring Regulatory Networks from Expression Data Using Tree-Based Methods**. *PLoS ONE*, **5**(9), September 2010. 58, 78

[125] Aapo Hyvärinen and Stephen M. Smith. **Pairwise Likelihood Ratios for Estimation of non-Gaussian Structural Equation Models**. *J. Mach. Learn. Res.*, **14**(1):111–152, January 2013. 20

[126] Trey E. Ideker, Vesteinn Thorsson, and Richard M. Karp. **Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design**. In *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific Press, 2000. 13

[127] International Human Genome Sequencing Consortium. **Initial sequencing and analysis of the human genome**. *Nature*, **409**:860–921, 2001. 4

[128] Pratik Jaluria, Konstantinos Konstantopoulos, Michael Betenbaugh, and Joseph Shiloach. **A perspective on microarrays: current applications, pitfalls, and potential uses**. *Microbial Cell Factories*, **6**(1):4, 2007. 11

[129] Dominik Janzing, Patrik O. Hoyer, and Bernhard Schlkopf. **Telling cause from effect based on high-dimensional observations.** In Johannes Frnkranz and Thorsten Joachims, editors, *ICML*, pages 479–486. Omnipress, 2010. 20

[130] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. **Genome-wide mapping of in vivo protein-DNA interactions.** *Science*, **316**(5830):1497–1502, June 2007. 14

[131] Hidde De Jong. **Modeling and simulation of genetic regulatory systems: A literature review**. *Journal of Computational Biology*, **9**:67–103, 2002. 15, 20, 71

[132] Christoph Kaleta, Anna Goehler, Stefan Schuster, Knut Jahreis, Reinhard Guthke, and Swetlana Nikolajewa. **Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence Analysis**. *BMC Systems Biology*, **4**(1), 2010. 14

[133] Guy Karlebach and Ron Shamir. **Modelling and analysis of gene regulatory networks**. *Nature Reviews Molecular Cell Biology*, **9**(10):770–780, 2008. 15

[134] S. Kauffman. **Homeostasis and differentiation in random genetic control networks.** *Nature*, **224**(5215):177–178, October 1969. 71

[135] Lutz Kilian. **Structural Vector Autoregressions**. CEPR Discussion Papers 8515, C.E.P.R. Discussion Papers, August 2011. 64

[136] E. M. Kleinberg. **An overtraining-resistant stochastic modeling method for pattern recognition**. *Ann. Statist.*, **24**(6):2319–2349, 12 1996. 58

[137] Naohiro Kobayashi K FAU - Inohara, Lorraine D Inohara N FAU - Hernandez, Jorge E Hernandez LD FAU - Galan, Gabriel Galan JE FAU - Nunez, Charles A Nunez G FAU - Janeway, Ruslan Janeway CA FAU - Medzhitov, Richard A Medzhitov R FAU - Flavell, and Flavell RA. **RICK/Rip2/CARDIAK mediates signalling for receptors of the innate and adaptive immune systems. PG - 194-9**. 153

[138] Ron Kohavi and George H. John. **Wrappers for Feature Subset Selection**. *Artif. Intell.*, **97**(1-2):273–324, December 1997. 16

[139] Kevin Kontos. *Gaussian Graphical Model Selection for Gene Regulatory Network Reverse Engineering and Function Prediction*. PhD thesis, Université Libre de Bruxelles, 2009. 37

[140] Kevin Kontos and Gianluca Bontempi. **Nested q-Partial Graphs for Genetic Network Inference from "Small n, Large p" Microarray Data.** In Mourad Elloumi, Josef Kng, Michal Linial, Robert F. Murphy, Kristan Schneider, and Cristian Toma, editors, *BIRD*, **13** of *Communications in Computer and Information Science*, pages 273–287. Springer, 2008. 52

[141] Kevin Kontos and Gianluca Bontempi. **An Improved Shrinkage Estimator to Infer Regulatory Networks with Gaussian Graphical Models**. In *Proceedings of the 2009 ACM Symposium on Applied Computing*, SAC '09, pages 793–798, New York, NY, USA, 2009. ACM. 52

[142] James T Kost and Michael P McDermott. **Combining dependent P-values**. *Statistics and Probability Letters*, **60**(2):183 – 190, 2002. 101

[143] R Kushwah, J R Oliver, J Wu, Z Chang, and J Hu. **Elf3 regulates allergic airway inflammation by controlling dendritic cell-driven T cell differentiation.** *Journal of immunology (Baltimore, Md. : 1950)*, **187**(9):4639–4653, November 2011. 153

[144] Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. **Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root?** *Journal of Econometrics*, **54**(1-3):159–178, 00 1992. 63

[145] P. Larrañaga, H. Karshenas, C. Bielza, and R. Santana. **A Review on Evolutionary Algorithms in Bayesian Network Learning and Inference Tasks**. *Information Sciences*, 2013. 52, 55

[146] L. Lauritzen. *Graphical Models*. 2009. 37, 46, 51

[147] Joel L Lebowitz. **Boltzmann's entropy and time's arrow**. *Physics today*, **46**:32–32, 1993. 31

[148] Sophie Lebre. **Inferring Dynamic Genetic Networks with Low Order Independencies**. *Statistical Applications in Genetics and Molecular Biology*, **8**(1):9, 2009. 81, 117

[149] Olivier Ledoit and Michael Wolf. **Improved estimation of the covariance matrix of stock returns with an application to portfolio selection**. *Journal of Empirical Finance*, **10**(5):603 – 621, 2003. 52

[150] Hans-Georg Leng X FAU - Muller and Muller HG. **Time ordering of gene coexpression. PG - 569-84**. 154

[151] Zheng Li, Ping Li, Arun Krishnan, and Jingdong Liu. **Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis.** *Bioinformatics (Oxford, England)*, **27**(19):2686–2691, October 2011. 68, 74

[152] Shoudan Liang, Stefanie Fuhrman, and Roland Somogyi. **REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures**. In *Pacific Symposium on Biocomputing*, **3**, pages 18–29, 1998. 71

[153] Elliott H. Lieb and Jakob Yngvason. **A Fresh Look at Entropy and the Second Law of Thermodynamics**. March 2000. 31

[154] Wei Keat Lim, Kai Wang, Celine Lefebvre, and Andrea Califano. **Comparative analysis**

of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**(13):i282–i288, 2007. 11

[155] MIGUEL LOPES AND GIANLUCA BONTEMPI. **Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data**. *Frontiers in Genetics*, **4**(303), 2013. 79, 157

[156] MIGUEL LOPES, BURAK KUTLU, MICHELA MIANI, CLAUS H. BANG-BERTHELSEN, JOACHIM STRLING, FLEMMING POCIOT, NATHAN GOODMAN, LEE HOOD, NILS WELSH, GIANLUCA BONTEMPI, AND DECIO L. EIZIRIK. **Temporal profiling of cytokine-induced genes in pancreatic -cells by meta-analysis and network inference**. *Genomics*, **103**(4):264 – 275, 2014. 136, 150

[157] MIGUEL LOPES, PATRICK MEYER, AND GIANLUCA BONTEPI. **Estimation of temporal lags for the inference of gene regulatory networks from time series**. In *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning,*, pages 19–26, 2012. 79

[158] AURELIE C. LOZANO, NAOKI ABE, YAN LIU 0002, AND SAHARON ROSSET. **Grouped graphical Granger modeling for gene expression regulatory networks discovery.** *Bioinformatics*, **25**(12), 2009. 74, 80

[159] HELMUT LÜTKEPOHL. *New Introduction to Multiple Time Series Analysis*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007. 19, 20, 59, 60, 61, 62, 64, 65

[160] HELMUT LUTKEPOHL AND HANS-EGGERT REIMERS. **Granger-causality in cointegrated VAR processes - the case of the term structure**. *Economics Letters*, **40**(3):263–268, 1992. 66

[161] HELMUT LTKEPOHL. **Testing for Causation Between Two Variables in Higher-Dimensional VAR Models**. In HANS SCHNEEWEI AND KLAUSF. ZIMMERMANN, editors, *Studies in Applied Econometrics*, Contributions to Economics, pages 75–91. Physica-Verlag HD, 1993. 65

[162] AVIV MADAR, ALEX GREENFIELD, ERIC VANDEN-EIJNDEN, AND RICHARD BONNEAU. **DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator**. *PLoS ONE*, **5**(3):e9803, 03 2010. 81

[163] PAUL MAGWENE AND JUNHYONG KIM. **Estimating genomic coexpression networks using first-order conditional independence**. *Genome Biology*, **5**(12):R100, 2004. 24

[164] RAMAMURTHY MANI, ROBERT, JOHN L. HARTMAN, GURI GIAEVER, AND FREDERICK P. ROTH. **Defining genetic interaction**. *Proceedings of the National Academy of Sciences*, **105**(9):3461–3466, March 2008. 9

[165] CHRISTOPHER D. MANNING, PRABHAKAR RAGHAVAN, AND HINRICH SCHÜTZE. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. 88

[166] DANIEL MARBACH, JAMES C COSTELLO, ROBERT KÜ FFNER, NICOLE M VEGA, ROBERT J PRILL, DIOGO M CAMACHO, KYLE R ALLISON, MANOLIS KELLIS, JAMES J COLLINS, GUSTAVO STOLOVITZKY, THE DREAM5 CONSORTIUM, AND YVAN SAEYS. **Wisdom of crowds for robust gene network inference**. *Nature Methods*, **9**(8):796–804, 2012. 23, 72, 73, 85, 155

[167] DANIEL MARBACH, ROBERT J. PRILL, THOMAS SCHAFFTER, CLAUDIO MATTIUSSI, DARIO FLOREANO, AND GUSTAVO STOLOVITZKY. **Revealing strengths and weaknesses of methods for gene network inference.** *Proceedings of the National Academy of Sciences of the United States of America*, **107**(14):6286–6291, April 2010. 72, 73

[168] DIMITRIS MARGARITIS AND SEBASTIAN THRUN. **Bayesian Network Induction via Local Neighborhoods**. In S.A. SOLLA, T.K. LEEN, AND K. MÜLLER, editors, *Advances in Neural Information Processing Systems 12*, pages 505–511. MIT Press, 2000. 56

[169] ADAM A. MARGOLIN, ILYA NEMENMAN, KATIA BASSO, ULF KLEIN, CHRIS WIGGINS, GUSTAVO STOLOVITZKY, RICCARDO D. FAVERA, AND ANDREA CALIFANO. **ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context**. *BMC Bioinformatics*, **7**(Suppl 1):S7+, October 2005. 76

[170] FLORIAN MARKOWETZ AND RAINER SPANG. **Inferring cellular networks – a review**. *BMC Bioinformatics*, **8**:S5, 2007. 15, 20, 24, 71, 105

[171] Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, Franois Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles**. *Nucleic Acids Research*, 2013. 14

[172] V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic acids research*, **34**(Database issue):D108–10, January 2006. 14, 146

[173] J McCarthy JV FAU - Ni, V M Ni J FAU - Dixit, and Dixit VM. **RIP2 is a novel NF-kappaB-activating and cell death-inducing kinase. PG - 16968-75**. 153

[174] John H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, Maryland, USA, second edition, 2009. 163

[175] William J. McGill. **Multivariate information transmission.** *Trans. of the IRE Professional Group on Information Theory (TIT)*, **4**:93–111, 1954. 33

[176] Robert McGrath and Gregory Meyer. **When effect sizes disagree: The case of r and d.** *Psychological Methods*, **11**(4):386–401, December 2006. 139

[177] Christopher Meek. **Causal Inference and Causal Explanation with Background Knowledge**. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 53

[178] Christopher Meek. **Strong Completeness and Faithfulness in Bayesian Networks**. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 411–418, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 47

[179] Patrick E. Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. **Information-theoretic inference of large transcriptional regulatory networks.** *EURASIP journal on bioinformatics & systems biology*, 2007. 16, 75

[180] Fabrice Moore, Najib Naamane, Maikel L Colli, Thomas Bouckenooghe, Fernanda Ortis, Esteban N Gurzov, Mariana Igoillo-Esteve, Chantal Mathieu, Gianluca Bontempi, Thomas Thykjaer, Torben F Ø rntoft, and Decio L Eizirik. **STAT1 is a master regulator of pancreatic beta cell apoptosis and islet inflammation.** *The Journal of biological chemistry*, **286**(2):929–941, 2011. 7, 136

[181] Noel G. Morgan and Sarah J. Richardson. **Enteroviruses as causative agents in type 1 diabetes: loose ends or lost cause?** *Trends in Endocrinology and Metabolism*, **25**(12):611 – 619, 2014. 6

[182] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Meth*, **5**(7):621–628, 7 2008. 11

[183] Nitai D. Mukhopadhyay and Snigdhansu Chatterjee. **Causality and pathway search in microarray time series experiment**. *Bioinformatics*, **23**(4):442–449, February 2007. 80

[184] Stanley A. Mulaik. *Linear Causal Modeling with Structural Equations*. Chapman and Hall/CRC, 1 edition, June 2009. 17, 18, 20, 41, 42, 44

[185] Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002. 48, 67

[186] Geeta J. Narlikar, Hua-Ying Fan, and Robert E. Kingston. **Cooperation between Complexes that Regulate Chromatin Structure and Transcription**. *Cell*, **108**(4):475 – 487, 2002. 9

[187] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003. 47

[188] I. Nemenman, W. Bialek, and R. D. R. Van Steveninck. **Entropy and information in neural spike trains: Progress on the sampling problem**. *Physical Review E*, **69**(5):56111, 2004. 31, 34

[189] A Y Ng, P Waring, S Ristevski, C Wang, T Wilson, M Pritchard, P Hertzog, and

I KOLA. **Inactivation of the transcription factor Elf3 in mice results in dysmorphogenesis and altered differentiation of intestinal epithelium.** *Gastroenterology*, **122**(5):1455–1466, May 2002. 153

[190] M W NICHOLAS AND KELLY NELSON. **North, South, or East? Blotting Techniques**. *J Invest Dermatol*, **133**(7):e10, July 2013. 10

[191] G. NOLTE, A. ZIEHE, V. NIKULIN, A. SCHLÖGL, N. KRÄMER, T. BRISMAR, AND K.-R. MÜLLER. **Robustly Estimating the Flow Direction of Information in Complex Physical Systems**. *Physical Review Letters*, **100**:234101, June 2008. 68

[192] JORDAN R OLIVER, RAHUL KUSHWAH, AND JIM HU. **Multiple roles of the epithelium-specific ETS transcription factor, ESE-1, in development and disease**. *Lab Invest*, **92**(3):320–330, March 2012. 153

[193] C. OLSEN, P. E. MEYER, AND G. BONTEMPI. **On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information.** *EURASIP journal on bioinformatics & systems biology*, 2009. 76

[194] CATHARINA OLSEN, GIANLUCA BONTEMPI, FRANK EMMERT-STREIB, JOHN QUACKENBUSH, AND BENJAMIN HAIBE-KAINS. **Relevance of different prior knowledge sources for inferring gene interaction networks**. *Frontiers in Genetics*, **5**(177), 2014. 14

[195] CATHARINA OLSEN, KATHLEEN FLEMING, NIALL PRENDERGAST, RENEE RUBIO, FRANK EMMERT-STREIB, GIANLUCA BONTEMPI, BENJAMIN HAIBE-KAINS, AND JOHN QUACKENBUSH. **Inference and validation of predictive gene networks from biomedical literature and gene expression data**. *Genomics*, **103**(56):329 – 336, 2014. 71

[196] RAINER OPGEN-RHEIN AND KORBINIAN STRIMMER. **Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data**. In *In Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB 2006*, pages 12–13, 2006. 74

[197] RAINER OPGEN-RHEIN AND KORBINIAN STRIMMER. **From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data**. *BMC Systems Biology*, **1**(1), 2007. 77

[198] F ORTIS, N NAAMANE, D FLAMEZ, L LADRIÈRE, F MOORE, D A CUNHA, M L COLLI, T THYKJAER, K THORSEN, T F ORNTOFT, AND D L EIZIRIK. **Cytokines interleukin-1beta and tumor necrosis factor-alpha regulate different transcriptional and alternative splicing networks in primary beta-cells.** *Diabetes*, **59**(2):358–374, February 2010. 7, 136, 150, 153

[199] M.R. OSBORNE, B. PRESNELL, AND B.A. TURLACH. **A new approach to variable selection in least squares problems**. *IMA journal of numerical analysis*, **20**(3):389, 2000. 39

[200] QUN PAN, OFER SHAI, LEO J. LEE, BRENDAN J. FREY, AND BENJAMIN J. BLENCOWE. **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet*, **40**(12):1413–1415, December 2008. 7

[201] LIAM PANINSKI. **Estimation of Entropy and Mutual Information**. *Neural Comput.*, **15**(6):1191–1253, June 2003. 34

[202] LIAM PANINSKI AND MASANAO YAJIMA. **Undersmoothed Kernel Entropy Estimators.** *IEEE Transactions on Information Theory*, **54**(9):4384–4388, 2008. 34

[203] TREVOR PARK AND GEORGE CASELLA. **The Bayesian Lasso**. *Journal of the American Statistical Association*, **103**(482):681–686, 2008. 39

[204] CHRISTOPHER C PATTERSON, GISELA G DAHLQUIST, EVA GYURUS E, ANDERS GREEN, AND GYULA SOLTESZ. **Incidence trends for childhood type 1 diabetes in Europe during 1989-2003 and predicted new cases 2005-20: a multicentre prospective registration study**, June 2009. 6, 150, 153

[205] JUDEA PEARL. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. 49

[206] JUDEA PEARL. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. 17, 18, 20, 41, 45, 46, 47, 48, 50, 52, 53, 55

[207] JUDEA PEARL. **The Causal Foundations of Structural Equation Modeling**, 2010. 20

[208] JUDEA PEARL. **The Do-Calculus Revisited**. *CoRR*, **abs/1210.4852**, 2012. 18, 45, 50

[209] JUDEA PEARL AND AZARIA PAZ. *Graphoids: A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department, 1985. 47

[210] HELEN PEARSON. **Genetics: What is a gene?** *Nature*, **441**(7092):398–401, May 2006. 7

[211] C A PENFOLD AND D L WILD. **How to infer gene networks from expression profiles, revisited**. *Interface Focus*, **1**(6):857–870, December 2011. 15, 71

[212] HANCHUAN PENG, FUHUI LONG, AND CHRIS DING. **Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.** *IEEE transactions on pattern analysis and machine intelligence*, **27**(8):1226–38, August 2005. 16, 26, 75

[213] PETER C.B. PHILLIPS AND PIERRE PERRON. **Testing for a Unit Root in Time Series Regression**. Cowles Foundation Discussion Papers 795R, Cowles Foundation for Research in Economics, Yale University, 1986. 63

[214] T. PRAMILA, W. WU, S. MILES, W.S. NOBLE, AND L.L. BREEDEN. **The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle.** *Genes Dev*, **20**(16):2266–78, 2006 (data accessible at NCBI GEO database (Edgar et al., 2002), accession GSE4987). 23, 115

[215] PAURUSH PRAVEEN AND HOLGER FRHLICH. **Boosting probabilistic graphical model inference by incorporating prior knowledge from multiple sources**. *PloS one*, **8**(6):e67410, 2013. 14

[216] ROBERT J. PRILL, DANIEL MARBACH, JULIO SAEZ-RODRIGUEZ, PETER K. SORGER, LEONIDAS G. ALEXOPOULOS, XIAOWEI XUE, NEIL D. CLARKE, GREGOIRE ALTAN-BONNET, AND GUSTAVO STOLOVITZKY. **Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges**. *PLoS ONE*, **5**(2):e9202, 02 2010. 21, 73

[217] J. ROSS QUINLAN. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. 57

[218] T REGAD AND M K CHELBI-ALIX. **Role and fate of PML nuclear bodies in response to interferon and viral infections.** *Oncogene*, **20**(49):7274–7286, October 2001. 153

[219] ALEX RHEE, RAYMOND CHEONG, AND ANDRE LEVCHENKO. **The application of information theory to biochemical signaling systems**. *Physical Biology*, **9**(4):045011, 2012. 31

[220] THOMAS RICHARDSON. **A Discovery Algorithm for Directed Cyclic Graphs.** In ERIC HORVITZ AND FINN VERNER JENSEN, editors, *UAI*, pages 454–461. Morgan Kaufmann, 1996. 20

[221] HERBERT ROBBINS. **An Empirical Bayes Approach to Statistics**, 1956. 52

[222] BRUCE A ROSA, YUHUA JIAO, SOOKYUNG OH, BERONDA L MONTGOMERY, WENSHENG QIN, AND JIN CHEN. **Frequency-based time-series gene expression recomposition using PRIISM.** *BMC Syst Biol*, **6**(1):69, 2012. 68

[223] RANADIP PAL SAAD HAIDER. **Boolean network inference from time series data incorporating prior biological knowledge**. *BMC Genomics*, (Suppl 6):S9, 2012. 74

[224] YVAN SAEYS, IÑAKI INZA, AND PEDRO LARRAÑAGA. **A Review of Feature Selection Techniques in Bioinformatics**. *Bioinformatics*, **23**(19):2507–2517, September 2007. 16

[225] MARILYN SAFRAN, IRINA DALAH, JUSTIN ALEXANDER, NAOMI ROSEN, TSIPPI INY STEIN, MICHAEL SHMOISH, NOAM NATIV, IRIS BAHIR, TIRZA DONIGER, HAGIT KRUG, ALEXANDRA SIROTA-MADI, TSVIYA OLENDER, YARON GOLAN, GIL STELZER, ARYE HAREL, AND DORON LANCET. **GeneCards Version 3: the human gene integrator.** *Database : the journal of biological databases and curation*, **2010**(0), August 2010. 144, 146

[226] F. SAMBO, B. DI CAMILLO, AND G. TOFFOLO. **CNET: an algorithm for reverse engineering of causal gene networks**. In *In Bioinformatics Methods for Biomedical Complex Systems Applications. 8th Workshop on Network Tools and Applications in Biology NETTAB2008*, pages 134–136. National Research Council, Milan, Italy, 2008. 74

[227] Michael A. Savageau and Eberhard O. Voit. **Recasting nonlinear differential equations as S-systems: a canonical nonlinear form**. *Mathematical Biosciences*, **87**(1):83 – 115, 1987. 71

[228] Juliane Schaefer, Rainer Opgen-Rhein, and Korbinian Strimmer. **Reverse Engineering Genetic Networks using the GeneNet Package**. *R News*, **6/5**:50–53, December 2006. 77

[229] Juliane Schäfer and Korbinian Strimmer. **A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics**. *Statistical Applications in Genetics and Molecular Biology, The Berkeley Electronic Press*, **4**(1), 2005. 51, 52

[230] Thomas Schaffter, Daniel Marbach, and Dario Floreano. **GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods**. *Bioinformatics*, **27**(16):2263–2270, 2011. wingx. 23, 115, 116

[231] T. Schreiber. **Measuring information transfer**. *Physical review letters*, **85**(2):461–464, 2000. 66

[232] D. Shalon, S. J. Smith, and P. O. Brown. **A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.** *Genome Research*, **6**(7):639–645, July 1996. 11

[233] C. E. Shannon. **A mathematical theory of communication**. *Bell system technical journal*, **27**, 1948. 31

[234] Teppei Shimamura, Seiya Imoto, Rui Yamaguchi, Andre Fujita, Masao Nagasaki, and Satoru Miyano. **Recursive regularization for inferring gene networks from time-course gene expression profiles**. *BMC Systems Biology*, **3**(1):41, 2009. 80

[235] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. **A Linear Non-Gaussian Acyclic Model for Causal Discovery**. *J. Mach. Learn. Res.*, **7**:2003–2030, December 2006. 20

[236] Ali Shojaie and George Michailidis. **Discovering graphical Granger causality using the truncating lasso penalty.** *Bioinformatics*, **26**(18), 2010. 80

[237] Ilya Shpitser and Judea Pearl. **Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models**. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1219–1226. AAAI Press, 2006. 50

[238] Daniel Simola, Chantal Francis, Paul Sniegowski, and Junhyong Kim. **Heterochronic evolution reveals modular timing changes in budding yeast transcriptomes**. *Genome Biology*, **11**(10):R105, 2010. 23, 115

[239] Michael L. Simpson, Chris D. Cox, and Gary S. Sayler. **Frequency domain analysis of noise in autoregulated gene circuits.** *Proc Natl Acad Sci U S A*, **100**(8):4551–4556, April 2003. 68

[240] Christopher A Sims. **Macroeconomics and Reality**. *Econometrica*, **48**(1):1–48, January 1980. 63

[241] Amit Singhal. **Modern information retrieval: A brief overview**. *IEEE Data Eng. Bull.*, **24**(4):35–43, 2001. 86

[242] M Skelin, Rupnik, and A Cencic. **Pancreatic beta cell lines and their applications in diabetes mellitus research**. *ALTEX*, **27(2)**:105–113, 2010. 136

[243] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000. 18, 47, 52, 53

[244] Peter Spirtes. **Directed Cyclic Graphical Representations of Feedback Models**. *CoRR*, **abs/1302.4982**, 2013. 20

[245] James H. Stock and Mark W. Watson. **Vector Autoregressions**. *Journal of Economic Perspectives*, **15**(4):101–115, 2001. 63, 64

[246] Gustavo Stolovitzky, Robert J. Prill, and Andrea Califano. **Lessons from the DREAM2 Challenges: A Community Effort to Assess Biological Network Inference**. *Annals of the New York Academy of Sciences*, **1158**(1):159–195, March 2009. 72, 85, 88

[247] B. D. Strahl and C. D. Allis. **The language of covalent histone modifications.** *Nature*, **403**(6765):41–45, January 2000. 9

[248] Carolin Strobl, James Malley, and Gerhard Tutz. **An Introduction to Recursive Partitioning: Rationale, Application and Characteristics**

of Classification and Regression Trees, Bagging and Random Forests, 2009. 57, 58

[249] PATRICK SUPPES. *A Probabilistic Theory of Causality*. Amsterdam,North-Holland Pub. Co., 1970. 19, 64

[250] T.A. SWANSON, S.I. KIM, M.J. GLUCKSMAN, AND M. LIEBERMAN. *Biochemistry, Molecular Biology, and Genetics*. Board review series. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2010. 7, 8, 9

[251] CHEEMENG TAN, FAISAL REZA, AND LINGCHONG YOU. **Noise-limited frequency signal transmission in gene circuits.** *Biophys J*, **93**(11):3753–61, 2007. 68

[252] JESPER TEGNER, STEPHEN K. YEUNG, JEFF HASTY, AND JAMES J. COLLINS. **Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.** *Proceedings of the National Academy of Sciences of the United States of America*, **100**(10):5944–5949, May 2003. 13

[253] THE ENCODE PROJECT CONSORTIUM. **An integrated encyclopedia of DNA elements in the human genome**. *Nature*, **489**(7414):57–74, September 2012. 9

[254] ROBERT TIBSHIRANI. **Regression Shrinkage and Selection Via the Lasso**. *Journal of the Royal Statistical Society, Series B*, **58**:267–288, 1994. 39

[255] RYAN J. TIBSHIRANI. **The lasso problem and uniqueness**. *Electron. J. Statist.*, **7**:1456–1490, 2013. 39

[256] GAPER TKAIK AND ALEKSANDRA M WALCZAK. **Information transmission in genetic regulatory networks: a review**. *Journal of Physics: Condensed Matter*, **23**(15):153102, 2011. 31

[257] HIRO Y. TODA AND TAKU YAMAMOTO. **Statistical inference in vector autoregressions with possibly integrated processes**. *Journal of Econometrics*, **66**(1-2):225–250, 1995. 20, 66

[258] JOHN A TODD. **Etiology of Type 1 Diabetes**, April 2010. 6

[259] IOANNIS TSAMARDINOS, CONSTANTIN ALIFERIS, ALEXANDER STATNIKOV, AND ER STATNIKOV. **Algorithms for Large Scale Markov Blanket Discovery**. In *In The 16th International FLAIRS Conference, St*, pages 376–380. AAAI Press, 2003. 49

[260] IOANNIS TSAMARDINOS, CONSTANTIN ALIFERIS, ALEXANDER STATNIKOV, AND ER STATNIKOV. **Algorithms for Large Scale Markov Blanket Discovery**. In *In The 16th International FLAIRS Conference, St*, pages 376–380. AAAI Press, 2003. 56

[261] IOANNIS TSAMARDINOS, CONSTANTIN F. ALIFERIS, AND ALEXANDER STATNIKOV. **Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations**. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 673–678, New York, NY, USA, 2003. ACM. 56

[262] IOANNIS TSAMARDINOS, LAURA E. BROWN, AND CONSTANTIN F. ALIFERIS. **The Max-min Hill-climbing Bayesian Network Structure Learning Algorithm**. *Mach. Learn.*, **65**(1):31–78, October 2006. 78

[263] CATALINA TUDOR, CARL SCHMIDT, AND K VIJAY-SHANKER. **eGIFT: Mining Gene Information from the Literature**. *BMC Bioinformatics*, **11**(1):418, 2010. 144

[264] SERGEY TULYAKOV, STEFAN JAEGER, VENU GOVINDARAJU, AND DAVID S. DOERMANN. **Review of Classifier Combination Methods.** In SIMONE MARINAI AND HIROMICHI FUJISAWA, editors, *Machine Learning in Document Analysis and Recognition*, **90** of *Studies in Computational Intelligence*, pages 361–386. Springer, 2008. 72

[265] EUGENE P. VAN SOMEREN, LODEWYK F. A. WESSELS, MARCEL J.T. REINDERS, AND ERIC BACKER. **Genetic Network Modeling**. *Pharmacogenomics*, **3**(4):507–525, 2002. 14, 15, 71, 73, 79

[266] THOMAS VERMA AND JUDEA PEARL. **Causal networks: semantics and expressiveness.** In ROSS D. SHACHTER, TOD S. LEVITT, LAVEEN N. KANAL, AND JOHN F. LEMMER, editors, *UAI*, pages 69–78. North-Holland, 1988. 45, 47

[267] THOMAS S. VERMA AND JUDEA PEARL. **Equivalence and synthesis of causal models**. In P. P. BONISSONE, M. HENRION, L. N. KANAL, AND J. F. LEMMER, editors, *UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–268, North Holland, 1990. Elsevier Science Publishers B.V. 48

[268] H. WANG, L. QIAN, AND E. DOUGHERTY. **Inference of gene regulatory networks using S-system: a**

unified approach. *IET systems biology*, **4**(2):145–156, March 2010. 71

[269] Tong Wang and Jie Yang. **A heuristic method for learning Bayesian networks using discrete particle swarm optimization**. *Knowledge and Information Systems*, **24**(2):269–281, 2010. 55

[270] Zhong Wang, Mark Gerstein, and Michael Snyder. **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet*, **10**(1):57–63, January 2009. 11

[271] J. D. Watson and F. H. C. Crick. **Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid**. *Nature*, **171**(4356):737–738, April 1953. 3

[272] X Wen, S Fuhrman, G S Michaels, D B Carr, S Smith, J L Barker, and R Somogyi. **Large-scale temporal gene expression mapping of central nervous system development**. *PNAS*, **95**(1):334–339, 1998. 154

[273] Halbert White, Karim Chalak, and Xun Lu. **Linking Granger Causality and the Pearl Causal Model with Settable Systems.** In Florin Popescu and Isabelle Guyon, editors, *NIPS Mini-Symposium on Causality in Time Series*, **12** of *JMLR Proceedings*, pages 1–29. JMLR.org, 2011. 67

[274] Halbert White and Xun Lu. **Granger Causality and Dynamic Structural Systems**. *Journal of Financial Econometrics*, **8**(2):193–243, 2010. 67

[275] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, et al. **Identification of genes periodically expressed in the human cell cycle and their expression in tumors**. *Molecular biology of the cell*, **13**(6):1977–2000, 2002. 74

[276] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. 1990. 51

[277] Frank Wilcoxon. **Individual Comparisons by Ranking Methods**. *Biometrics Bulletin*, **1**(6):80–83, December 1945. 162

[278] A. Wille and P. Bühlmann. **Low-Order Conditional Independence Graphs for Inferring Genetic Networks**. *Statistical Applications in Genetics and Molecular Biology*, **5**(1), 2006. 24

[279] Anja Wille, Philip Zimmermann, Eva Vranova, Andreas Furholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelic, Peter von Rohr, Lothar Thiele, Eckart Zitzler, Wilhelm Gruissem, and Peter Buhlmann. **Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana**. *Genome Biology*, **5**(11):R92+, 2004. 24

[280] Jeremy E. Wilusz, Hongjae Sunwoo, and David L. Spector. **Long noncoding RNAs: functional surprises from the RNA world**. *Genes & Development*, **23**(13):1494–1504, July 2009. 5, 9

[281] Sewall Wright. **Correlation and causation**. *Journal of Agricultural Research*, **10**:557–585, 1921. 41

[282] J Wu, R Duan, H Cao, D Field, C M Newnham, D R Koehler, N Zamel, M A Pritchard, P Hertzog, M Post, A K Tanswell, and J Hu. **Regulation of epithelium-specific Ets-like factors ESE-1 and ESE-3 in airway epithelial cells: potential roles in airway inflammation.** *Cell research*, **18**(6):649–663, June 2008. 153

[283] Tong T. Wu and Kenneth Lange. **Coordinate Descent Algorithms for Lasso Penalized Regression**. *The Annals of Applied Statistics*, **2**(1):224–244, 2008. 39

[284] K. Y. Yeung, K. M. Dombek, K. Lo, J. E. Mittler, J. Zhu, E. E. Schadt, R. E. Bumgarner, and A E Raftery. **Construction of regulatory networks using expression time-series data of a genotyped population.** *Proceedings of the National Academy of Sciences of the United States of America*, **108**(48):19436–19441, November 2011. 74

[285] Paul Yin X FAU - Krikorian, Thomas Krikorian P FAU - Logan, Vilmos Logan T FAU - Csizmadia, and V Csizmadia. **Induction of RIP-2 kinase by proinflammatory cytokines is mediated via NF-kappaB signaling pathways and involves a novel feed-forward regulatory mechanism. PG - 251-9 LID - 10.1007/s11010-009-0226-y [doi]**. 153

[286] P Ylipaasto, B Kutlu, S Rasilainen, J Rasschaert, K Salmela, H Teerijoki, O Korsgren, R Lahesmaa, T Hovi, D L Eizirik, T Otonkoski, and M Roivainen. **Global profiling of coxsackievirus- and cytokine-induced gene expression in human pancreatic islets**. *Diabetologia*, **48**(8):1510–22, 2005. 7, 150

[287] Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, and Erich D. Jarvis. **Advances to Bayesian network inference for generating causal networks from observational biological data**. *Bioinformatics*, **20**(18):3594–3603, December 2004. 78

[288] L. Yu and H. Liu. **Feature selection for high-dimensional data: a fast correlation-based filter solution**. In *Proceedings of the International Conference on Machine Leaning*, **20**, page 856, 2003. 74

[289] Ming Yuan, Ming Yuan, Yi Lin, and Yi Lin. **Model selection and estimation in regression with grouped variables**. *Journal of the Royal Statistical Society, Series B*, **68**:49–67, 2006. 40

[290] G.U. Yule. *An Investigation Into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades*. Royal Statistical Society, 1899. 20

[291] Mahdi Zamanighomi, Mostafa Zamanian, Michaelk Kimber, and Zhengdao Wang. **Gene regulatory network inference from perturbed time-series expression data via ordered dynamical expansion of non-steady state actors**. *bioRxiv*, 2014. 13

[292] Z. G. Zhang, Y. S. Hung, S. C. Chan, W. C. Xu, and Y. Hu. **Modeling and identification of gene regulatory networks: A Granger causality approach.** In *ICMLC*, pages 3073–3078. IEEE, 2010. 80

[293] Peng Zhao and Bin Yu. **On Model Selection Consistency of Lasso**. *J. Mach. Learn. Res.*, **7**:2541–2563, December 2006. 39

[294] Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. **TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach**. *BMC Bioinformatics*, **11**(1):154, 2010. 74, 79

[295] Cunlu Zou and Jianfeng Feng. **Granger causality vs. dynamic Bayesian network inference: a comparative study**. *BMC Bioinformatics*, **10**(1):122+, 2009. 74

[296] Hui Zou. **The Adaptive Lasso and Its Oracle Properties**. *Journal of the American Statistical Association*, **101**(476):1418–1429, 2006. 39, 40

[297] Hui Zou and Trevor Hastie. **Regularization and variable selection via the Elastic Net**. *Journal of the Royal Statistical Society, Series B*, **67**:301–320, 2005. 38, 39, 40

[298] Chotirat ann Ratanamahatana. **Feature selection for the naive bayesian classifier using decision trees**. *Applied Artificial Intelligence*, **17**:2003. 57