PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Distances and Stability in biological Network Theory

Roberto Visintainer

Advisor:

Dott. Giuseppe Jurman

Fondazione Bruno Kessler

March 2013

To my Grandparents

Acknowledgments

I want to especially thank Kristina Lerman and Steve Horvath and their groups at USC and UCLA who welcomed me and gave me many valuable advices for my Ph.D. work.
None of this would have been possible without the friends and colleagues MPBA group at FBK, especially Cesare and Giuseppe who patiently guided me making these last 4 years such an enriching period both professionally and personally.
My special gratitude goes to my family who always supported me during the difficult moments in spite of the fact that I was never really able to explain them much about my network-related problems.

Abstract

In this thesis we introduce, define and quantitatively assess the stability of the algorithms for the reconstruction of networks. We will focus on theory, development and implementation of operative procedures and algorithms for the assessment of stability in complex networks for biological systems, with gene regulatory networks as the key example. A major issue affecting network inference is indeed the high variability of network reconstruction and network topology inferred after data perturbation, different parameter choices and alternative methods. Network stability will thus be used to measure reliability of inferred topology, also obtaining confidence intervals for the outcomes. The methods will be employed to introduce a new approach to reproducibility in the study of complex networks. It will also be coupled with statistical machine learning models, in order to integrate feature selection and network inference within a pathway profiling approach. The evaluation of similarity between networks will be the first and central operative procedure of the developed pipelines, the key point being the identification of distances that can compare network structures improving over classical measures based on the confusion matrix, too coarse for this task. A combination of spectral and edit distances especially tailored for biological networks will be investigated and applied to several high-throughput biological datasets of different nature and with different tasks in oncogenomics, neurogenomics and exposomics.

Keywords

Network Comparison, Network Distances, Stability, Reproducibility

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Reproducibility, *i.e.*, the possibility of independently repeating a suite of experiments obtaining the same (or very similar) outcome of the original study, is a key ingredient of the scientific method. In the last few years, the need for reproducibility has become a major task also in very young disciplines such as computational biology and bioinformatics, where the relevant impact of noise and the paucity of data represent daily obstacles to overcome in warranting a completely reproducible pipeline to be set and shown [66, 138]. Among the several aspects included under the umbrella definition of reproducibility, this thesis is mainly concerned with providing a level of confidence to associate to the experiments' results. Namely, we aim at quantitatively define a degree of (in)stability to the biological network inference tasks, which is the core of the systems biology, the meeting point of complex network science, computational biology and statistical machine learning. Complex networks (graph structures) appear at all levels of the cellular organization, as a mathematically efficient representation of the interactions taking place among the basic cell elements, at all scales, from the gene to the metabolic and signaling level. Since the knowledge of the mechanisms underlying the cellular processes requires a paradigm shift from the reductionist approach of separately studying all ingredients to a

complexity-aware overview of the net of their mutual relations, the amount of research activities aimed at reconstructing such networks from various biological signals has skyrocketed in the last decade [16]. Moving even one step further, the concept of stability as the continuous dependence of the inference algorithm result from perturbations of the original data is of particular interest because of the ever growing diffusion of two novel research directions stemming from the network reconstruction theory. The former is the differential network analysis methods, where the emphasis of detecting the features discriminating two conditions or two phenotypes is moving form the gene to the pathway (and thus the network) level [65, 31] and the latter is the integration of the biological network with socioeconomic and contact networks describing people's behavior in order to construct a brand new network medicine approach [95, 16].

As anticipated in the previous paragraph, the problem of inferring a biological network structure starting from a set of high-throughput measurements (*e.g.* gene expression arrays or digital gene expression from Next Generation Sequencing data) has been positively answered by a huge number of deeply different solutions published in the literature in the last fifteen years, ranging from purely deterministic (algebraic or analytic) to purely probabilistic (Bayesian). In this thesis, we also propose a novel reconstruction method (called RegnANN) based on artificial neural networks, which we prove to be a good compromise between performance and stability [56]. Nonetheless, network reconstruction suffers from being a underdetermined problem, being the number of interactions highly larger than the number of independent measurements [40]: thus any algorithm has to look for a compromise between accuracy and feasibility, allowing simplifications that inevitably mine the precision of the final outcome, for instance including a relevant number of false positive links [76]. This makes the inference problem "a daunting task" [18], not only in terms of devising an effective

algorithm, but also in terms of quantitatively interpreting the obtained results. In general, the reconstruction accuracy is far from being optimal in many situations with the presence of several pitfalls [103], related to both the methods and the data [60], with the extreme situation of many link prediction being statistically equivalent to random guesses [116]. In particular, the size (and the quality) of the available data play a critical role in the inference process, as widely acknowledged [94, 53, 105]. All these considerations support deeming network reconstruction a still unsolved problem [135].

Despite the ever rising number of available algorithms, only recently efforts have been carried out towards an objective comparison of network inference methods also highlighting current limitations [4, 83] and relative strengths and disadvantages [98]. Among those, it is worthwhile mentioning the international DREAM challenge [100], whose key result in the last edition advocated integration of predictions from multiple inference methods as an effective strategy to enhance performances taking advantage from the different algorithms' complementarity [40]. Nevertheless, the algorithm uncertainty has been so far assessed only in terms of performance, i.e. distance of the reconstructing network from the ground truth, wherever available, while not much has been instead investigated with respect to the stability of the methods. This can be of particular interest when no gold standard (ground truth network) is available for the given problem, and thus there is no chance to evaluate the algorithm's accuracy, leaving the stability as the sole rule of thumb for judging the reliability of the obtained network. Here we propose to tackle the issue by quantifying inference variability with respect to data perturbation, and, in particular, data subsampling. If a portion of data is randomly removed before inferring the network, the resulting graph is likely to be different from the one reconstructed from the whole dataset and, in general, different subsets of

data would generate different networks. Thus, in the spirit of applying reproducibility principles to this field, one has to accept the compromise that the inferred/non inferred links are just an estimation, lying within a reasonable probability interval. In brief, we aim at proposing a set of four indicators allowing the researcher to quantitatively evaluate the reliability of the inferred/non-inferred links. In detail, we quantitatively assess, for a given ratio of removed data and for a give number of resampling, the mutual distances among all inferred networks and their distances to the network generated by the whole dataset, with the idea that, the smaller the average distance, the stabler the network. Moreover, we provide a ranked list of the stablest links and nodes, where the rank is induced by the variability of the link weight and the node degree across the generated networks, the less variable being the top ranked.

Last but not least, thorough the whole stability pipeline the major ingredient is represented by availability of a consistent network metric expressing the distance between two graphs sharing the same nodes but a different wiring. The part of network theory dealing with the assessment of the similarity of two networks is called network comparison. Comparison methods are essential with dynamic networks to measure differences between two consecutive network states and then model the whole series, for instance when investigating the changes of a protein-protein interaction network during a biological process such as a disease. The theory of network comparison is based on the variety of similarity measures, whose taxonomy is essentially parted into two major branches: the indirect methods of feature-based measures and the direct methods making use of a suitable distance. Although fruitful insights can be drawn by indirect methods, a distance must be employed whenever a quantitative assessment of the differences between two elements is required. Traditional choices are members of the family of the edit distances, where the minimum number of link operations

(deletion and insertion) for transforming one topology into the other is evaluated, and the family of spectral distances, where the difference of the eigenvalues distribution of one the connectivity matrices of the networks is taken into account. To cope with the different pros and cons of both edit and spectral similarity, we propose here the HIM distance [75] which is the product metric of the spectral Ipsen-Mikhailov and the edit Hamming distance: the HIM distance is the base of the whole aforementioned stability framework.

Biological applications of the HIM distance and of the stability indicators are shown in the last chapter, where a number of tasks in exposomics, oncogenomics and neurogenomics are presented and discuss, as examples of how these newly introduced algorithms can be an effective tools for the researcher in the network branch of the systems biology.

Overall, the structure of the thesis goes as follows: after Chapter 2 collecting background material and notation, we show in Chapter 3 a comparative review of spectral distances for network comparison. Chapter 4 is devoted to the definition of the novel HIM distance together with its properties, while Chapter 5 is the core of the thesis where the stability indicators are introduced and discussed with some examples of applications. Finally in Chapter 6 we extensively show a number of biological applications on several omics realms. We conclude drawing conclusions in Chapter 7.

Chapter 2

Background and Notation

The representation of complex systems in terms of networks allows the formalization of the system agents and their interactions. By means of the properties of the underlying graph it is possible to describe and analyze the system itself. For instance the study of the power supply system of a big city using network theory could give insights about weakness points in the system and avoid possible failures.

2.1 Networks

2.1.1 Definitions

Any network can be formally represented as a mathematical entity called graph. A graph consists of a number N of *nodes*, also called *vertices* that can be finite or infinite and E *edges*, *links* or *arrows* that connect a couple of vertices representing an interaction ($N \in \mathbb{N}\{\infty\}$). For any network G, its topology consists of the set $V(G) = \{v_1, ..., v_n\}$ of its nodes and the set $E(G) = \{e_1 = (v_{i_1}, v_{j_1}), \cdots e_E = (v_{i_E}, v_{j_E})\}$ of its edges, neglecting here weights and directions. Different types of graph sharing the same topology are displayed in Figure 2.1. If there exist an edge connecting two nodes x

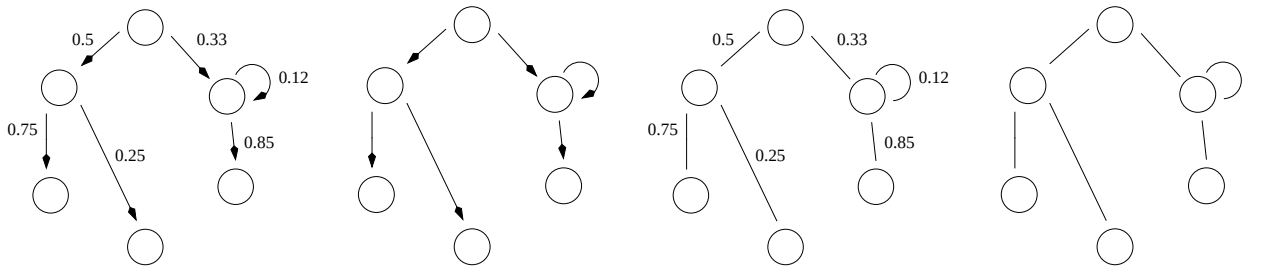


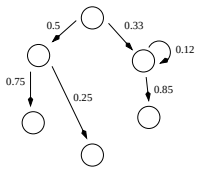
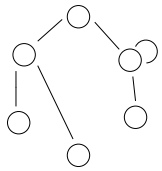
Figure 2.1: Network types

and y we say that they are *neighbors*, we can identify a set of neighbors for each node.
Links can be *bidirectional* or *unidirectional*, this basic feature determining whether the graph is

- **directed** (digraph): contains exclusively unidirectional links
- **undirected**: contains exclusively bidirectional links
- **mixed**: can contain both unidirectional and bidirectional links.

Graphically edges are depicted as arrows to symbolize directed links, lines or double-headed arrows for undirected ones. Only undirected graphs will be used hereafter. From the definition of graph it follows that any link can connect two nodes, but also a self connection is possible: an edge from a vertex to itself originates a *loop* in the network. Another feature of the interactions that can be carried by the edges is their intensity or weight, in this case we have a weighted network. For instance the weight of a link could be used to convey the information about the number of passengers moving from an airport to another in a transportation network. Formally, a *weighted network* $G(V, E, W)$ can be formalized as a graph in which links (x, y) are associated to a number called weight of the link $w(x, y)$ so that if $w(x, y) = 0$ then $(x, y) \notin E$ and $w(x, y) \neq 0$ if $w(x, y) \in E$.

Table 2.1: Adjacency matrices for the weighted directed network can be written in two alternative ways (1) with sign indicating direction (2) asymmetric, with the (positive) value only in the entry (i, j) to represent the connection $i \to j$, see Fig. 2.1 and its topology; nodes ordering is clockwise starting from the top node.

Network	Adjacency matrix
	$\begin{pmatrix} 0 & 0.33 & 0 & 0 & 0 & 0.5 \\ (-0.33) & 0.12 & 0.85 & 0 & 0 & 0 \\ 0 & (-0.85) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & (-0.25) \\ 0 & 0 & 0 & 0 & 0 & (-0.75) \\ (-0.5) & 0 & 0 & 0.25 & 0.75 & 0 \end{pmatrix}$
	$\begin{matrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{matrix}$

2.1.2 Connectivity Matrices

A widely used way to represent graphs is by means of matrices especially the *adjacency* matrix.

Adjacency Matrix

The adjacency matrix A is defined as an $N \times N$ squared matrix in which each entry a_{ij} corresponds to the link between the nodes i and j. In particular, for an unweighted link a_{ij} will be 1 when the link is present $((i, j) \in V)$ and 0 otherwise, see 2.1.
A is a very important and useful tool in graph theory, it is in fact enough to understand many of its basic topological characteristics.

- If A is symmetric, i.e. $A(h, k) = A(k, h) \forall h, k \in V$, then the graph is undirected.

- If the diagonal of A has all entries equal to 0, i.e. $A(h,h) = 0 \forall h \in 1,...,n$ there are no self-loops in the graph.

For a weighted link we can define a the matrix of weights of G as $W = (w(x,y))_{x,y \in V}$. The weights matrix can alone completely describe the topology and the characteristics of a graph, in this case we talk about *weighted adjacency matrix* as shown in 2.1. If A is in the form: $A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$, where B is a $p \times q$ matrix, we have a **bipartite** graph, a graph in which the nodes can be classified into two groups N_1 with $|N_1| = p$ and N_2 with $|N_2| = q$. A link (i,j) exists if and only if i and j belong to different groups. Another specific configuration of A is the block diagonal matrix: $A = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}$, where B_1 and B_2 are $p \times p$ and $q \times q$ matrices respectively. Also in this case we have a subdivision of the nodes into two groups $N_1 with |N_1| = p$ and $N_2 with |N_2| = q$, but the links connect exclusively couples of nodes belonging to the same group forming two separate subgraphs. This kind of adjacency matrix, where the groups of nodes are in general upper bounded by the number of nodes, is called **disconnected** graph.

The entries in the diagonal a_{ij} can be different from zero if selfloops are allowed; if no self-loops occur we call the graph simple: thus simple graphs have adjacency metric with zero diagonal. In Table 2.1 we show two examples of adjacency matrices for two graphs whose representation reads the nodes clockwise starting from the top one. In general a graphical representation is not unique, in the sense that it depends on the actual labeling of the nodes and isomorphic graphs (identical graphs with permuted labels) share the same adjacency matrix. Similarly, graphical representations are not unique too, since node placement is arbitrary.

Degree

The degree of a node is a concept of crucial importance in graph theory since it is the measure of level of interaction of the node with its neighbors and consequently with the whole network. We define the **out-degree** $d_{out}(x)$ as the number of links that exit from node x. Similarly we refer to **in-degree** $d_{in}(x)$ as the number of links that point to x. Both the previous definitions are applied to directed graphs: for undirected graph the in- and out-degree coincide and thus the **degree** $d(x)$ indicates the number of links touching the node x itself. Following this definition for the majority of the authors the self loops are counted twice. We also define the $N \times N$ degree matrix D as the diagonal matrix with the degree of each node as entries. For instance the degree matrix of the bottom network in Table 2.1 is $D = \begin{pmatrix} 2 & & & & & \\ & 4 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 3 \end{pmatrix}$.

The **weighted degree** (also called **strength**) of a node x in an undirected network is defined as the sum of the weights of all the links touching x, so we have that $s(x) = \sum_{y \in V} w(x, y)$ where V is the set of neighbors of x.

Laplacian Matrix

The Laplacian matrix L of a graph is defined as the difference between the degree matrix and the adjacency matrix $L = D - A$. From the definition follows that for an unweighted undirected graph without loops (a simple graph), the sums of the rows and the columns of L are zero.
Two normalizations of the Laplacian matrix exist $\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$ and $\delta = D^{1/2} \mathcal{L} D^{-1/2}$, where I is the identity matrix and

$D^{-1/2}$ is the diagonal matrix with entries $-\frac{\delta_{ij}}{\sqrt{deg_j}}$. Their entries can explicitly written as:

$$\mathcal{L} = \begin{cases} 1 & \text{if } i = j \text{ and } \deg_i \neq 0 \\ -\frac{1}{\sqrt{\deg_i \deg_j}} & \text{if } ij \text{ is an edge} \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta = \begin{cases} 1 & \text{if } i = j \text{ and } \deg_i \neq 0 \\ -\frac{1}{\deg_j} & \text{if } (i,j) \in V \text{ is an edge} \\ 0 & \text{otherwise} \end{cases}$$

Other kinds of networks have been described in the literature, but will not be used here. In labeled graphs, nodes are classified by functions from some subsets of the integers to the vertices or edges. Hypergraphs instead are characterized by links that can connect any number of vertices, while in multigraphs a couple of nodes can be connected by any number of links.

2.1.3 Spectrum

The eigenvalues of a matrix $M \in \mathbb{C}^{n\times n}$ are the n roots of its characteristic polynomial $p(z) = det(zI - M)$. The set of these roots is called the *spectrum* and is denoted by $\lambda(M)$. If $\lambda(M) = \lambda_1, \ldots, \lambda_n$ then it follows that

$$det(a) = \lambda_1, \lambda_2, \cdots, \lambda_n.$$

Moreover, if we define the *trace* of A by

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

then the $tr(A) = \lambda_1 + \cdots + \lambda_n$. this follows by looking at the coefficient of z in the characteristic polynomial.

If $\lambda \in \lambda(A)$ then the nonzero vectors $x \in \mathbb{C}^n$ that satisfy

$$Ax = \lambda x$$

are referred to as *eigenvectors*. More precisely, x is a *right eigenvector* for λ if $Ax = \lambda x$ and a left eigenvector if $x^H A = \lambda x^H$. Unless otherwise stated, "eigenvector" means "right eigenvector" [55].

An undirected and unweighted graph has symmetric real connectivity matrices and therefore real eigenvalues and a complete set of orthonormal eigenvectors. Also, for each eigenvalue, its algebraic multiplicity coincides with its geometric multiplicity. Since A has zero diagonal, its trace and hence the sum of the eigenvalues is zero. Moreover, L is positive semidefinite and singular, so the eigenvalues are $0 = \mu_0 \leq \mu_1 \leq \cdots \leq \mu_{n1}$ and their sum (the trace of L) is twice the number of edges. Finally, the eigenvalues of L lie in the range $[0, 2]$. While the connectivity matrices depend on the vertex labeling, the spectrum is a graph invariant. Two graphs are called isospectral or cospectral if the corresponding connectivity matrices of the graphs have equal multisets of eigenvalues. Isospectral graphs need not be isomorphic, but isomorphic graphs are always isospectral. Moreover it can be proved that the spectrum of the adjacency matrix of a bipartite graph is symmetric with respect to 0, *i.e.* if α is an eigenvalue of A then also $-\alpha$ is an eigenvalue. Network classification in terms of their spectrum is still an open problem [144, 150, 151]: however, a first attempt to (qualitative) network classification in terms of graph spectra can be found in [12, 13] by Banerjee.

Cauchy-Lorentz distribution

The Cauchy-Lorentz distribution is a continuous probability distribution with probability distribution function PDF given by:

$$f(x; x_0, \gamma) = \frac{1}{\pi}\left(\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right),$$

where x_0 indicates the peak of the distribution (also called the mode) of the distribution, and γ specifies half the width of the PDF at half the maximum height: see the graphical trend of the Cauchy Lorenz distribution in Figure 2.1.3.

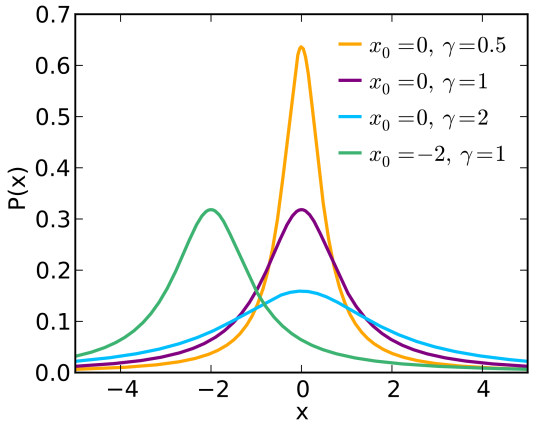


Figure 2.2: Examples of Cauchy-Lorentz distributions with different parameters.

2.1.4 A minimal example

Consider the two networks $I_1, I_2 \in \boldsymbol{N}$ with corresponding adjacency matrices A^{I_1}, A^{I_2} shown in Fig. 2.3 and 2.4.
The corresponding Laplacian matrices and eigenvalues are

$$L^{I_1} = \begin{pmatrix} 3 & -1 & 0 & 0 & -1 & 0 & 0 & -1 \\ -1 & 3 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 2 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 2 & -1 & -1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 3 & -1 \\ -1 & -1 & 0 & 0 & 0 & 0 & -1 & 3 \end{pmatrix} \quad \text{spec}(L^{I_1}) = \begin{bmatrix} 0 \\ 0.657077 \\ 1 \\ 2.529317 \\ 3 \\ 4 \\ 4 \\ 4.813607 \end{bmatrix}$$

$$A^{I_1} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

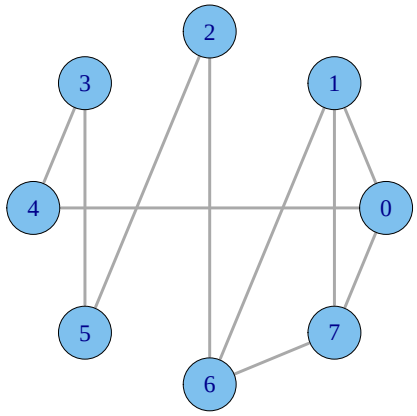


Figure 2.3: Adjacency matrix and graphical representation of I_1

$$L^{I_2} = \begin{pmatrix} 3 & -1 & 0 & 0 & 0 & -1 & -1 & 0 \\ -1 & 3 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & -1 & -1 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 2 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & -1 & 2 \end{pmatrix} \qquad \mathrm{spec}(L^{I_2}) = \begin{bmatrix} 0 \\ 0 \\ 0.340321 \\ 1.145088 \\ 3 \\ 3 \\ 3.854912 \\ 4.659679 \end{bmatrix}$$

From the above spectra, we can compute the corresponding Cauchy-Lorentz distributions $\rho_{I_{\{1,2\}}}(\omega, \overline{\gamma})$, where $\overline{\gamma} = 0.4450034$: their plots are shown in Fig. 2.5.

$$A^{I_2} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

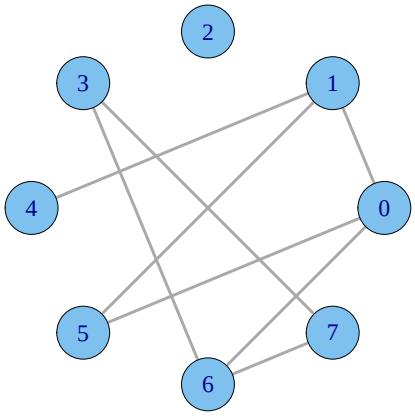
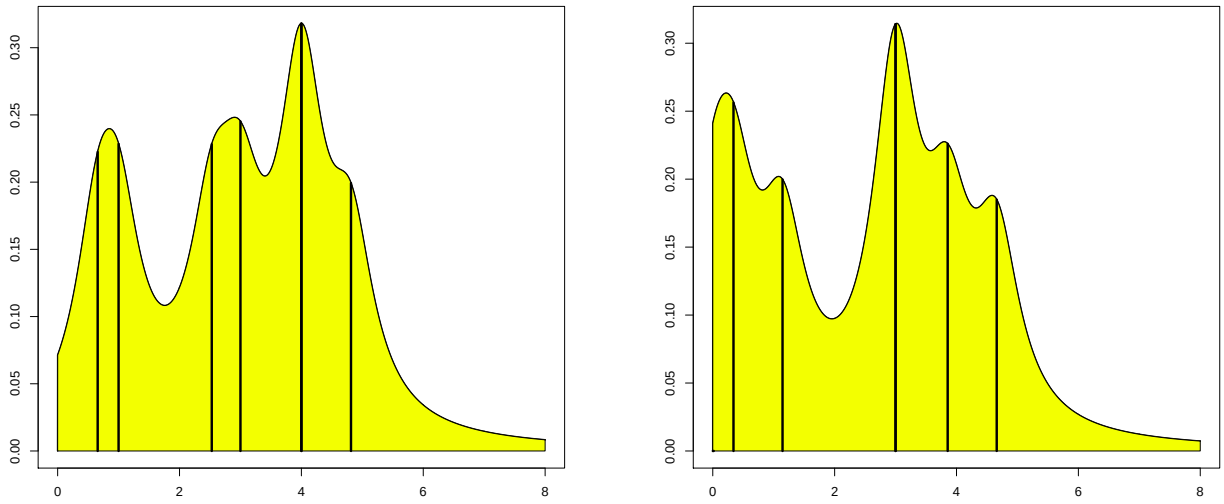


Figure 2.4: Adjacency matrix and graphical representation of I_2

2.2 Biological Networks

Citing Barabasi in [16], "We will never understand the workings of a cell if we ignore the intricate networks through which its proteins and metabolites interact with each other". In fact, all elements of a cell, from the genes in the DNA to the molecules involved in the signal transduction mechanisms, are deeply interconnected at various levels: all these elements and their connections are described by all the structures known as biological networks. The need for adopting a novel approach to mine the underlying knowledge is nowadays shared by the entire community of researches, as well as the need for a common new language to benefit from contributions from different disciplines [90].

For an exhaustive description of the biological networks, we refer to [158,



$\rho_{I_1}(\omega, \overline{\gamma})$ $\rho_{I_2}(\omega, \overline{\gamma})$

Figure 2.5: Lorentzian distribution of the Laplacian spectra for I_1 and I_2. Vertical lines indicate eigenvalues.

153, 27]; hereafter we recall some basic facts and properties.
Networks in biology can be grouped under a few major categories:

- Gene Regulatory (or Transcriptional) Network: it is the structure representing the mutual interactions (RNA and protein expression products) within a cell of a collection of DNA segments through their RNA and protein expression products), thus regulating the rates at which genes in the network are transcribed into mRNA. Some of the interacting factors serve only to activate other genes, and they are called the transcription factors.

- Proteinprotein interaction network: it is the structure (called interactome) collecting all the binding occurring between proteins in a organism.

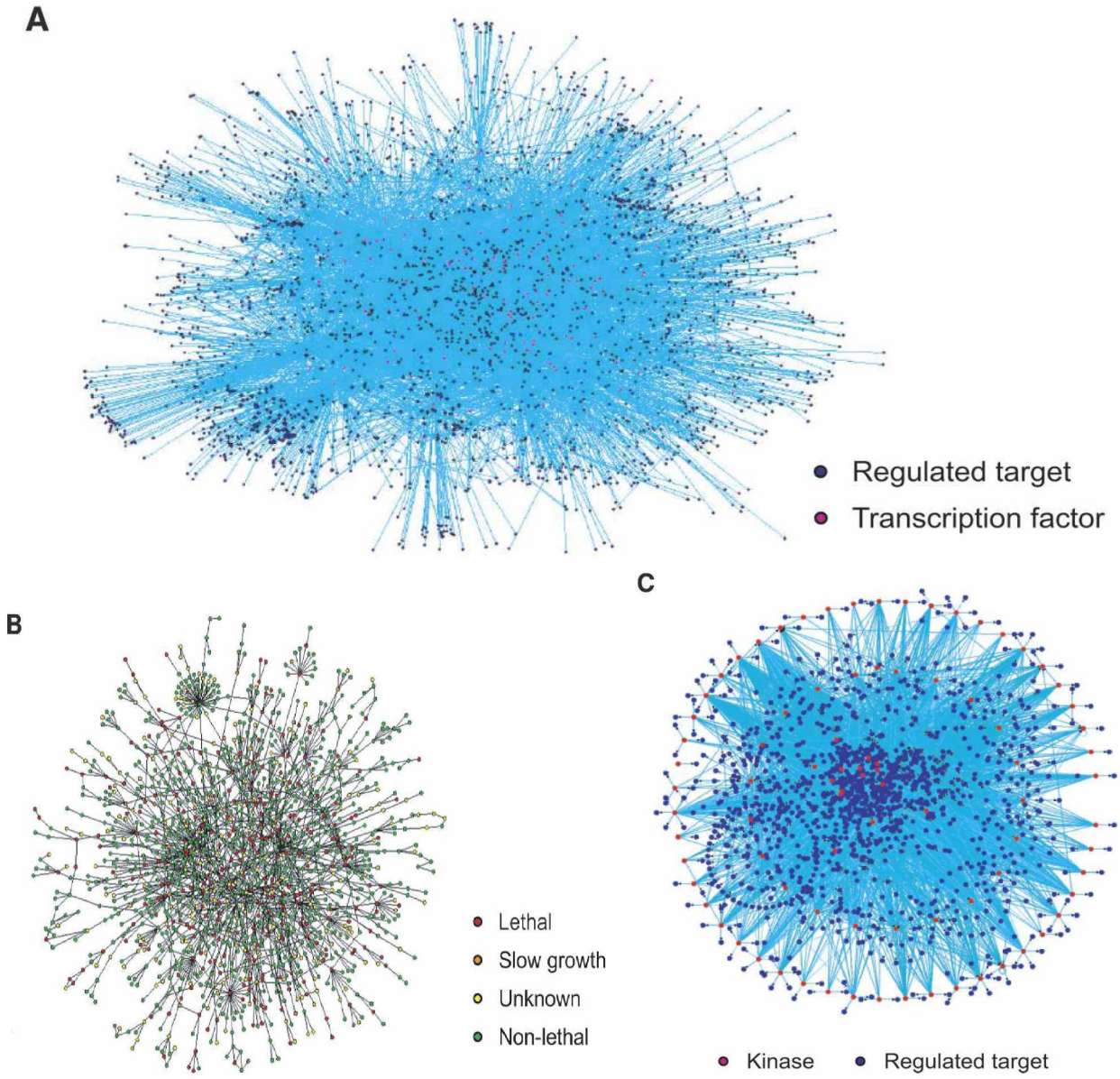
- Protein phosphorylation network collects all the regulation of proteins

by phosphorylation.

- Metabolic interaction network (or metabolic pathway): includes the chemical reactions of metabolism and the regulatory interactions that guide these reactions, thus collecting all metabolic and physical processes that determine the physiological and biochemical properties of a cell.

- Signalling network: it is the network of reactions that govern how a cell responds to its environment, together with the corresponding dynamic flow through the network (transduction) (*e.g*, from a receptor to a transcription factor that modifies expression of a gene).

A graphical display of the five above categories is shown in Fig. 2.2, originally included in [158].
Although the above networks are very diverse and heterogeneous, they all share a few key characteristics. One of the most powerful empirical rules derived by biological observations is that their topology is sparse: there is a small constant number of edges per node, much smaller than the total number of nodes. For instance, genes are regulated by a constant number of other genes (2-4 in bacteria, 5-10 in eukaryotes). Recent studies have shown that the frequency distribution of connectivity of nodes in biological networks tends to be long tailed, similar to a power-law distribution. Thus, biological networks are modeled according to a scale-free distribution: $P(k) = k^{-\gamma}$, where k is the degree (number of connections) and γ is some network-specific constant. The scale-free nature of gene networks yields the emergence of hubs (highly connected nodes) which are central in the network and are responsible for a large amount of overall regulation. Thus, the rest of the nodes are connected by very short paths, yielding overall short longest-path between nodes. This handful of highly connected nodes also support network integrity, making networks


A
Regulated target
Transcription factor
B
Lethal
Slow growth
Unknown
Non-lethal
C
Kinase
Regulated target

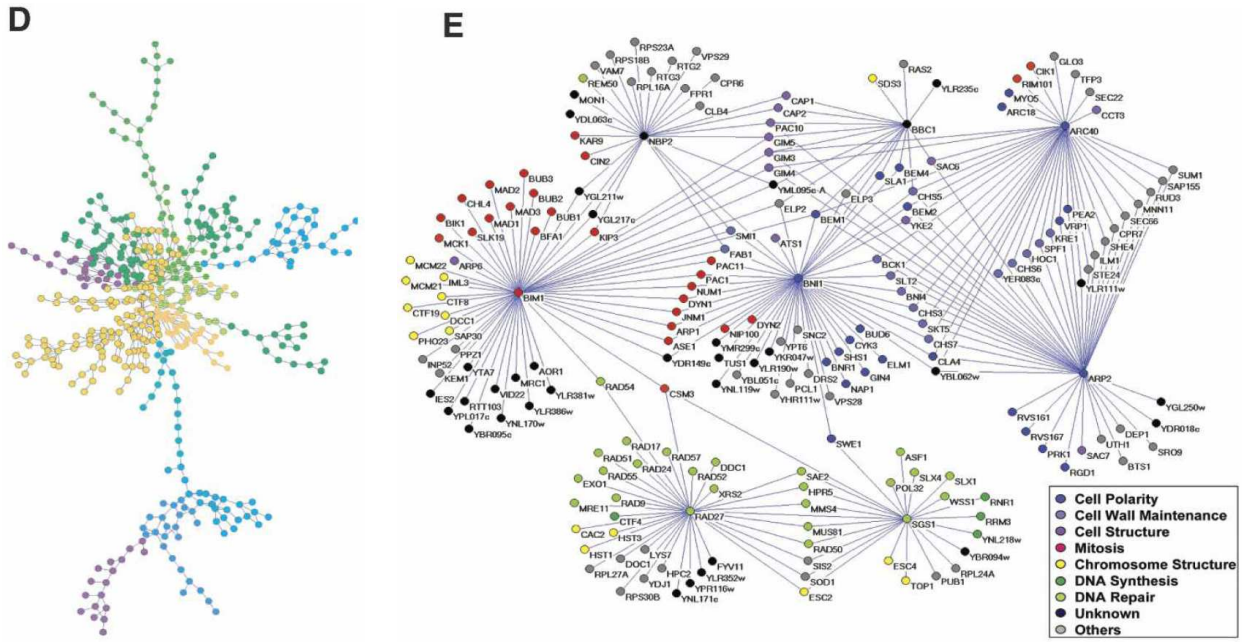


Figure 2.6: Examples of the five major biological networks. (A) A yeast transcription factor-binding network, composed of known transcription factor-binding data collected with large-scale ChIPchip and small-scale experiments. This figure was generated with the program Pajek [39]. (B) A yeast proteinprotein interaction network, containing proteinprotein interactions identified by yeast two-hybrid and protein complexes identified by affinity purification and mass spectrometry [17]. (Reprinted by permission from Macmillan Publishers Ltd: Nature [69], 2001.) Nodes are colored according to the mutant phenotype. (C) A yeast phosphorylation network comprised primarily of in vitro phosphorylation events identified using protein microarrays [117]. The figure was generated with Osprey 1.2.0. [25]. (D) An *E.Coli* metabolic network with 574 reactions and 473 metabolites colored according to their modules (Reprinted by permission from Macmillan Publications Ltd: Nature [58], 2005). (E) A yeast genetic network constructed with synthetic lethal interactions using SGA analysis on eight yeast genes (From [139]; reprinted with permission from AAAS). Nodes are colored according to their YPD cellular roles [taken from [158]].

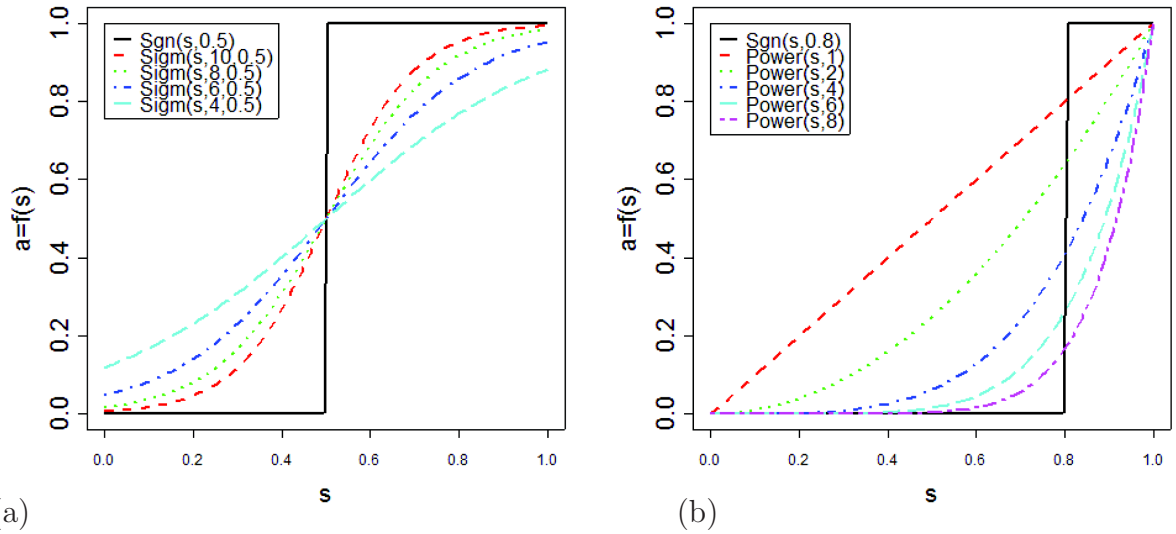
robust against random failures but exceedingly vulnerable upon targeted attack. Biological networks are very robust to fluctuations of their parameter values and there are strong indications that only specific topologies can guarantee such robustness. In fact, the resistance to noise is one of the main effects of the intrinsic robustness of networks to random fluctuations (for instance, of the concentration of regulators) and it is an important feature also when considering the modelling process. This is a fundamental characteristic as the input to the modelling process are observations of a biological phenomenon that are typically very noisy. As observed by Wuchty and colleagues in [153] all these properties are biologically grounded by the fact that many mutations have little or no phenotypic effect, which is coherent with the occurrence of genes that either cannot propagate their failure or whose function can be taken care of by different part is of the network. On the other hand, the presence of genes supporting multiple signaling and thus responsible for widespread changes upon their failure proves the crucial role of hubs.

2.3 Network Inference

As observed by Hurley and coworkers in [64], in the last five years the number of published algorithms for reconstructing a biological network from high-throughput measurements has grown exponentially, and they have helped unveiling significant biological findings in several species, from simple organisms to humans. The nature of the proposed algorithms is very heterogeneous, ranging from algebra, to differential equations to probability: see for instance [40, 102, 111] for some comparative reviews. However, no single method has emerged as the best performing across a wide range of tasks, as shown for instance by the outcome of the various editions of the DREAM challenge [131, 132, 116, 100, 115]; in particular, the main

conclusion drawn from in the last edition is that, in average, the integration of results coming from different methods can be an effective strategy [99].

In this thesis, we will mainly deal with two kinds of methods: the former aims at detecting interactions as nodes' coexpression, while the latter tries to spot also indirect dipendencies. As described in [102] and proved on a wide range of situations in [3], coexpression networks are biologically sound structures in describing complex interactions. They are constructed by computing a similarity score for each pair of genes (as weighted networks), or reduced to unweighted graphs after thresholding the similarity above a certain value. The underlying rationale, called the guilt-by-association heuristic, is the assumption that if two genes show similar expression profiles, they are supposed to follow the same regulatory regime, *i.e.*, coexpression is a reasonable approximation of coregulation. The Weighted Gene Coexpression Network Analysis (WGCNA) and the Topological Overlap Matrix (TOM) approaches described hereafter follow this line, and different correlation measure can be used within its framework (see Sec. 2.3.1). However, coexpression networks cannot distinguish direct from indirect dependencies based on the similarity of expression patterns: for example, a graph of three nodes X, Y, Z mutually connected by coexpression can match different regulatory schemes $X \to Y \to Z$, $X \to (Y, Z)$ or even $W \to (X, Y, Z)$ for an external node W. To deal with this issue other methods have been developed: among these, Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) and Context Likelihood of Relatedness are probably the most widely used by researchers worldwide. We provide in the following section a brief description of ARACNE and CLR, together with the description of a novel method called Reverse Engineering Gene Networks by Artificial Neural Networks (RegnANN) aimed at detecting indirect interactions with higher stability.



(a) (b)

Figure 2.7: Adjacency functions for different parameter values. a) Sidgmoid and signum adjacency functions. b) Power and signum adjacency functions. The value of the adjacency function (y-axis) is plotted as a function of the similarity (co-expression measure). Note that the adjacency function maps the interval $[0, 1]$ into $[0, 1]$. [61, 154]

2.3.1 Weighted Gene Coexpression Network Analysis

WGCNA [154, 88] is a general framework for "soft" thresholding that weights each connection by a number in $[0, 1]$. In gene coexpression networks, each gene corresponds to a node. With the aim of retrieving the adjacency matrix of the network starting from the data, one needs first to define a measure of similarity between the expression profile of two genes. In general the similarity measure s quantifies the level of connection between the two measured gene profiles. Applying s to any possible couple (i, j) of genes in the dataset we obtain the $n \times n$ matrix $S = [s_{ij}]$. The next step is to transform the S matrix into an adjacency matrix $A = [a_{ij}]$ that encodes the actual connection strength between each pair of nodes. To perform this task one can use an adjacency function which transforms the co-expression similarities into connection strengths. The parameters of this function are derived both from statistical and biological criteria.

In the WGCNA pipeline at this point the resulting adjacency matrix is used to define a measure of node distance used to define network modules through a clustering phase. At this point various intramodular and intermodular features can be computed such as for example the intramodular connectivity that helps determine the significance of a module [61].
In this thesis we chose to use as function s the absolute value of the Pearson correlation $s_{ij} = |cor(i,j)|$ or the Maximal Information Coefficient measure (MIC). The only constraint for the similarity measure is that to be bounded in $[0,1]$. The adjacency function is a monotonically increasing function that maps the interval $[0,1]$ into $[0,1]$. We can divide the adjacency functions into two main families: soft thresholding and hard thresholding functions; as the names suggest the former functions produce weighted adjacency matrices while the second ones produce (binary) unweighted adjacency matrices. The most widely used adjacency function is the signum function that applies a hard threshold to the similarity values. The application of this function implies the very delicate choice of the parameter τ so that:

$$a_{ij} = \text{sign}(a_{ij}, \tau) \equiv \begin{cases} 1 & if \quad s_{ij} \geq \tau \\ 0 & if \quad s_{ij} < \tau \end{cases}$$

It is obvious that an erroneous choice of the parameter τ can lead to a loss of information, since, for instance setting $\tau = 0.7$ means that a value of $cor = 0.69$ would lead to no link at all in the final adjacency matrix. To avoid hard thresholding, in [154] two soft thresholding methods are proposed: the sigmoid function

$$a_{ij} = \text{sigmoid}(s_{ij}, \alpha, \tau_0) \equiv \frac{1}{1 + e^{-\alpha(a_{ij} - \tau_0)}} \ ,$$

with parameters τ_0 and α and the power adjacency function

$$a_{ij} = \text{power}(s_{ij}, \beta) \equiv |s_{ij}|^{\beta} \ ,$$

with the only parameter β. For the experiments in this thesis we decided to use the power adjacency function since the parameter β can be chosen to approximate the sigmoid function. Another advantage of the use of the power function is that if s_{ij} factors so that $s_{ij} = s_i s_j$ then $a_i j$ factors as well, $a_{ij} = a_i a_j$ with $a_i = (s_i)^\beta$. In [154] is shown that the two adjacency functions produce very similar results provided that the parameters are chosen following the scale-free topology criterion. The most critical step in this approach is the choice of the parameter τ or β depending on the choice of the functions. The choice of the parameters determines the sensitivity and specificity of the pairwise connection strengths. For example if τ is set too low we could incur in too many false positive links in the matrix reconstruction because of the effect of noisy data. On the other hand, if τ is set to high we will have an adjacency matrix too sparse, and thus we lose important information about the structure of the connections. In order to solve this problem several approaches have been applied in the literature to threshold the significance level of the correlation instead of the correlation coefficient itself. The significance level of a correlation coefficient can be estimated by using the Fisher transformation. Thus thresholding a correlation coefficient is replaced by thresholding the corresponding p-value. Finally, instead of focusing on the significance of the correlation or the network size, we propose to pick the threshold by making use of the fact that despite significant variation in their individual constituents and pathways, metabolic networks have been found to display approximate scale free topology [70, 119, 61].

2.3.2 Topological Overlap Matrix

The topological overlap of two nodes reflects their relative interconnectedness: although this method is only indirectly involving co-expression, we still list TOM under the relevance network umbrella definition. In partic-

ular the Topological Overlap of the couple of nodes (i, j) is:

$$\omega_{i,j} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$ and $k_i = \sum_u a_{iu}$ is the node connectivity. In case of hard thresholding we have that l_{ij} equals the number of common neighbors of the nodes i, j that are connected. The Topological Overlap $\omega(i, j)$ equals one if the node with lower connectivity is connected with a set of nodes that are also neighbors of the other node and i and j are directly connected. On the other hand we have that $\omega(i, j)$ is zero in case that i and j are disconnected and they share no neighbors. The formula of TOM is generalizable to weighted adjacency matrices just using the weighted $0 \leq a_{ij} \leq 1$ in the formula above. Moreover since $l_{ij} \leq min(\sum_{u \neq j} a_{iu}, \sum_{u \neq i} a_{ju})$ it follows that $l_{ij} \leq min(k_i k_j) - a_{ij}$ therefore if $0 \leq a_{ij} \leq 1$ then $0 \leq l_{ij} \leq 1$. The topological overlap matrix $\Omega = [\omega_{ij}]$ is a similarity measure [78] since it is non-negative and symmetric [154].

2.3.3 Aracne

Aracne is a method originally written to cope with the complexity typical of the regulatory networks of the mammalian cells. It is anyways able to address more general deconvolution problems such as transcriptional and metabolic networks. This technique has been designed especially to avoid the problem of false positive which affects the great part of algorithms based on co-expression. Applying the Data Processing Inequality (DPI), Aracne can remove the majority of indirect links [101, 109, 35]. In this thesis we used the algorithm implementation provided on Bioconductor [104] and the default tolerance values for DPI were used. As often happens, Aracne makes use of a hard thresholding for the binarization of the resulting adjacency matrix. As many other methods, ARACNE relies on the definition of a threshold for the binarization of the adjacency matrix.

In absence of a good heuristic for defining such threshold, on the synthetic data-sets we will adopt the area under the curve (AUC) as performance metric.

2.3.4 CLR

CLR is based on the mutual information score and can be seen as an evolution of the class of the relevance network algorithms [46] designed to predict the relations between transcription factors and target genes. The evolution of CLR stands in an additional step of background correction added to the phase of mutual information estimation. At first for each gene a statistical likelihood of the mutual information score is computed with respect to its network contest. Then for each couple Transcription Factor-Target Gene, the mutual information score is compared to the context likelihood of both the elements and turned into a z-score. In this thesis we used the implementation presented in [104]. As in the case of ARACNE, in absence of a good heuristic for defining a binarization threshold for the inference of the adjacency matrix, on the synthetic data-sets we will adopt the area under the curve (AUC) as performance metric.

2.3.5 RegnANN

RegnANN [56] reconstructs networks through an ensemble of feed-forward multilayer perceptrons. Each member of the ensemble is essentially a multi-variable regressor (one to many) trained using an input expression matrix to learn the relationships (correlations) among a target gene and all the other genes. Formally, let us consider the multilayer perceptron as in Fig: 2.8 (right):1 input neuron I, 1 layer of H hidden units and 1 layer of K output units. Indicating with g the activation function of each unit and $w_{h,k}$ the weights associated with the links between the output layers and

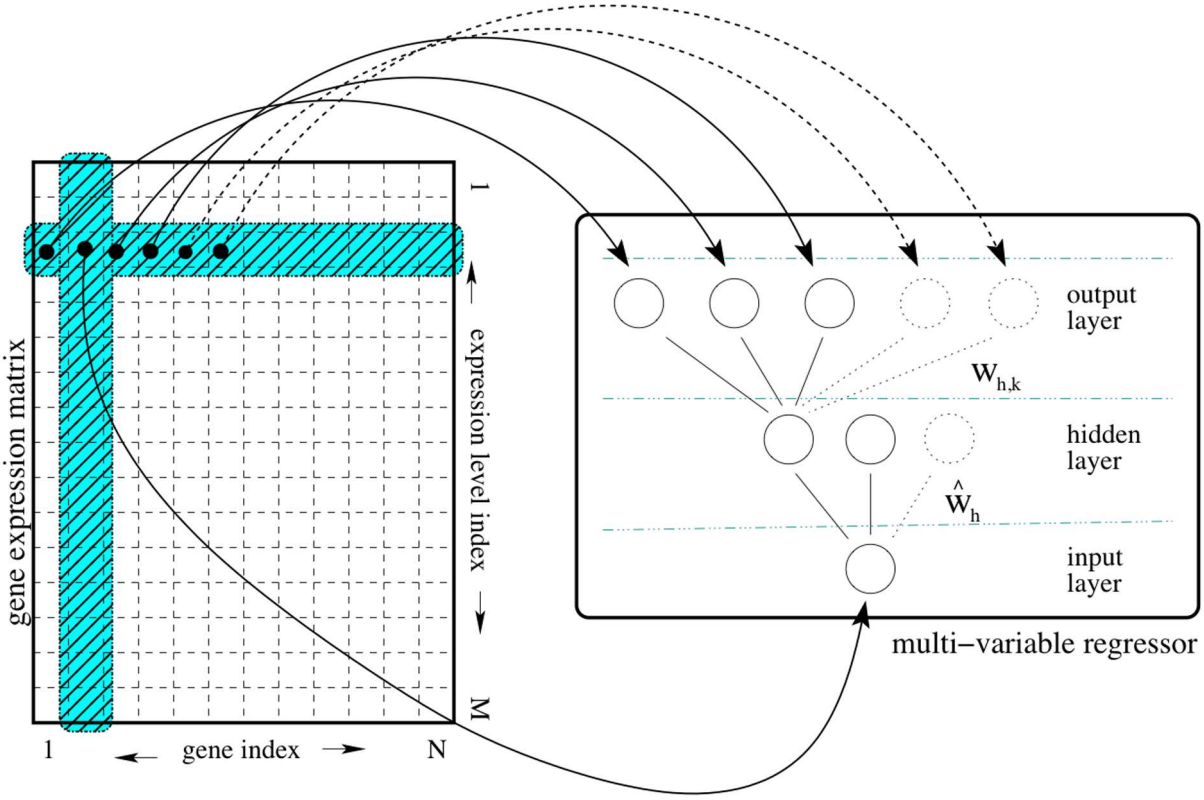


Figure 2.8: The ad hoc procedure proposed to build the training input/output patterns starting from a gene expression matrix.Each input pattern corresponds to the expression value for the selected gene of interest.

the hidden layer and with $\hat{w}_h$ the weights of the links between input neuron and hidden layer, the value O_k for the output unit k can be calculated as follows:

$$O_k = g\left(\sum_{h=1}^{H} w_{h,k} \cdot g(\hat{w}_h \cdot I)\right)$$

The value $O_k$is the inferred interaction between the corresponding gene k and the gene associated with the input neuron I. We proceed in determining the interactions among genes separately and then we join the information to form the overall gene network. From each row of the gene expression matrix we build a set of input and output patterns used to train with back-propagation [22] a selected multilayer perceptron. Each

input pattern corresponds to the expression value for the selected gene of interest. The output pattern is the row-vector of expression values for all the other genes for the given row in the gene expression matrix (Figure 2.8). By cycling through all the rows in the matrix, each regressor in the ensemble is trained to learn the correlations among one gene and all the others. Repeating the same procedure for all the columns in the expression matrix, the ensemble of multi-variable regressors is trained to learn the correlations among all the genes. The procedure of learning separately the interactions among genes is very similar to the one presented in [127], where the authors propose to estimate the neighborhood of each gene (the correlations among one gene and all the others) independently and then joining these neighborhoods to form the overall network, thus reducing the problem to a set of identical atomic optimizations.

We build N (one for each of the N genes in the network) multilayer perceptrons with one input node, one layer of hidden nodes and one layer of $N-1$ output nodes, adopting the hyperbolic tangent as activation function. The input node takes the expression value of the selected gene rescaled in $[-1, 1]$. The number of hidden nodes is set to the square root of the number of inputs by the number of outputs. This value is to be considered a rule of thumb granting enough hidden units to solve the regression problem and allowing dynamical adaptation of the structure of RegnANN to the size of the biological network under study. The output layer provides continuous output values in the range $[-1, 1]$.

The algorithm of choice for training each multi-layer perceptron is the back-propagation algorithm [22]. The back-propagation is a standard algorithm for learning feed-forward multilayer perceptrons that essentially looks for the minimum of the error function in the weight space using the method of gradient descent. The error function is defined as the difference between the output of each neuron in the multilayer perceptron and its expected value.

The back-propagation algorithm starts with the forward-propagation of the input value in the multilayer perceptron, followed by the backward propagation of the errors from the output layer toward the input neuron. The algorithm corrects the weight values according to the amount of error each unit is responsible for. Formally, the weight values at learning epoch τ are updated as follows:

$$\Delta w^{(\tau)} = -\eta \nabla E + \mu \Delta w^{(\tau-1)}$$

To keep the notation simple w refers to both the weights associated with the links between the output layers and the hidden layer and with the weights of the links between input neuron and hidden layer. ∇E refers to the gradient of the error in weight space. η is the learning rate;μ is the momentum.

Although back-propagation is essentially a heuristic optimization method and alternatives such as Bayesian neural network learning [108] have more sound theoretical basis, in the proposed multi-variable regression schema the simple back-propagation algorithm allows us to design a far less complex system. This is due to how Bayesian neural network learning handles the regression problem. As indicated in [107]: Networks are normally used to define models for the conditional distribution of a set of target values given a set of input values.[...]. For regression and logistic regression models, the number of target values is equal to the number of network outputs. This implies that in the case of Bayesian learning an extra procedure is required to discretize the target values from the continuous range [-1,1] and that for each ensemble member the layer of output neurons ($N-1$ in the case of back-propagation) has to be translated into a matrix of neurons of size $(N-1) \times T$, where T is the number of desired target values. Accordingly, also the hidden layer becomes a matrix of neurons, each one with its own set of parameters. Thus, in the context of multivariable regression,

adopting back-propagation allows us to design a lower complexity inference system limiting issues related to high dimensional settings. Once the ensemble is trained, the topology of the gene regulatory network is obtained by applying a second procedure. Considering each gene in the network separately, we pass a value of 1 to the input neuron of the correspondent multilayer perceptron, consequently recording its output values. The continuous output values in the range [-1,1] represent the expected normalized expression values for the other genes (its neighborhood). This procedure basically aims at verifying the correlation between the input gene and all the others: assuming the input gene maximally expressed (the value 1), an output value of 1 indicates that the correspondent gene will be also maximally expressed, thus indicating perfect correlation between the two genes. An output value of -1 indicates that the correspondent gene will be maximally under-expressed: perfect anti-correlation of the two genes. Thus, the continuous output values in the range [-1,1] are interpretable in terms of positive correlation (> 0), anti-correlation (< 0) and no-correlation (0). By cycling this procedure through all the ensemble members in the regression system, we obtain N (one for each of the N genes in the network) vectors of length $N - 1$ of continuous values in [-1,1]. The correlation matrix is obtained by correctly joining the N vectors. It is important to note that all the values of the diagonal of the adjacency matrix are equal to 0 by construction: this procedure does not allow discovering of gene self correlation (regulation) patterns, but only correlation patterns among different genes. Finally, the adjacency matrix of the sought gene network is obtained by thresholding the correlation coefficients.

2.4 Correlation Measures

2.4.1 Pearson

In statistics the Pearson correlation index between two variables is a measure of the linearity between the variables and it is computed as the ratio of their covariance and the product of the respective standard deviations. Given two variables x and y their Pearson correlation is thus defined as follows:

$$\rho_{xy} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

where σ_{xy} is the covariance of the variables while σ_x and σ_y the two standard deviations. The value of ρ_{xy} ranges in $[-1, 1]$: when ρ_{xy} is greater than 0 the two variables are said to be directly correlated, if ρ_{xy} is smaller than 0 then x and y are inversely correlated. If ρ_{xy} equals 0 then the variables are uncorrelated. Pearson indexes of n variables can be collected in a squared matrix of dimension $[n \times n]$ which will be symmetric and with the diagonal equal to 1 since $\rho_{ij} = \rho_{ji}$ and $\rho_{ii} = \frac{\sigma_{ii}}{\sigma_i^2}$.

2.4.2 Biweight Midcorrelation

In order to overcome the problem of outlaiers in Pearson correlation in [152] is proposed the bicorrelation which is considered to be a good alternative to the standard correlation. Such algorithm was also applied in [128] by Song and coworkers proving that, using as reference the gene ontology enrichment, the bicorrelation coupled with TOM performs better than a MIC based approach in the detecting of submodules.
To define the biweight midcorrelation of two variables $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$ we first define u_a and u_b with $i = 1, \ldots, m$:

$$u_a = \frac{x_a - med(x)}{9mad(x)} u_b = \frac{y_b - med(y)}{9mad(y)} \tag{2.1}$$

where $med(x)$ stands for the median of x and $mad(x)$ is the median absolute deviation of x this allow us to define the weight w_a of x_a as:

$$w_a^{(x)} = (1 - u_a^2)^2 I(1 - |u_a|) \tag{2.2}$$

the indicator $I(1|u_a|)$ equals 1 if $1|u_a| > 0$ and 0 otherwise. Therefore, w_a ranges from 0 to 1. It also decreases as x_a gets away from $med(x)$, stays at 0 when x_a differs from $med(x)$ by more than $9mad(x)$. Given that we can define the analogous weight for u_b we can define the biweight midcorrelation for the variables x and y as:

$$BiCor(x,y) = \frac{\sum_{a=1}^{m}(x_a - med(x))w_a^{(x)}(y_a - med(y))w_a^{(y)}}{\sqrt{\sum_{b=1}^{m}[(x_b - med(x))w_b^{(x)}]^2}\sqrt{\sum_{c=1}^{m}[(y_c - med(y)w_c^{(y)})]^2}} \tag{2.3}$$

In this thesis a modified version of *bicorrelation* implemented in WGCNA (R package [88, 154]) is used. Setting a coherent thereshold θ one can say that a value of $BiCor(x,y) > \theta$ indicates that the genes described by the variables x and y are similarly expressed. [86]

2.4.3 Maximal Information Coefficient

Maximal Information Coefficient (MIC) is one of the five similarity measures between variables originally introduced as the MINE (Maximal Information-based Nonparametric Exploration) in the paper [120] based on the intuition that if two variables are somehow bond by a relationship then it is possible to encapsulate their scatterplot within a grid. The calculation of MIC consists in the exploration of all the possible subdivisions of the scatterplot up to the maximum resolution of the grid. This resolution is dependent from the samplesize of the considered data (See Fig:2.9). For every pair of integers (x, y) the largest possible mutual information is computed by applying an $x - by - y$ grid to the scatterplot of the two variables. In order

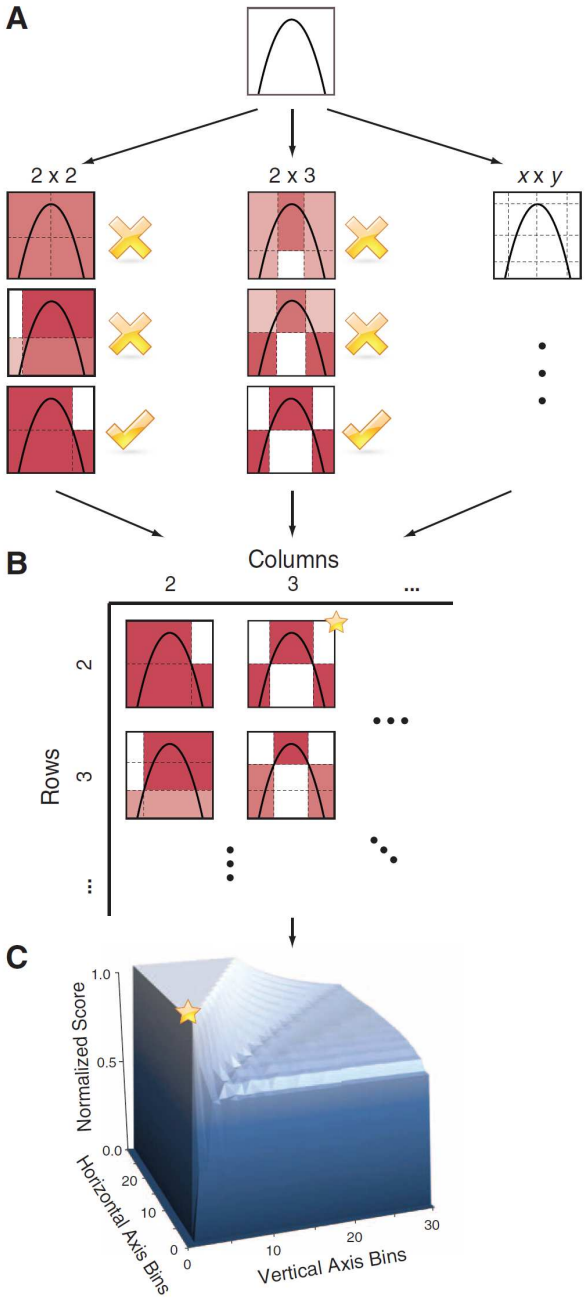


Figure 2.9: Computing MIC (A) For each pair (x, y), the MIC algorithm finds the x-by-y grid with the highest induced mutual information. (B) The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized score. (C) The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface. In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score, and the star in (C) marks that grid's corresponding location on the surface. [taken from [120]]

to obtain values comparable within grids of different dimensions we normalize the measured values obtaining the normalized values in the range [0,1]. We define the characteristic matrix $M = (m_{x,y})$, where $m_{x,y}$ is the highest normalized mutual information achieved by any $x - by - y$ grid, and the corresponding MIC statistic as the maximum value in M (Fig. 2.9, B and C). More formally, for a grid G, let I_G denote the mutual information of the probability distribution induced on the boxes of G, where the probability of a box is proportional to the number of data points falling inside the box. The (x, y)-th entry $m_{x,y}$ of the characteristic matrix equals $\max I_G / \log \min x, y$, where the maximum is taken over all x-by-y grids G. MIC is the maximum of $m_{x,y}$ over all the ordered pairs (x, y) such that $xy < B$, where B is a function of sample size; we usually set $B = n^{0.6}$. Every entry of M falls in the range $[0, 1]$, and so MIC does as well. MIC is also symmetric [i.e., $MIC(X, Y) = MIC(Y, X)$] due to the symmetry of mutual information. Since I_G depends only on the rank order of the data, MIC is invariant under order-preserving transformations of the axes. Notably, although mutual information is used to quantify the performance of each grid, MIC is not an estimate of mutual information. To calculate M, we would ideally optimize over all possible grids. For computational efficiency, we instead use a dynamic programming algorithm that optimizes over a subset of the possible grids and appears to approximate well the true value of MIC in practice. [120]

2.5 Resampling Techniques

2.5.1 Bootstrap

Bootstrap methods depend on the notion of a bootstrap sample. Let F be the empirical distribution of the observed values $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, so $f \to (x_1^*, x_2^*, \ldots, x_n^*)$ where $(x_1^*, x_2^*, \ldots, x_n^*)$ are a randomized or resampled

version of $(x_1, x_2, \ldots, x_n)$. Thus we might have $x_1^* = x_7, x_2^* = x_4, \ldots, x_n^* = x_6$. The bootstrap dataset or bootstrap resample $(x_1^*, x_2^*, \ldots, x_n^*)$ consists of members of the original dataset $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, with some of the samples taken zero, one or more times. Now suppose we wish to estimate a parameter of interest $\theta = t(F)$ on the basis of $\boldsymbol{x}$. For this purpose we can calculate an estimate $\hat{\theta} = s(\boldsymbol{x})$ from $\boldsymbol{x}$. Now we can calculate how accurate $\hat{\theta}$ is by computing:

$$\hat{\theta}^* = s(\boldsymbol{x}^*) \tag{2.4}$$

where the quantity $s(\boldsymbol{x}^*)$ is the result of applying the same function $s(\cdot)$ to $\boldsymbol{x}^*$ as we applied to $\boldsymbol{x}$.

So the bootstrap estimate of the *standard error* $se_{\hat{F}}(\hat{\theta})$ is an estimate that uses the $\hat{F}$ function in place of the unknown distribution F. So the bootstrap estimate of $se_{\hat{F}}(\hat{\theta})$ is defined by

$$se_{\hat{F}}(\hat{\theta}^*) \tag{2.5}$$

In practice the bootstrap estimate of $se_F(\hat{\theta})$ is the standard error of $\hat{\theta}$ for data sets of size n randomly sampled from $\hat{F}$ [45, 37].

In particular in this thesis the bootstrap empirical distribution is widely used to compute the confidence intervals for the presented results. We compute 95% *bootstrap confidence intervals* of the data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ by producing 1000 bootstrap resampling $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{1000})$ of the data and setting the confidence intervals as the lower and upper bounds that contain 950 of the means $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \ldots, \boldsymbol{\mu}_{1000})$ of the resamples.

2.5.2 Cross validation

Random cross validation is technique commonly used to honestly predict how a statistical analysis will work on an independent dataset. It is widely used in machine learning where the prediction is the main goal and typically one wants to know how a specific prediction model will behave on

an independent validation dataset. In particular it is interesting to predict what the accuracy of the system is on data different from the ones used for the test.

Suppose we have a prediction model with one or more unknown parameters. Typically a fitting process is used to find the set of parameters for which the best classification accuracy is reached. If we now apply the same model with the best computed parameters to an unknown validation dataset we find that the model does not fit the new data as well as for the training set. This phenomenon is known as overfitting and is very common when the training set is not big enough.

Cross-validation is a technique aimed at obtaining a plausible estimation of the accuracy of the classification model even without having an explicit independent validation data-set. One of the most used cross-validation setups is the **k-fold cross-validation** which consists of a preliminary partition of the whole dataset in k groups of the same number of samples. The training and test phase is performed k times and each time a different one of the k groups is discarded from the training-set to be used as test-set. The collection of k accuracy values obtained in this way is considered an honest prediction of the accuracy of the model. The reliability of the cross-validation depends mainly on the numerosity of the samples and the chosen k.

Chapter 3

Quantitative Network Comparison

In this chapter we review and benchmark a class of methods that tackle the problem of structural comparison between networks, with particular attention to the biological case, *e.g.* gene regulatory and protein networks. As mentioned in Section 2.1 a complete network is defined in the literature [23] as a graph with a structure that dynamically evolves in time. In terms of structure the term "complex" was introduced by Strogatz to identify non "regular" networks like chains, grids, lattices and fully-connected graphs. We can think of the extreme complex network as a completely random graph. In real applications the observable networks lie in between the two extremes, normally more on the random side [133] [97].
The problem of network comparison has been tackled in many different fields over the last years. A number of solutions have been proposed with a wide variety of approaches ranging from statistical physics to machine learning [43] [54]. In this chapter we present a brief review of classes of methods designed to solve the problem of comparing structure between networks. [1] [110] [93] [84] [34]

3.1 Global and Local Distances

A main discriminant factor among approaches is their globality or locality. The former takes into account the overall structure of the network, for instance using a function of the eigenvalues of one of the connectivity matrices of the graph of the network. Measures in this family are called spectral distances; by definition these distances can not distinguish isospectral graphs. The latter set of distances are also known as edit-like and they give a quantitative measure of the diversity between two networks as a function of the number of link operations needed to transform a graph into the other. Even if different weight (cost) strategies can be applied to make this approach more sophisticated its is focus on single-link variations overlooking the overall structure of the network.
Cost-based functions stem from the parallel theory of graph alignment: the edit distance and its variants use the minimum cost of transformation of one graph into another by means of the usual edit operations insertion and deletion of links. Other classical network comparison measures are those based on the confusion matrix, such as the pairs Precision/Recall or Sensitivity/Specificity, or the F-score. However, all these measures evaluate only the number of detected/undetected links, without considering the difference between the global structure of the inferred and the real topology: deep structural differences can occur with the same confusion matrix. Again, all these measures are local, because, for each link, only the structure of its neighborhood gives a contribution to the distance value, while the structure of the whole topology is not considered.

To overcome the locality issue in network comparison, a few global distances have been proposed: among them, the family of structural measures are particularly relevant. The label "structure-based" distance groups all other measures that do not rely on cost functions or characteristic features.

Structural analysis is of central importance in computational biology [84]. Structure and structural properties of networks have been studied in a wide variety of fields in science [23, 1, 110] ranging from statistical physics to machine learning [43, 54]. One notable example in this family is the recently proposed use of Ihara ζ-functions for network volume measurements [124, 125]. Remarkably, equivalence of some structure-based distance and the edit distance has been proven [29].

The family of spectral measures, which is investigated in this paper, is also part of the group of structure-based distances. As the name suggests, their definition is based on (functions of) the spectrum of one of the possible connectivity matrices of the network, *i.e.* its set of eigenvalues. Although the idea of using spectral measures for network comparison is quite recent, the theory of graph spectra started in the early 50's and since then many of its aspects have been deeply studied [32, 146], including a first classification of networks [14]. The spectral theory has been also recently applied to biological networks [13, 15], where the properties of being scale-free (the degree distribution following a power law) and small-world (most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps) are particularly evident. The idea of using spectral measures for network comparison is instead only recent and it relies on similarity measures that are functions of the network eigenvalues. However, it is important to note that, because of the existence of isospectral networks, all these measures are indeed distances between classes of isospectral graphs: they are relatively rare (especially in real networks) and qualitatively similar [59]. Estimates (also asymptotic) of the eigenvalues distribution are available for complex networks [121].

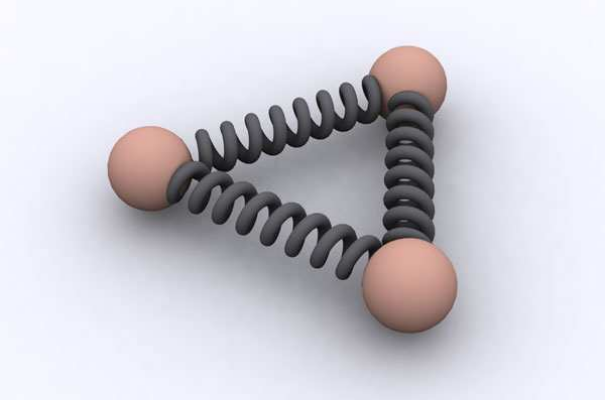
3.2 Spectral Similarity Measures

As mentioned in Section 2.1 we propose a short review of a set of similarity measures based on the graph spectra analysis.
The first distance we take into consideration was originally introduced in [10, 68] as a measure of a graph's spectrum. Pincombe in [114] was the first to use D1 as an intra-graph distance to analyze changes in time-series of graphs. Here we consider G and H two graphs both having N nodes and let $\lambda_0 = 0 \leq \lambda_1 \leq \cdots \leq \lambda_{N-1}, \mu = 0 \leq \mu_1, \leq \cdots \leq \mu_{N-1}$ be the respective Laplacian spectra. Once we set the parameter $k \leq N$, the distance is defined as:

$$d_1(G,H) = \begin{cases} \sqrt{\dfrac{\sum_{i=N-k}^{N-1} (\lambda_i - \mu_i)^2}{\sum_{i=N-k}^{N-1} \lambda_i^2}} & \text{if } \sum_{i=N-k}^{N-1} \lambda_i^2 \leq \sum_{i=N-k}^{N-1} \mu_i^2 \\ \sqrt{\dfrac{\sum_{i=N-k}^{N-1} (\lambda_i - \mu_i)^2}{\sum_{i=N-k}^{N-1} \mu_i^2}} & \text{if } \sum_{i=N-k}^{N-1} \mu_i^2 < \sum_{i=N-k}^{N-1} \lambda_i^2 \end{cases} \tag{3.1}$$

Being the D1 non-negative, symmetric and separated we can indeed say that it is a metric.
A more recent spectral distance was presented by Ipsen and Mikhailov in [67] with the aim of reconstructing a graph starting from its spectrum making use of a stochastic process of mutations and selection. The idea behind D2 is to consider the N-nodes network as the composition of an N physical elements structure bonded by springs. In this model every element

Figure 3.1: Representation of the physical network model of D2 distance.



has the same mass, the springs have identical elastic properties and the pattern of connections is set by the adjacency matrix of the considered network see Fig. 3.1. The described dynamical system is described by the set of N differential equations

$$\ddot{x}_i + \sum_{j=1}^{N} A_{ij}(x_i - x_j) = 0 \quad \text{for } i = 0, \cdots, N-1 \ .$$

The vibrational frequencies ω_i are given by the eigenvalues of the Laplacian matrix of the network: $\lambda_i = \omega_i^2$, with $\lambda_0 = \omega_0 = 0$. For this reason in [32], the Laplacian spectrum is called vibrational spectrum. The spectral density for a graph as the sum of Lorentz distributions is defined as

$$\rho(\omega) = K \sum_{i=1}^{N-1} \frac{\gamma}{(\omega - \omega_k)^2 + \gamma^2}$$

where γ is the common width, the parameter which specifies the half-width at half-maximum (HWHM), equal to half the interquartile range. K is the normalization constant solution of

$$\int_0^\infty \rho(\omega) \mathrm{d}\omega = 1 \ .$$

Then the spectral distance ϵ between two graphs G and H with densities $\rho_G(\omega)$ and $\rho_H(\omega)$ can then be defined as

$$\epsilon(G,H) = \sqrt{\int_0^\infty [\rho_G(\omega) - \rho_H(\omega)]^2 \mathrm{d}\omega} \ . \tag{3.2}$$

Note that the two above integrals can be explicitly computed through the relation $\int \frac{1}{1+x^2}\mathrm{d}x = \arctan(x)$.

A simpler measure D3 was introduced in [157] for graph matching, using the graph edit distance as the reference baseline. The authors compute the spectrum associated to the classical adjacency matrix, laplacian matrix, signless Laplacian matrix $|L| = D + A$, and normalized Laplacian ($\mathcal{L}$) matrix. They also introduce two more functions: the path length distribution and the heat kernel h_t. The heat kernel is related to the Laplacian by the equation

$$\frac{\partial h_t}{\partial t} = -Lh_t \ ,$$

so that

$$h_t(u,v) = \sum_{i=0}^{N-1} e^{-\lambda_i t}\phi_i(u)\phi_i(v) \ ,$$

where λ_i are the Laplacian eigenvalues and ϕ_i the corresponding eigenvectors. For $t \to 0$, $h_t \to I - Lt$, while when $t \to \infty$ then $h_t \to e^{-\lambda_{N-1}t}{\phi_{N-1}}^T\phi_{N-1}$. By varying t different representations can be obtained, from the local ($t \to 0$) to the global ($t \to \infty$) structure of the network. Moreover, if $D_k(u,v)$ is the number of paths of length k between nodes u and v, the following identity holds:

$$h_t(u,v) = e^{-t}\sum_{i=0}^{N^2-1} D_k(u,v)\frac{t^k}{k!} \ ,$$

which allows the explicit computation of the path length distribution:

$$D_k(u,v) = \sum_{i=0}^{N-1}(1-\lambda_i)^k\phi_i(u)\phi_i(v) \ .$$

The proposed distance is just the Euclidean distance between the vectors of (ordered) eigenvalues (for a given matrix M) for the two networks being compared:

$$d_M(G,H) = \sqrt{\sum_{i=0}^{N-1} \left(\lambda_i^{(G,M)} - \lambda_i^{(H,M)}\right)^2}\,, \tag{3.3}$$

where $\lambda_{(T,M)}$ are the eigenvalues of the graph T w.r.t. the matrix M, where M is either a connectivity matrix, or the heat kernel matrix or the path length matrix. As a final observation, the authors claim that the heat kernel matrix has the highest correlation with the edit distance, while the adjacency matrix has the lowest.

A similar formula D4 is proposed in [33] as the squared Euclidean (L_2) between the vectors of the Laplacian matrix:

$$d(G,H) = \sum_{i=0}^{N-1} \left(\lambda_i^{(G,L)} - \lambda_i^{(H,L)}\right)^2\,. \tag{3.4}$$

The next and last two measures are based on the concept of spectral distribution.

The distance D5 is introduced in [48], aiming at comparing Internet networks topologies. Let f_λ be the (normalized Laplacian) eigenvalued distribution, and $\mu(\lambda)$ a weighting function and define a generic distance between graphs G and H as follows

$$d_{\mu,p}(G,H) = \int_\lambda \mu(\lambda)\left(f_{\lambda,G}(\lambda) - f_{\lambda,H}(\lambda)\right)^p \mathrm{d}\lambda\,.$$

The weighting function is then defined as $\mu(\lambda) = (1-\lambda)^4$, an approximation of the graph irregularity as defined in [32], while the usual Euclidean metric is chosen, so that $p = 2$: the exact formula thus reads

$$d(G,H) = \int_\lambda (1-\lambda)^4 \left(f_{\lambda,G}(\lambda) - f_{\lambda,H}(\lambda)\right)^2 \mathrm{d}\lambda\,. \tag{3.5}$$

Calculating the eigenvalues of a large (even sparse) matrix is computationally expensive; an approximated version is also proposed, based on estimation of the distribution f of eigenvalues by means of pivoting and Sylvester's Law of Inertia, used to compute the number of eigenvalues that fall in a given interval. To estimate the distribution K equally spaced bins in the range $[0, 2]$ are used, so that a weighted spectral distribution measure for a graph G can be defined for an integer $n > 0$ as follows:

$$\omega_n(G) = \sum_{k \in K} (1-k)^n f(\lambda = k) \ .$$

The generic formula can be now specialized to:

$$d_n(G, H) = \sum_{k \in K} (1-k)^n (f_G(\lambda = k) - f_H(\lambda = k))^2 \ , \tag{3.6}$$

a family of metrics parametrized by the integer N. The last spectral measure D6 in this review was presented in [12] and it employs two different divergence measures, Kullback-Leibler and Jensen-Shannon. The Kullback-Leibler divergence measure is defined on two probability distributions p_1, p_2 of a discrete random variable X as

$$\mathrm{KL}(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)} \ .$$

The Kullback-Leibler divergence measure is not a metric, because is not symmetric and it does not satisfy the triangle inequality. To overcome this problem, the author consider the Jensen-Shannon measure, which in some sense is the symmetrization of KL:

$$\mathrm{JS}(p_1, p_2) = \frac{1}{2}\mathrm{KL}\left(p_1, \frac{p_1 + p_2}{2}\right) + \frac{1}{2}\mathrm{KL}\left(p_2, \frac{p_1 + p_2}{2}\right) \ .$$

With this definition, the square root of JS is a metric. Thus, if f is the (normalized Laplacian) spectral probability distribution, a distance between two networks can be defined as

$$d(G, H) = \sqrt{\mathrm{JS}(f_G, f_H)} \ . \tag{3.7}$$

3.2.1 Benchmarking Experiments

Here we describe the use of the distances summarized in Tab. 3.1 for the comparison of network topologies. To such aim, we constructed three synthetic benchmark datasets, detailed hereafter. All simulations have been performed within the R statistical environment [118]. Throughout all simulations, we kept, for each distance, the parameter values as in the reference paper wherever possible, *e.g.*, $\gamma = 0.08$ for the scale of the Lorentz distribution in D2; the heat diffusion kernel in D3; the time $t = 3.5$ for the kernel in distance D3. For D1 we choose to use the $\lfloor \frac{N}{2} \rfloor$ largest eigenvalues.

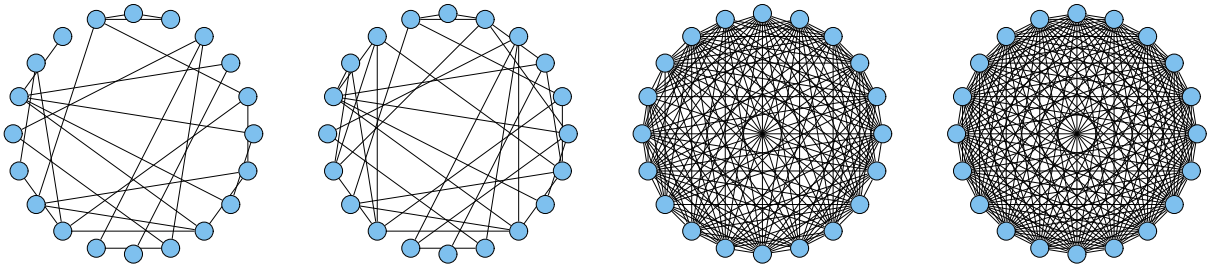
3.2.2 Data Description

The simulated topologies are generated within the R statistical environment [118] by means of the simulator provided by the package *netsim* [41, 42], producing networks that mimic the principal characteristics of transcriptional regulatory networks. The simulator takes into account the scale-free distribution of the connectivity and constructs networks whose clustering coefficient is independent of the number of nodes in the network. All random graphs are generated by keeping the default values of *netsim* for the structural parameters.

In the first experiment we consider a random network A on N vertices and we compare it with the full connected network with the same number of nodes F, the complementary network $\overline{A}$ and a matrix A_p obtained from A by modifying (inserting/deleting) about the $p\%$ of the nodes. For smoothing purposes, the process is repeated b times to obtain the first benchmarking dataset $\mathcal{B}_1(b, N, p)$. An instance of this benchmark dataset is shown in Fig. 3.2. In Tab. 3.2 we show the average on $b = 50$ instances of the number of nodes of the starting matrix A and the perturbed matrix A_5. Because of the small number of links in the original matrix, the 5%

Table 3.1: Spectral graph distances

Distance	Formula	Equation	Ref.
D1	$d_k(G,H) = \begin{cases} \sqrt{\frac{\sum_{i=N-k}^{N-1}(\lambda_i-\mu_i)^2}{\sum_{i=N-k}^{N-1}\lambda_i^2}} & \text{if } \sum_{i=N-k}^{N-1}\lambda_i^2 \leq \sum_{i=N-k}^{N-1}\mu_i^2 \\ \sqrt{\frac{\sum_{i=N-k}^{N-1}(\lambda_i-\mu_i)^2}{\sum_{i=N-k}^{N-1}\mu_i^2}} & \text{if } \sum_{i=N-k}^{N-1}\mu_i^2 < \sum_{i=N-k}^{N-1}\lambda_i^2 \end{cases}$	(3.1)	[114]
D2	$\epsilon(G,H) = \sqrt{\int_0^\infty [\rho_G(\omega)-\rho_H(\omega)]^2 \mathrm{d}\omega}$	(3.2)	[67]
D3	$d_M(G,H) = \sqrt{\sum_{i=0}^{N-1}\left(\lambda_i^{(G,M)}-\lambda_i^{(H,M)}\right)^2}$	(3.3)	[157]
D4	$d(G,H) = \sum_{i=0}^{N-1}\left(\lambda_i^{(G,L)}-\lambda_i^{(H,L)}\right)^2$	(3.4)	[33]
D5e	$d(G,H) = \int_\lambda (1-\lambda)^4 \left(f_{\lambda,G}(\lambda)-f_{\lambda,H}(\lambda)\right)^2 \mathrm{d}\lambda$	(3.5)	[48]
D5a	$d_n(G,H) = \sum_{k\in K}(1-k)^n (f_G(\lambda=k)-f_H(\lambda=k))^2$	(3.6)	[48]
D6	$d(G,H) = \sqrt{\mathrm{JS}(f_G,f_H)}$	(3.7)	[12]



A A_5 $\overline{A}$ F

Figure 3.2: Benchmark Dataset $\mathcal{B}_1(b, 25, 5)$: the original graph A, the perturbed graph A_5, the complemental graph $\overline{A}$ and the fully connected graph F.

perturbation mostly reflects in links insertion. On average, the density of the original graph A can be expressed by the relation $l \simeq 1.7N - 5$, where l is the number of links and N the number of vertices.

In the second experiment we simulate a time-series of T networks on N nodes starting from a randomly generated graph S_1, where each successive element S_i of the series is generated from its ancestor S_{i-1} by randomly modifying $p\%$ of the links. Again $b = 50$ instances of the series are created and collected into the second benchmarking dataset $\mathcal{B}_2(b, T, N, p)$. With this strategy, the number of existing links is increasing with the series index, being the original adjacency matrix almost sparse. The starting

Table 3.2: Number of links in the original matrix A, in the fully connected matrix F (maximum number of links for the given dimension) and in the perturbed matrix A_5, expressed as mean ± standard deviation on 50 replicates.

N	**F**	**A**	**A_5**
10	45	13.4±2.0	13.1±2.3
20	190	29.0±3.6	36.6±5.2
50	1225	79.3±7.4	131.8±4.2
100	4950	164.5±13.6	388.2±12.1

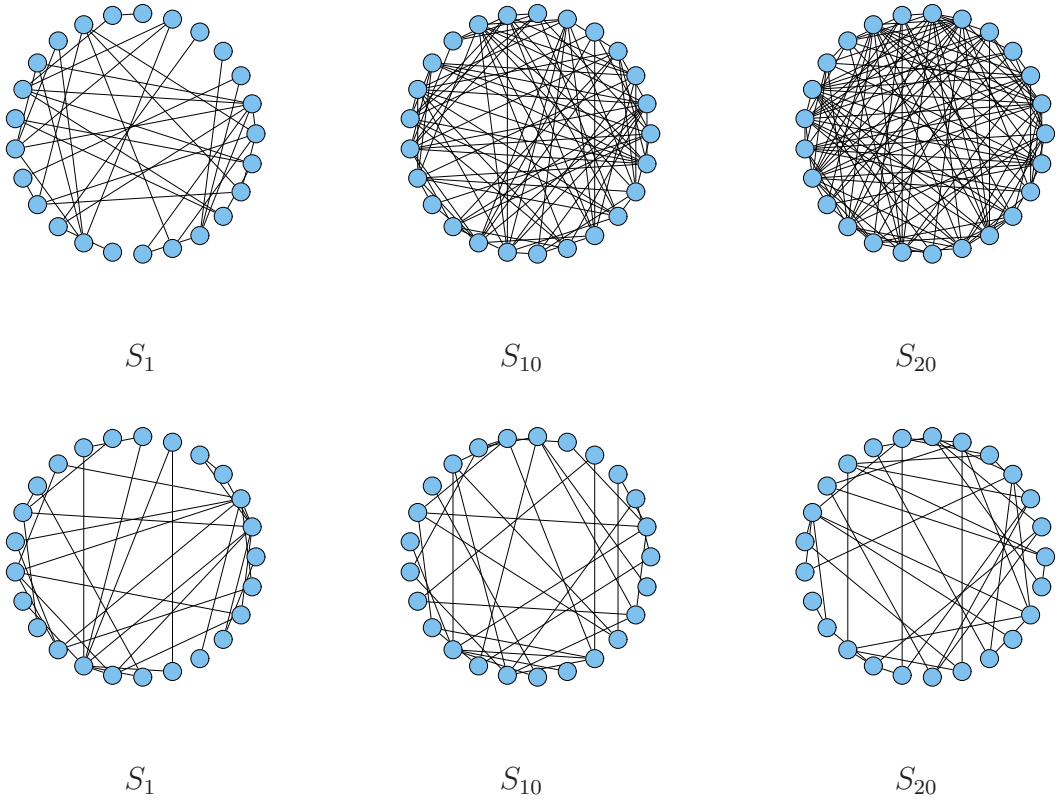


Figure 3.3: Benchmark Datasets $\mathcal{B}_2(b, 20, 25, 5)$ (upper row) and $\mathcal{B}_3(b, 20, 25, 5, 5)$ (lower row): the original graph S_1 (first element of the series), the tenth element S_{10} of the series and the final graph S_{20}.

matrix S_1 has on average 38.1±5.2 nodes, while the last element of the series S_{20} has 132.3±8.2.

The third experiment is based on a benchmark dataset $\mathcal{B}_3(b, T, N, nd, na)$. Starting from $\mathcal{B}_2(b, T, N, p)$, different perturbations are applied: each successive element S_i of the series is generated from its ancestor S_{i-1} by randomly deleting nd links and adding na links. By construction, the number of existing links for all elements of the series is constant. Three elements of the benchmarking datasets $\mathcal{B}_2$ and $\mathcal{B}_3$ are shown in Fig. 3.3.

Table 3.3: Results of the experiments on the first benchmarking dataset. For each measure D1-D6 and number of network vertices N, we report the values of the distances between the network A and the networks A_5, $\overline{A}$ and F in terms of the minimum (m), mean (μ) $\pm$ standard deviation and maximum (M) on the 50 replicates. Values of D5 are in 10^{-3}.

N	D	A_5			$\overline{A}$			F		
		m	$\mu \pm \sigma$	M	m	$\mu \pm \sigma$	M	m	$\mu \pm \sigma$	M
10	1	0.025	0.108 ± 0.053	0.197	0.085	0.982 ± 0.383	1.564	0.424	1.324 ± 0.350	1.811
10	2	0.215	0.319 ± 0.052	0.403	0.47	0.857 ± 0.174	1.066	1.434	1.563 ± 0.04	1.635
10	3	0	0.067 ± 0.074	0.294	0.006	0.415 ± 0.39	1.83	0.028	0.472 ± 0.402	1.925
10	4	0	2.182 ± 1.01	4.533	14.33	151.8 ± 71.5	328.1	336	470.4 ± 61.7	598
10	5	0	0.941 ± 0.603	1.844	0.092	3.635 ± 2.340	8.907	0.518	4.112 ± 2.306	9.491
10	6	0.102	0.169 ± 0.039	0.259	0.192	0.386 ± 0.084	0.507	0.431	0.507 ± 0.04	0.552
20	1	0.037	0.194 ± 0.069	0.342	2.117	2.768 ± 0.379	3.71	2.455	3.038 ± 0.372	4.006
20	2	0.202	0.284 ± 0.049	0.381	1.025	1.091 ± 0.034	1.165	1.538	1.55 ± 0.008	1.568
20	3	0.044	0.154 ± 0.132	0.577	0.588	1.04 ± 0.333	2.05	0.643	1.103 ± 0.336	2.123
20	4	1.812	15.9± 6.5	28.5	2584	3658 ± 420	4761	4898	5531 ± 243	6146
20	5	0.358	0.836 ± 0.503	2.459	2.416	3.623 ± 6.441	1.041	2.439	3.654 ± 6.45	1.036
20	6	0.135	0.207 ± 0.04	0.323	0.581	0.772 ± 0.879	0.077	0.652	0.767 ± 0.83	0.05
50	1	0.389	0.504 ± 0.072	0.606	6.676	8.057 ± 0.784	9.064	6.924	8.288 ± 0.771	9.253
50	2	0.275	0.344 ± 0.042	0.437	1.152	1.195 ± 0.025	1.228	1.533	1.54 ± 0.005	1.549
50	3	0.668	1.186 ± 0.313	1.77	2.078	3.356 ± 0.647	4.428	2.138	3.423 ± 0.649	4.497
50	4	138	237 ± 48	353	83850	92670 ± 3078	97710	102700	107300 ± 1613	110000
50	5	0.888	1.875 ± 0.541	2.765	2.379	3.993 ± 0.847	5.42	2.379	3.992 ± 0.849	5.42
50	6	0.435	0.559 ± 0.0751	0.711	1.372	1.481 ± 0.061	1.597	1.183	1.277 ± 0.063	1.39
100	1	0.804	0.977 ± 0.076	1.086	13.55	16.07 ± 1.032	17.6	13.77	16.28 ± 1.027	17.8
100	2	0.451	0.506 ± 0.025	0.544	1.215	1.264 ± 0.019	1.293	1.524	1.533 ± 0.004	1.543
100	3	2.116	3.606 ± 0.665	4.64	4.506	6.723 ± 0.992	8.166	4.566	6.79 ± 0.995	8.238
100	4	1784	2161 ± 136	240	842900	861200 ± 9575	880600	915800	925100 ± 4880	935900
100	5	1.645	2.787 ± 0.525	3.589	2.602	3.941 ± 0.630	4.824	2.602	3.941 ± 0.631	4.824
100	6	0.933	1.102 ± 0.074	1.204	2.07	2.229 ± 0.088	2.397	1.694	1.839 ± 0.078	1.997

3.2.3 Results

In Exp. 1 the six distances D1-D6 were applied on 4 instances of $\mathcal{B}_1(50, N, 5)$ for $N = 10, 20, 25, 100$ and distances between the original graph A and the three companion matrices F, $\overline{A}$ and A_p were computed. Results are collected in Tab. 3.3.

Distance D4 spans a considerably wider range than other measures, due to the absence of the square root in the comparison of the Laplacian spectra, while D5 is restricted into a very small interval. The same distance D4 also shows a high dependency on the dimension of the considered matrices and the number of the links (see Tab. 3.3).

The best stability in terms of the relative standard deviation σ/μ is reached by D2 and D4. Furthermore, D2, differently from all other measures, is almost independent of the number of vertices. Finally, D6 is the only

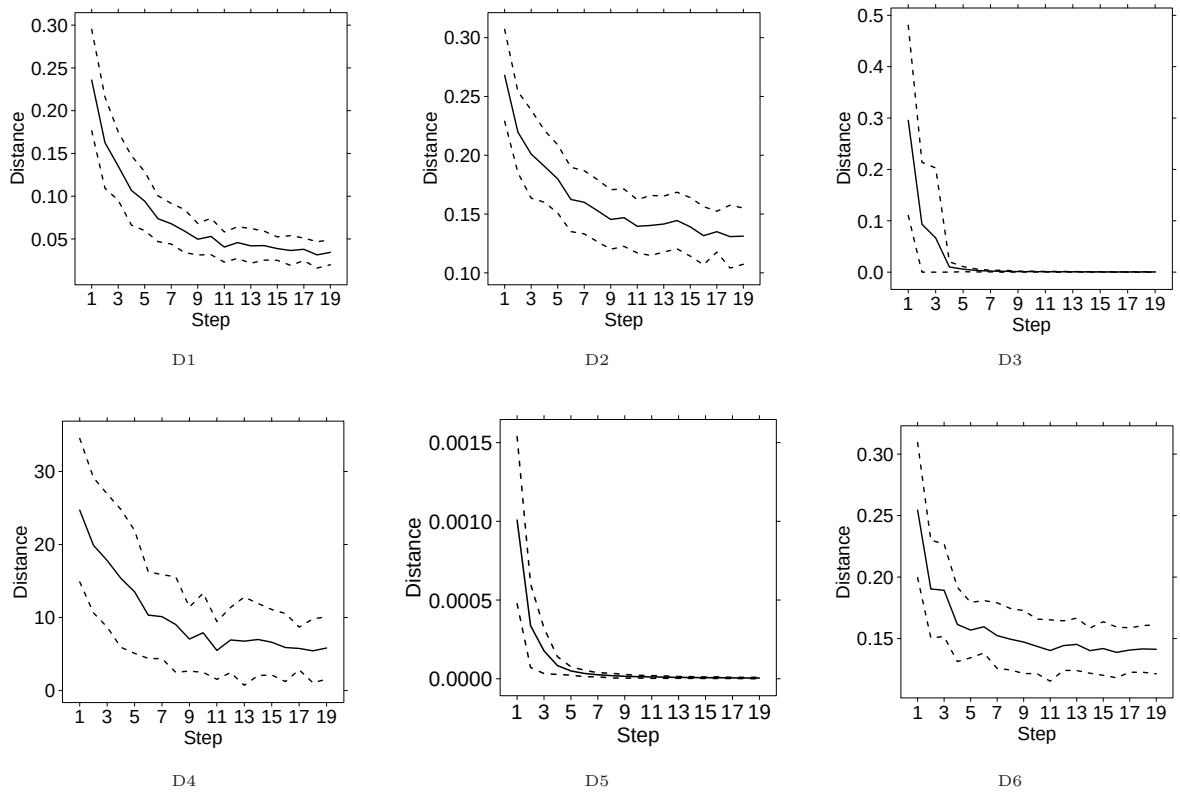


Figure 3.4: Plots of the distances of consecutive elements of the series for the dataset $\mathcal{B}_2(50, 25, 5)$. Solid line: mean over the $b = 50$ replicates; dashed lines: minimum and maximum over the $b = 50$ replicates.

measure that, in the cases with $N > 10$, gives a lower distance for F than for $\overline{A}$.

The summary plots in Fig. 3.4 display results of Exp. 2 on the benchmark dataset $\mathcal{B}_2(20, 20, 25, 5)$. Distances between consecutive elements (S_i, S_{i+1}) of the series (defined Step i) were computed: results are averaged on the 50 replicates. For all D1-D6, distance decreases for increasing steps, although on different ranges (as already pointed out for Experiment 1) and with different widths for the confidence intervals. D3 and D5 decrease more quickly for initial steps, so they are less useful when comparing large networks.

To better highlight similarities and differences among the distances regard-

less of their ranges of values, we also computed their mutual correlations and plotted the mutual scatter plots in Fig. 3.5. All correlation values are quite high, ranging from 0.8225 to 0.9970: D3 and D5 are mutually strongly correlated, but they tend to separate from the other distances, as evidenced both from the global correlation values and the scatter plot profiles distancing from the panel diagonals.

The Experiment 3 was performed on the benchmark dataset $\mathcal{B}_3(50, 25, 5, 5)$, and the results are reported in two figures matching those of Exp. 2. Since the difference between consecutive pairs of elements of the series is quite

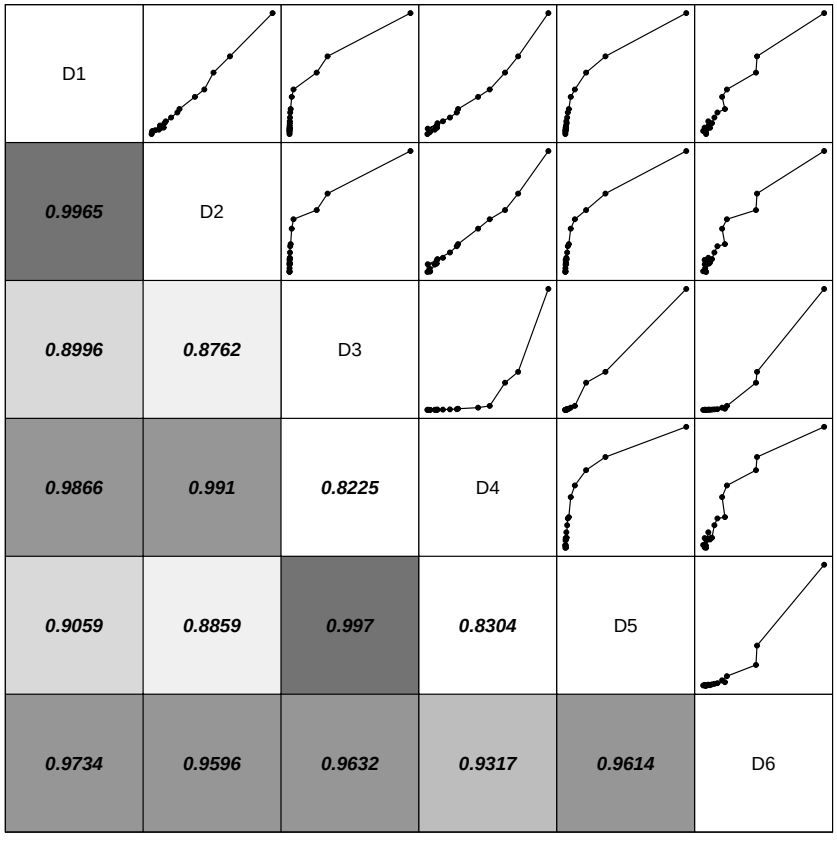


Figure 3.5: Mutual scatterplots (upper triangle) and correlation values (lower triangle) for the Exp. 2.

similar throughout all the steps, as expected all distances show a nearly constant trend as shown in Fig. 3.6.
The oscillations around the mean value are nevertheless strongly varying among different measures, as evidenced by Fig. 3.7. In particular, distance D3 is anticorrelated to all distances but D5; furthermore only in 4 cases out of 15 we obtain a correlation value higher than 0.7, with again $D1$, $D2$, $D4$ and $D6$ forming a group of more similar behaviour.
Possible hierarchy of the six distances was explored by clustering. Two dendrograms are built for Exp. 2 and Exp. 3 by using the *hclust* package in R and shown in Fig. 3.8. The clusters have average linkage and the correlation distance $\mathrm{cd}(\cdot,\cdot) = 1 - \mathrm{Corr}(\cdot,\cdot)$ is used as the dissimilarity measure. Although there is an appreciable coherence among measures on macroscopic trends, when downscaling to microscopic trends correlations get much looser. Distances $D1$, $D2$, $D4$, $D6$ seem to group together, while $D3$ has a more erratic behaviour. Finally, a wide difference in the range of values occurs in the cluster heights between the two experiments: the homogeneous macroscopic situation of Exp. 2 has a narrower height span than the microscopic case in Exp. 3.

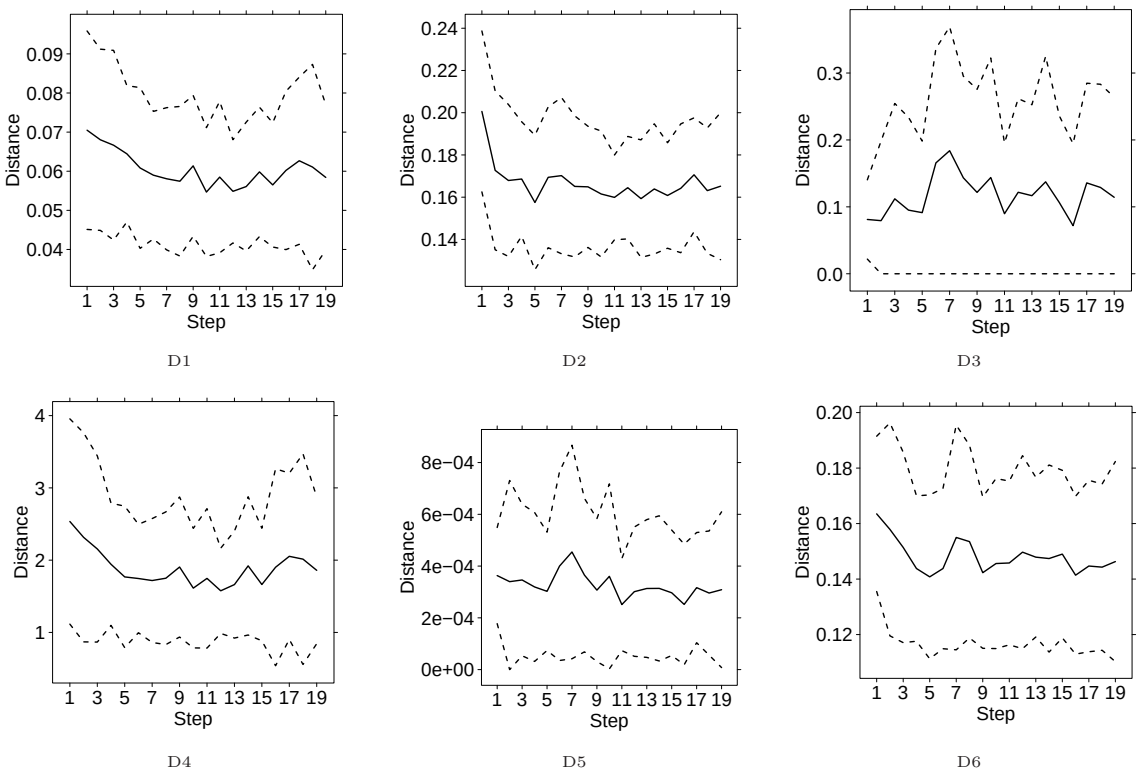


Figure 3.6: Plots of the distances of consecutive elements of the series for the dataset $\mathcal{B}_3(50, 25, 5, 5)$. Solid line: mean over the $b = 50$ replicates; dashed lines: minimum and maximum over the $b = 50$ replicates.

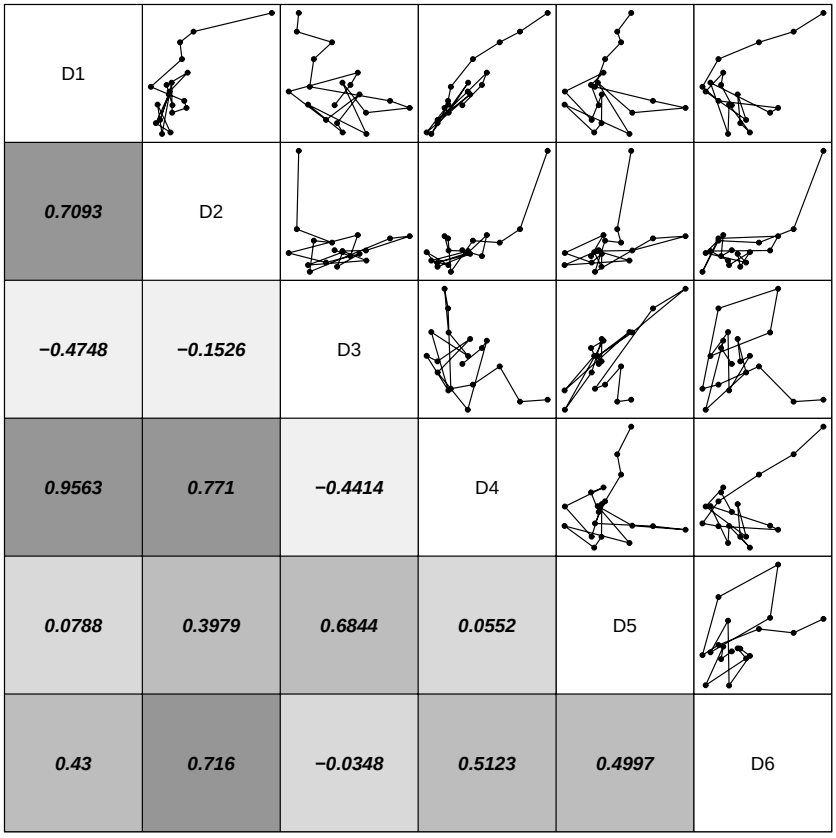
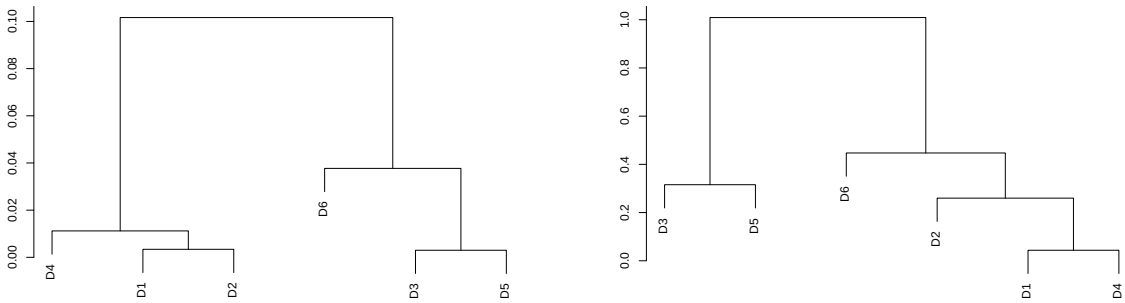


Figure 3.7: Mutual scatterplots (upper triangle) and correlation values (lower triangle) for the Exp. 3.



Experiment 2 Experiment 3

Figure 3.8: Cluster dendrograms with average linkage and correlation distance of D1-D6 for the two Experiments 2 and 3.

Chapter 4

HIM, Hamming - Ipsen-Mikhailov Distance

In the previous chapter we outlined the main families of algorithms for the network comparison, their flows and advantages. We focused on two of the most common families: edit-like and spectral distances. In order to combine the strength of the two approaches and try to correct their limitations we propose here a product metric called HIM (Hamming Ipsen-Mikhailov) with both global and local characteristics.

4.1 Definition

The HIM distance [75] is a metric for network comparison combining an edit distance (Hamming [143, 44]) and a spectral one (Ipsen-Mikhailov [67]). As discussed in [74], edit distances are local, that is they focus only on the portions of the network interested by the differences in the presence/absence of matching links. Spectral distances evaluate instead the global structure of the compared topologies, but they cannot distinguish isomorphic or isospectral graphs, which can correspond to quite different conditions within the biological context. Their combination into the HIM distance represents an effective solution to the quantitative evaluation of

network differences.

Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two simple networks on N nodes, described by the corresponding adjacency matrices A_1 and A_2, with $a_{ij}^{(1)}, a_{ij}^{(2)} \in \mathcal{F}$, where $\mathcal{F} = \mathbb{F}_2 = \{0, 1\}$ for unweighted graphs and $\mathcal{F} = [0, 1]$ for weighted networks. Denote then by $\mathbb{I}_N$ the identity $N \times N$ matrix $\mathbb{I}_N = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & \cdots & & \\ 0 & 0 & \cdots & 1 \end{pmatrix}$, by $\mathbb{1}_N$ the unitary $N \times N$ matrix with all entries equal to one and by $_N$ the null $N \times N$ matrix with all entries equal to zero. Finally, denote by $\mathcal{E}_N$ the empty network with N nodes and no links (with adjacency matrix $_N$) and by $\mathcal{F}_N$ the undirected full network with N nodes and all possible $N(N-1)$ links (whose adjacency matrix is $\mathbb{1}_N - \mathbb{I}_N$).

The definition of the Hamming distance is the following:

$$\text{Hamming}(\mathcal{N}_1, \mathcal{N}_2) = \sum_{1 \le i \ne j \le N} |A_{ij}^{(1)} - A_{ij}^{(2)}| \ .$$

To guarantee independence from the network dimension (number of nodes), we normalize the above function by the factor $\overline{\eta} = \text{Hamming}(\mathcal{E}_N, \mathcal{F}_N) = N(N-1)$:

$$H(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} |A_{ij}^{(1)} - A_{ij}^{(2)}| \ . \tag{4.1}$$

When $\mathcal{N}_1$ and $\mathcal{N}_2$ are unweighted networks, $H(\mathcal{N}_1, \mathcal{N}_2)$ is just the fraction of different matching links (over the total number $N(N-1)$ of possible links) between the two graphs. In all cases, $H(\mathcal{N}_1, \mathcal{N}_2) \in [0, 1]$, where the lower bound 0 is attained only for identical networks $A_1 = A_2$ and the upper bound 1 is reached whenever the two networks are complementary $A_1 + A_2 = \mathbb{1}_N - \mathbb{I}_N = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ & \cdots & & \\ 1 & 1 & \cdots & 0 \end{pmatrix}$.

Among spectral distances, we consider the Ipsen-Mikhailov distance IM which has been proven to be the most robust in a wide range of situations [74]. We recall here the main characteristic of the IM distance introduced in [67] as a tool for network reconstruction from its Laplacian spectrum,

the definition of the Ipsen-Mikhailov metric follows the dynamical interpretation of a N–nodes network as a N–atoms molecule connected by identical elastic strings, where the pattern of connections is defined by the adjacency matrix of the corresponding network. The dynamical system is described by the set of N differential equations

$$\ddot{x}_i + \sum_{j=1}^{N} A_{ij}(x_i - x_j) = 0 \quad \text{for } i = 0, \cdots, N-1 \ . \tag{4.2}$$

We recall that the Laplacian matrix L of an undirected network is defined as the difference between the degree D and the adjacency A matrices $L = D - A$, where D is the diagonal matrix with vertex degrees as entries. L is positive semidefinite and singular [32, 9, 130, 140], so its eigenvalues are $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{N-1}$. The vibrational frequencies ω_i for the network model in Eq. 4.2 are given by the eigenvalues of the Laplacian matrix of the network: $\lambda_i = \omega_i^2$, with $\lambda_0 = \omega_0 = 0$. The spectral density for a graph as the sum of Lorentz distributions is defined as

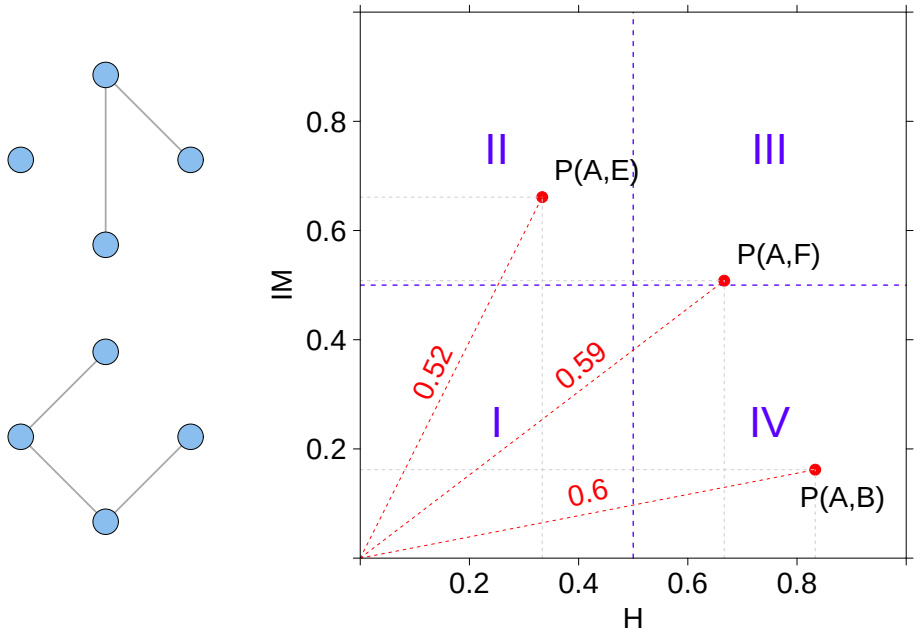
$$\rho(\omega, \gamma) = K \sum_{i=1}^{N-1} \frac{\gamma}{(\omega - \omega_i)^2 + \gamma^2} \ ,$$

where γ is the common width and K is the normalization constant defined as

$$K = \frac{1}{\gamma \sum_{i=1}^{N-1} \int_0^\infty \frac{\mathrm{d}\omega}{(\omega - \omega_i)^2 + \gamma^2}} \ ,$$

so that $\int_0^\infty \rho(\omega, \gamma)\mathrm{d}\omega = 1$. The scale parameter γ specifies the half-width at half-maximum, which is equal to half the interquartile range. Then the spectral distance ϵ_γ between two graphs G and H on N nodes with densities $\rho_G(\omega, \gamma)$ and $\rho_H(\omega, \gamma)$ can then be defined as

$$\epsilon_\gamma(G, H) = \sqrt{\int_0^\infty \left[\rho_G(\omega, \gamma) - \rho_H(\omega, \gamma)\right]^2 \mathrm{d}\omega} \ . \tag{4.3}$$



(a) (b)

Figure 4.1: An example of HIM distance. (a) Network A (top) and Network B (bottom); (b) Representation of the HIM distance in the Ipsen-Mikhailov and Hamming distance space between networks A versus B, E and F, where F is the fully connected network and E is the empty one.

The highest value of ϵ_γ is reached, for each N, when evaluating the distance between $\mathcal{E}_N$ and $\mathcal{F}_N$. Defining $\overline{\gamma}$ as the (unique) solution of

$$\epsilon_\gamma(\mathcal{E}_N, \mathcal{F}_N) = 1 \ ,$$

we can now define the normalized Ipsen-Mikahilov distance as

$$\mathrm{IM}(G,H) = \epsilon_{\overline{\gamma}}(G,H) = \sqrt{\int_0^\infty \left[\rho_G(\omega,\overline{\gamma}) - \rho_H(\omega,\overline{\gamma})\right]^2 \mathrm{d}\omega} \ ,$$

so that $\mathrm{IM}(G,H) \in [0,1]$ with upper bound attained only for $(G,H) = (\mathcal{E}_N, \mathcal{F}_N)$. Finally, the HIM distance is defined as the product metric of the normalized Hamming distance H and the normalized Ipsen-Mikhailov

IM distance, normalized by the factor $\sqrt{2}$ to set its upper bound to 1:

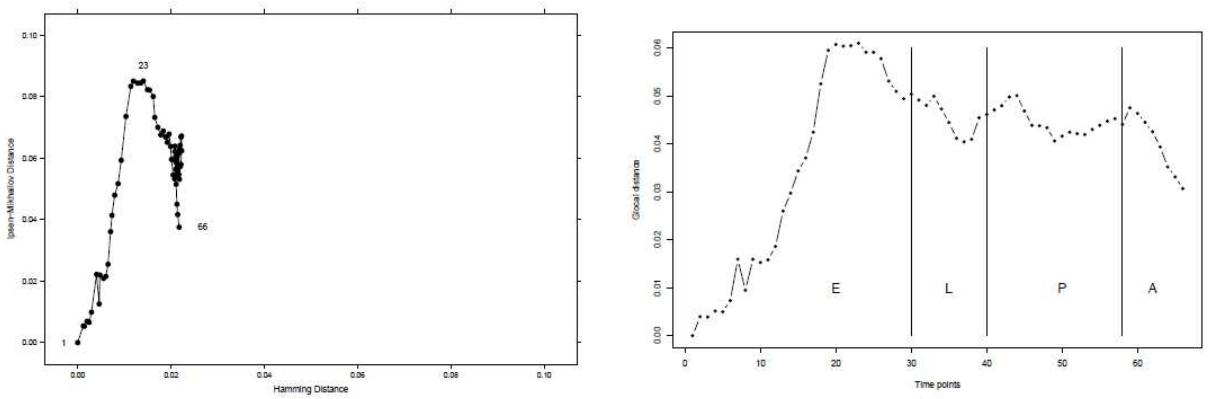
$$HIM(N_1, N_2) = \frac{1}{\sqrt{2}}\sqrt{\mathrm{H}(N_1, N_2)^2 + \mathrm{IM}(N_1, N_2)^2}$$

We can represent the HIM distance in the $[0, 1] \times [0, 1]$ Hamming/Ipsen-Mikhailov space, where a point $P(x, y)$ represents the distance between two networks N_1 and N_2 whose coordinates are $x = \mathrm{H}(N_1, N_2)$ and $y = \mathrm{IM}(N_1, N_2)$ and the norm of P is $\sqrt{2}$ times the HIM distance $HIM(N_1, N_2)$. The same holds for weighted networks, provided that the weights range in $[0, 1]$. In Fig. 4.1 we provide an example of this representation of the HIM distance between networks of four nodes. Roughly splitting the Hamming/Ipsen-Mikhailov space into four main zones I,II,III,IV as in Figure 4.1, we can say that two networks whose distances correspond to a point in zone I are quite close both in terms of matching links and of structure, while those falling in the zone III are very different with respect to both characteristics. Networks corresponding to a point in zone II have many common links, but their structure is rather different, while a point in zone IV indicates two networks with few common links, but with similar structure. Full mathematical details about the HIM distance and its two components H and IM are available in [75].

4.2 A Biological Example

In [82], the authors used the Keller algorithm to infer the gene regulatory networks of Drosophila melanogaster from a time series of gene expression data measured during its full life cycle. They selected 66 time points during the developmental cycle, spanning across four different stages (Embryonic time points $1 - 30$, Larval t.p. $31 - 40$, Pupal t.p. $41 - 58$, Adult t.p. $59 - 66$), following the dynamics of 588 gene ontological groups and then constructing a time series of inferred networks N_i . Hereafter we evaluate

the structural differences between N_i and the initial network N_1, as measured by the glocal distance: the resulting plot is displayed in Figure 4.2. The largest variations, both between consecutive terms and with respect to the initial network N_1 , occur in the embrional stage (E). In particular, it is interesting to note that the dynamics of the networks move N_i away from N_1 until time points 23, then the following terms start getting closer again to N_1 in terms of glocal distance: such behaviour was detected also in the original paper, but only qualitatively, while the introduced metrics can provide a quantitative assessment of the occurring differences. Finally, it can be appreciated the different range of the two distances: while Hamming distance ranges between 0 and 0.0223, the Ipsen-Mikhailov distance has 0.0851 as its maximum, indicating an higher variability of the networks in terms of structure rather than matching links.



(a) (b)

Figure 4.2: (a) Evolution of distances of the D. melanogaster network time series in the Hamming/Ipsen-Mikhailov space and (b) evolution of glocal distances of the D. melanogaster network along 66 time points in the 4 stages Embryonic (E), Larval (L), Pupal (P) and Adult (A)

4.3 Module Preservation.

In this Section we test the similarity of the behavior of the HIM distance with some other employed measures already known in the literature. In many biological applications it is interesting to study the variation of modules structure across different gene networks. In particular to identify the effects of a clinical condition on a certain pathway one could determine if its connectivity structure is still preserved. As Langfelder and coworkers point out in [87] the fact that a module is non-preserved can either prove a real biological difference between same tissues under the same condition (*e.g.*, sex specific modules) or being the product of uninteresting data outliers. An immediate approach to evaluate module preservation is to consider just the overlap in the module membership. Of course this procedure overlooks the point that the nature of connection pattern within the modules are of great functional importance. Thus, cross-tabulation based methods often miss structural factors important to determine whether a module is in fact preserved or not. In case of non preserved modules applying cross-tabular approaches one can just state that the set of genes in the reference module can not be found in any of the identified test set modules. It is impossible to make any assertion about the presence of the module in the test set irrespectively to the module detection parameter setting and procedure. Module preservation analysis have important applications, *e.g.* as shown in [87] the wiring of apoptosis genes in a human cortical network differs from that in chimpanzees. They propose an approach based on several module preservation statistics that do not need a true module assignment in the test set. The statistics are identified and characterized by the type of inherent network representation. Some preservation statistics apply to generic networks uniquely defined by an adjacency matrix, some others are defined just for correlation networks in which each value

is the pairwise correlation value between numerical variables. They show how the use of aggregation of different statistics allows the construction of summary module preservation measures. The statistics we will mainly consider here are the $Z_{summary}$ and $medianRank_{summary}$ described in detail in Appendix A.1 and defined as

$$Z_{summary} = \frac{Z_{connectivity} + Z_{density}}{2} \tag{4.4}$$

$$medianRank_{summary} = \frac{medianRank_{density} + medianRank_{connectivity}}{2} \tag{4.5}$$

In what follows we compare the statistic methods presented in [87] with the HIM distance by testing them on four gene co-expression network applications already presented in [149, 52, 122, 80, 112]:

- Preservation of cholesterol biosynthesis pathway in mouse tissues
- Comparison of human and chimpanzee brain networks
- Preservation of selected KEGG pathways between human and chimpanzee brain networks
- Sex differences in mouse liver networks.

4.3.1 Data

Multi-tissues Mice Data

Liver gene expression data from 135 female mice were used for this analysis. The F2 intercross used, the animal husbandry and physiological trait measurement details are described in detail in [149, 52]. Genotyping was conducted by ParAllele (Affymetrix, Santa Clara, California, United States) using the molecular inversion probe (MIB) multiplex and involved over

Table 4.1: Statistics Description Summary

Name of statistic	Eigen. dec. of Conn. Matrix	SVD of Expr. Data	Depends on N	Uses Perm. Test	Comp. Speed	Ave. $\|cor\|$ with other stats
Ipsen	Yes	No	Yes	No	$\propto N$	0.534
Hamm	No	No	No	No	Fast	0.451
HIM	Yes	No	Yes	No	$\propto N$	0.576
ZsummQ	No	Yes	Yes	Yes	Slow	0.533
Zsumm	No	Yes	Yes	Yes	Slow	0.563
Zdens	No	Yes	Yes	Yes	Slow	0.617
Zconn	No	Yes	Yes	Yes	Slow	0.550
OsummQ	No	Yes	No	No	Fast	0.501
Osumm	No	Yes	No	No	Fast	0.572
Odens	No	Yes	No	No	Fast	0.538
Oconn	No	Yes	No	No	Fast	0.481

1,300 SNPs, genomic DNA was isolated from kidney [52]. RNA preparation and array hybridizations were performed at Rosetta Inpharmatics. The platform for microarray analysis is the custom ink-jet microarrays (Agilent Technologies [Palo Alto, California, United States], [122]). It contains 2,186 control probes and 23,574 noncontrol oligonucleotides. RNA was extracted from livers, reverse transcribed and labeled with either $Cy3$ or $Cy5$ fluorochromes. Purified $Cy3$ or $Cy5$ complementary RNA was hybridized to at least two microarray slides and scanned. Arrays were quantified on the basis of spot intensity relative to background, adjusted for experimental variation between arrays using average intensity over multiple channels, and fit to an error model to determine significance (type I error). Gene expression is reported as the ratio of the mean log10 intensity (mlratio) relative to the pool derived from 150 mice randomly selected from the F2 population.

Several **data-filtering** steps were taken in order to minimize noise in the

gene expression dataset for the experiments about the module detection and preservation. First, preliminary evidence showed major differences in gene expression levels between sexes among the F2 mice used, and therefore only female mice were used for the analysis. Only those mice with complete phenotype, genotype, and array data were used, this gave a final experimental sample of 135 female. To reduce the computational burden and to possibly enhance the signal in our data, we used only the 8,000 most-varying female liver genes in our preliminary network construction. For module detection, we limited our analysis to the 3,600 most-connected genes because our module construction method and visualization tools cannot handle larger datasets at this point. By definition, module genes are highly connected with the genes of their module (i.e., module genes tend to have relatively high connectivity). Thus, for the purpose of module detection, restricting the analysis to the most-connected genes should not lead to major information loss. Since the network nodes in our analysis correspond to genes as opposed to probesets, we eliminated multiple probes with similar expression patterns for the same gene. Specifically, the 3,600 genes were examined, and where appropriate, gene isoforms and genes containing duplicate probes were excluded by using only those with the highest expression among the redundant transcripts. This final filtering step yielded a count of 3,421 genes for the experimental network construction [52].

Human and Chimpanzee Brains Data

The dataset used for network construction consisted of 36 Affymetrix (Santa Clara, CA) HGU95Av2 microarrays surveying gene expression with 12,625 probe sets in three adult humans and three adult chimpanzees across six matched brain regions: Brocas area, anterior cingulate cortex, primary visual cortex, prefrontal cortex, caudate nucleus, and cerebellar vermis [80]. After eliminating probes with sequence differences between the species, all

arrays were scaled to the same average intensity, and quantile normalization was performed. Four thousand probe sets were selected for network analysis based on high variance in human brain relative to a nonneural tissue (lung). From these, 2,241 probe sets with the highest connectivity were clustered on the basis of TOM (see Section 2.3.1) to identify modules of coexpressed genes.

Functional Annotation of Hub Genes and Modules. GenMAPP 2.0 http://www.genmapp.org was used to search among hub genes and modules for enrichment of functional categories of genes defined by the Gene Ontology Consortium [7] http://www.geneontology.org. The significance of each enriched category was also assessed on the basis of differential connectivity between humans and chimpanzees [112].

4.3.2 Results

First we considered the data relative to the male and female liver: in particular we want to evaluate the preservation of gene modules from the female tissue versus the male one [52]. In figure 4.4 we show the plot of the 12 modules individuated by WGCNA represented in three Module-Size vs. Preservation-Measure spaces, $medianRank$, $Z_{summary}$ and HIM measure respectively. Here we focus on the parallelism and correspondence between the Network-Statistics (a and b) and Spectral based measures (c). We can notice a high overall agreement for the majority of the modules; in particular it is interesting to highlight how the light-yellow and salmon modules are clearly the least preserved for the HIM measure confirming the results obtained with the Network-Statistics based measures, while the cyan module results borderline for all the measures. High agreement is also reached for the group of the five biggest modules brown, black, green, blue and red especially between $Z_{summary}$ and HIM. There is a general low agreement for the remaining four modules.

In figure 4.5 we present two boxplot relative to the data about the cholesterol pathway data in eight different mouse tissues and in particular we show how much the pathway is preserved using as reference the tissue indicated on the y axis and using as test the one on the x axis. In the top plot are depicted all the mutual HIM distances computed between the TOM networks, while in the bottom one are represented the results produced with $Z_{summary}$ method. As expected the first one results symmetric and with diagonal equal to one because of the properties of the distance. In general the results are different between the two methods and it is interesting to compare the values with the data representation of figure 4.3. HIM distance better underlines the similarity between the samples from same tissue across the two genders, while apart for the Liver tissue with female reference and male test the $Z_{summary}$ never top ranks the comparison between the same tissue. The Liver tissue shows a particular behaviour also considering the HIM preservation measure that hightlights how the male and female liver tissues are structurally similar just to eachother while they are different from all the others. Also the muscle tissue shows a similar but less evident characteristic while brain and adipose tissues show a high structural similarity also across the two sexes.
Finally in figure 4.6 we present a splom graph of the correlation values between the results of each of the considered measures. Here we considered $Z_{summary}$ ans $medianrank_{summary}$ both used to assess the level of preservation of the modules ($ZsummaryP$ and $MRsummaryP$) and also used to measure the quality of the modules ($ZsummaryQ$ and $MRsummaryQ$ A.2)
In conclusion the HIM distance not only shows a good agreement with more classical measures, but it also better points out some subtle differences between samples that other tested measures are not able to capture.

Table 4.2: Mean and Standard Error of Spearman correlations across all the datasets

	Ipsen	Hamming	Mod.Ipsen	Mean
Ipsen	1±0	0.422±0.103	0.981±0.008	0.49
Hamming	0.422±0.103	1±0	0.422±0.127	0.42
Mod.Ipsen	0.981±0.008	0.422±0.127	1±0	0.53
ZsummaryQuality	0.504±0.105	0.456±0.091	0.545±0.101	0.50
ZsummaryPreser.	0.397±0.11	0.531±0.114	0.441±0.123	0.53
Zdensity	0.407±0.096	0.602±0.127	0.452±0.087	0.57
Zconnectivity	0.406±0.123	0.491±0.107	0.466±0.147	0.51
MRsummaryQuality	0.485±0.112	0.416±0.116	0.537±0.122	0.44
MRsummaryPreser.	0.46±0.101	0.293±0.087	0.499±0.104	0.53
MRdensity	0.449±0.112	0.273±0.093	0.484±0.109	0.50
MRconnectivity	0.393±0.094	0.249±0.087	0.426±0.105	0.46

Figure 4.3: **Network representation of the Cholesterol biosynthesis gene module in the considered mouse tissues.** The module is here represented as a weighted signed correlation network where the nodes represent the genes from the GO category Cholesterol Biosynthetic Process. Module preservation techniques applied here allow the assessment of the similarity between these networks. Here we represent the connectivity pattern between the cholesterol biosynthesis genes in 4 different tissues from male and female mice. The thickness of the link represents the absolute value of correlation, while the colors red and green show positive correlation or anticorrelation respectively. The dimension of the nodes is proportional to their connectivity values, so the hubs of the module are represented by larger circles. This kind of plot shows how across the tissues there is a high resemblance between the module in male and female samples.

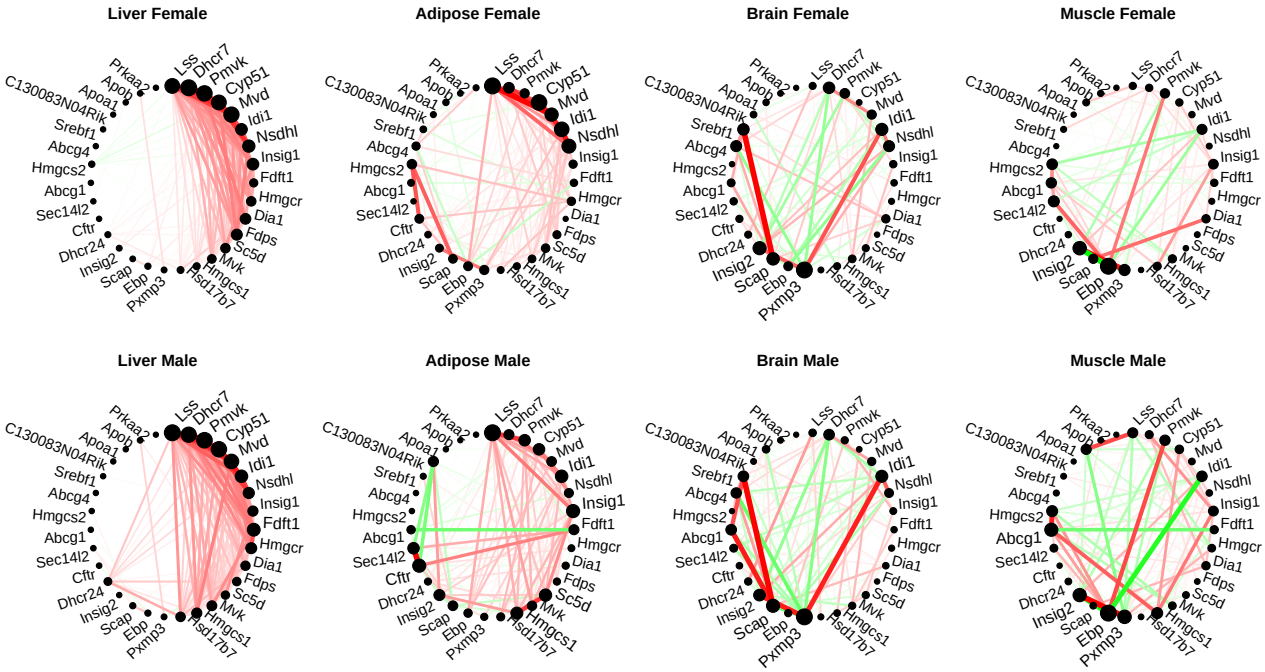
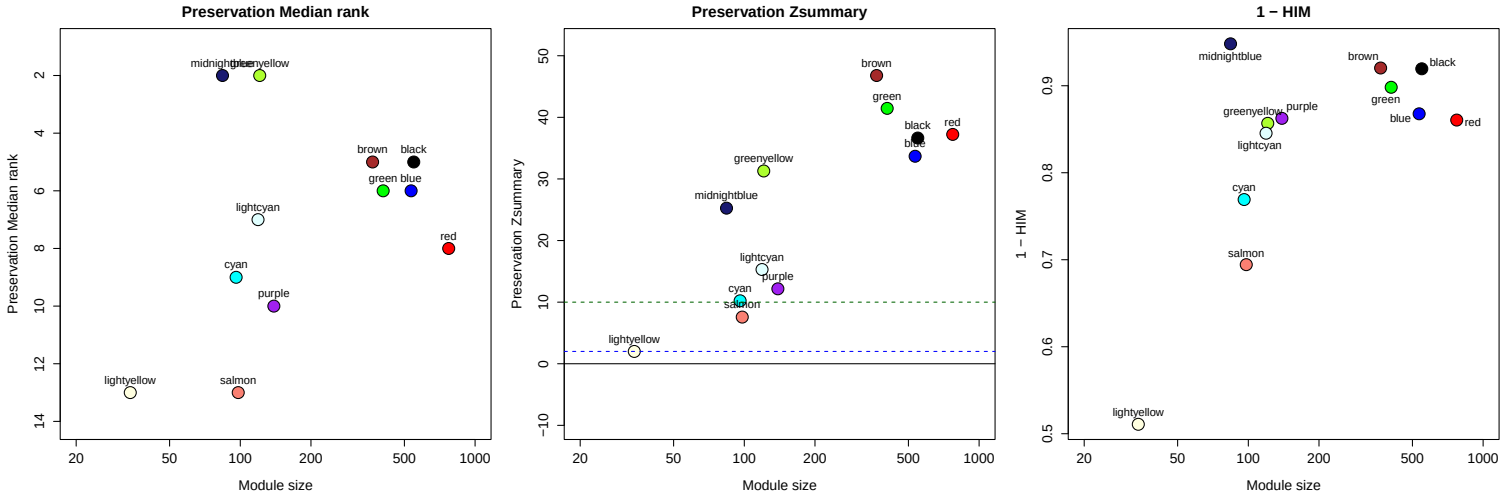


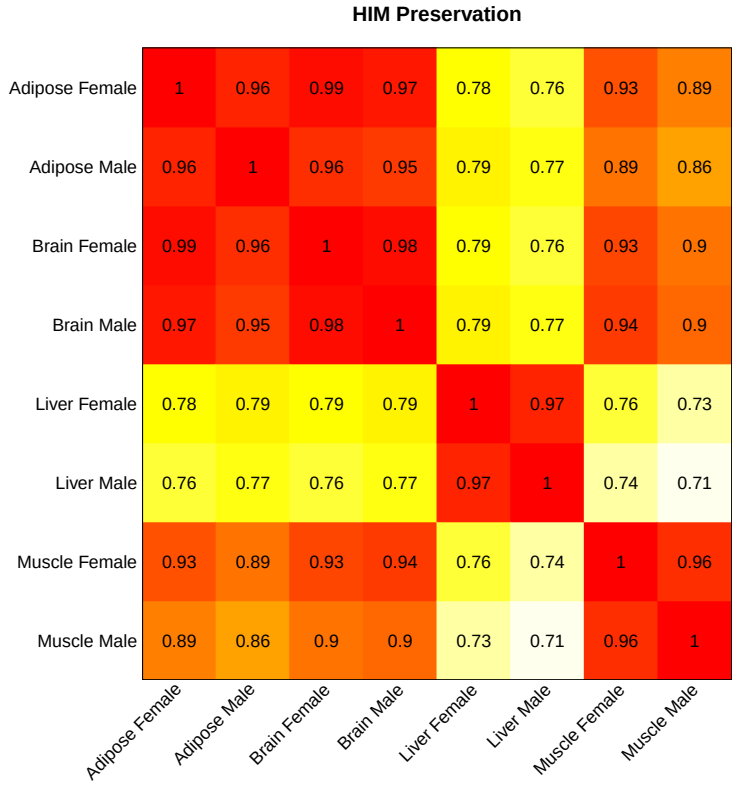
Figure 4.4: Preservation Measures: a) Median Rank, b) Zsummary, c) HIM based (1-HIM). 12 modules detected in female liver data in a Modul Size vs. Preservation plot.



(a) (b) (c)

Figure 4.5: A) HIM Preservation of the cholesterol pathway between the tissues. Z Summary Preservation of the cholesterol pathway between the tissues. B) On rows are presented the reference tissues and on columns the test tissues

A)



B)

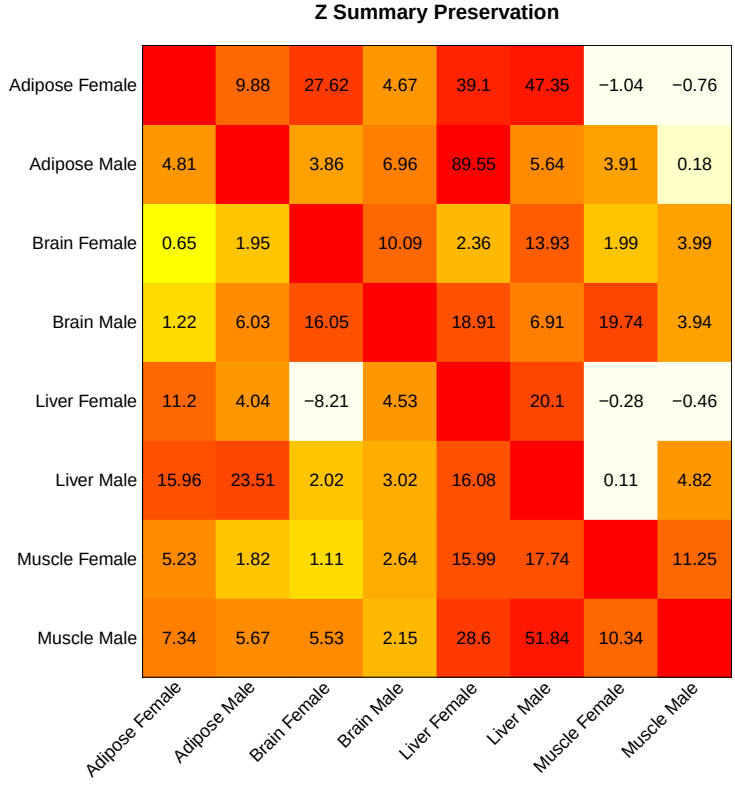
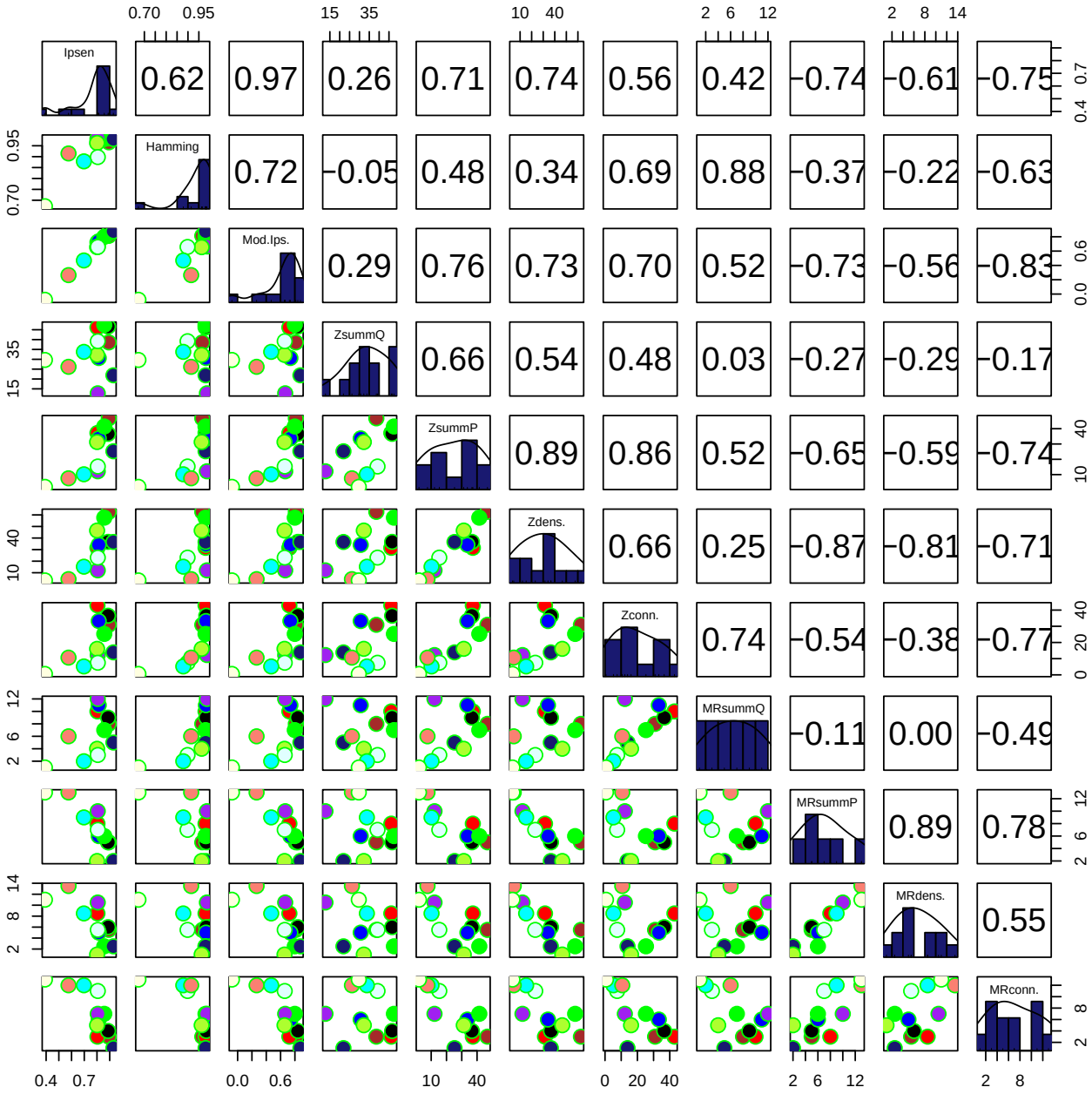


Figure 4.6: **Correlation between measures of female mouse liver module preservation in male data.** Correlation between the preservation measures of the 12 modules computed with the analized methods (Ipsen-Mikhailov (ϵ), Hamming (H), HIM (ϕ), $Z_{summaryQuality}$, $Z_{summaryPreservation}$, $Z_{density}$, $Z_{connectivity}$, $medianRank_{summaryQuality}$, $medianRank_{summaryPreservation}$, $medianRank_{density}$, $medianRank_{connectivity}$). Considering the plot as a matrix, lower triangular elements are depicted a pairplot for each couple of measures. Each circle represents one of the modules detected with WGCNA. On the diagonal we present a barplot of the distribution of the measures for each method. The upper triangular part of the plot reports the values for Spearman correlation.



Chapter 5

Stability

The network inference algorithm uncertainty has been so far assessed only in terms of performance, i.e. distance of the reconstructing network from the ground truth, wherever available, while not much has been instead investigated with respect to the stability of the methods. This can be of particular interest when no gold standard is available for the given problem, and thus there is no chance to evaluate the algorithm's accuracy, leaving the stability as the sole rule of thumb for judging the reliability of the obtained network. Here we propose to tackle the issue by quantifying inference variability with respect to data perturbation, and, in particular, data resampling (see Section 2.5).

5.1 Stability indicators

We introduce now four stability indicators that, together with a subsampling technique can be used to carry out the task of stability assessment on an inference algorithm. The scheme of such analysis is presented in Fig. 5.1.

1. Given a dataset D with s samples and p features, reconstruct (with a chosen algorithm ALG) the network N_D on the whole dataset D;

denote the p nodes of N_D by $x_1^D, \ldots, x_p^D$ and its edges' weight by a_{hk}^D, for $k, h = 1, \ldots, p$.

2. Choose two integers n, r with $n < s$ and $r \leq \binom{s}{n}$, and build a set $\mathcal{D}_{(n,r)} = \{D_1, \ldots D_r\}$ where D_i is a dataset built choosing n samples from D.

3. Reconstruct, by using the same algorithm ALG, the corresponding networks N_{D_i} on the subsampled data.

4. Compute the following indicators:

   - $I_1(n, r) = \{\text{HIM}(N_D, N_{D_i}) \colon i = 1, \ldots, r\}$
   - $I_2(n, r) = \{\text{HIM}(N_{D_i}, N_{D_j}) \colon i, j = 1, \ldots r, i \neq j\}$
   - $I_3(n, r) = \{a_{hk}^{D_i}\}$ for $i = 1, \ldots, r$ and $k, h = 1, \ldots, p$
   - $I_4(n, r) = \{\partial(x_h^{D_i})\}$ for $i = 1, \ldots, r$ and $h = 1, \ldots, p$ and ∂ the degree function.

5. For each set of values I_i compute the mean, the range (defined as the difference between maximum and minimum value) and the 95% studentized bootstrap confidence intervals [37] as implemented in the R package *boot* [30].

6. Comparative analysis of the statistics of the four indicators $I_1, \ldots I_4$ will describe the level of confidence (stability) in the network N_D, in its links and in its nodes.

The first two indicators concern the stability of the entire network, measuring the mutual distances of the networks inferred from the different replicates and their distances to the network constructed on the whole dataset. The other two indicators concern instead the stability (and thus the reliability) of the single nodes and links, in terms of mutual variability

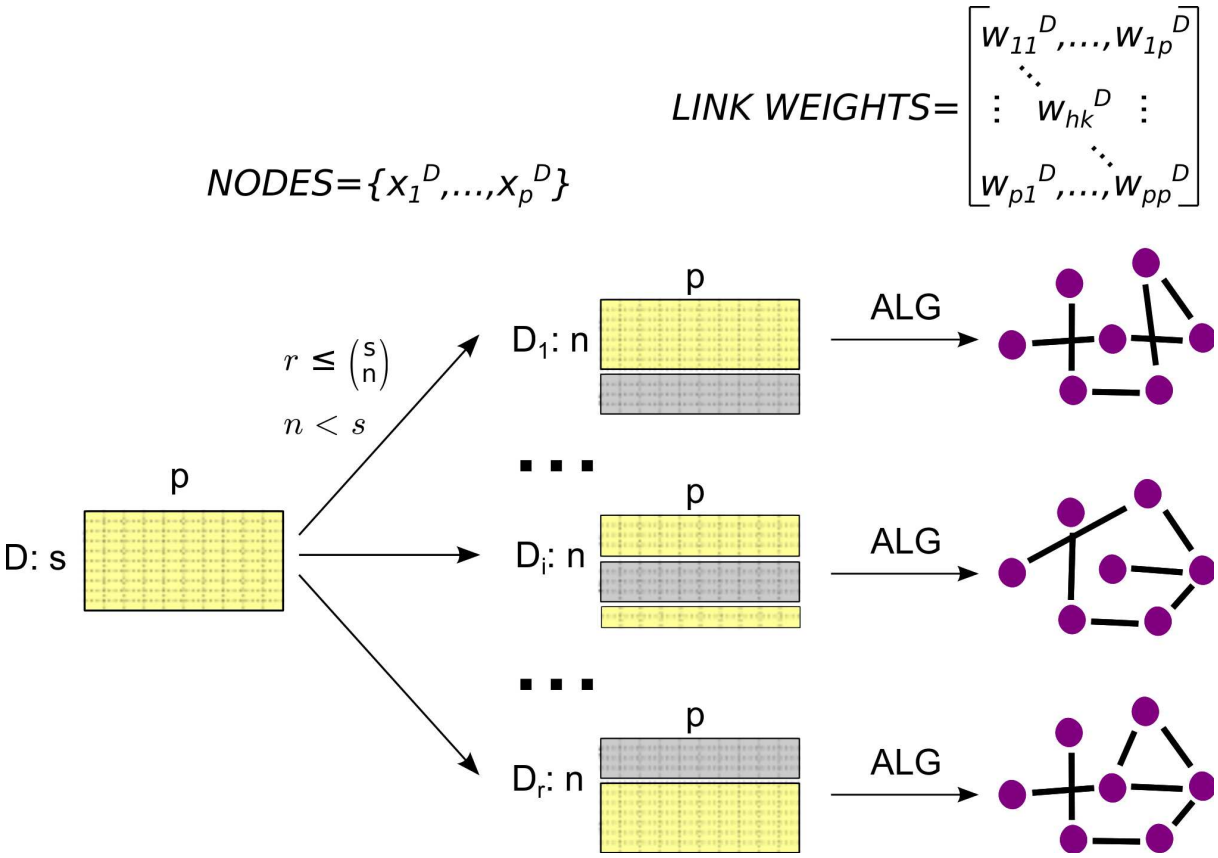


Figure 5.1: Scheme of a resampling framework applied on a dataset D made by p features and s samples. In this example the number of folds is r so that each subsample training set is made by n samples. r needs to be smaller than s choose n.

of their respective degree and weight. In particular, for all experiments on both synthetic and biological datasets we used $n = s - 1$, $r = 1$ [leave-one-out stability, LOO for short], and 20 different instances of k-fold cross validation (discarding the test portion) for $k = 2, 4, 10$ (denoted by $k2$, $k4$ and $k10$ in what follows), and thus $n = \lfloor \frac{s(k-1)}{k} \rfloor$ and $r = 20k$.

5.2 Reproducibility In Network Inference and Analysis

5.2.1 False Discovery Rate (FDR) effect on correlation networks

As a first experiment, we want to assess the different level of stability in a correlation network inferred by a set of synthetic high-throughput signals when the inference (absolute value of Pearson correlation) is computed with or without False Discovery Rate control (see for instance [72]). As the correlation measure, we use the classical (absolute) Pearson correlation of the WGCNA [61] and the novel correlation measure called Maximal Information Coefficient (MIC), component of the Maximal Information-based Nonparametric Exploration (MINE) statistics [120, 129, 106]. For a set of values $n < m$ and an adequate number of resampling $r = \min\{20, \binom{m}{n}\}$, compute the indicators $I_j(n, r)$ for $j = 1, \ldots, 4$ for all the used algorithms. We used the following pipeline to create the FDR-corrected correlation networks.

1. Let be D a dataset with m samples described by q features, and let $C(h, k) = |\mathrm{cor}(x_h, x_k)|$ where x_j is the j-th feature of D across the m samples and cor is a correlation measure.

2. Build the standard correlation network N_D using the rule $a_{hk} = C(h, k)$

3. Build the FDR controlled (at p-value $\wp = 10^{-z}$) correlation network $M_D^{\wp}$ using the rule

$$a_{hk} = \begin{cases} C(h,k) & \text{if } |F_D^z(h,k)| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

where the set F_z is defined as follows

$$F_D^z = \{\mathrm{cor}(\sigma_i(x_h), \tau_i(x_k)) \geq C(h,k) \colon \sigma_i, \tau_i \in S_m, i = 1, \ldots, \max\{10^z, m!\}\}$$

Synthetic Data Generation

As a synthetic benchmark for evaluating differences between Pearson and MIC correlation measures, and to assess the impact of the FDR filter on the construction of a correlation network, we built a dataset S consisting of 100 measurements (samples) of 20 variables (features) f_i, from which we constructed the corresponding correlation networks on 20 nodes. The dataset S was generated starting from its correlation matrix M_S, which was randomly generated with the following three constraints:

$$\mathrm{Corr}(f_i, f_j) \approx \begin{cases} 0.9 & \text{for } 1 \leq i \neq j \leq 5 \\ 0.7 & \text{for } 6 \leq i \neq j \leq 10 \\ 0.4 & \text{for } 11 \leq i \neq j \leq 16 \ , \end{cases}$$

for Corr the Pearson correlation. The correlation matrix M_S is plotted in Fig. 5.2: clearly, the correlation values in the three groups defined by the above constraints represent true relations between the variables, while all other smaller correlation values are due to the underlying random generation model for M_S.

Results

Starting from the dataset S we built five correlation networks, using MIC, absolute Pearson correlation without FDR correction (WGCNA) and absolute Pearson correlation with FDR correction, with p-values $\wp = 10^{-2}, 5 \cdot 10^{-3}, 10^{-4}$. The plots of the graphs for three of the networks are displayed in Fig. 5.3. As expected, while the WGCNA networks with highest FDR correction $\wp = 10^{-4}$ is discarding all links as not significant apart from the edges connecting the two disjoint sets of nodes $\{f_i \colon 1 \leq i \leq 5\}$ and $\{f_i \colon 6 \leq i \leq 11\}$ (the strongest correlations in the matrix M_S), WGNCA and MIC generates two fully connected networks with a majority of weak

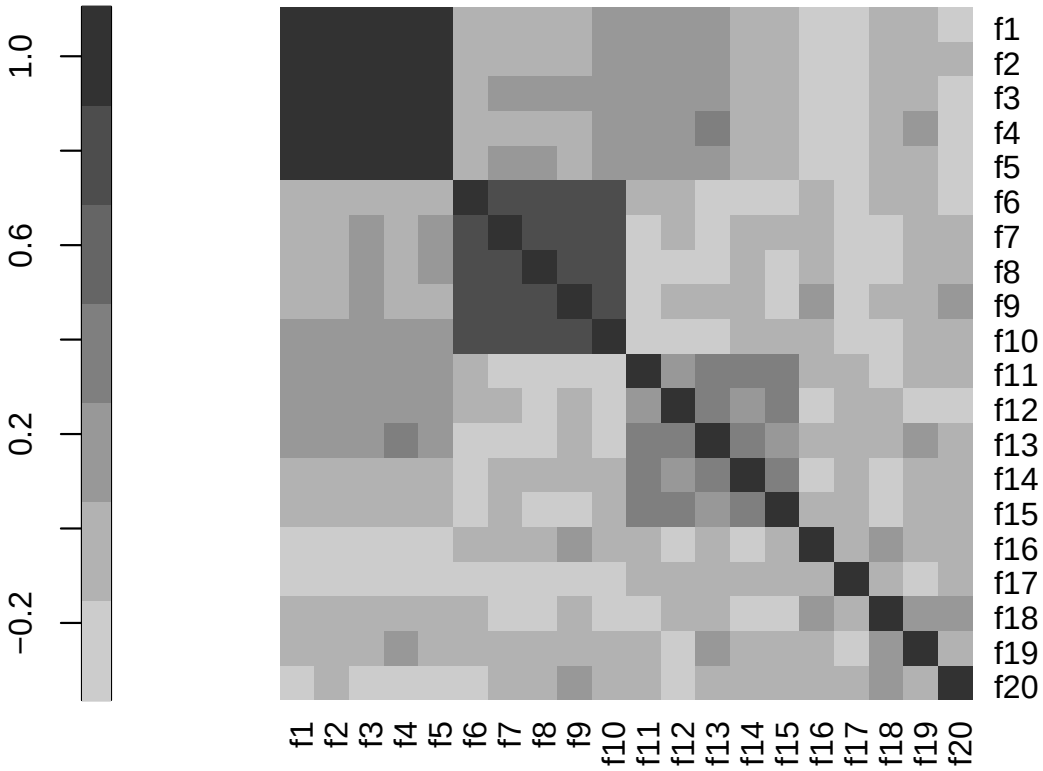
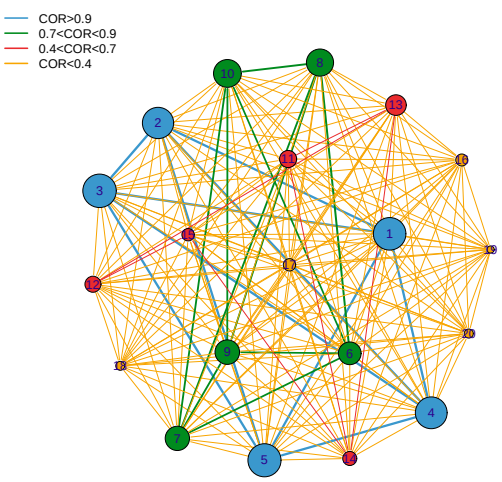


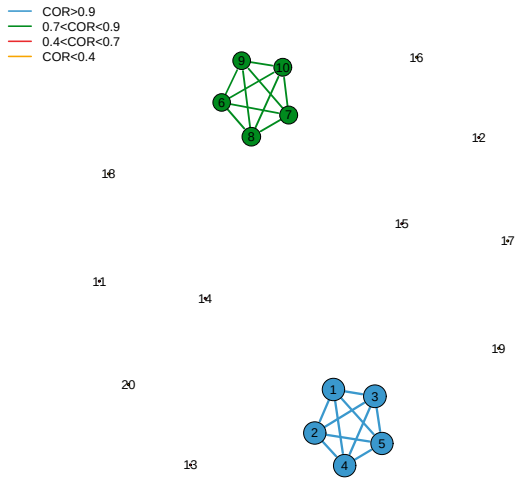
Figure 5.2: The correlation matrix M_S used to generate the synthetic dataset S

links. Then we computed the four indicators $I_1, \ldots I_4$ for all the five networks described above, in the setup described in Sec. 5.1. Main statistics for all the indicators I_1 and I_2 are reported in Tab. 5.1 and displayed in Fig. 5.4.

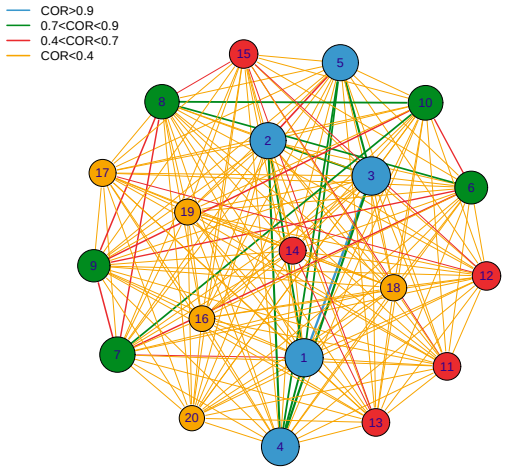
As expected, the ratio of the discarded data has a strong impact on both the indicators I_1 and I_2: in the leave-one-out case the indicators' values are close to zero regardless of the algorithm, while in the k-fold cross-validation case the stability is worsening for decreasing values of k, in terms of both mean and confidence intervals. This means that the networks inferred from a subset of data have larger distance both mutually and from the network reconstructed from the whole datasets, but also that these distances have larger variability. From the point of view of the different algorithms in-



(a) WGCNA



(b) WGCNA FDR 1e-4



(c) MIC

Figure 5.3: Correlation networks inferred by the dataset S using (a) absolute Pearson, (b) absolute Pearson with FDR correction at p-value 10^{-4} and (c) MIC. Node label i corresponds to feature f_i, node size is proportional to node degree and link colors identify different classes of link weights.

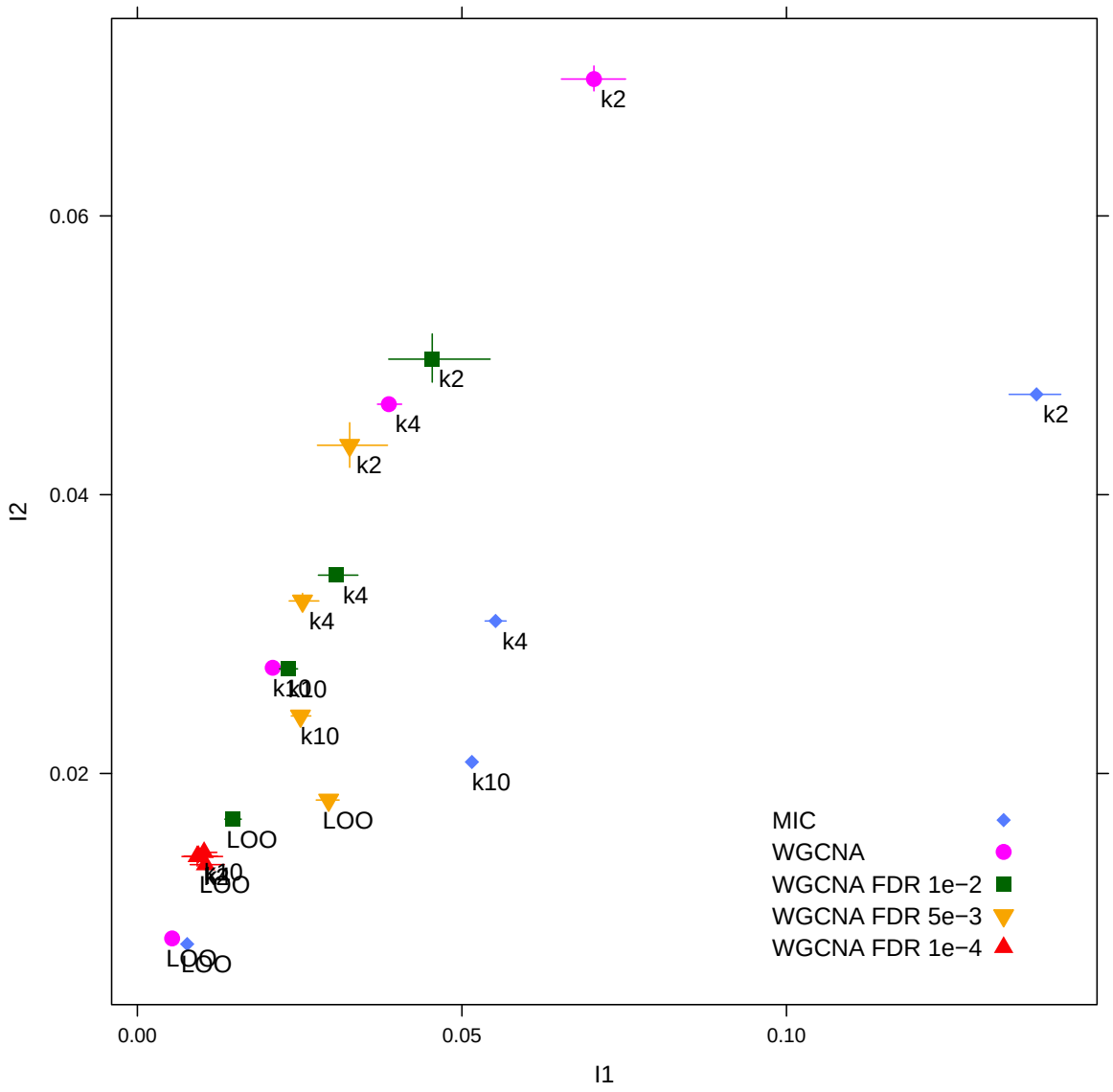


Figure 5.4: I_1 and I_2 stability indicators (mean and confidence intervals) for different instances of the WGCNA and MIC networks on the dataset S and for different values of data subsampling.

volved, the stricter the p-value in the FDR controlled WGCNA networks, the stabler the networks, with non controlled WGCNA and MINE as the worst performer in terms of stability. This is due to the fact that they are taking into account all possible correlation values, while most of the

Table 5.1: Statistics (mean, bootstrap confidence intervals and range) of the stability indicators I_1 and I_2 for different instances of the WGCNA and MIC networks on the dataset S and for different values of data subsampling.

ALG	k	I	mean	CI lower	CI upper	min	max
MIC	k10	I_1	0.052	0.051	0.052	0.041	0.067
MIC	k10	I_2	0.021	0.021	0.021	0.014	0.036
MIC	k2	I_1	0.139	0.134	0.142	0.112	0.158
MIC	k2	I_2	0.047	0.047	0.048	0.035	0.067
MIC	k4	I_1	0.055	0.054	0.057	0.040	0.071
MIC	k4	I_2	0.031	0.031	0.031	0.022	0.045
MIC	LOO	I_1	0.008	0.007	0.008	0.004	0.011
MIC	LOO	I_2	0.008	0.008	0.008	0.003	0.014
WGCNA	k10	I_1	0.021	0.020	0.022	0.011	0.040
WGCNA	k10	I_2	0.028	0.028	0.028	0.012	0.064
WGCNA	k2	I_1	0.070	0.065	0.076	0.037	0.108
WGCNA	k2	I_2	0.070	0.069	0.071	0.042	0.117
WGCNA	k4	I_1	0.039	0.037	0.041	0.020	0.062
WGCNA	k4	I_2	0.046	0.046	0.047	0.025	0.088
WGCNA	LOO	I_1	0.005	0.005	0.006	0.001	0.015
WGCNA	LOO	I_2	0.008	0.008	0.008	0.002	0.023
WGCNA FDR 1e-2	k10	I_1	0.023	0.022	0.025	0.007	0.074
WGCNA FDR 1e-2	k10	I_2	0.028	0.027	0.028	0.002	0.102
WGCNA FDR 1e-2	k2	I_1	0.045	0.039	0.054	0.014	0.107
WGCNA FDR 1e-2	k2	I_2	0.050	0.048	0.051	0.006	0.152
WGCNA FDR 1e-2	k4	I_1	0.031	0.028	0.034	0.010	0.069
WGCNA FDR 1e-2	k4	I_2	0.034	0.034	0.035	0.006	0.096
WGCNA FDR 1e-2	LOO	I_1	0.015	0.013	0.016	0.005	0.035
WGCNA FDR 1e-2	LOO	I_2	0.017	0.017	0.017	0.001	0.047
WGCNA FDR 5e-3	k10	I_1	0.025	0.024	0.027	0.004	0.054
WGCNA FDR 5e-3	k10	I_2	0.024	0.024	0.024	0.001	0.083
WGCNA FDR 5e-3	k2	I_1	0.033	0.028	0.038	0.008	0.070
WGCNA FDR 5e-3	k2	I_2	0.044	0.042	0.045	0.002	0.121
WGCNA FDR 5e-3	k4	I_1	0.025	0.023	0.028	0.006	0.056
WGCNA FDR 5e-3	k4	I_2	0.032	0.032	0.033	0.004	0.099
WGCNA FDR 5e-3	LOO	I_1	0.029	0.028	0.031	0.003	0.048
WGCNA FDR 5e-3	LOO	I_2	0.018	0.018	0.018	0.000	0.054
WGCNA FDR 1e-4	k10	I_1	0.010	0.009	0.012	0.000	0.053
WGCNA FDR 1e-4	k10	I_2	0.014	0.014	0.015	0.000	0.055
WGCNA FDR 1e-4	k2	I_1	0.009	0.007	0.013	0.001	0.031
WGCNA FDR 1e-4	k2	I_2	0.014	0.013	0.015	0.001	0.040
WGCNA FDR 1e-4	k4	I_1	0.009	0.007	0.012	0.001	0.049
WGCNA FDR 1e-4	k4	I_2	0.014	0.014	0.014	0.001	0.054
WGCNA FDR 1e-4	LOO	I_1	0.010	0.008	0.013	0.000	0.044
WGCNA FDR 1e-4	LOO	I_2	0.013	0.013	0.014	0.000	0.045

smaller values do not represent existing relations between variables, but they are rather a noise effect. As a first result then we showed that the use

of a FDR control procedure for correlation help stabilizing the inference procedure, improving the performance of a method already acknowledged as effective [3].

We move now on to discuss the stablest links and nodes in the three cases WGCNA, WGCNA FDR 1e-4 and MIC: in particular, in Tab. 5.2 and 5.3 we show the top-ranked links and nodes ordered for decreasing range over mean of their weights across all resampling $k4$. The results collected in the tables are consistent with the structure of the starting correlation matrix M_S and the behaviour of the inference algorithms. For the WGCNA case, the top 20 stablest links are those of the two fully connected subgroups $F_{1,5} = \{f_i : 1 \leq i \leq 5\}$ and $F_{6,10}\{f_i : 6 \leq i \leq 10\}$ with largest Pearson correlation values in M_S. The same applies to WGCNA FDR 1e-4 (and with approximately the same values of weight range over weight mean as for WGCNA), for which these 20 links are the only existing (see Fig. 5.3). Among the following ranked links in WGCNA, those belonging to the $F_{11,15} = \{f_i : 11 \leq i \leq 15\}$ group (whose correlation of about 0.3 was imposed as a constraint for M_S) are emerging, with a couple of exceptions, but with larger instability values (0.33-0.78 vs. 0.03-0.14). The remaining links are the unstablest, displaying Range/Mean values always larger than 0.83: they are the randomly correlated links of M_S. It is interesting to note that the MIC network, due to the nature of the MIC statistics aimed at detecting relations between variables other than linear, displays similar but not identical results: the values of Range/Mean are confined in a narrower interval, and, although many links belonging to the $F_{1,5}$ and $F_{6,10}$ groups are highly ranked, some of them can also be found in much lower positions of the standing.

Similar considerations hold for the ranking of the stablest nodes: for WGCNA, the top ranking nodes are the $F_{1,5}$ and the $F_{6,10}$ (with similar Range/Mean values), with those in $F_{11,15}$ come next, leaving the remain-

ing five as the most unstable, with higher Range/Mean values. These five nodes, on the contrary, are the stablest for WGCNA FDR 1e-4: in fact, they are not wired to any other node in any of the resampling, so their Range/Mean values are void. The nodes $F_{1,5} \cup F_{6,10}$ then follow in the ranking with small associated values, and the nodes $F_{11,15}$ close the standing with definitely higher values. In fact, although the nodes $F_{11,15}$ have degree zero in the WGCNA FDR 1e-4 inferred from the whole S, some links involving them exist in some of the resampling on the subset of data. To conclude with, in the MIC case again the ranking values span a much narrower range than the other two cases, and the obtained dwranking has most of the nodes in $F_{1,5}$ in top positions, while for the other nodes the relation with the structure of M_S is very weak.
Finally, the analogous tables for other ratios of the data subsampling schema (LOO, $k2$ and $k10$) are almost identical.

5.3 Inference Methods Comparison on Synthetic Data

We chose to analyze the performances of some of the most commonly used inference algorithms such as:

- Aracne *(ARA)* [101]
- Context likelihood of relatedness *(CLR)* [46]

and some novel ones like:

- RegnANN *(REG)* 2.3.5
- Maximum Information based *(MIC)* 2.4.3
- Bicorrelation method *(BIC)* 2.4.2

- Gene coexpression method *(COR)* 2.4.1
- Topological Overlap Matrix *(TOM)* 2.3.2
- We also considered the *BIC* and *COR* methods within a false discovery rate *(FDR)* control framework 5.2.1.

5.3.1 Synthetic Data

Data Generation

We chose to use a 20 nodes network with 5 regulators and 42 interactions were randomly generated as in Fig. 5.5. The kinetic model of the network was generated mimicking a biologically plausible one and in particular that of *Escherichia Coli* [126] (note that no self loops are present in the topology). A synthetic gene expression dataset was generated simulating 20 steady states levels of variations of the network, which were obtained by applying multifactorial perturbations to the original network. We simulate multifactorial perturbations by slightly increasing or decreasing the basal activation of all genes of the network simultaneously by different random amounts. We considered each experiment as a gene expression profile from a different patient. We chose to use the model of noise in microarrays that was used for the *DREAM4* challenges [116], which is similar to a mix of normal and log-normal noise. The benchmark data we used is made by 10 different generations of the synthetic gene expression from the same kinetic model. This benchmark was needed to evaluate also the stability of the tested inference methods. Both the generation of the topology and the generation of the dataset were performed with GeneNetWeaver [123].

Results

Here we show the performances of some inference algorithms both in terms of accuracy and stability. As distance measure we use the HIM combina-

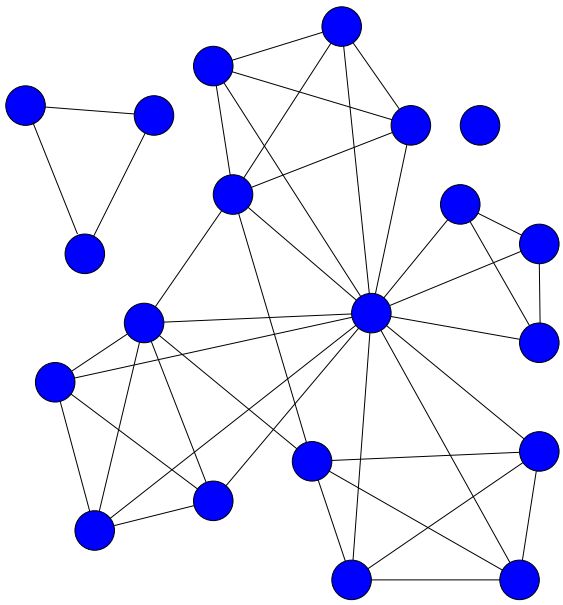


Figure 5.5: Random topology generated with GeneNetWeaver (20 nodes, 5 regulators, 42 links).

tion between the Ipsen-Mikhailov and the Hamming distance described in Section 4.

In Fig. 5.7 we can notice how the two classic inference algorithms Aracne and CLR clearly outperform the others, since the HIM distance between the inferred networks they produce and the gold standard is less than a half of the one produced with the other systems that do not make use of the *FDR* correction. The good performances of ARA and CLR can anyways be explained with the fact that these two algorithms solve the inference problem using a mutual information-based method and it was expected since they were the ones who best performed in the DREAM4 Network Inference Challenge in which GeneNetWeaver was used to generate the data. It is also interesting to see that the confidence intervals vary greatly

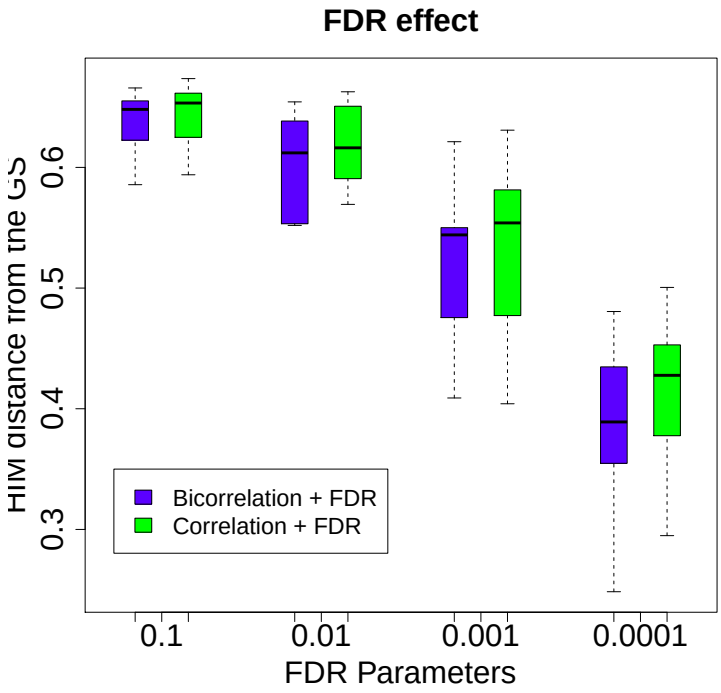


Figure 5.6: The effect of different FDR settings on accuracy and stability of network inference performed with correlation and bicorrelation.

among the algorithms and in particular RegnANN seems to be the most stable.

Moreover we can see how the use of *FDR* correction to the correlation and bicorrelation-based methods leads to a clear improvement in the accuracy of the inference, but the cost is an evident worsening in the stability of the performances. This phenomenon is clearly represented in Fig. 5.6 where the HIM distance is depicted against the *FDR* parameter. The increase of *FDR* leads to a better average accuracy, but also to a degradation of the stability of the result. It is safe to say that the use of the correction can be a very important tool, but it is also crucial to choose the best trade off between accuracy and stability. The degradation of the stability can be explained with the fact that the *FDR* correction, that practically implement a hard thresholding on the adjacency matrix, applied to noisy

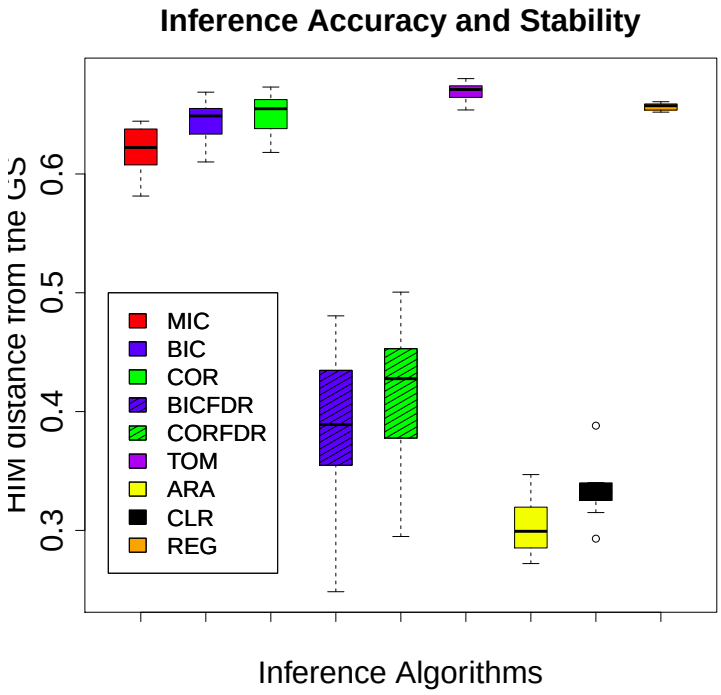


Figure 5.7: Performances of the 9 inference algorithm tested on synthetic dataset computed ad HIM distance from the gold standard (GS). FDR=10^{-4}

data can lead to a variable number of false negative that are reduced in a more conservative approach (lower *FDR* parameter).

5.3.2 *Escherichia Coli* Data

Data Description

We selected a sub-network *GS* of the gene regulatory network of *Escherichia Coli* made by 50 nodes and their 102 connections. *GS* includes 5 randomly chosen regulators *(arcA, rutR, gadE, gadX, gadW)* and their neighbors, as shown in Fig. 5.8 the links of the topology are directed, but we decided to consider them bidirectional. As for the synthetic example we simulate multifactorial perturbations by slightly modifying the basal activation of all genes of the network simultaneously by different random

amounts. We obtained 50 steady states simulations for the 50 nodes of the network; starting from these simulations, using GeneNetWeaver, we produced 10 different generations of the synthetic gene expression from the same kinetic model obtaining a benchmark for accuracy and stability analysis.

Results

The boxplot 5.10 highlights the huge difference in accuracy between Aracne and CLR and the other in favor of the two classic algorithms. The stability across the 10 data generations is very high for all the methods. In Fig. 5.9 we can see how the FDR correction influences the results for bicorrelation and correlation methods, as shown in 5.6 decreasing the FDR value the accuracy increases while we have a degradation in the stability of the results. Though the overall behavior is the same as in the synthetic data, the scale of the plot shows that the effect is much reduced in the case of *E.Coli* data with respect of the synthetic data.

Table 5.2: Top ranked links, ordered by weight range over weight mean across all 20 resampling of $k4$ 4-fold cross validation, for the three algorithms WGCNA, WGCNAFDR1e-4 and MIC

WGCNA		WGCNA FDR 1e-4		MIC	
$f_i - f_j$	Range/Mean	$f_i - f_j$	Range/Mean	$f_i - f_j$	Range/Mean
1 - 3	0.03	1 - 3	0.03	3 - 4	0.20
2 - 3	0.04	3 - 4	0.04	2 - 3	0.20
1 - 2	0.04	2 - 3	0.04	1 - 3	0.21
1 - 4	0.04	1 - 4	0.05	3 - 5	0.22
3 - 4	0.04	3 - 5	0.05	1 - 2	0.23
2 - 4	0.04	1 - 2	0.05	1 - 5	0.25
4 - 5	0.04	2 - 4	0.05	1 - 4	0.26
2 - 5	0.05	2 - 5	0.06	4 - 5	0.27
1 - 5	0.05	4 - 5	0.06	7 - 10	0.28
3 - 5	0.05	1 - 5	0.06	7 - 8	0.29
6 - 8	0.08	6 - 8	0.08	6 - 8	0.29
8 - 10	0.10	7 - 8	0.09	6 - 10	0.30
7 - 8	0.11	8 - 10	0.10	1 - 20	0.31
7 - 9	0.11	8 - 9	0.11	2 - 4	0.31
8 - 9	0.11	6 - 7	0.11	8 - 10	0.31
9 - 10	0.11	7 - 10	0.12	2 - 5	0.32
6 - 7	0.11	7 - 9	0.12	9 - 10	0.32
7 - 10	0.12	9 - 10	0.13	7 - 20	0.33
6 - 10	0.13	6 - 9	0.13	14 - 16	0.33
6 - 9	0.14	6 - 10	0.15	5 - 17	0.35
11 - 13	0.33			6 - 7	0.35
14 - 15	0.41			11 - 17	0.36
13 - 14	0.46			6 - 9	0.36
12 - 13	0.58			1 - 10	0.37
12 - 15	0.60			10 - 11	0.37
11 - 14	0.62			10 - 20	0.37
13 - 15	0.71			4 - 17	0.37
11 - 15	0.78			2 - 8	0.37
14 - 18	0.78			4 - 10	0.37
3 - 11	0.83			6 - 13	0.37
5 - 11	0.83			2 - 14	0.37
1 - 11	0.84			9 - 11	0.38
4 - 11	0.85			15 - 16	0.38
3 - 10	0.87			15 - 17	0.38
5 - 16	0.89			7 - 13	0.39
8 - 17	0.89			9 - 18	0.39
2 - 11	0.91			12 - 19	0.39
8 - 12	0.91			6 - 18	0.39
4 - 13	0.91			8 - 9	0.39
1 - 13	0.93			4 - 18	0.39
3 - 13	0.93			16 - 17	0.39
8 - 13	0.94			4 - 19	0.39
9 - 17	0.94			16 - 19	0.39
1 - 16	0.95			7 - 19	0.40
1 - 10	0.95			5 - 8	0.40
14 - 16	0.97			14 - 15	0.40
5 - 10	0.97			13 - 15	0.40
11 - 12	0.98			4 - 11	0.40
12 - 16	0.98			7 - 9	0.41
2 - 13	0.99			13 - 19	0.41

Table 5.3: Top ranked nodes, ordered by degree range over degree mean across all 20 resampling of $k4$ 4-fold cross validation, for the three algorithms WGCNA, WGCNA FDR 1e-4 and MIC. (*) indicates that Ratio and Mean are both zero.

WGCNA		WGCNA FDR 1e-4		MIC	
f_i	Range/Mean	f_i	Range/Mean	f_i	Range/Mean
4	0.17	16	0*	3	0.08
10	0.18	17	0*	19	0.08
3	0.20	18	0*	1	0.08
1	0.21	19	0*	4	0.09
9	0.23	20	0*	8	0.09
2	0.23	3	0.03	10	0.09
5	0.24	1	0.04	5	0.10
7	0.24	2	0.04	2	0.10
6	0.24	5	0.05	17	0.10
8	0.25	7	0.07	20	0.10
11	0.40	8	0.07	15	0.11
13	0.40	6	0.09	9	0.11
15	0.43	9	0.09	13	0.11
12	0.45	10	0.09	11	0.11
14	0.48	4	0.13	16	0.11
18	0.55	15	4.42	12	0.11
16	0.60	14	7.05	7	0.11
17	0.68	12	22.82	6	0.12
20	0.70	13	26.05	14	0.13
19	1.15	11	41.83	18	0.13

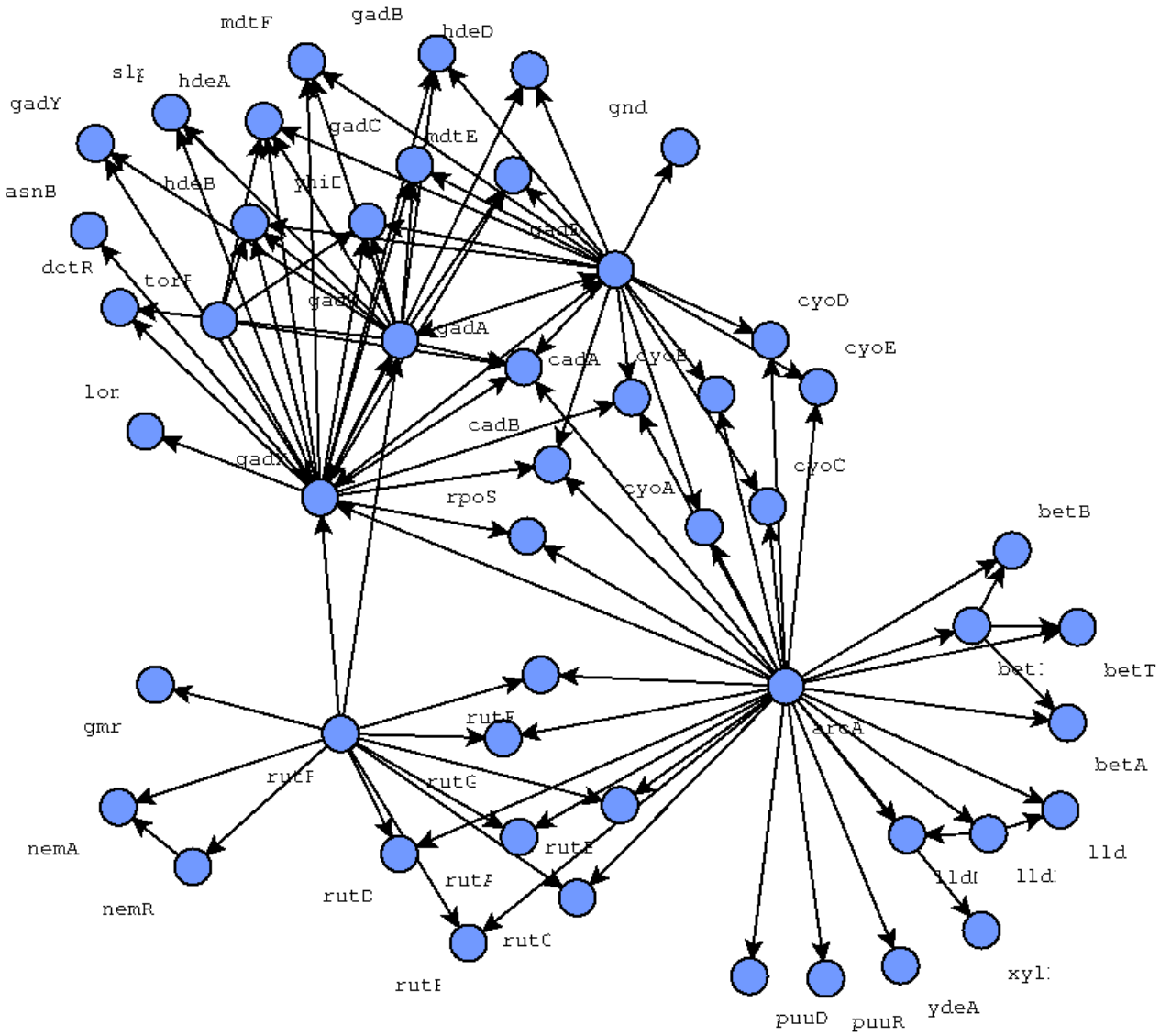


Figure 5.8: A subnetwork of *Escherichia Coli* consisting of 50 nodes and their 102 connections; in particular notice the connections involving the 5 regulators *(arcA, rutR, gadE, gadX, gadW)*.

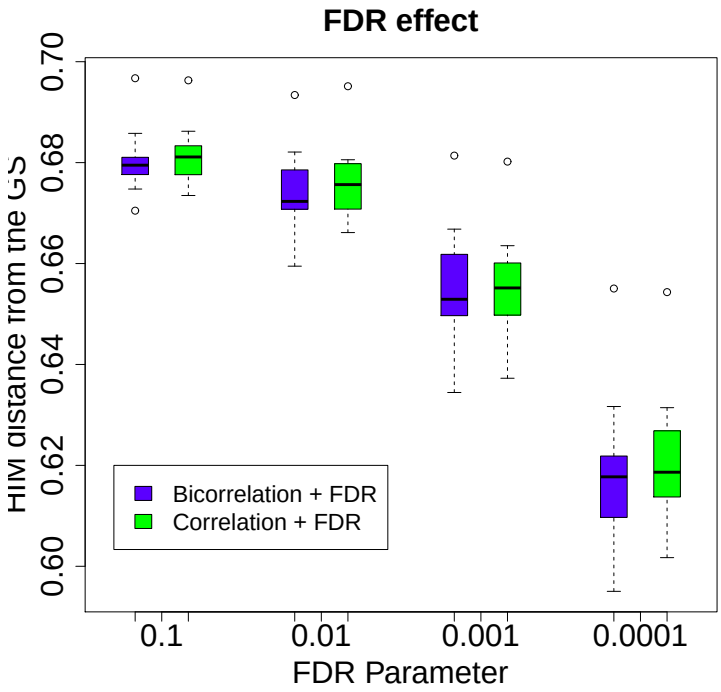


Figure 5.9: The effect of different FDR settings on accuracy and stability of network inference performed with correlation and bicorrelation.

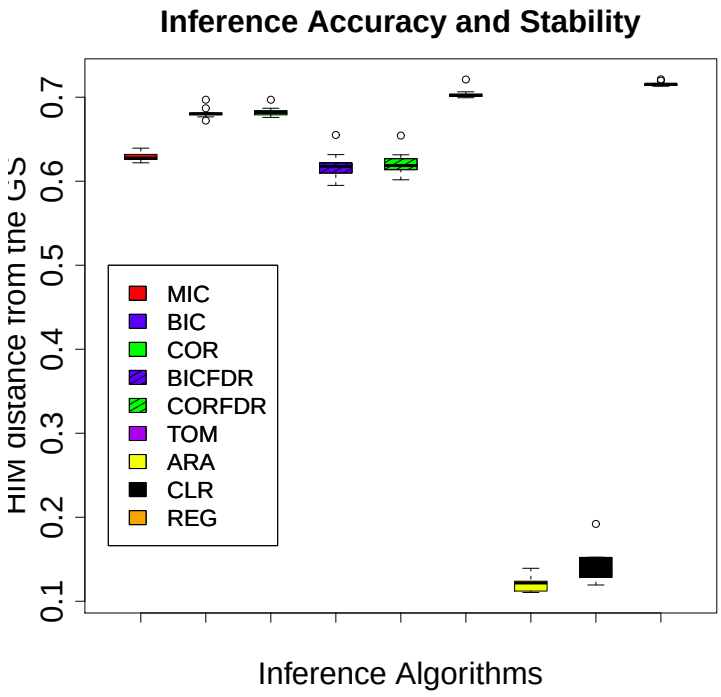


Figure 5.10: Performances of the 9 inference algorithm tested on the *E.Coli* subnetwork dataset computed ad HIM distance from the gold standard (GS). FDR=10^{-4}

Chapter 6

Differential Networking

In this chapter we present three applications of network comparison and stability assessment in the framework of (biological) differential network analysis.

6.1 Biological Network Comparison: a miRNA example

Investigating the relations connecting human microRNA (miRNA) and how they evolve in cancer has been recently a key topic for researcher in biology [147, 11], with hepatocellular carcinoma (HCC) as a notable example [89, 57]. In the following example, we use the stability indicators $I_1, \ldots, I_4$ on a recent miRNA microarray dataset with two phenotypes to highlight differences in the corresponding inferred networks. As reconstruction algorithm we use the Context Likelihood of Relatedness (CLR) approach [46], belonging to the relevance networks class of algorithms and generating undirected weighted graphs with weights bounded between zero and one. In particular, interactions are scored by using the mutual information between the corresponding gene expression levels coupled with an adaptive background correction step. Although suboptimal if the number

of variables is much larger than the number of variables, it was observed that CLR performs well in terms of prediction accuracy and some CLR predictions in literature were later experimentally validated [5].

Data description

We start out from the Hepatocellular Carcinoma dataset introduced in the paper [28] and later used in [71], publicly available at the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) at the accession number GSE6857. The dataset collects 482 tissue samples from 241 patients affected by hepatocellular carcinoma (HCC). For each patients, a sample from cancerous hepatic tissue and a sample from surrounding non-cancerous hepatic tissue are available, hybridized on the Ohio State University CCC MicroRNA Microarray Version 2.0 platform consisting of 11520 probes collecting expressions of 250 non-redundant human and 200 mouse microRNA (miRNA). After a preprocessing phase including imputation of missing values as in [141] and discarding probes corresponding to non-human (mouse and controls) miRNA, we end up with the dataset $\mathcal{HCC}$ of 240+240 paired samples described by 210 human miRNA, with the cohort consisting of 210 male and 30 female patients. We thus parted the whole dataset $\mathcal{HCC}$ into four subsets combining the sex and disease status phenotypes, collecting respectively the cancer tissue for the male patients (MT), the cancer tissue for the female patients (FT) and the corresponding two datasets including the non cancer tissues (MnT, FnT).

Results

Using the CLR algorithm we first generated the four networks inferred from the whole sets of data and corresponding to the combinations of the two binary phenotypes: a portrait of the resulting graphs is depicted in Fig. 6.2, discarding links whose weight is smaller than 0.1. As a first ob-

servation, the four networks have a different structure, for instance the tumoral tissues graphs being more connected than the controls and the female graphs more than the corresponding male ones (see for instance the density values in Fig. 6.2). In particular, their mutual HIM distances are reported in Tab. 6.1, together with the corresponding two-dimensional scaling plot, showing that the networks corresponding to the female patients (and, in particular, the one inferred from cancer tissue) are notably different from those arising from the subset of data for the male patients. We then computed the stability indicators I_1 and I_2 in the setup described

MnT	FT	FnT	
0.0412	0.0858	0.0235	MT
	0.1265	0.0618	MnT
		0.0684	FT

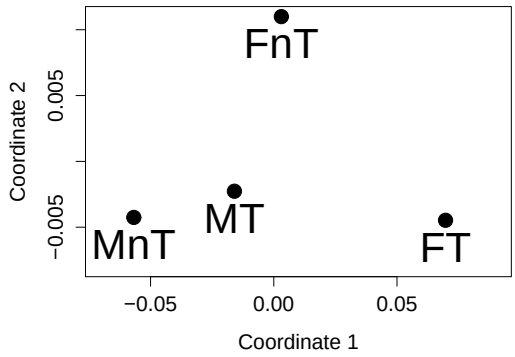


Figure 6.1: Mutual HIM distances for the four CLR inferred networks MT, MnT, FT, FnT reconstructed from the whole corresponding subsets and corresponding 2D multidimensional scaling plot.

in Sec. 5.1, and the corresponding statistics are collected and displayed in Tab. 6.1 and Fig. 6.3.

It is immediately evident the different sample size impact on the network stability: the networks corresponding to male patients have smaller values for I_1 and I_2 (and thus they are much stabler) than the corresponding female counterparts, and this effect is even stronger than the one due to the ratio of the chosen subsets of data: the leave-one-out stability for FT and FnT is worse than k10 and k4 stability for MT and MnT. On the other

hand, while control and cancer networks display similar level of stability in the male networks at all levels of subsampling ratio, in the female group the network associated to the controls is much stabler than the matching control networks, and this is evident when the size of the subset used for inference gets smaller, in particular for $k = 2$.

Table 6.1: Statistics (mean, bootstrap confidence intervals and range) of the stability indicators I_1 and I_2 for the CLR inferred networks on the datasets MT, MnT, FT, FnT, for different values of data subsampling.

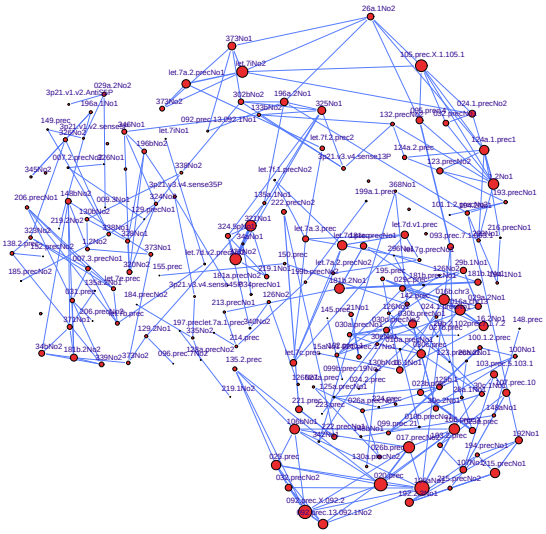
PROBL	k	I	mean	lower	upper	min	max
FT	k10	I_1	0.040	0.037	0.044	0.002	0.177
FT	k10	I_2	0.054	0.054	0.055	0.000	0.256
FT	k2	I_1	0.069	0.056	0.082	0.006	0.154
FT	k2	I_2	0.089	0.084	0.093	0.005	0.250
FT	k4	I_1	0.057	0.049	0.066	0.004	0.190
FT	k4	I_2	0.078	0.076	0.080	0.003	0.305
FT	LOO	I_1	0.022	0.016	0.032	0.002	0.093
FT	LOO	I_2	0.032	0.030	0.035	0.001	0.143
FnT	k10	I_1	0.032	0.029	0.035	0.002	0.093
FnT	k10	I_2	0.045	0.044	0.045	0.000	0.179
FnT	k2	I_1	0.094	0.071	0.117	0.006	0.257
FnT	k2	I_2	0.119	0.113	0.124	0.006	0.391
FnT	k4	I_1	0.062	0.054	0.072	0.005	0.203
FnT	k4	I_2	0.080	0.078	0.082	0.003	0.307
FnT	LOO	I_1	0.022	0.017	0.027	0.003	0.048
FnT	LOO	I_2	0.030	0.028	0.032	0.001	0.094
MT	k10	I_1	0.011	0.010	0.013	0.001	0.048
MT	k10	I_2	0.016	0.016	0.016	0.001	0.092
MT	k2	I_1	0.040	0.033	0.051	0.003	0.146
MT	k2	I_2	0.051	0.048	0.054	0.003	0.218
MT	k4	I_1	0.024	0.020	0.029	0.002	0.099
MT	k4	I_2	0.033	0.032	0.033	0.001	0.148
MT	LOO	I_1	0.002	0.002	0.002	0.000	0.018
MT	LOO	I_2	0.003	0.003	0.003	0.000	0.030
MnT	k10	I_1	0.009	0.008	0.010	0.001	0.034
MnT	k10	I_2	0.013	0.013	0.013	0.001	0.061
MnT	k2	I_1	0.033	0.026	0.041	0.003	0.104
MnT	k2	I_2	0.037	0.035	0.039	0.002	0.158
MnT	k4	I_1	0.018	0.015	0.022	0.001	0.067
MnT	k4	I_2	0.025	0.024	0.026	0.001	0.102
MnT	LOO	I_1	0.002	0.002	0.002	0.000	0.009
MnT	LOO	I_2	0.003	0.003	0.003	0.000	0.016

Finally, to show how to use indicators I_3 and I_4 to extract information about stability of some interesting links, we first rank all links according to

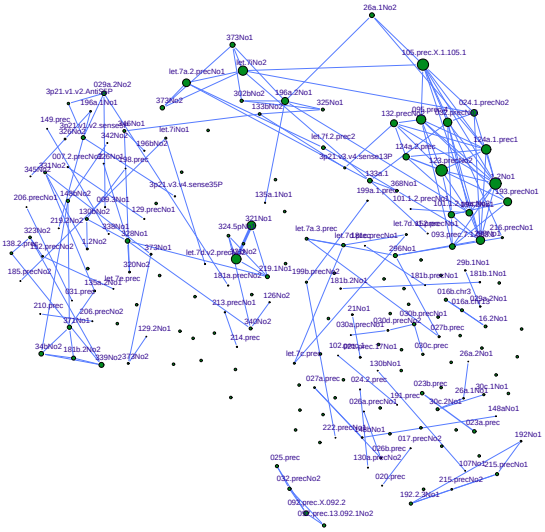
their weight Range/Mean value for all the four cases MT, MnT, FT, FnT, and then we point out six links worth a comment, listed in Tab. 6.2. The link (a) is top ranking in all four cases as expected, since *hsa-mir_321No1* and *hsa-mir_321No2* denote essentially the same miRNA (identical or with very similar sequences, [6]. The same applies to the links (b) and (c), but in these cases the stability of these two links in the FnT network is not as good as in the other three cases, probably due to the presence of noise in the data. The link (d) is interesting because of the difference of its stability between the male and the female networks, indicating a link probably associated to sex rather than HCC. The behavior of link (e) is even more singular: it is one of the stablest links for the FT network, while is not even picked up as a link by CLR in the FnT network. Finally, link (f) is a very well known connection in literature, strongly associated to cancer [147, 24, 51] as confirmed by its high stability in the MT and FT networks only.

Table 6.2: Position in the weight Range/Mean ranking in the four cases MT, MnT, FT, FnT for six miRNA-miRNA links.

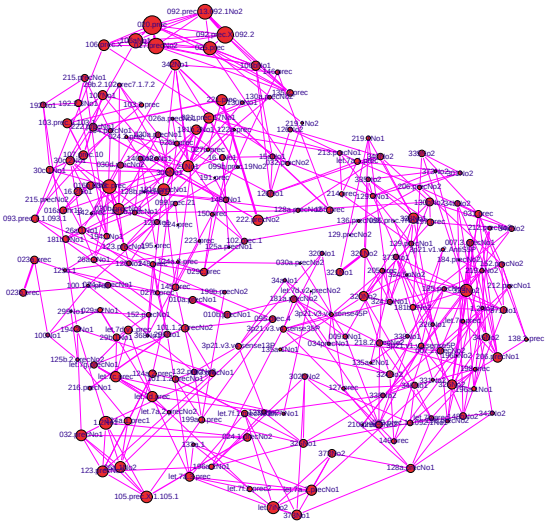
id	hsa-mir_idx1	hsa-mir_idx2	MT	MnT	FT	FnT
(a)	321No1	321No2	1	1	9	2
(b)	016b.chr3	16.2No1 3	12	15	309	
(c)	021.prec.17No1	21No1	27	5	2	921
(d)	219.1No1	321No2	2	6	1903	314
(e)	326No1	342No2	132	1017	3	-
(f)	192.2.3No1	215.precNo1	4	300	4	3340



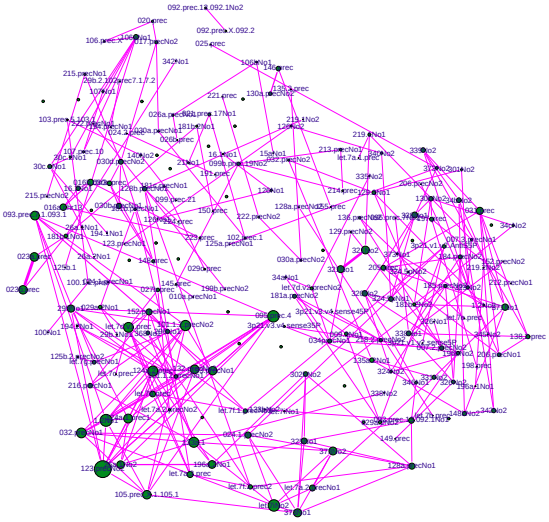
(a) MT, $d = 0.0153$



(b) MnT, $d = 0.0092$



(c) FT, $d = 0.0206$



(d) FnT, $d = 0.0121$

Figure 6.2: CLR networks (and corresponding density values) inferred from the 4 subsets (a) Male Tumoral (MT) (b) Male not Tumoral (MnT) (c) Female Tumoral (FT) and (d) Female non Tumoral (FnT) of the datasets $\mathcal{HCC}$. Links are thresholded at weight 0.1, node position is fixed across the four networks, node dimension is proportional to the degree and edge width is proportional to link weight.

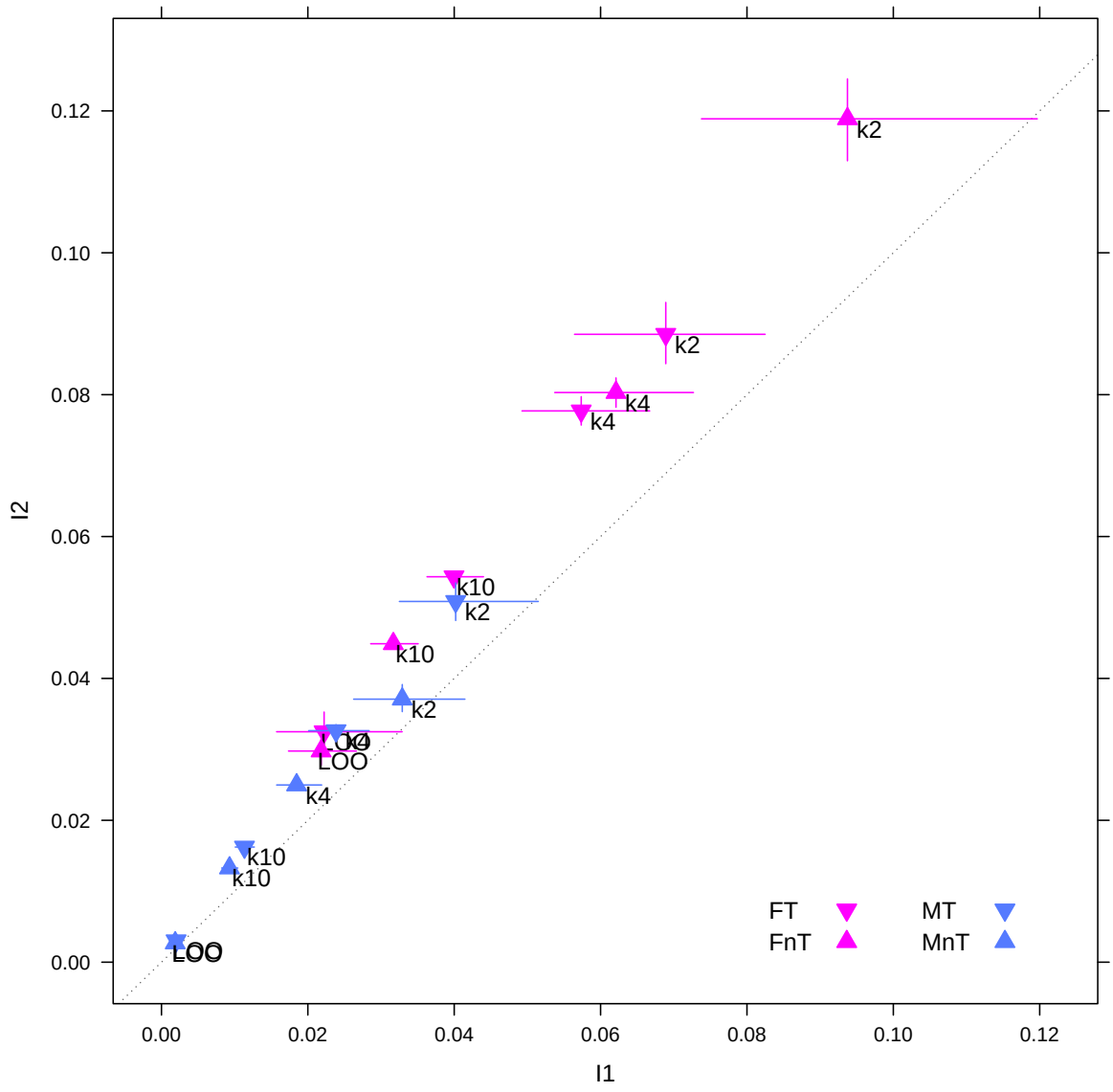


Figure 6.3: I_1 and I_2 stability indicators (mean and confidence intervals) of CLR inferred networks for different values of data subsampling on the four subgroups Male Tumoral (MT), Male not Tumoral (MnT), Female Tumoral (FT) and Female non Tumoral (FnT) of the datasets $\mathcal{HCC}$.

6.2 Sources of Variability in Pathway Profiling

We apply the HIM distance within a framework which includes a set of network medicine tools (see schema in Fig. 6.4). The framework is adapted from a computational pipeline for benchmarking feature selection algorithms, enrichment procedure and network inference methods [19]. Here we discuss the main modules composing the framework.

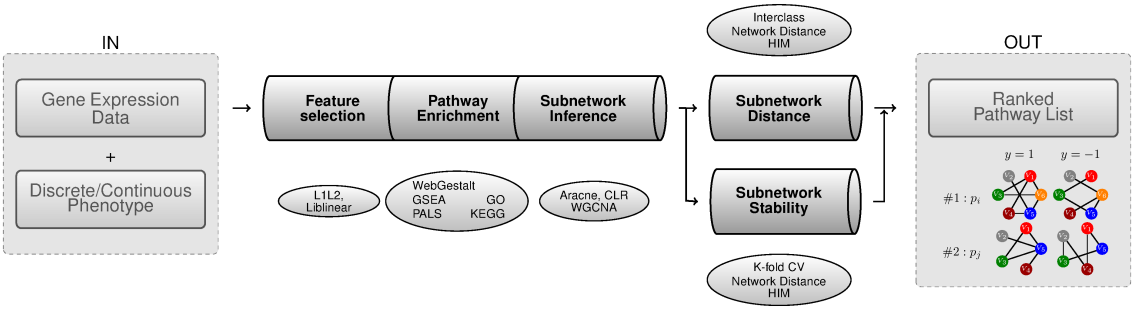


Figure 6.4: The general scheme of the HIM framework. Algorithms and tools used in the PD study are listed in ovals.

$\mathcal{M}$ *module.* In the first step, the most relevant features are selected by means of a predictive model $\mathcal{M}$ according to a proper Data Analysis Protocol (DAP), as proposed in [138]. For $\mathcal{M}$, we consider first the $\ell_1\ell_2$ regularization algorithm with double optimization [38], which can be tuned to give a minimal set of discriminative genes or larger sets including correlated genes and it is based on the optimization principle presented in [159]. The $\ell_1\ell_2$ DAP is implemented in two stages organized as nested loops of 10-fold cross-validation [20]. The first stage identifies the minimal set of relevant variables in terms of prediction error; starting from the minimal list, the second one selects the family of nested lists of relevant variables for increasing values of linear correlation. As alternative model, we consider Liblinear, a linear Support Vector Machine (SVM) classifier specifically designed for large datasets [47]. In particular, the classical dual optimization problem with L2-SVM loss function is solved with a coordinate descent

method. For our experiment we adopt the ℓ_2-regularized penalty term and the module of the weights for ranking purposes within a 100×3-fold cross validation schema. We build a model for increasing feature sublists where the feature ranking is defined according to the importance for the classifier. We choose the model, and thus the top ranked features, by balancing classifier accuracy and signature stability [73].

$\mathcal{E}$-$\mathcal{D}$ module. Enrichment procedures ($\mathcal{E}$) are knowledge-based pathway analysis methods, which exploit the information stored as in public repositories ($\mathcal{D}$), such as the Gene Ontology (GO) [136] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [77], both used in this study. For each term, GO provides a level of evidence, in total 22 evidences grouped in six categories. We used levels: IMP, inferred from mutant phenotype; IGI, from genetic interaction; IPI, from physical interaction; ISS, from sequence similarity; IDA, from direct assay; IEP, from expression pattern; IEA, from electronic annotation [137].
The two knowledge bases GO and KEGG were used in combination with three enrichment methods. As they can be categorized according to the underlying algorithm [63, 81], we considered WebGestalt as representative of the Singular Enrichment Analysis family, GSEA for the Gene Set Enrichment Analysis one, and the Pathways and Literature Strainer (PaLS) for the Modular Enrichment Analysis category [63].

WebGestalt is an online gene set analysis toolkit [155] taking as input a list of relevant genes or probesets. It adopts the hypergeometric test to evaluate functional category enrichment and performs a multiple test adjustment (the default method is the one from [21]). The user may choose different significance levels and the minimum number of genes belonging to the selected functional groups.
GSEA [134] first performs a correlation analysis between the features and the phenotype defining a ranking on the feature list. Secondly GSEA de-

termines whether the members of given gene sets are randomly distributed in the obtained ranked feature list or primarily found at the top or bottom. It thus calculates enrichment scores considering separately pathways over-represented at the top and at the bottom of the ranked list. We used the *GSEA Preranked* tool, feeding the gene list of top-ranked genes according to model $\mathcal{M}$ into the GSEA enrichment engine. In our framework we thus consider only the positively scoring gene sets of the preranked list output, which includes also genes that are highly discriminant and down-regulated in cases vs controls.

PaLS [2] takes a list or a set of lists of genes (or protein identifiers) and shows which ones share the same GO terms or KEGG pathways, following a criterion based on a threshold t percentage set by the user. The tool provides as output those functional groups that are shared at least by the $t\%$ of the selected genes. PaLS is aimed at easing the biological interpretation of results from studies of differential expression and gene selection, without assigning any statistical significance to the final output.

$\mathcal{N}$ module. We adopted three different subnetwork reconstruction algorithms $\mathcal{N}$: the Weighted Gene Co-Expression Networks Analysis (WGCNA) algorithm [61], the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (see Section 2.3.3) [101], and the Context Likelihood of Relatedness (CLR) (see Section 2.3.4) approach [46]. WGCNA is based on the idea of using (a function of) the absolute correlation between the expression of a couple of genes across the samples to define a link between them.

Procedure. The typical analysis considers a collection of n subjects, each described by a p-dimensional vector x of measurements. Each sample is associated with a phenotype label, e.g. $y = \{1, -1\}$, assigning it to a class, in a classification task. Hence the dataset is defined as $n \times p$ expression data matrix X, where $p \gg n$, and Y vector of labels. The output of a model

$\mathcal{M}$ is a gene signature $g_1, ..., g_k$ containing the k most discriminant features. We move the focus of the analysis from single genes to functionally related pathways by applying an enrichment algorithm $\mathcal{E}$, with reference to a knowledge base $\mathcal{D}$ such as KEGG, to explore known information on molecular interaction networks [77], or GO, to explore functional characterization and biological annotation. We retrieve for each gene g_i the corresponding whole pathway $p_i = \{h_1, ..., h_t\}$, where the genes $h_j \neq g_i$ not necessarily belong to the original signature $g_1, ..., g_k$. Extending the analysis to all the h_j genes of the pathway allows us to explore functional interactions that would otherwise get lost. For each pathway p_i, networks $N_{p_i,y}$ are reconstructed separately on data from the different classes, limiting the inference to the sole genes belonging to the pathway p_i in order to avoid the problem of intrinsic underdeterminacy of the task. As an additional caution against underdeterminacy, in our experiments we limit the analysis to pathways having more than 4 nodes and less than 1000 nodes. In summary, a real-valued adjacency matrix is inferred from X for each class y, for each model $\mathcal{M}$, for each enrichment tool $\mathcal{E}$, for each source of information $\mathcal{D}$, for each pathway p_i, and for each subnetwork inference algorithm $\mathcal{N}$. In the framework, the quantitative assessment of network differences is the key step for evaluating the impact of each component. As outlined in subsection 4.1, we use the HIM distance to detect the most disrupted pathways and to evaluate the stability of the network reconstruction.

6.3 HIM Framework on Biological datasetata

6.3.1 Children susceptibility to air pollution

The first dataset (GSE7543) collects data of children living in two regions of the Czech Republic with different air pollution levels ([145]): 23 children

recruited in the polluted area of Teplice and 24 children living in the cleaner area of Prachatice. Blood samples were hybridized on Agilent Human 1A 22k oligonucleotide microarrays. After normalization we retained 17564 features.

Experimental Results

The SRDA analysis of the air pollution dataset was performed within a 100×5-fold cross validation (CV) schema, producing a gene signature, characterizing the molecular differences between children in Teplice (polluted) and Prachatice (not polluted). The signature consists of 50 probesets, corresponding to 43 genes, achieving 76% accuracy.
The enrichment analysis on the signature allowed a functional characterization of the relevant genes, identifying 11 enriched ontologies in GO (listed in Appendix Table 6.3). We then constructed the corresponding WGCN network for the 11 selected pathways for both cases and controls.
Table 6.3 lists the 11 enriched pathways identified during the analysis of the air pollution dataset and the total number of the genes belonging to each pathway. The list is ranked by the normalized Ipsen-Mikhailov distance $\hat{\epsilon}$ (see Section 3.2): the top elements of the list are the most disrupted pathways between the two conditions. The pathways listed in Table 6.5 are a subset of those reported in Table 6.3.
Most of these pathways concern the *developmental* processes: this GO class contains ontologies especially related to the development of skeletal and nervous systems (GO:0001501 and GO:0007399) that undergo a rapid and constant growth in children. Other enriched terms are related to the capacity of an organism to defend itself (i.e *response to wounding*, GO:0009611 and *inflammatory response*, GO:0006954), to the regulation of the cell death (i.e. *negative regulation of apoptosis*, GO:0043066), the *multicellular organismal process*, GO:0032501, the *glycerlolipid metabolic*

process, GO:0046486, the response to external stimuli (i.e. *inflammatory response*, *response to wounding*) and to the locomotion (i.e. GO:0040011, GO:0007626).

Table 6.3: Air Pollution Experiment: pathways corresponding to mostly discriminant genes $g_1, ..., g_k$ ranked by the normalized Ipsen-Mikhailov distance $\hat{\epsilon}$. The number of genes belonging to the pathway is also provided.

Pathway	$\hat{\epsilon}$	# Genes
GO:0043066	0.257	21
GO:0001501	0.149	89
GO:0009611	0.123	16
GO:0007399	0.093	252
GO:0016787	0.078	718
GO:0005516	0.076	116
GO:0007275	0.076	453
GO:0006954	0.048	180
GO:0005615	0.038	417
GO:0007626	0.000	5
GO:0006066	0.000	8

Table 6.4 provides the subset of Agilent probesets (together with their corresponding Gene Symbol and GO pathway) belonging to the signature $g_1, ..., g_k$ and having a non zero value of the differential node degree Δd. Since the Δd score is computed as the difference between the weighted degree in the two classes, the top elements in Table 6.4 are those whose number of interactions varies most between the two conditions.

In Table 6.5 we report the most biologically relevant pathways, ranked for decreasing normalized Ipsen-Mikhailov distance $\hat{\epsilon}$, which provides a measure of the structural distance between the networks inferred for the two classes. The most disrupted pathway is GO:0043066, *i.e. apoptosis* followed by GO:0001501 *i.e. skeletal development*. Since the children under

Table 6.4: Air Pollution Experiment: list of Agilent probesets in the signature with their corresponding Entrez Gene Symbol ID and GO pathway. The list is ranked according to the decreasing absolute value of the differential node degree Δd.

Agilent ID	Gene Symbol	Pathway	Δd
4701	NRGN	GO:0007399	-2.477
12235	DUSP15	GO:0016787	-1.586
8944	CLC	GO:0016787	-1.453
3697	ITGB5	GO:0007275	-1.390
4701	NRGN	GO:0005516	-1.357
12537	PROK2	GO:0006954	1.069
13835	OLIG1	GO:0007275	0.834
11673	HOXB8	GO:0007275	-0.750
16424	FKHL18	GO:0007275	-0.685
13094	DHX32	GO:0016787	-0.575
8944	CLC	GO:0007275	0.561
14787	MATN3	GO:0001501	0.495
15797	CXCL1	GO:0006954	0.467
15797	CXCL1	GO:0005615	0.338
11302	MYH1	GO:0005516	-0.194
15797	CXCL1	GO:0007399	0.131

study are undergoing natural development, especially physical changes of their skeleton, the high differentiation between cases and controls of the GO:0001501 and the involvement of pathway GO:0007275 *i.e. developmental process* is biologically very sound. Another relevant pathway is GO:0006954, representing the response to infection or injury caused by chemical or physical agents. Several genes included in GO:0005516, (*i.e. calmodulin binding*) bind or interact with calmodulin, that is a calcium-binding protein involved in many essential processes, such as inflammation, apoptosis, nerve growth, and immune response. This is a key pathway that is linked with all the above mentioned terms as well as to GO:0007399, *i.e.*

Table 6.5: Air Pollution Experiment: most important pathways ranked by the normalized Ipsen-Mikhailov distance $\hat{\epsilon}$. The Entrez gene symbol ID is also provided for the selected probesets $g_1, ..., g_k$ in the corresponding pathway.

Pathway Code	$\hat{\epsilon}$	Gene Symbol
GO:0043066	0.257	
GO:0001501	0.149	MATN3
GO:0007399	0.093	NRGN
GO:0016787	0.078	DHX32, CLC
GO:0005516	0.076	MYH1
GO:0007275	0.076	FKHL18, HOXB8, OLIG1
GO:0006954	0.048	PROK2

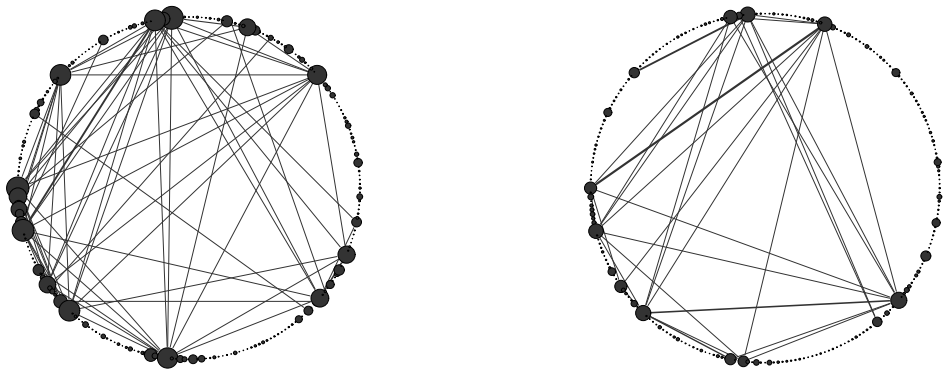
nervous system development, being one of the most stimulated pathways together with GO:0001501.

As described in Section 6.2 the pipeline also provides a score Δd of the variation of the number of interactions for $g_1, ..., g_k$. The full list is provided in Appendix Table 6.4, here we discuss a subset of the most biologically relevant genes.

FKHL18, HOXB8, PROK2, DHX32, MATN3 are directly involved in the development. CLC is a key element in the inflammation and immune system. OLIG1 is a transcription factor that works in the oligodendrocytes within the brain. NRGN binds calcium and is a target for thyroid hormones in the brain. Finally, MYH1 encodes for myosin that is a major contractile protein that forms striated, smooth and non-muscle cells. MYH1 isoforms show expression that is spatially and temporally regulated during development.

Figure 6.5 shows the network of the GO:0007399 pathway, related to the nervous system development in the two cohorts. It is clear that several connections among the genes within this pathway are missing in the subjects living in the polluted area (Teplice). Therefore the nervous system devel-

opment in these children is potentially at risk compared to those living in the not polluted city (Prachatice).



(a) Prachatice (b) Teplice

Figure 6.5: Networks of the pathway GO:0007399 (*nervous system development*) for Prachatice children (a) compared with Teplice children (b). Node diameter is proportional to the degree, and edge width is proportional to connection strength (estimated correlation).

6.3.2 Alzheimer's Disease

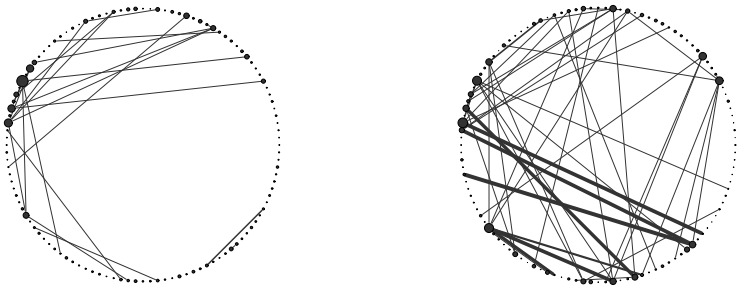
For AD we analyzed two GEO datasets: GSE9770 and GSE5281 ([92, 91]). The first includes 74 controls and 34 samples from non-demented patients with AD (since it is the earliest AD diagnosed, we will label it as early hereafter) and the second is composed of 74 controls and 80 samples from patients with late onset AD. The samples were extracted from six brain regions, differently susceptible to the disease: entorhinal cortex (EC), hippocampus (HIP), middle temporal gyrus (MTG), posterior cingulate cortex (PC), superior frontal gyrus (SFG) and primary visual cortex (VCX). The latter is known to be relatively spared by the disease, therefore we did not consider the samples within the VCX region. Overall, we analyzed 62 controls and 29 AD samples for GSE9770 and 62 controls and 68 AD samples for GSE5281. Biological data were hybridized on Affymetrix HG-U133Plus2.0 platform, estimating the expression of 54713 probesets for each sample.

Experimental Results

Classification and feature selection via $\ell_1\ell_2$, performed within a 9-fold nested CV schema for AD early and 8-fold for AD late, gives respectively 90% accuracy and 95% with 50 probesets for both cases.

We apply in the AD case the same network analysis strategy as in the PD experiment inferring for both cases and controls 51 selected pathways for early stage AD and 34 for late stage AD. The full list of reconstructed pathways is reported in Table 6.7. In Table 6.6 we summarize the main findings discussed hereafter.

Similarly to the PD analysis, we attempt a comparative analysis of the outcome for early and late stage AD having characterized the functional alteration of pathways for the two AD stages and comment the most mean-



(a) early AD patients (b) controls

Figure 6.6: Networks of the pathway GO:0019787 for AD early development patients (a) compared with healthy subjects (b). Node diameter is proportional to the degree, and edge width is proportional to connection strength (estimated correlation).

ingful results from the biological viewpoint.

Four common pathways were identified: GO:0019226 *i.e. transmission of nerve impulse*, GO:0008015 *i.e. blood circulation*, GO:0000267 *i.e. cell fraction* and GO:0042598 *i.e. vesicular fraction*.

The majority of pathways characterizing early stages of AD are related to the nervous system, and the blood. Among the nervous system related pathways the most damaged are: GO:0007399 *i.e. nervous system development*, GO:0007417 *i.e. central nervous system development*, GO:0042391 *i.e. regulation of membrane potential*, GO:0042552 *i.e. myelination*, GO:0050877 *i.e. neurological system process*, GO:0001508 *i.e. regulation of action potential* and GO:0019226 *i.e. transmission of nerve impulse*.

The majority of the pathways characterizing late stage AD are related to the cell, to the nervous system and to the response of the organism to var-

Table 6.6: AD: most important pathways ranked by normalized Ipsen-Mikhailov distance $\hat{\epsilon}$. The Entrez gene symbol ID is also provided for the selected probesets $g_1, ..., g_k$ in the corresponding pathway. In bold, common pathways between early and late stage AD.

	Pathway Code	$\hat{\epsilon}$	Gene Symbol
AD early	**GO:0042598**	0.21	
	GO:0019787	0.16	UBE2D3
	GO:0007417	0.10	MPB
	GO:0001508	0.14	
	GO:0051246	0.15	UBE2D3
	GO:0016874	0.12	UBE2D3
	GO:0004842	0.11	UBE2D3
	GO:0005768	0.08	EGFR
	GO:0016567	0.07	UBE2D3
	GO:0050877	0.06	
	GO:0042552	0.05	
	GO:0008015	0.04	
	GO:0042391	0.04	
	GO:0007399	0.04	NTRK2
	GO:0046982	0.03	EGFR
	GO:0006633	0.02	PTGDS
	GO:0019226	0.00	
	GO:0000267	0.00	
AD late	GO:0040012	0.36	SNCA
	GO:0042598	0.23	
	GO:0019226	0.12	
	GO:0030334	0.10	
	GO:0045892	0.09	SPEN
	GO:0042493	0.06	SNCA
	GO:0042127	0.05	
	GO:0008283	0.04	CAT
	GO:0005215	0.03	XK
	GO:0008217	0.03	HBD
	GO:0007601	0.03	
	GO:0007268	0.03	
	GO:0007610	0.03	
	GO:0008289	0.03	
	GO:0008015	0.02	
	GO:0016564	0.02	SPEN, ATXN1
	GO:0008284	0.02	
	GO:0008285	0.02	EIF2AK1
	GO:0020037	0.02	EIF2AK1, CAT, HBD
	GO:0000267	0.00	
	GO:0050890	0.00	

ious stimuli, see Table 6.6 and 6.7. Among the pathways centered on the cell, mentioned in descending order based on the numerosity of the genes, there are: GO:0008283 *i.e. cell proliferation*, GO:0008283 *i.e. negative regulation of cell proliferation*, GO:0008284 *i.e. positive regulation of cell proliferation*, GO:0042127 *i.e. regulation of cell proliferation*, GO:0030334 *i.e. regulation of cell migration*. The pathways related to the nervous system are: GO:0007268 *i.e. synaptic transmission*, GO:0007610 *i.e. behavior*, GO:0050890 *i.e. cognition*. Other relevant nodes are those related to the transcription regulation (GO:0016564, GO:0045892), the visual per-

ception (GO:0007601), and the heme and lipid binding (i.e. GO:0020037, GO:0008289).

The genes characterizing the early stage AD are reported in Table 6.6 and 6.8. UBE2D3 is an ubiquitin, targeting abnormal or short-lived proteins for degradation. It is a member of the E2 ubiquitin-conjugating enzyme family. This enzyme functions in the ubiquitination of the tumor-suppressor protein p53. It is also involved in several signaling pathways (BMP, TGF-β, TNF-α/NF-kB and in the immune system), in the protein processing in the endoplasmatic reticulum. PTGDS is an enzyme that catalyzes the conversion of prostaglandin H2 (PGH2) to postaglandin D2 (PGD2). It functions as a neuromodulator as well as a trophic factor in the central nervous system and it is also involved in smooth muscle contraction/relaxation and is a potent inhibitor of platelet aggregation. This gene is preferentially expressed in brain. Quantifying the protein complex of PGD2 and TTR in CSF may be useful in the diagnosis of AD, possibly in the early stages of the disease ([96]). EGFR is a transmembrane glycoprotein that is a member of the protein kinase superfamily. This protein is a receptor for members of the epidermal growth factor family that binds to epidermal growth factor. Binding of the protein to a ligand induces receptor dimerization and tyrosine autophosphorylation and leads to cell proliferation. This gene is involved in several pathways related to signaling, some type of cancer, to the cell proliferation, migration and adhesion and to the axon guidance. It is expressed in pediatric brain tumors ([113]). NTRK2 is member of the neurotrophic tyrosine receptor kinase (NTRK) family. This kinase is a membrane-bound receptor that upon neurotrophin binding phosphorylates itself and members of the MAPK pathway. Signalling through this kinase leads to cell differentiation. Mutations in this gene have been associated with obesity and mood disorders. SNPs in this gene is associated with AD ([36]).

The genes associated to late stage AD are listed in Table 6.6 and 6.9. Even if SNCA is a known hallmark for PD, it also known to be expressed in late-onset familial AD ([142]). Other relevant genes are: SPEN, EIF2AK1, CAT, HBD, ATXN1, XK. The first gene a hormone inducible transcriptional repressor. Repression of transcription by this gene product can occur through interactions with other repressors by the recruitment of proteins involved in histone deacetylation or through sequestration of transcriptional activators. SPEN is involved in the Notch signaling pathway that is important for cell-dell communication since it involves gene regulation mechanisms that control multiple cell differentiation processes (*i.e. neuronal function and development, stabilization of arterial endothelial fate and angiogenesis, cardiac valve homeostasis*) during embryonic and adult life. EIF2AK1 acts at the level of translation initiation to downregulate protein synthesis in response to stress, therefore it seems to have a protective role diminishing the overproduction of proteins such as SNCA or beta amyloid. CAT encodes for catalase a key antioxidant enzyme in the bodies defense against oxidative stress, therefore it act against the oxidative stress present in the brain of AD patients. This gene together with EIF2AK1 seems to fight against the disease. HBD like, HBB commented in subsection 6.3.3, could display the same role ([8]). ATXN1 i s involved in the autosomal dominant cerebellar ataxias (ADCA), an heterogeneous group of neurodegenerative disorders characterized by progressive degeneration of the cerebellum brain stem and spinal cord. Therefore, because of specific characteristics of these diseases (like the affected brain areas and the characteristics of the movement disorders), it might as well play a role in AD. Finally, mutations of XK have been associated with McLeod syndrome an X-linked recessive disorder characterized by abnormalities in the neuromuscular and hematopoietic systems.

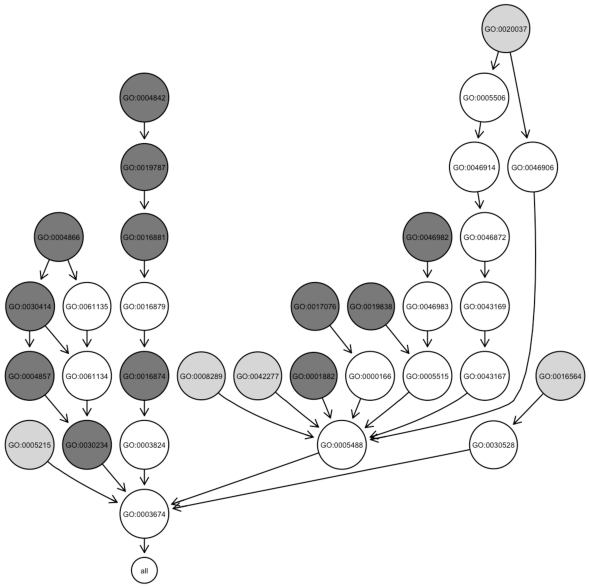
Table 6.7 reports the most discriminant pathways for the two AD stages

as selected by the presented pipeline, ranked by decreasing normalized $\hat{\epsilon}$ distance. Table 6.6 summarizes the main results here detailed in Table 6.7, 6.8 and 6.9. The common pathways are: GO:0019226 *i.e. transmission of nerve impulse*, GO:0008015 *i.e. blood circulation*, GO:0000267 *i.e. cell fraction* and GO:0042598 *i.e. vesicular fraction*. The relevance of blood circulatory system in AD has already been highlighted in [26] and references therein.
Figure 6.7 visualizes the enriched pathways in the Molecular Function and Biological Process domains. Despite only 4 pathways were found as common between early and late AD, it is easy to note that the majority of selected pathways belong to common GO classes.

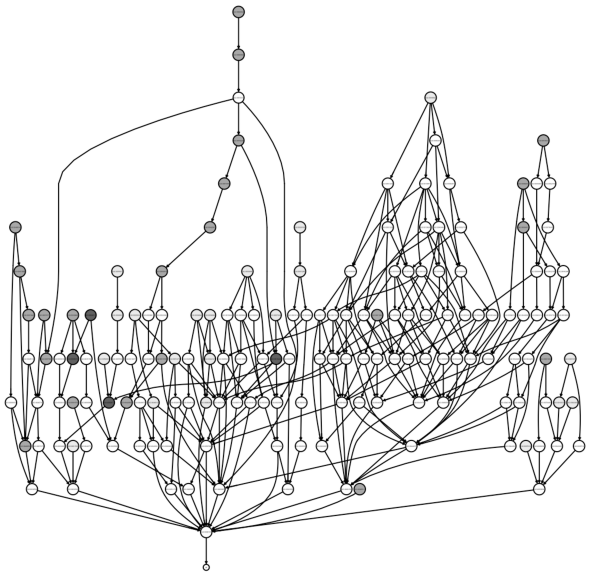
Tables 6.8 and 6.9 provide details of the network analysis results on early and late stage AD, respectively. The elements of the two signatures having non zero Δd are listed for decreasing absolute value of the differential node degree score, thus giving top positions to genes that change most the interaction network between the case/control condition.
Table 6.8 reports the most disrupted probesets within the early stage AD, ranked according to the differential node degree Δd. We note that the most disrupted gene is HBB, within *regulation of blood vessel size* and *regulation of blood vessels*. Table 6.9 reports the most disrupted genes within the late stage AD, ranked according to the differential node degree Δd. The majority of such genes (SPEN, SNCA, EIF2AK1, ELF1, CAT, ATXN1, HBD) belong to *regulation of locomotion*, *transcription repressor activity*, *response to drug* and *heme binding*.

Table 6.7: AD Experiment: selected pathways for early (left) and late (right) stage corresponding to mostly discriminant genes $g_1, ..., g_k$ ranked by the normalized Ipsen-Mikhailov distance $\hat{\epsilon}$. The number of genes belonging to the pathway is also provided. In bold, the common pathways.

AD early			AD late		
Pathway	$\hat{\epsilon}$	# Genes	Pathway	$\hat{\epsilon}$	# Genes
GO:0048514	0.22	22	GO:0040012	0.36	9
GO:0042598	0.21	16	**GO:0042598**	0.23	16
GO:0016881	0.19	109	**GO:0019226**	0.12	27
GO:0019787	0.16	116	GO:0030334	0.10	93
GO:0019725	0.16	14	GO:0045892	0.09	218
GO:0051246	0.15	121	GO:0009968	0.06	107
GO:0001508	0.14	31	GO:0042493	0.06	160
GO:0006631	0.14	171	GO:0050877	0.06	31
GO:0030234	0.13	29	GO:0042127	0.05	140
GO:0016874	0.12	735	GO:0009725	0.05	47
GO:0004842	0.11	368	GO:0042277	0.05	63
GO:0007417	0.10	199	GO:0015630	0.05	99
GO:0012505	0.10	216	GO:0008283	0.04	785
GO:0050880	0.09	26	GO:0005819	0.04	142
GO:0048471	0.08	263	GO:0008217	0.03	106
GO:0005792	0.08	409	GO:0005626	0.03	68
GO:0005768	0.08	490	GO:0000165	0.03	94
GO:0004857	0.08	57	GO:0005215	0.03	685
GO:0031982	0.07	34	GO:0007268	0.03	377
GO:0016567	0.07	206	GO:0007601	0.03	402
GO:0008217	0.07	105	GO:0008289	0.03	285
GO:0001666	0.07	225	GO:0007610	0.03	84
GO:0030141	0.06	69	GO:0008284	0.02	507
GO:0050877	0.06	31	GO:0001503	0.02	171
GO:0042552	0.05	36	GO:0007243	0.02	220
GO:0001568	0.05	79	GO:0008285	0.02	578
GO:0048511	0.04	49	**GO:0008015**	0.02	103
GO:0016023	0.04	108	GO:0016564	0.02	380
GO:0007399	0.04	806	GO:0020037	0.02	265
GO:0008015	0.04	103	GO:0051270	0.00	9
GO:0042391	0.04	67	GO:0010033	0.00	44
GO:0031410	0.03	482	GO:0050890	0.00	31
GO:0046982	0.03	364	GO:0050953	0.00	24
GO:0006633	0.02	109	**GO:0000267**	0.00	5
GO:0045121	0.02	136			
GO:0004866	0.02	194			
GO:0008366	0.00	22			
GO:0019228	0.00	19			
GO:0006873	0.00	10			
GO:0042592	0.00	25			
GO:0001974	0.00	28			
GO:0019226	0.00	27			
GO:0001944	0.00	4			
GO:0048771	0.00	12			
GO:0048856	0.00	20			
GO:0019838	0.00	85			
GO:0017076	0.00	11			
GO:0030414	0.00	42			
GO:0001882	0.00	8			
GO:0000267	0.00	4			
GO:0031090	0.00	6			



(a) MF



(b) BP

Figure 6.7: GO subgraphs for Alzheimer's early and late stage (Molecular Function and Biological Processes domains). Selected nodes are represented in light gray, gray and dark gray for late, early and common nodes.

Table 6.8: AD Experiment (early): list of Affymetrix probesets in the early stage signature with their corresponding Entrez Gene Symbol and GO pathway. The list is ranked according to the decreasing absolute value of the differential node degree Δd.

Affy Probeset ID	Gene Symbol	Pathway	Δd
209116_x_at	HBB	GO:0050880	1.670
209116_x_at	HBB	GO:0008217	1.445
211748_x_at	PTGDS	GO:0006633	1.273
240383_at	UBE2D3	GO:0016874	-1.165
240383_at	UBE2D3	GO:0019787	-0.703
201061_s_at	STOM	GO:0045121	-0.662
240383_at	UBE2D3	GO:0051246	-0.613
201983_s_at	EGFR	GO:0046982	-0.476
221795_at	NTRK2	GO:0007399	-0.262
212226_s_at	PPAP2B	GO:0001568	0.259
201983_s_at	EGFR	GO:0005768	-0.256
211696_x_at	HBB	GO:0050880	-0.224
209072_at	MBP	GO:0008366	0.166
211696_x_at	HBB	GO:0008217	-0.149
212187_x_at	PTGDS	GO:0006633	-0.139
201185_at	HTRA1	GO:0019838	0.124
240383_at	UBE2D3	GO:0004842	0.120
209072_at	MBP	GO:0007417	0.113
240383_at	UBE2D3	GO:0016567	-0.047

Table 6.9: AD Experiment (late): list of Affymetrix probesets in the late stage signature with their corresponding Entrez Gene Symbol and GO pathway. The list is ranked according to the decreasing absolute value of the differential node degree Δd.

Affy Probeset ID	Gene Symbol	Pathway	Δd
201996_s_at	SPEN	GO:0016564	1.590
211546_x_at	SNCA	GO:0040012	1.410
211546_x_at	SNCA	GO:0042493	1.310
201996_s_at	SPEN	GO:0045892	1.246
217736_s_at	EIF2AK1	GO:0020037	-1.066
201005_at	CD9	GO:0008285	0.725
210943_s_at	LYST	GO:0015630	0.706
204466_s_at	SNCA	GO:0042493	0.461
207827_x_at	SNCA	GO:0040012	0.434
206698_at	XK	GO:0005215	0.433
209184_s_at	IRS2	GO:0008283	0.208
212420_at	ELF1	GO:0016564	-0.203
207827_x_at	SNCA	GO:0042493	0.201
205592_at	SLCA4A1	GO:0005215	0.180
211922_s_at	CAT	GO:0008283	0.173
211922_s_at	CAT	GO:0020037	-0.094
203231_s_at	ATXN1	GO:0016564	-0.073
217736_s_at	EIF2AK1	GO:0008285	-0.072
204466_s_at	SNCA	GO:0040012	0.048
206834_at	HBD	GO:0008217	0.045
206834_at	HBD	GO:0020037	0.019

6.3.3 Parkinson's Disease

A gene expression dataset of PD was considered to test the HIM framework [156]. PD is a neurodegenerative disorder that impairs the motor skills at the onset and the cognitive and the speech functions successively. The biological samples consisted of whole substantia nigra tissue from 11 PD patients and 18 healthy controls. Gene expression was measured by the Affymetrix HG-U133A platform, available at Gene Expression Omnibus (GEO) as GSE20292. Data were normalized with the *rma* algorithm in the R Bioconductor *affy* package with a custom CDF adopting the most up-to-date platform annotation and Entrez identifiers (from BrainArray: http://brainarray.mbni.med.umich.edu, v. 14.1.0, ENTREZG), while the enrichment phase was performed using HUGO gene symbol identifiers.

Results and Discussion

The $\ell_1\ell_2$ feature selection identified 458 discriminant genes giving an average prediction performance of 80.8%, while Liblinear selected the top-500 genes with an accuracy of 80% (95% boostrap Confidence Interval: (0.78,0.83)) and a stability of 0.70. The two lists have only 119 genes in common. As the feature selection method is the starting point of our analysis, to limit its impact we employed two approaches from the same family of regularization methods: both classifiers adopt a ℓ_2-regularization penalty term combined with different loss functions and, for $\ell_1\ell_2$, with an alternative regularization term. We used similar model selection protocols, both ensuring that results are not affected by selection-bias. We first compare together the impact of the different sources of variability, but will come back later to difference in pathways when the model choice is the only difference.

In general, the number of significantly enriched pathways varied greatly

Table 6.10: Number $(n)m$ of pathways found for the network inference step for different combinations of model $\mathcal{M}$, knowledge-base $\mathcal{D}$, and enrichment $\mathcal{E}$. n: all networks (unfiltered); m: filtered networks, having more than 5 and less than 1000 genes on HG-U133A with non-null intra-class variance. Intersections ℓL, $\mathcal{E}_{3\cap}$ and $\mathcal{E}_{2\cap}$ are respectively defined as $\ell\text{L} := \ell_1\ell_2 \cap \text{Liblinear}$, $\mathcal{E}_{2\cap} := \text{WebGestalt} \cap \text{PaLS}$, $\mathcal{E}_{3\cap} := \text{WebGestalt} \cap \text{GSEA} \cap \text{PaLS}$.

$\mathcal{M}$	$\mathcal{D}$	$\mathcal{E}$			$\mathcal{E}_{3\cap}$	$\mathcal{E}_{2\cap}$
		WebGestalt	GSEA	PaLS		
$\ell_1\ell_2$	GO	(114) 92	(7) 7	(381) 331	(0) 0	(39) 30
	KEGG	(43) 43	(2) 2	(71) 71	(2) 2	(43) 43
Liblinear	GO	(83) 45	(0) 0	(404) 356	(0) 0	(21) 12
	KEGG	(56) 55	(1) 1	(77) 77	(1) 1	(56) 55
ℓL	GO	(21) 8	(0) 0	(272) 225	(0) 0	(5) 1
	KEGG	(21) 20	(0) 0	(45) 45	(0) 0	(21) 20

depending on the model $\mathcal{M}$, the enrichment $\mathcal{E}$ and the knowledge-base $\mathcal{D}$, as reported in Table 6.10. For this metric, the main source of variation is the choice of enrichment procedure, followed by the reference knowledge-base. For $\mathcal{M}$=$\ell_1\ell_2$ and globally for GO and KEGG, we found 157, 452 and 9 enriched pathways for WebGestalt, PaLS and GSEA respectively and similarly for $\mathcal{M}$=Liblinear, 139, 481 and 1. No GO terms were found as common to all three enrichment methods for $\mathcal{M}$=$\ell_1\ell_2$ and $\mathcal{D}$ = GO, but if we limit the observation to WebGestalt and PaLS we found a small overlap (39 GO terms). Similar result were found for $\mathcal{M}$ = Liblinear, but with only 21 GO terms shared between WebGestalt and PaLS. For $\mathcal{D}$ = KEGG, two pathways are common to the three enrichment algorithms for $\ell_1\ell_2$ and one for Liblinear. Excluding the GSEA algorithm, we found a significant overlap: 43 pathways for $\ell_1\ell_2$ and 56 for Liblinear. In general, WebGestalt and PaLS provide results which are closer than those provided by GSEA both in terms of number of retrieved pathways. Also, more numerous enrichment lists were found for GO rather than KEGG, but with a smaller

overlap.

The HIM distance between networks separately inferred for cases and controls was computed for all combinations of choices for $\mathcal{M}, \mathcal{E}, \mathcal{D}, \mathcal{N}$; a full landscape is available in Supplementary Fig. 6.8. Including all choices of $\mathcal{M}, \mathcal{E}, \mathcal{D}, \mathcal{N}$, H ranges in [0.002, 0.431] and IM in [0.005, 0.703], respectively with medians $med_{\mathrm{H}} = 0.044$, $med_{\mathrm{IM}} = 0.133$ and variances $var_{\mathrm{H}} = 0.001$ and $var_{\mathrm{IM}} = 0.016$. The choice of the feature selection method has a limited effect, with respect to the variation found across $\mathcal{E}$ and $\mathcal{D}$. A remarkable difference in number of pathways and in variability is found between $\mathcal{E}$ = GSEA vs PaLS (see Table 6.10). In general, we observe that structural changes (variability in the IM component) have more impact than differences in rewiring (variability along the H axis).

As an example of HIM analysis, we considered $\mathcal{N}$=Aracne, with $\mathcal{D}$=GO and all models $\mathcal{M}$ (Fig. 6.10): two clusters are identified, with one cluster prevalently along the IM coordinate. Considering the distribution of the HIM distances Fig.6.9, the threshold HIM=0.15 (equidistant from the two centroids located by *kmeans* at HIM=0.06 and HIM=0.25) was used to define a separation surface in the HIM maps in Fig. 6.10(a). We found that the distribution of pathway cardinality is skewed towards smaller networks (less than 100 nodes) within the threshold, and instead almost equally distributed above threshold, as shown in Fig. 6.10(b).

From now on, we focus our analysis on the subset of most disrupted pathways (MDPs). Given the strong variability due to $\mathcal{N}$, MDPs are defined as the pathways whose HIM distance between the network inferred on cases and that on controls is larger than a threshold $\tau = 0.05$ for all network inference methods. As shown in Table 6.11, the incidence of MDPs increases approximately twofold if we weaken the MDP definition to HIM$> \tau$ for at least one reconstruction method $\mathcal{N}$.

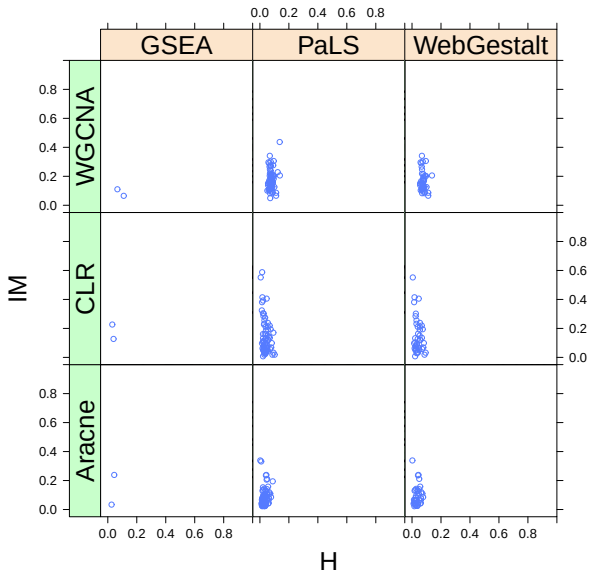
The HIM analysis reveals strong differences between the reconstruction

Table 6.11: Summary of most disrupted pathways retrieved by WebGestalt and PaLS.

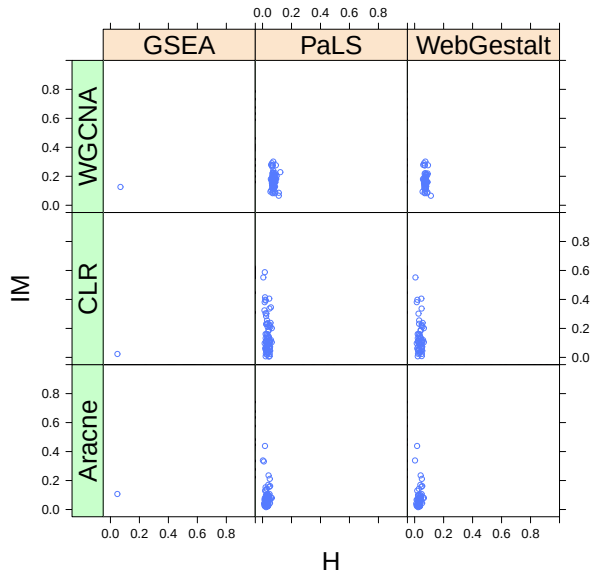
$\mathcal{M}$	$\mathcal{D}$	at least one $\mathcal{N}$	all $\mathcal{N}$(*)
$\ell_1\ell_2$	GO	30	18 (60%)
	KEGG	43	21 (49%)
Liblinear	GO	12	6 (50%)
	KEGG	55	21 (38%)

(*) incidence of MDPs (HIM $> \tau = 0.05$)

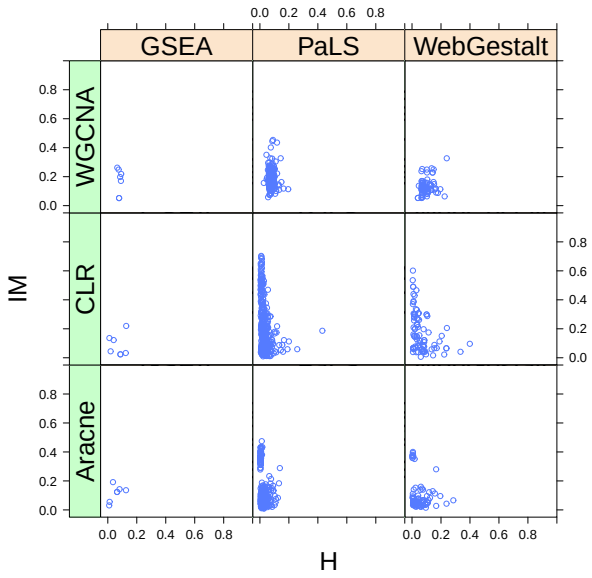
methods, with an increasing fraction of MDPs for WGCNA over CLR and Aracne (see Fig. 6.11(a) and (b) for $\mathcal{M}$ = Liblinear and $\mathcal{D}$ = KEGG). We conclude that the variability due to the choice of reconstruction methods should be seriously taken into account in network medicine studies.



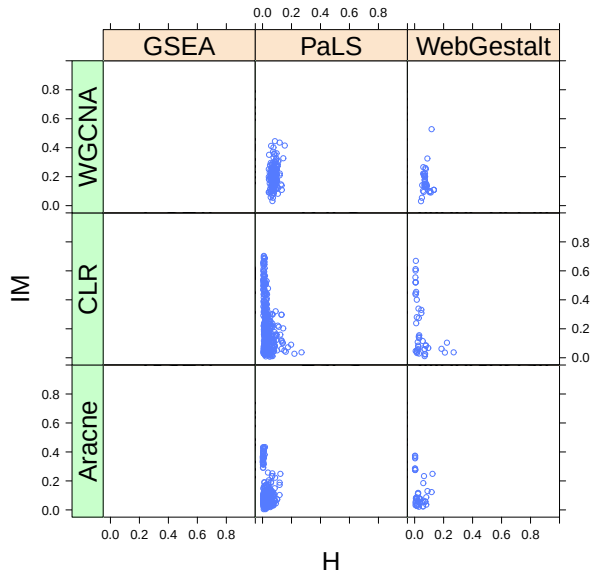
(a) $\ell_1\ell_2$ and KEGG



(b) Liblinear and KEGG

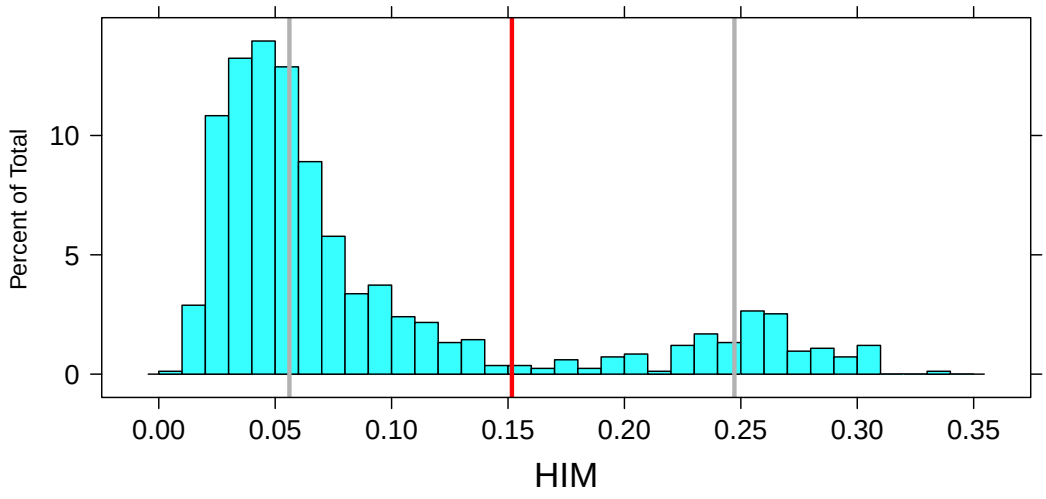


(c) $\ell_1\ell_2$ and GO

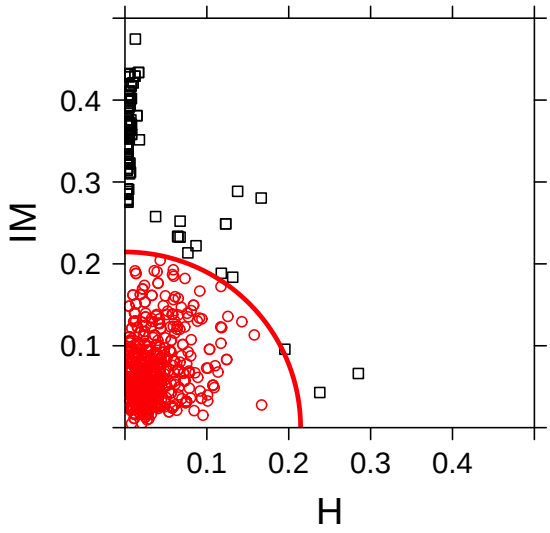


(d) Liblinear and GO

Figure 6.8: HIM maps for all combinations of $\mathcal{M}, \mathcal{D}, \mathcal{E}$ and $\mathcal{N}$. Subplot (c) is reproduced in the main paper as Figure 6.8(d).

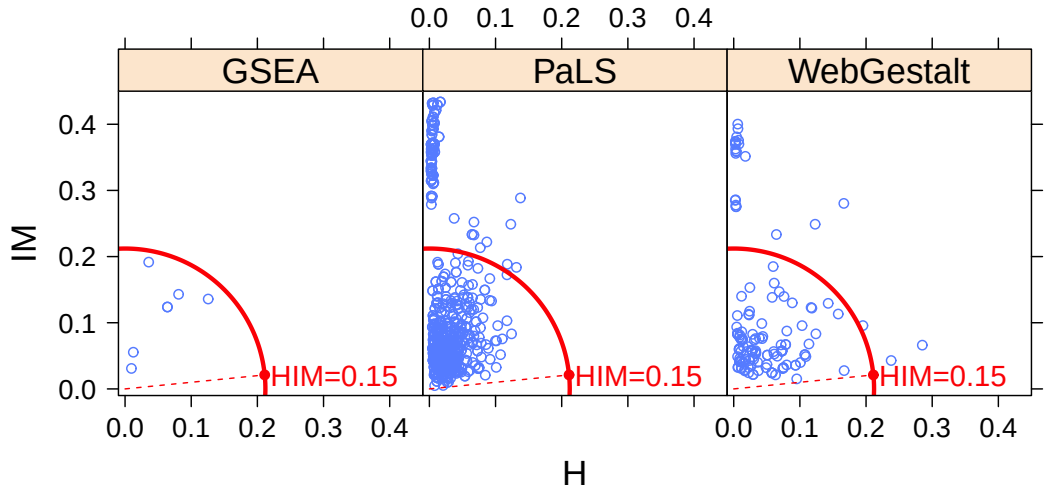


(a)

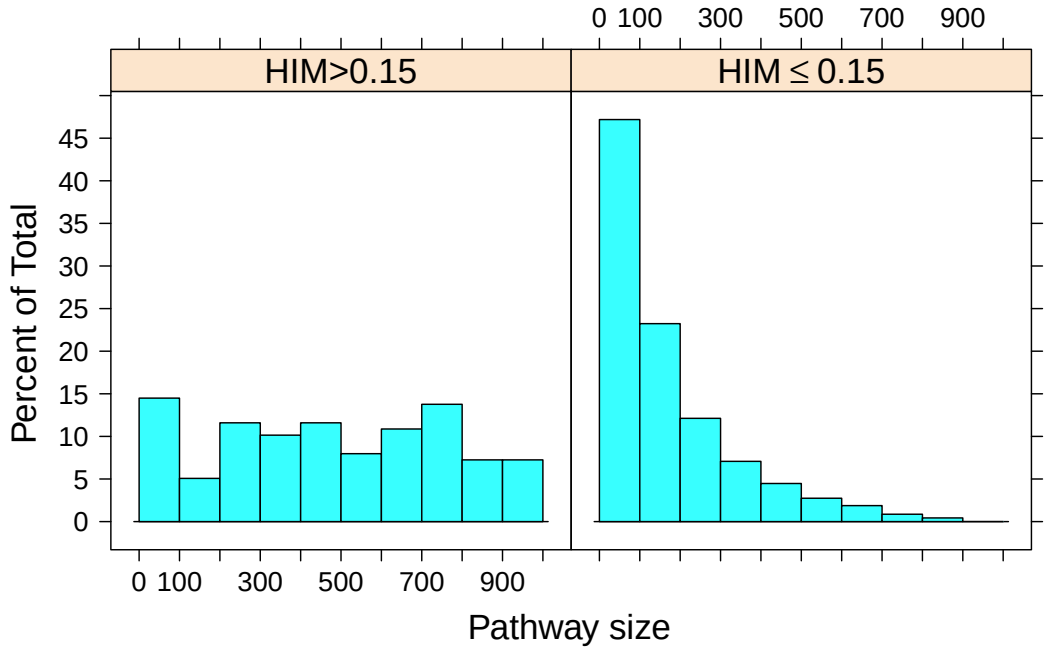


(b)

Figure 6.9: Distance distribution for $\mathcal{N}$ = Aracne and $\mathcal{D}$ = GO (all enrichment methods and all models). (a) Distribution of the HIM distance. Gray line: *kmeans* centroids (HIM ≈ 0.056 and HIM ≈ 0.247). Red line: chosen threshold HIM ≈ 0.152, equidistant from the two centroids. (b) HIM map of the two centroids. Red line: HIM $= 0.15$.

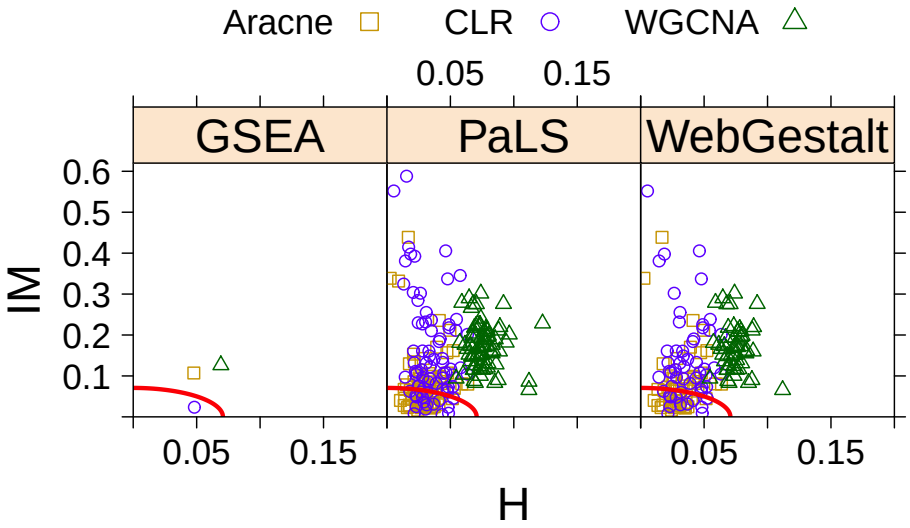


(a)

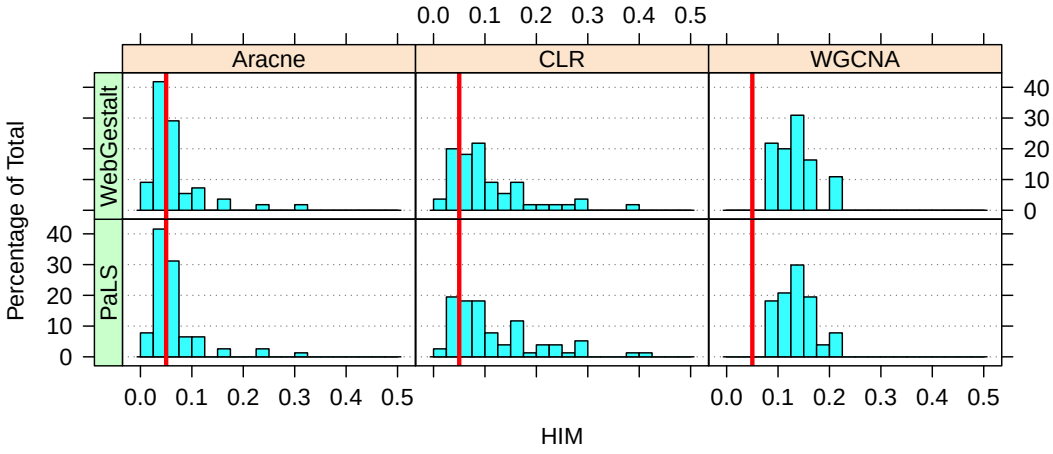


(b)

Figure 6.10: Distance distribution for $\mathcal{N}$ = Aracne and $\mathcal{D}$ = GO. (a) HIM maps distance for different $\mathcal{E}$ methods. Red line corresponds to threshold HIM = 0.15 separating two clusters. (b) Histograms of pathway cardinality below and above threshold.



(a)



(b)

Figure 6.11: HIM plots for $\mathcal{M}$ = Liblinear and $\mathcal{D}$ = KEGG, for all enrichment methods. Symbols indicate enrichment methods: Aracne (squares), CLR (circles), WGCNA (triangle). Red line: the threshold $\tau = 0.05$ defining MDPs. (a) HIM maps grouped by $\mathcal{E}$. Each pathway is inferred by the three methods $\mathcal{N}$ as detailed in the legend on top of the figure. (b) Trellis displays for histogram plots of HIM distance distribution conditioned for WebGestalt and PaLS and the three subnetwork inference algorithms $\mathcal{N}$.

Similarly, given a fixed common threshold on HIM, we also found that MDP lists can significantly differ for different $\mathcal{M}$ models, or reference ontologies: as shown in Supplementary Tables 6.12 and 6.13 for GO terms and KEGG pathways respectively, different methods may select rather different list of MDPs pathways. However, when all methods agree, the biological significance can be high. The *Amyotrophic Lateral Sclerosis* (ALS) KEGG 05014 pathway is the only MDP selected by all three enrichment tools $\mathcal{E}$. This finding is of biological interest as both ALS and PD are neurodegenerative diseases severely affecting the skeletal muscles, and sharing significant features, mainly at the mitochondrial level.

To complete this prototype network medicine study, we quantified the difference between networks as separately inferred from PD patients and controls for the ALS pathway (Fig. 6.12). For WGCNA, higher correlation links were found for PD cases (Fig 6.12 (a-b)). We also computed the stability of reconstruction in terms of HIM distance between m=100 networks for a 2/3 subsampling and that on all data, given a class. We found that stability depends on $\mathcal{N}$, the networks inferred with ARACNE being the most diverse between cases and controls, as shown in Fig. 6.12(c). The variability between individual networks due to resampling can be severe (Fig. 6.13). We replicated the stability analysis by using the Leave-one-out schema: in terms of distribution we found less striking differences but a similar behavior, as shown in Fig. 6.14. By projecting the information on the HIM map (Fig. 6.12(d)), it is clear that for WGCNA the variability on the PD network results due both to structural changes as well as to link weight differences. On the other hand, for ARACNE, the changes regard mostly the IM component.

We also compared the HIM ranks of the MDPs found for both WebGestalt and PaLS, for fixed $\mathcal{D}$ = KEGG, listed in Tab: 6.13. Five highly disrupted pathways were found as shared between WebGestalt and PaLS (bold en-

tries in Tab: 6.13). In particular, the *Pathogenic E. coli infection* KEGG pathway 05130 is top ranked for both models, which is consistent with recent findings of increased presence of the gram negative bacteria *E. coli* in sigmoid mucosa samples from patients with PD compared to controls [49]. Further effects of exposure to *E. coli* bacterial products in PD has also been reported [148]. Indeed, stronger correlations were found for PD cases in the case/control pair of WGCNA networks for KEGG 05130 (Fig. 6.14(a-b)). The HIM stability analysis for the same pathway found ARACNE as the most unstable method, in particular for the PD cases, 6.14(c,d)). Leave-One-Out estimates lower instability levels, but confirms that stability between classes is more similar for CLR rather than for Aracne and WGCNA and that WGCNA is the most stable method on the PD dataset for this pathway (Fig. 6.15(e,f)).

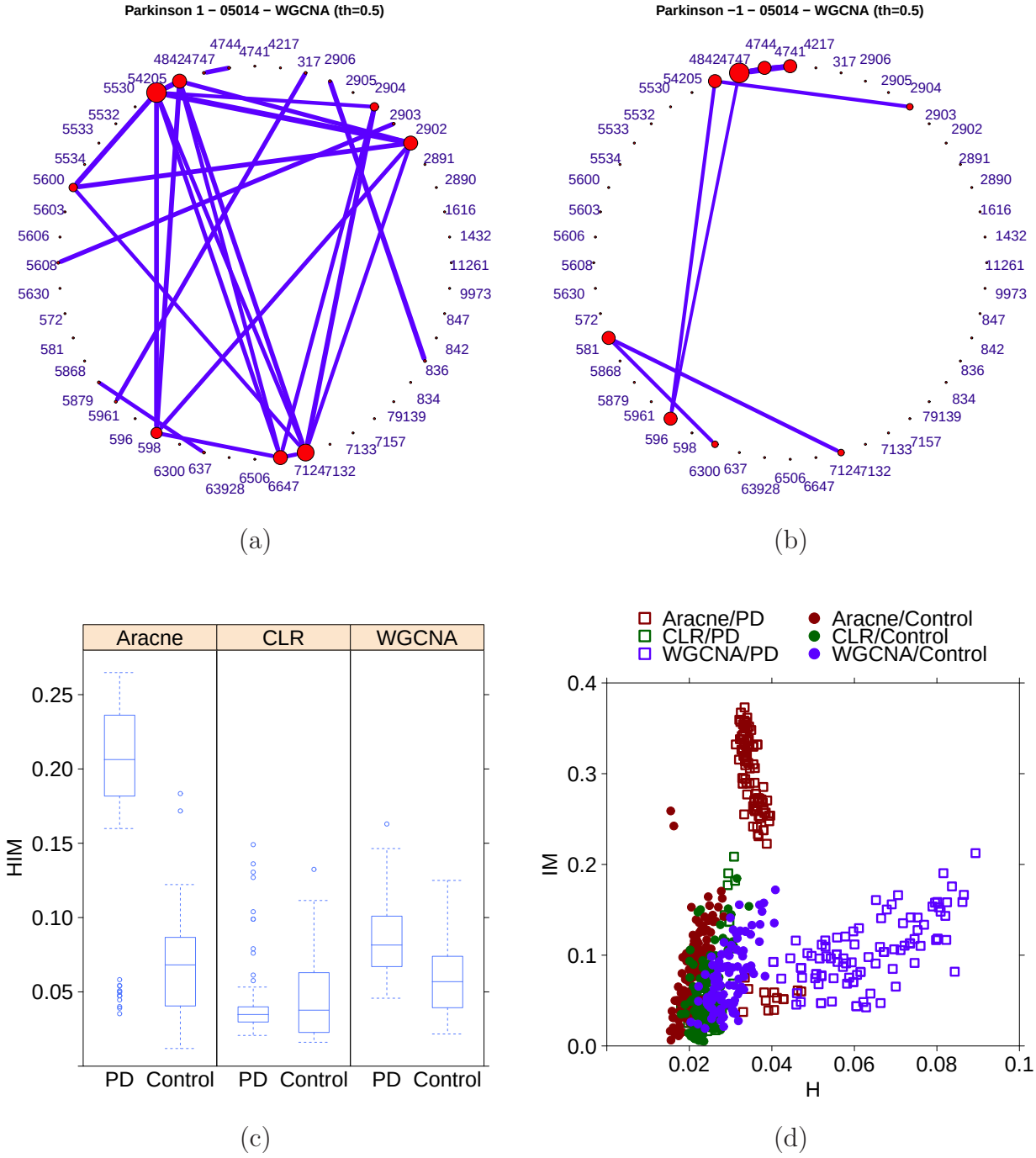
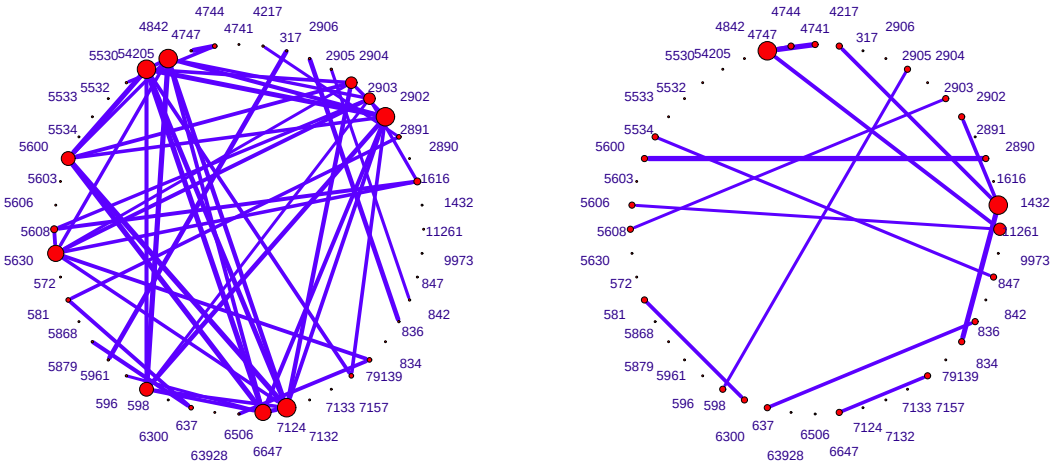
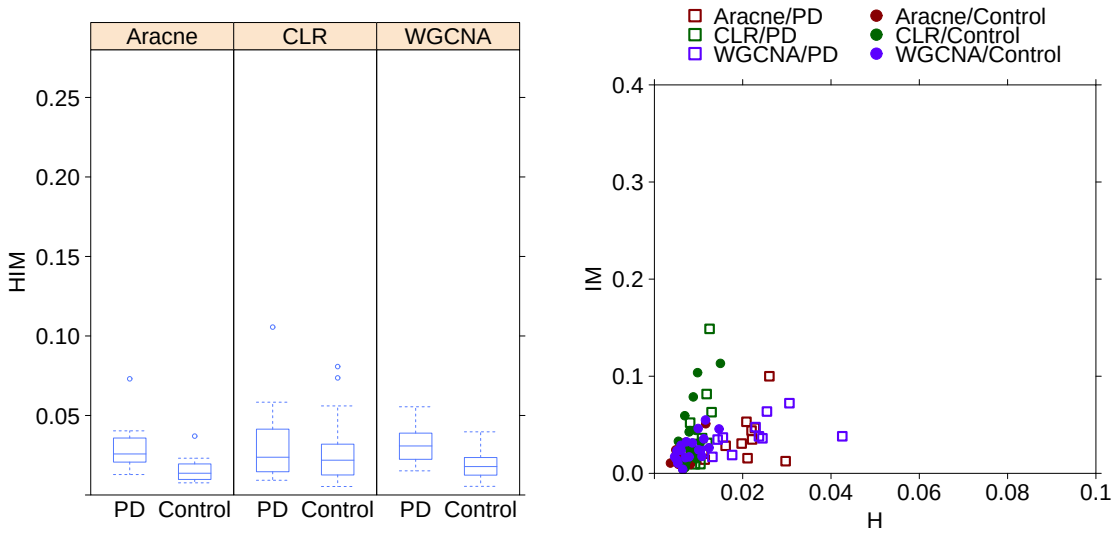


Figure 6.12: Network analysis of the *ALS* KEGG pathway (as defined by PALS). (a-b): Networks were separately inferred by WGCNA for the PD patients (a) and controls (b). The networks are thresholded at edge weight 0.5 for graphic purposes. Node labels represent Entrez IDs. (c): Boxplots of the HIM stability distribution ($m = 100$ replicates as defined in Subsection 4.1) comparing PD patients and controls separately for the 3 inference methods $\mathcal{N}$. (d): HIM map of all m comparisons.



(a) (b)

Figure 6.13: Variability of networks on the *ALS* KEGG pathway, defined by PALS, inferred by WGCNA on PD samples, for m=100 replicates and 2/3 resampling. The two network instances have (a) smallest HIM and (b) largest HIM from the network inferred on all samples (shown in the main paper, Fig: 6.11(a)). Only links of weight > 0.5 are displayed. Node labels represent Entrez ID.



(a) (b)

Figure 6.14: Leave-One-Out stability of the *ALS* KEGG pathway (as defined by PALS). (a): Boxplots of the Leave-One-Out HIM stability distribution comparing PD patients and controls separately for the 3 inference methods $\mathcal{N}$. (b): HIM map of all m_{+1} and m_{-1} comparisons. Different colors are used for the three $\mathcal{N}$.

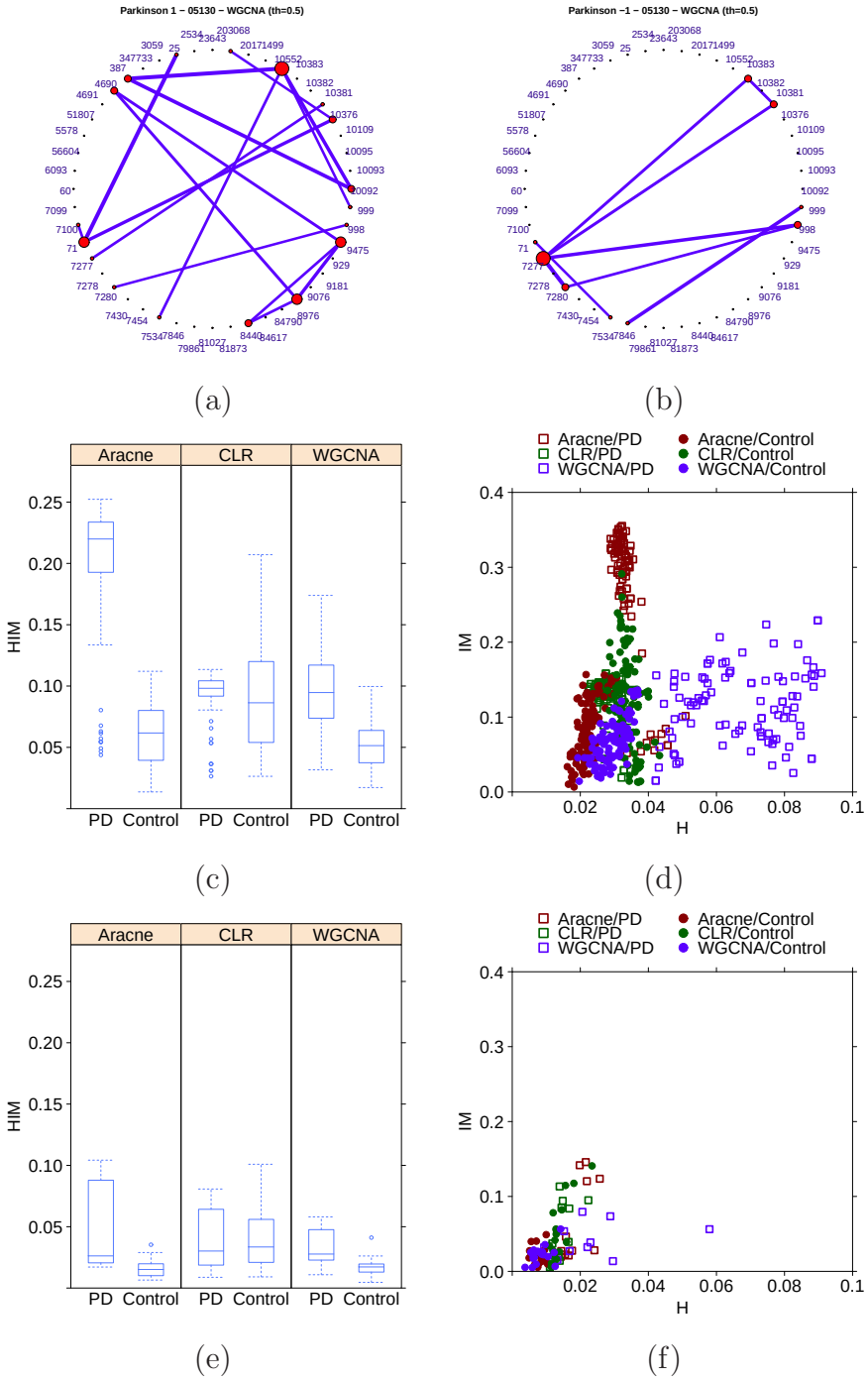


Figure 6.15: Network analysis of the *Pathogenic E. coli infection* KEGG pathway (as defined by PALS). (a-b): Networks were separately inferred by WGCNA for the PD patients (a) and controls (b). The networks are thresholded at edge weight 0.5 for graphic purposes. Node labels represent Entrez ID. (c): Boxplots of the HIM stability distribution ($m = 100$ replicates as defined in the main paper, Subsection 2.1) comparing PD patients and controls separately for the 3 inference methods $\mathcal{N}$. (d): HIM map of all m comparisons. Different colors are used for the three $\mathcal{N}$. (e): Boxplots of the HIM Leave-One-Out stability distribution comparing PD patients and controls separately for the 3 inference methods $\mathcal{N}$. (f): HIM map of all m_{+1} and m_{-1} comparisons. Different colors are used for the three $\mathcal{N}$.

Table 6.12: Summary of GO terms in MDPs common between WG and PaLS, for both $\mathcal{M}$ models. GO terms are sorted for decreasing HIM median (computed over $\mathcal{E}$ and $\mathcal{N}$). Bold fonts identify the GO terms shared by models.

$\ell_1\ell_2$			Liblinear		
ID	Term name	HIM	ID	Term name	HIM
GO:0031966	Mitochondrial membrane	0.272	GO:0005783	Endoplasmic reticulum	0.256
GO:0005739	Mitochondrion	0.261	GO:0042127	Regulation of cell proliferation	0.252
GO:0005743	Mitochondrial inner membrane	0.148	GO:0016973	Poly(A)+ mRNA export from nucleus	0.192
GO:0046961	Proton-transporting ATPase activity, rotational mechanism	0.126	GO:0015629	Actin cytoskeleton	0.115
GO:0042802	Identical protein binding	0.112	GO:0006469	Negative regulation of protein kinase activity	0.098
GO:0007018	Microtubule-based movement	0.110	**GO:0005747**	**Mitochondrial respiratory chain complex I**	0.081
GO:0048487	Beta-tubulin binding	0.110			
GO:0045202	Synapse	0.109			
GO:0000502	Proteasome complex	0.107			
GO:0005753	Mitochondrial proton-transporting ATP synthase complex	0.106			
GO:0015986	ATP synthesis coupled proton transport	0.105			
GO:0042734	Presynaptic membrane	0.103			
GO:0005747	**Mitochondrial respiratory chain complex I**	0.081			
GO:0015078	Hydrogen ion transmembrane transporter activity	0.080			
GO:0008137	NADH dehydrogenase (ubiquinone) activity	0.065			
GO:0015992	Proton transport	0.064			
GO:0006120	Mitochondrial electron transport, NADH to ubiquinone	0.061			
GO:0005874	Microtubule	0.057			

Table 6.13: Summary of KEGG pathways in MDPs common between WebGestalt and PaLS, for both $\mathcal{M}$ models. KEGG pathways are sorted for decreasing HIM median (computed over $\mathcal{E}$ and $\mathcal{N}$). Bold fonts identify the KEGG pathways shared by models $\mathcal{M}$.

$\ell_1\ell_2$			Liblinear		
ID	Pathway name	HIM	ID	Pathway name	HIM
01100	**Metabolic pathways**	**0.239**	04630	Jak-STAT signaling pathway	0.281
05130	**Pathogenic Escherichia coli infection**	**0.169**	**01100**	**Metabolic pathways**	**0.239**
03050	Proteasome	0.162	**05130**	**Pathogenic Escherichia coli infection**	**0.169**
04910	**Insulin signaling pathway**	**0.162**	04623	Cytosolic DNA-sensing pathway	0.163
00620	Pyruvate metabolism	0.140	**04910**	**Insulin signaling pathway**	**0.162**
05213	**Endometrial cancer**	**0.133**	00330	Arginine and proline metabolism	0.158
00310	Lysine degradation	0.119	03030	DNA replication	0.134
00240	Pyrimidine metabolism	0.114	**05213**	**Endometrial cancer**	**0.133**
05110	Vibrio cholerae infection	0.105	05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.123
00270	Cysteine and methionine metabolism	0.105	05212	Pancreatic cancer	0.117
05120	Epithelial cell signaling in Helicobacter pylori infection	0.098	04912	GnRH signaling pathway	0.109
00230	Purine metabolism	0.096	05210	Colorectal cancer	0.099
00562	Inositol phosphate metabolism	0.096	04662	B cell receptor signaling pathway	0.090
04140	Regulation of autophagy	0.096	05332	Graft-versus-host disease	0.086
05014	Amyotrophic lateral sclerosis (ALS)*	0.094	04660	T cell receptor signaling pathway	0.084
00600	Sphingolipid metabolism	0.092	04520	Adherens junction	0.083
00020	Citrate cycle (TCA cycle)	0.091	04310	Wnt signaling pathway	0.079
05218	Melanoma	0.074	04621	NOD-like receptor signaling pathway	0.074
00051	Fructose and mannose metabolism	0.073	**04722**	**Neurotrophin signaling pathway**	**0.070**
04722	**Neurotrophin signaling pathway**	**0.070**	05214	Glioma	0.065
00980	Metabolism of xenobiotics by cytochrome P450	0.064	04370	VEGF signaling pathway	0.064

**Note: This is the only pathway shared across all enrichment methods $\mathcal{E}$.*

Chapter 7

Conclusions

Network medicine and differential network analysis requires that a fair degree of trust be assigned to networks built from omics data in order to develop reliable network signatures of disease. Variability in network reconstruction and pathway profiling can be injected by different sources, from noise in the data to choices in network modeling; moreover, underdeterminacy from limited sample sizes is also a major issue, given that the ratio between network dimension (number of nodes) and the number of available data to infer interactions has a key role for the stability of the inferred structure [40]. In this thesis we proposed a solution for the assessment of stability and quality of network reconstruction which is quantitative (and thus reproducible) and consistent as shown by the outcomes of the biological applications. The aim here is to provide the researchers with an effective tool to compare either the inference algorithms or the investigated dataset. In particular, we introduced a suite of four stability indicators for assessing the variability of network reconstruction algorithm as functions of a data subsampling procedure. Two indicators are based on a measure of a normalized distance between networks and they are global, giving a confidence measure on the whole inferred dataset, while the other two are local, associating a reliability score to the network nodes

and detected links. They are of particular interest when no gold standard is known for the studied task, so they can work as a substitute for the algorithm accuracy. The proposed approach is extensively tested on a broad range of biological applications from high-throughput data to practically demonstrate its use in various research tasks.

Empirical quantitativeness in this framework is provided by the use of the novel Hamming-Ipsen-Mikhailov (HIM) network distance to evaluate differences between graphs. The HIM distance captures both local (link occurrence) and global (spectral) structural differences between the investigated graph, avoiding the pitfalls affecting its components when separately considered. HIM metric is consistent with more classical network similarity approaches, but it is able to better capture finer differences, while avoiding unwanted behaviors affecting other distances. Furthermore, we showed how HIM can be effectively used to turn qualitative considerations (for instance, on dynamic network evolution) into quantitative ones, thus available for objective comparisons.

Finally, we also introduced the novel inference method RegnANN, based on Artificial Neural Networks, aimed at effectively detect higher order relations between graph nodes (*e.g.*, genes in a transcriptional network): this method proved to achieve reconstruction performances comparable with those reached by more classical algorithm, but, in average, showing a better stability.

We conclude with the remark that most of the shown applications were computationally very intensive and thus not feasible on a standard workstation: we thus made intensive use of the HPC facility at FBK, the Linux KORE cluster endowed with more than 700 cores and 200 TB disk space.

Bibliography

[1] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47, 2002.

[2] A. Alibés, A. Cañada, and R. Díaz-Uriarte. PaLS: filtering common literature, biological terms and pathway information. *Nucleic Acids Res*, 36(Web Server issue):W364–W367, 2008.

[3] J.D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao. Comparing Statistical Methods for Constructing Large Scale Gene Networks. *PLoS ONE*, 7(1):e29348, 2012.

[4] G. Altay and F. Emmert-Streib. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, 26(14):1738–1744, 2010.

[5] J. Ambroise, A. Robert, B. Macq, and J.-L. Gala. Transcriptional Network Inference from Functional Similarity and Expression Data: A Global Supervised Approach. *Statistical Applications in Genetics and Molecular Biology*, 11(1):Article 2, 2012.

[6] V. Ambros, B. Bartel, D.P. Bartel, C.B. Burge, J.C. Carrington, X. Chen, G. Dreyfuss, S.R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3):277–279, 2003.

[7] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genet*, 25:25–29, 2000.

[8] H. Atamna and K. Boyle. Amyloid-beta peptide binds with heme to form a peroxidase: relationship to the cytopathologies of alzheimer's disease. *Proc Natl Acad Sci U S A*, 103(9):3381–6, 2006.

[9] F.M. Atay, T. Bıyıkoğlu, and J. Jost. Network synchronization: Spectral versus statistical properties. *Physica D Nonlinear Phenomena*, 224:35–41, 2006.

[10] S. Bacle. *Extremal metrics on graphs and manifold.* PhD thesis, McGill University, 2005.

[11] S. Bandyopadhyay, R. Mitra, U. Maulik, and M.Q. Zhang. Development of the human cancer microRNA network. *Silence*, 1:6, 2010.

[12] A. Banerjee. Structural distance and evolutionary relationship of networks. arXiv:0807.3185v2 [q-bio.QM], 2009.

[13] A. Banerjee and J. Jost. Spectral plots and the representation and interpretation of biological data. *Theory in Biosciences*, 126(1):1431–7613 (Print) 1611–7530 (Online), 2007.

[14] A. Banerjee and J. Jost. Spectral plot properties: towards a qualitative classification of networks. *Networks and heterogeneous media*, 3(2):395–411, 2008.

[15] A. Banerjee and J. Jost. Graph spectra as a systematic tool in computational biology. *Discrete Appl. Math.*, 157(10):2425–2431, 2009.

[16] A. L. Barabási. The network takeover. *Nature Physics*, 8, January 2012.

[17] A. L. Barabsi and E. Bonabeau. Scale-free networks. *Scientific American*, 288(60-69), 2003.

[18] A. Baralla, W.I. Mentzen, and A. de la Fuente. Inferring Gene Networks: Dream or Nightmare? *Annals of the New York Academy of Science*, 1158:246–256, 2009.

[19] A. Barla, G. Jurman, R. Visintainer, M. Squillario, M. Filosi, S. Riccadonna, and C. Furlanello. *Springer Handbook of Bio-/Neuroinformatics*, chapter A machine learning pipeline for discriminant pathways identification. Springer, 2012. ISBN:978-3-642-30573-3. In press.

[20] A. Barla, S. Mosci, L. Rosasco, and A. Verri. A method for robust variable selection with significance assessment. In M. Verleysen, editor, *Proc. ESANN 2008*, pages 83–88, 2008.

[21] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*, 57(1):289–300, 1995.

[22] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.

[23] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Phys. Rep.*, 424(4–5):175–308, 2006.

[24] C.J. Braun, X. Zhang, I. Savelyeva, S. Wolff, U.M. Moll, T. Schepeler, T.F. Ørntoft, C.L. Andersen, and M. Dobbelstein. p53-Responsive

MicroRNAs 192 and 215 Are Capable of Inducing Cell Cycle Arrest. *Cancer Research*, 68(24):10094–10104, 2008.

[25] B. J. Breitkreutz, C. Stark, and M. Tyers. Osprey: a network visualization system. *Genome biology*, 3(12), 2002.

[26] D.A. Brown, A. Forward, S. Marsh, and M.P. Caulfield. Antagonist discrimination between ganglionic and ileal muscarinic receptors. *British Journal of Pharmacology*, 120(S1):444–446, 1997.

[27] M. Buchanan, G. Caldarelli, P. De Los Rios, F. Rao, and M. Vendruscolo, editors. *Networks in Cell Biology*. Cambridge University Press, 2010.

[28] A. Budhu, H.-L. Jia, M. Forgues, C.-G. Liu, D. Goldstein, A. Lam, K. A. Zanetti, Q.-H. Ye, L.-X. Qin, C. M. Croce, Z.-Y. Tang, and X. W. Wang. Identification of Metastasis-Related MicroRNAs in Hepatocellular Carcinoma. *Hepatology*, 47(3):897–907, 2008.

[29] H. Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18:689–694, 1997.

[30] A. Canty and B.D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2012. R package version 1.3-5.

[31] Frederic Chibon. Cancer gene expression signatures the rise and fall? *European Journal of Cancer*, in press, 2013.

[32] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[33] F. Comellas and J. Diaz-Lopez. Spectral reconstruction of complex networks. *Physica A*, 387:6436–6442, 2008.

[34] A.P. Cootes, S.H. Muggleton, and M.J.E. Sternberg. The Identification of Similarities between Biological Networks: Application to the Metabolome and Interactome. *J. of Mol. Biol.*, 369:1126–1139, 2007.

[35] T.M. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[36] A. Cozza, E. Melissari, P. Iacopetti, V. Mariotti, A. Tedde, B. Nacmias, A. Conte, S. Sorbi, and S. Pellegrini. Snps in neurotrophin system genes and alzheimer's disease in an italian population. *J Alzheimers Dis*, 15(1):61–70, 2008.

[37] A.C. Davison and D.V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, 1997.

[38] C. De Mol, S. Mosci, M. Traskine, and A. Verri. A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data. *J Comput Biol*, 16(5):677–690, 2009.

[39] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)*. Cambridge University Press, January 2005.

[40] R. De Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, 2010.

[41] B. Di Camillo. *netsim: Gene network simulator*, 2007. R package version 1.1.

[42] B. Di Camillo, G. Toffolo, and C. Cobelli. A Gene Network Simulator to Assess Reverse Engineering Algorithms. *Ann. N.Y. Acad. Sci.*, 1158:125–142, 2009.

[43] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes. Critical phenomena in complex networks. *Rev. Mod. Phys.*, 80(4):1275–1335, 2008.

[44] E.R. Dougherty. Validation of gene regulatory networks: scientific and inferential. *Briefings in Bioinformatics*, 12(3):245–252, 2010.

[45] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, May 1994.

[46] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins, and T.S. Gardner. Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology*, 5(1):e8, 2007.

[47] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J Lin. LIBLINEAR: A library for large linear classification. *J Mach Learn Res*, 9:1871–1874, 2008.

[48] D. Fay, H. Haddadi, A.W. Moore, R. Mortier, S. Uhlig, and A. Jamakovic. A weighted spectrum metric for comparison of Internet topologies. *SIGMETRICS Perform. Eval. Rev.*, 37(3):67–72, 2009.

[49] C. B. Forsyth, K. M. Shannon, J. H. Kordower, R. M. Voigt, M. Shaikh, J. A. Jaglin, J. D. Estes, H. B. Dodiya, and A. Keshavarzian. Increased Intestinal Permeability Correlates with Sigmoid Mucosa alpha-Synuclein Staining and Endotoxin Exposure Markers in Early Parkinson's Disease. *PLoS ONE*, 6(12):e28032, 2011.

[50] T. Fuller, A. Ghazalpour, J. Aten, T. Drake, A. Lusis, and S. Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 2008.

[51] S.A. Georges, M.C. Biery, S. Kim, J.M. Schelter, J. Guo, A.N. Chang, A.L. Jackson, M.O. Carleton, P.S. Linsley, M.A. Cleary, and B.N. Chau. Coordinated Regulation of Cell Cycle Transcripts by p53-Inducible microRNAs, miR-192 and miR-215. *Cancer Research*, 68(24):10105–10112, 2008.

[52] A. Ghazalpour, S. Doss, B. Zhang, C. Plaisier, S. Wang, E.E. Schadt, A. Thomas, T.A. Drake, A.J. Lusis, and S Horvath. Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight. *PLoS Genet*, 2(8):e130, 2006.

[53] J. Gillis and P. Pavlidis. The role of indirect connections in gene networks in predicting function. *Bioinformatics*, 27(13):1860–1866, 2011.

[54] A. Goldenberg, A.X. Zheng, S.E. Fienberg, and E.M. Airoldi. A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2009.

[55] Gene H. Golub and Charles F. Van Loan. *Matrix computations (2nd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1993.

[56] M. Grimaldi, R. Visintainer, and G. Jurman. RegnANN: Reverse Engineering Gene Networks using Artificial Neural Networks. *PLoS ONE*, 6(12):e28646, 2011.

[57] Z. Gu, C. Zhang, and J. Wang. Gene regulation is governed by a core network in hepatocellular carcinoma. *BMC Systems Biology*, 6(1):32, 2012.

[58] R Guimerà and L Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[59] W.H. Haemers and E. Spence. Enumeration of cospectral graphs. *Eur. J. Comb.*, 25(2):199–211, 2004.

[60] F. He, R. Balling, and A.-P. Zeng. Reverse engineering and verification of gene networks: Principles, assumptions, and limitations of present methods and future perspectives. *J Biotechnol*, 144(3):190–203, 2009.

[61] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer New York Dordrecht Heidelberg London, 2011.

[62] S. Horvath and J. Dong. Geometric Interpretation of Gene Coexpression Network Analysis. *PLoS Comput. Biol.*, 4(8):e1000117, 2008.

[63] D.W. Huang, B.T. Sherman, and R.A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, 2009.

[64] D. Hurley, H. Araki, Y. Tamada, B. Dunmore, D. Sanders, S. Humphreys, M. Affara, S. Imoto, K. Yasuda, Y. Tomiyasu, K. Tashiro, C. Savoie, V. Cho, S. Smith, S. Kuhara, S. Miyano, D. S. Charnock-Jones, E. J. Crampin, and C. G. Print. Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. *Nucleic Acids Research*, 40(6):2377–2398, March 2012.

[65] T. Ideker and N. J. Krogan. Differential network biology. *Molecular Systems Biology*, 8(1), January 2012.

[66] J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2):149–155, January 2008.

[67] M. Ipsen and A.S. Mikhailov. Evolutionary reconstruction of networks. *Phys. Rev. E*, 66(4):046109, 2002.

[68] D. Jakobson and I. Rivin. Extremal metrics on graphs, I. *Forum Math.*, 14(1):147–163, 2002.

[69] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, may 2001.

[70] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, oct 2000.

[71] J. Ji, J. Shi, A. Budhu, Z. Yu, M. Forgues, S. Roessler, S. Ambs, Y. Chen, P.S. Meltzer, C.M. Croce, L.-X. Qin, K. Man, C.-M. Lo, J. Lee, I.O.L. Ng, J. Fan, Z.-Y. Tang, H.-C. Sun, and X.W. Wang. MicroRNA Expression, Survival, and Response to Interferon in Liver Cancer. *New England Journal of Medicine*, 361:1437–1447, 2009.

[72] Y. Jiao, K. Lawler, G. Patel, A. Purushotham, A.F. Jones, A. Grigoriadis, A. Tutt, T. Ng, and A.E. Teschendorff. DART: Denoising Algorithm based on Relevance network Topology improves molecular pathway activity inference. *BMC Bioinformatics*, 12(1):403, 2011.

[73] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Algebraic Comparison of Partial Lists in Bioinformatics. *PLoS ONE*, 7(5):e36540, 2012.

[74] G. Jurman, R. Visintainer, and C. Furlanello. An introduction to spectral distances in networks. *Frontiers in Artificial Intelligence and Applications*, 226:227–234, 2011.

[75] G. Jurman, R. Visintainer, S. Riccadonna, M. Filosi, and C. Furlanello. A glocal distance for network comparison. arXiv:1201.2931 [math.CO], 2012.

[76] A. Kamburov, U. Stelzl, and R. Herwig. Intscore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Research*, first published online May 30, 2012.

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

[78] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9th edition, March 1990.

[79] M. P. Keller, Y. Choi, P. Wang, D. B. B. Davis, M. E. Rabaglia, A. T. Oler, D. S. Stapleton, C. Argmann, K. L. Schueler, S. Edwards, H. A. Steinberg, N. E. Chaibub, R. Kleinhanz, S. Turner, M. K. Hellerstein, E. E. Schadt, B. S. Yandell, C. Kendziorski, and A. D. Attie. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome research*, 18(5):706–716, May 2008.

[80] P. Khaitovich, B. Muetzel, X. She, M. Lachmann, I. Hellmann, J. Dietzsch, S. Steigele, H. Do, G. Weiss, W. Enard, F. Heissig, T. Arendt,

K. Nieselt-Struwe, E.E. Eichler, and S. Pääbo. Regional Patterns of Gene Expression in Human and Chimpanzee Brains. *Genome Res*, 14(8):1462–1473, 2004.

[81] P. Khatri, M. Sirota, and A.J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol*, 8(2):e1002375, 2012.

[82] M. Kolar, L. Song, A. Ahmed, and E.P. Xing. Estimating time-varying networks. *Ann. Appl. Stat.*, 4(1):94–123, 2010.

[83] A. Krishnan, A. Giuliani, and M. Tomita. Indeterminacy of Reverse Engineering of Gene Regulatory Networks: The Curse of Gene Elasticity. *PLoS ONE*, 2(6):e562, 2007.

[84] V. Lacroix, L. Cottret, P. Thébault, and M.F. Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5(4):594–617, 2008.

[85] P. Langfelder and S. Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC Sys Biol*, 1:54, 2007.

[86] P. Langfelder and S. Horvath. Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11):1–17, 2012.

[87] P. Langfelder, R. Luo, M. Oldham, and S. Horvath. Is My Network Module Preserved and Reproducible? *PLoS Comput Biol*, 7(1):e1001057, 2011.

[88] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559+, 2008.

[89] P. T.-Y. Law and N. Wong. Emerging roles of microRNA in the intracellular signaling networks of hepatocellular carcinoma. *Journal of Gastroenterology and Hepatology*, 26(3):437–449, 2011.

[90] Y. Lazebnik. Can a biologist fix a radio?–Or, what I learned while studying apoptosis. *Cancer cell*, 2(3):179–182, September 2002.

[91] W. S. Liang, T. Dunckley, T. G. Beach, A. Grover, D. Mastroeni, K. Ramsey, R. J. Caselli, W. A. Kukull, D. Mckeel, J. C. Morris, C. M. Hulette, D. Schmechel, E. M. Reiman, J. Rogers, and D. A. Stephan. Neuronal gene expression in non-demented individuals with intermediate Alzheimer's Disease neuropathology. *Neurobiology of Aging*, In Press, Corrected Proof, 2010.

[92] W. S. Liang, E. M. Reiman, J. Valla, T. Dunckley, T. G. Beach, A. Grover, T. L. Niedzielko, L. E. Schneider, D. Mastroeni, R. Caselli, W. Kukull, J. C. Morris, C. M. Hulette, D. Schmechel, J. Rogers, and D. A. Stephan. Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11):4441–4446, March 2008.

[93] D. Liben-Nowell. *An Algorithmic Approach to Social Networks*. PhD thesis, Massachusetts Institute of Technology, 2005.

[94] B.A. Logsdon and J. Mezey. Gene Expression Network Reconstruction by Convex Feature Selection when Incorporating Genetic Perturbations. *PLoS Computational Biology*, 6(12):e1001014, 2010.

[95] J. Loscalzo and A. L. Barabasi. Systems biology and the future of medicine. *Wiley interdisciplinary reviews. Systems biology and medicine*, 3(6):619–627, November 2011.

[96] M. A. Lovell, B. C. Lynn, S. Xiong, J. F. Quinn, J. Kaye, and W. R. Markesbery. An aberrant protein complex in csf as a biomarker of alzheimer disease. *Neurology*, 70(23):2212–8, 2008.

[97] B. MacArthur, R.J. Sánchez-García, and J. Anderson. Symmetry in complex networks. *Discrete Appl. Math.*, 156(18):3525–3531, 2008.

[98] P. Madhamshettiwar, S. Maetschke, M. Davis, A. Reverter, and M. Ragan. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine*, 4(5):41, 2012.

[99] D. Marbach, J.C. Costello, R. Kuffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, M. Kellis, J.J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nat Methods*, 9(8):796–804, 2012.

[100] D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, 107(14):6286–6291, 2010.

[101] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(7):S7, 2006.

[102] F. Markowetz and R. Spang. Inferring cellular networks–a review. *BMC bioinformatics*, 8 Suppl 6(Suppl 6):S5+, 2007.

[103] P. Meyer, L.G. Alexopoulos, T. Bonk, A. Califano, C.R. Cho, A. de la Fuente, D. de Graaf, A.J. Hartemink, J. Hoeng, N.V. Ivanov, H. Koeppl, R. Linding, D. Marbach, R. Norel, M.C. Peitsch, J.J. Rice, A. Royyuru, F. Schacherer, J. Sprengel, K. Stolle, D. Vitkup,

and G. Stolovitzky. Verification of systems biology research in the age of collaborative competition. *Nature Biotechnology*, 29(9):811–815, 2011.

[104] P. Meyer, F. Lafitte, and G. Bontempi. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics*, 9(1):461+, 2008.

[105] M.A. Miller, X.-J. Feng, G. Li, and H.A. Rabitz. Identifying Biological Network Structure, Predicting Network Behavior, and Classifying Network State With High Dimensional Model Representation (HDMR). *PLoS ONE*, 7(6):e37664, 2012.

[106] Nature Biotechnology. Finding correlations in big data. *Nature Biotechnology*, 30(4):334–335, 2012.

[107] R. Neal. *Bayesian learning for neural networks.* PhD thesis, Department of Computer Science, University of Toronto., 1995.

[108] R. M. Neal. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics).* Springer, 1 edition, August 1996.

[109] I. Nemenman, G.S. Escola, W.S. Hlavacek, P.J. Unkefer, C.J. Unkefer, and M.E. Wall. Reconstruction of Metabolic Networks from High-Throughput Metabolite Profiling Data. *Ann NY Acad Sci*, 1115:102–115, 2007.

[110] M.E.J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256, 2003.

[111] C. J. Oates and S. Mukherjee. Network Inference and Biological Dynamics. *Ann. Appl. Stat.*, 6(3):1209 – 1235, 2012.

[112] M.C. Oldham, S. Horvath, and D.H. Geschwind. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A*, 103(47):179730–17978, 2006.

[113] A. Patereli, G. A. Alexiou, K. Stefanaki, M. Moschovi, I. Doussis-Anagnostopoulou, N. Prodromou, and O. Karentzou. Expression of epidermal growth factor receptor and her-2 in pediatric embryonal brain tumors. *Pediatr Neurosurg*, 46(3):188–92, 2010.

[114] B. Pincombe. Detecting changes in time series of network graphs using minimum mean squared error and cumulative summation. In W. Read and A.J. Roberts, editors, *Proceedings of the 13th Biennial Computational Techniques and Applications Conference, CTAC-2006*, volume 48 of *ANZIAM J.*, pages C450–C473, 2007.

[115] R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and G. Stolovitzky. Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge. *Sci. Signal.*, 4(189):mr7+, September 2011.

[116] R.J. Prill, D. Marbach, J. Saez-Rodriguez, P.K. Sorger, L.G. Alexopoulos, X. Xue, N.D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLoS ONE*, 5(2):e9202, 02 2010.

[117] J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. R. McCartney, M. C. Schmidt, N. Rachidi, S.-J. Lee, A. S. Mah, L. Meng, M. J. R. Stark, D. F. Stern, C. D. Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. F. Predki, and M. Snyder. Global analysis of protein phosphorylation in yeast. *Nature*, 438(7068):679–84, December 2005.

[118] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[119] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, aug 2002.

[120] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting novel associations in large datasets. *Science*, 6062(334):1518–1524, 2011.

[121] G.J. Rodgers, K. Austin, B. Kahng, and D. Kim. Eigenvalue spectra of complex networks. *Journal of Physics A: Mathematical and General*, 38(43):9431, 2005.

[122] E.E. Schadt, S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, P.S. Linsley, M. Mao, R.B. Stoughton, and S.H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–302, 2003.

[123] T. Schaffter, D. Marbach, and D. Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011. wingx.

[124] O. Shanker. Defining dimension of a complex network. *Mod. Phys. Lett. B*, 21(6):321–326, 2007.

[125] O. Shanker. Graph zeta function and dimension of complex network. *Mod. Phys. Lett. B*, 21(11):639–644, 2007.

[126] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31:64–68, 2002.

[127] L. Song, M. Kolar, and E. P. Xing. KELLER: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–i136, June 2009.

[128] L. Song, P. Langfelder, and S. Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13(1):328, 2012.

[129] T. Speed. A Correlation for the 21st Century. *Science*, 6062(334):1502–1503, 2011.

[130] D.A. Spielman. Spectral Graph Theory: The Laplacian (Lecture 2). Lecture notes, 2009.

[131] G. Stolovitzky, D. Monroe, and A. Califano. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22, December 2007.

[132] G. Stolovitzky, R. J. Prill, and A. Califano. Lessons from the dream2 challenges. *Annals of the New York Academy of Sciences*, 1158(1):159–195, 2009.

[133] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[134] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A

knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545–15550, 2005.

[135] G. Szederkenyi, J. Banga, and A. Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Systems Biology*, 5(1):177, 2011.

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

[137] The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(suppl 1):D258–D261, 2004.

[138] The MAQC Consortium. The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*, 28(8):827–838, 2010.

[139] A.H. Tong. Systematic genetic analysis with ordered arrays of yeast deletion mutants, 2001.

[140] R. Tönjes and B. Blasius. Perturbation Analysis of Complete Synchronization in Networks of Phase Oscillators. arXiv:0908.3365, 2009.

[141] O.G. Troyanskaya, M. Cantor, G. Sherlock, P.O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[142] D. W Tsuang, R. G. Riekse, K. M Purganan, A. C. David, T. J. Montine, G. D. Schellenberg, E. J. Steinbart, E. C. Petrie, T. D. Bird, and J. B. Leverenz. Lewy body pathology in late-onset familial

alzheimer's disease: a clinicopathological case series. *J Alzheimers Dis*, 9(3):235–42, 2006.

[143] K. Tun, P. Dhar, M. Palumbo, and A. Giuliani. Metabolic pathways variability and sequence/networks comparisons. *BMC Bioinformatics*, 7(1):24, 2006.

[144] E.R. van Dam and W.H. Haemers. Which graphs are determined by their spectrum? *Linear Algebra Appl.*, 373:241–272, 2003.

[145] D. M. van Leeuwen, M. P. Peter, J. M. Hendriksen, A. Boorsma, M. H. M. van Herwijnen, R. W. H. Gottschalk, M. Kirsch-Volders, L. E. Knudsen, R. J. Sram, E. Bajak, J. H. M. van Delft, and J. C. S. Kleinjans. Genomic analysis suggests higher susceptibility of children to air pollution . *BMC Bioinformatics*, 29(5):977983, 2008.

[146] P. Van Mieghem. *Graph Spectra for Complex Networks*. Cambridge University Press, 2011.

[147] S. Volinia, M. Galasso, S. Costinean, L. Tagliavini, G. Gamberoni, A. Drusco, J. Marchesini, N. Mascellani, M.E. Sana, R. Abu Jarour, C. Desponts, M. Teitell, R. Baffa, R. Aqeilan, M.V. Iorio, C. Taccioli, R. Garzon, G. Di Leva, M. Fabbri, M. Catozzi, M. Previati, S. Ambs, T. Palumbo, M. Garofalo, A. Veronese, A. Bottoni, P. Gasparini, C.C. Harris, R. Visone, Y. Pekarsky, A. de la Chapelle, M. Bloomston, M. Dillhoff, L.Z. Rassenti, T.J. Kipps, K. Huebner, F. Pichiorri, D. Lenze, S. Cairo, M.-A. Buendia, P. Pineau, A. Dejean, N. Zanesi, S. Rossi, G.A. Calin, C.-G. Liu, J. Palatini, M. Negrini, A. Vecchione, A. Rosenberg, and C.M. Croce. Reprogramming of miRNA networks in cancer and leukemia. *Genome Research*, 20(5):589–599, 2010.

[148] M. Vos, G. Esposito, J.N. Edirisinghe, S. Vilain, D.M. Haddad, J.R. Slabbaert, S. Van Meensel, O. Schaap, B. De Strooper,

R. Meganathan, V.A. Morais, and P. Verstreken. Vitamin K2 Is a Mitochondrial Electron Carrier That Rescues Pink1 Deficiency. *Science*, 336(6086):1306–1310, 2012.

[149] S. Wang, N. Yehya, E.E. Schadt, T.A. Drake, and A.J. Lusis. Genetic and Genomic Analysis of Fat Mass Trait with Complex Inheritance Reveals Marked Sex Specificity. *PLoS Genet*, 2(2):e15, 2006.

[150] W. Wang and C.-X. Xu. A sufficient condition for a family of graphs being determined by their generalized spectra. *Eur. J. Combin.*, 27:826–840, 2006.

[151] W. Wang and C.-X. Xu. On the asymptotic behavior of graphs determined by their generalized spectra. *Discrete Math.*, 310:70–76, 2010.

[152] Rand R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Statistical Modeling and Decision Science. Academic Press, 2nd edition, December 2004.

[153] S. Wuchty, E. Rasasz, and A. L. Barbarasi. The architecture of Biological Networks, 2003.

[154] B. Zhang and S. Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[155] B. Zhang, S. Kirov, and J. Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 33(Web Server issue):W741–W748, 2005.

[156] Y. Zhang, M. James, F.A. Middleton, and R.L. Davis. Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, en-

ergy metabolism, and signaling pathways, and suggests novel disease mechanisms. *Am J Med Genet B Neuropsychiatr Genet*, 137B(1):5–16, 2005.

[157] P. Zhu and R.C. Wilson. A study of graph spectra for comparing graphs. In W. Clocksin, A. Fitzgibbon, and P. Torr, editors, *Proceedings of the 16-th British Machine Vision Conference*, pages 2833–2841, 2005.

[158] X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024, May 2007.

[159] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J R Stat Soc B*, 67(Part 2):301–320, 2005.

Appendix A

Module Preservation: Measures and Results

A.1 Module Preservation Measures

Because preservation statistics measure different aspects of module preservation, their results may not always agree. We find it useful to aggregate different module preservation statistics into composite preservation statistics. Composite preservation statistics also facilitate a fast evaluation of many modules in multiple networks. We define several composite statistics. For correlation networks based on quantitative variables, the 4 density preservation statistics are summarized by $Z_{density}$ A.2, the 3 connectivity based statistics are summarized by $Z_{connectivity}$ A.3, and all individual Z statistics are summarized by $Zsummary$ defined as follows

$$Z_{summary} = \frac{Z_{connectivity} + Z_{density}}{2} \tag{A.1}$$

$$Z_{density} = median(Z_{meanCor}, Z_{meanAdj}, Z_{propVarExpl}, Z_{meanKME}). \tag{A.2}$$

$$Z_{connectivity} = median(Z_{cor.kIM}, Z_{cor.Adj}, Z_{cor.kME}, Z_{cor.kMEall}, Z_{cor.cor}). \tag{A.3}$$

The Z statistics often depends on the module size (i.e. the number of nodes in a module). This fact reflects the intuition that it is more significant

to observe that the connectivity patterns among hundreds of nodes are preserved than to observe the same among say only 5 nodes. Having said this, there will be many situations when the dependence on module size is not desirable, e.g., when preservation statistics of modules of different sizes are to be compared. In this case, we recommend to either focus on the observed values of the individual statistics or alternatively to summarize them using the composite module preservation statistic *medianRank*.
We define the *medianRank* as an alternative rank-based measure that relies on observed preservation statistics rather than the permutation Z statistics. For each statistic a, we rank the modules based on the observed values $obs_a^{(q)}$. Thus, each module is assigned a rank $rank_a^{(q)}$ for each observed statistic. We then define the median density and connectivity ranks

$$\begin{aligned} medianRank.density^{(q)} &= median_{a \in Densitystatistics}(rank_a^{(q)}) \\ medianRank.connectivity^{(q)} &= median_{a \in Connectivitystatistics}(rank_a^{(q)}) \end{aligned} \tag{A.4}$$

The *medianRank* is useful for comparing relative preservation among multiple modules: a module with lower median rank tends to exhibit stronger observed preservation statistics than a module with a higher median rank. Since *medianRank* is based on the observed preservation statistics (as opposed to Z statistics) we find that it is much less dependent on module size [87].

$$medianRank_{summary} = \frac{medianRank.density + medianRank.connectivity}{2} \tag{A.5}$$

Module preservation statistics for general networks Here we describe module preservation statistics that can be used to determine whether a module that is present in a reference network (with adjacency $A^{[ref]}$) can also be found in an independent test network (with adjacency $A^{[test]}$). Specifically, assume the vector $Cl^{[ref]}$ encodes the module assignments in

the reference network. Thus $Cl_i^{[ref]} = q(q \in \{1, \ldots, Q^{[ref]}\})$ if node i is assigned to module q. We reserve the label $Cl = 0$ for nodes that are not assigned to any module. For a given module q with n_q nodes, the $n_q \times n_q$ module adjacency matrices are denoted $A^{[ref](q)}$ and $A^{[test](q)}$ in the reference and test networks, respectively. We propose network concepts that can be useful for determining whether a module q(found in the reference network) is preserved in the test network.

Intuitively, one may call a module q preserved if it has a high density in the test network. We define the *meanadjacency* for module q as the module density in the test network,

$$meanAdj^{[test](q)} = density^{[test](q)} = mean(vectorizeMatrix(A^{[test](q)})) \tag{A.6}$$

Connectivity preservation statistics quantify how similar connectivity of a given module is between a reference and a test network. For example, module connectivity preservation can mean that, within a given module q, nodes with a high connection strength in the reference network also exhibit a high connection strength in the test network. This property can be quantified by the correlation of intramodular adjacencies in reference and test networks. Specifically, if the entries of the first adjacency matrix $A^{[ref](q)}$ are correlated with those of the second adjacency matrix $A^{[test](q)}$ then the adjacency pattern of the module is preserved in the second network. Therefore, we define the *adjacencycorrelation* of the module q network as

$$cor.Adj^{(q)} = cor(vectorizeMatrix(A^{[ref](q)}), vectorizeMatrix(A^{[test](q)})) \tag{A.7}$$

High $cor.Adj^{(q)}$ indicates that adjacencies within the module q in the reference and test networks exhibit similar patterns. If module q is preserved in the second network, the highly connected hub nodes in the reference network will often be highly connected hub nodes in the test network. In other

words, the intramodular connectivity $kIM^{[ref](q)}$ in the reference network should be highly correlated with the corresponding intramodular connectivity $kIM^{[test](q)}$ in the test network. Thus, we define the correlation of intramodular connectivities,

$$cor.kIM^{(q)} = cor(kIM^{[ref](q)}, kIM^{[test](q)}), \tag{A.8}$$

where $kIM^{[ref](q)}$ and $kIM^{[test](q)}$ are the vectors of intramodular connectivities of all nodes in module q in the reference and test networks, respectively.

Module preservation statistics for correlation networks The specific nature of correlation networks allows us to define additional module preservation statistics. The underlying information carried by the sign of the correlation can be used to further refine the statistics irrespective of whether a signed or unsigned similarity is used in network construction [87]. To simplify notation, we define

$$\begin{aligned} r_{ij}^{[ref]} &= cor(x_i^{[ref]}, x_j^{[ref]}) \\ r_{ij}^{[test]} &= cor(x_i^{[test]}, x_j^{[test]}) \end{aligned} \tag{A.9}$$

We will use the notation $r_{ij}^{[ref](q)}$ for the correlation matrix restricted to the nodes in module q. We define the mean correlation density of module q as

$$meanCor^{[test](q)} = mean\{vectorizeMatrix(sign(r_{ij}^{[ref](q)})r_{ij}^{[test](q)})\}. \tag{A.10}$$

Thus the correlation measure of module preservation is the mean correlation in the test network multiplied by the sign of the corresponding correlations in the reference network. We note that a correlation that has the same sign in the reference and test networks increases the mean, while a correlation that changes sign decreases the mean. Because the preservation statistic keeps track of the sign of the corresponding correlation in the reference network, we call it the mean sign-aware correlation.

To measure the preservation of connectivity patterns within module q between the reference and test networks, we define a correlation-based measure *cor.cor* similar to the *cor.Adj* statistic

$$cor.cor^{(q)} = cor(vectorizeMatrix(r^{[ref](q)}), cvectorizeMatrix(z^{[test](q)})) \tag{A.11}$$

Eigennode summarizes a correlation module and provides a measure of module membership. Many module construction methods lead to correlation network modules comprised of highly correlated variables. For such modules one can summarize the corresponding module vectors using the first principal component denoted by $E^{(q)}$, referred to as the module eigennode (ME) or (in gene co-expression networks) the module eigengene. For example, the gene expression profiles of a given co-expression module can be summarized with the module eigengene [62, 85, 79]. To visualize the meaning of the module eigengene, consider the heat map in Figure 5A. Here rows correspond to genes inside a given module and columns correspond to microarray samples. The module eigennode $E^{(q)}$ can be used to define a quantitative measure of module membership [62] of node i in module q:

$$kME_i^{(q)} = cor(x_i, E^{(q)}), \tag{A.12}$$

where x_i is the profile of node i. The module membership $kME_i^{(q)}$ lies in $[-1, 1]$ and specifies how close node i is to module q. $kME_i^{(q)}$ is also referred to as module eigengene-based connectivity [52, 50]. Both intramodular network concepts and inter modular network concepts can be used to study the preservation of network modules. By measuring how these network concepts are preserved from a reference network to a test network, one can define network module preservation statistics as described below [87].

Eigennode-based density preservation statistics. The concept of the module eigennode also gives rise to several preservation statistics that in effect measure module density, or, from a different point of view, how well

the eigennode represents the whole module. In [87] is proven that the proportion of variance explained (PVE) can also be calculated as mean squared kME value:

$$propVarExpl^{[test](q)} = mean_{i \in \mathcal{M}_q}\{(kME_i^{[test](q)})\}, \tag{A.13}$$

where $E^{[test](q)}$ is the eigennode of module q in the test network. The $meansign - awaremodulemembership$ is defined as:

$$meanKME^{[test](q)} = mean_{i \in \mathcal{M}_q}\{sign(kME_i^{[ref](q)})kME_i^{[test](q)})\} \tag{A.14}$$

Eigennode-based connectivity preservation statistics. Intuitively, if the internal structure of a module is preserved between a reference and a test network , we expect that a variable with a high module membership in the reference network will have a high module membership in the test network as well; conversely, variables with relatively low module membership in the reference network should also have a relatively low module membership in the test network. In other words, intramodular hubs in the reference network should also be intramodular hubs in the test network. For a given module q we define the $cor.kME^{(q)}$ statistic as

$$cor.kME^{(q)} = cor_{i \in \mathcal{M}_q}(kME_i^{[ref](q)}, kME_i^{[test](q)})$$

where the correlation runs only over variables that belong to module q. We also define an analogous statistic by correlating the module membership of all network variables in the reference and test networks:

$$cor.kMEall^{(q)} = cor(kME_i^{[ref](q)}, kME_i^{[test](q)})$$

A.2 Statistics for module quality assessment

An important use of module preservation statistics is to define measures of module quality (or robustness), which may inform the module definition.

For example, to measure how robustly a module is defined in a given correlation network, one can use resampling techniques to create reference and test sets from the original data and evaluate module preservation across the resulting networks. Thus, any module preservation statistic naturally gives rise to a module quality statistic by applying it to repeated random splits (interpreted as reference and test set) of the data. By averaging the module preservation statistic across multiple random splits of the original data, one arrives at a module quality statistic. [87] Implementing the above mentined idea we indicate the two quality control measures $ZsummQ$ and $MRsummQ$ that refer to $Z_{summary}$ A.1 and $medianrank_{summary}$ A.5 respectively.

A.3 Additional Results

Table A.1: Preservation of female mouse liver modules in male data ref 1 test 2 (corr method: Spearman)

	Ipsen	Hamming	Mod.Ipsen
Ipsen	1.00	0.62	0.97
Hamming	0.62	1.00	0.72
Mod.Ipsen	0.97	0.72	1.00
ZsummaryQuality	0.26	-0.05	0.29
ZsummaryPreserv.	0.71	0.48	0.76
Zdensity	0.74	0.34	0.73
Zconnectivity	0.56	0.69	0.70
MRsummaryQuality	0.42	0.88	0.52
MRsummaryPreserv.	-0.74	-0.37	-0.73
MRdensity	-0.61	-0.22	-0.56
MRconnectivity	-0.75	-0.63	-0.83

Table A.2: Preservation of human brain modules in chimpanzee brains and vice versa ref 1 test 2 (corr method: Spearman)

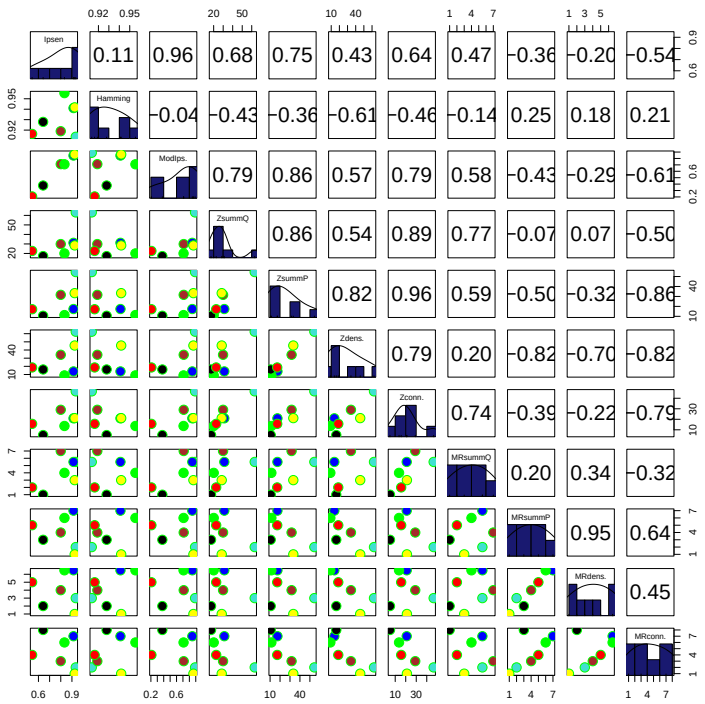
	Ipsen	Hamming	Mod.Ipsen
Ipsen	1.00	0.11	0.96
Hamming	0.11	1.00	-0.04
Mod.Ipsen	0.96	-0.04	1.00
ZsummaryQuality	0.68	-0.43	0.79
ZsummaryPreser.	0.75	-0.36	0.86
Zdensity	0.43	-0.61	0.57
Zconn	0.64	-0.46	0.79
MRsummaryQuality	0.47	-0.14	0.58
MRsummaryPreser.	-0.36	0.25	-0.43
MRdensity	-0.20	0.18	-0.29
MRconnectivity	-0.54	0.21	-0.61

Table A.3: Preservation of human brain modules in chimpanzee brains and vice versa ref 2 test 1 (corr method: Spearman)

	Ipsen	Hamming	Mod.Ipsen
Ipsen	1.00	0.11	0.96
Hamming	0.11	1.00	-0.04
Mod.Ipsen	0.96	-0.04	1.00
ZsummaryQuality	0.21	-0.68	0.32
ZsummaryPreser.	0.21	-0.68	0.32
Zdensity	0.14	-0.79	0.29
Zconn	0.79	-0.25	0.86
MRsummaryQuality	0.58	-0.47	0.72
MRsummaryPreser.	0.50	-0.11	0.67
MRdensity	0.46	-0.14	0.64
MRconnectivity	0.18	0.00	0.23

Figure A.1: A) Preservation of human brain modules in chimpanzee brains (corr method: Spearman). B) Preservation of chimpanzee brain modules in human brains (corr method: Spearman)

A)



B)

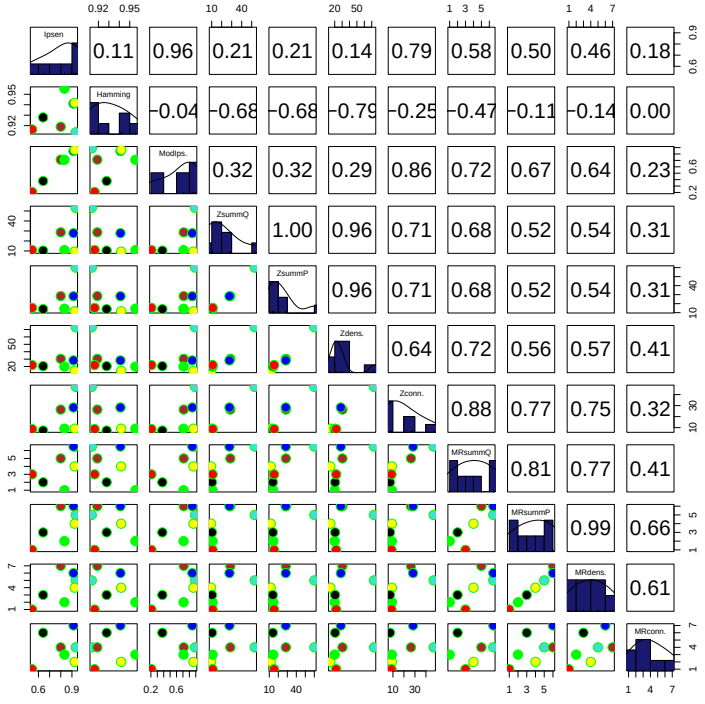
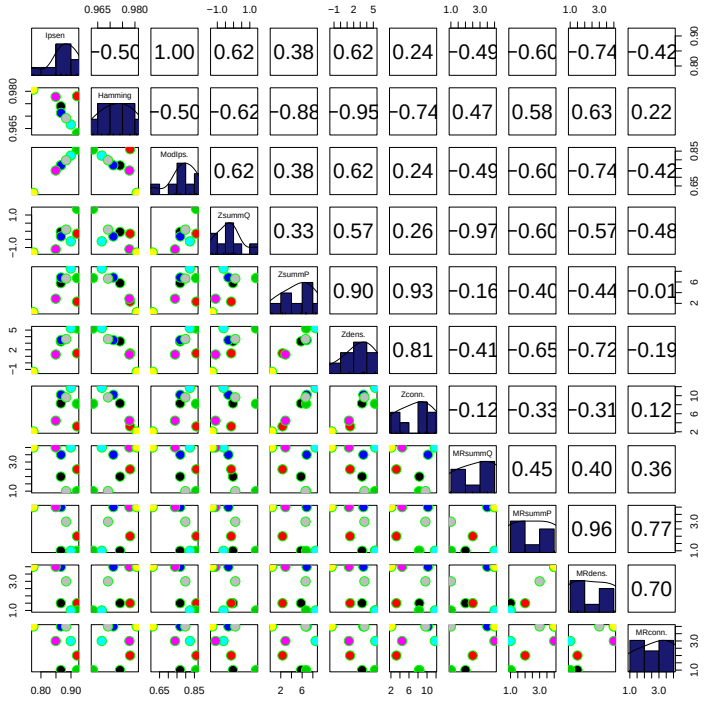


Figure A.2: A) Preservation of KEGG pathways between human and chimpanzee data using human as reference and chimp as test (corr method: Spearman). B) Preservation of KEGG pathways between human and chimpanzee data using chimp as reference and human as test (corr method: Spearman)

A)



B)

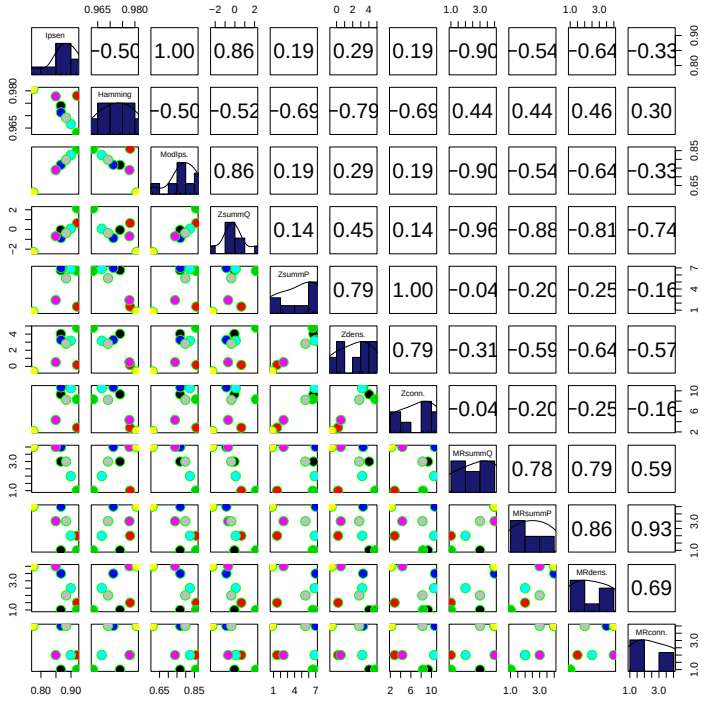


Figure A.3: Preservation of Cholesterol Biosynthesis Process module among 8 tissue/gender cobinations in F2 mice (corr method: Spearman)

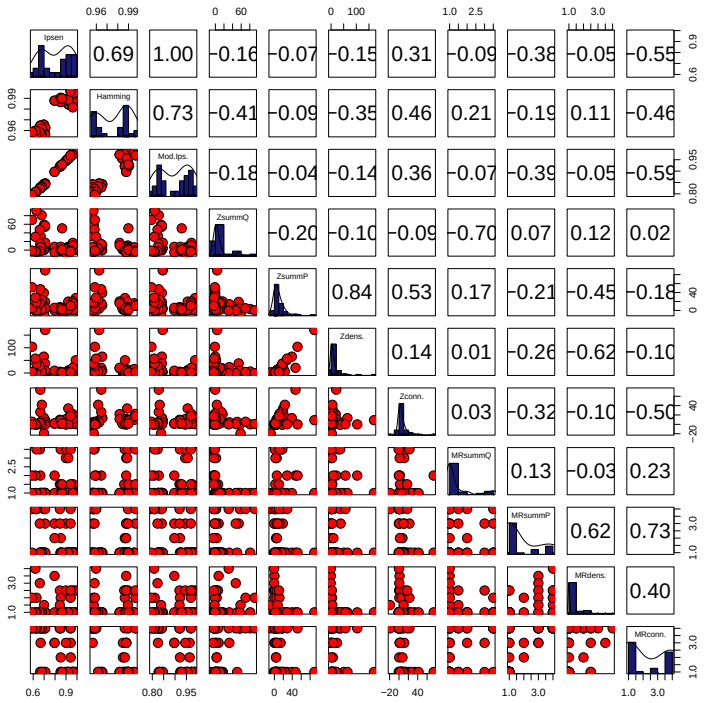


Table A.4: Preservation of KEGG pathways between human and chimp data ref 1 test 2 (corr method: Spearman)

	Ipsen	Hamming	Mod.Ipsen
Ipsen	1.0	-0.50	1.00
Hamming	-0.50	1.0	-0.50
Mod.Ipsen	1.00	-0.50	1.0
ZsummaryQuality	0.62	-0.62	0.62
ZsummaryPreser.	0.38	-0.88	0.38
Zdensity	0.62	-0.95	0.62
Zconnectivity	0.24	-0.74	0.24
MRsummaryQuality	-0.49	0.47	-0.49
MRsummaryPreser.	-0.60	0.58	-0.60
MRdensity	-0.74	0.63	-0.74
MRconnectivity	-0.42	0.22	-0.42

Table A.5: Preservation of KEGG pathways between human and chimp data ref 2 test 1 (corr method: Spearman)

	Ipsen	Hamming	Mod.Ipsen
Ipsen	1.00	-0.50	1.00
Hamming	-0.50	1.00	-0.50
Mod.Ipsen	1.00	-0.50	1.00
ZsummaryQuality	0.86	-0.52	0.86
ZsummaryPreser.	0.19	-0.69	0.19
Zdensity	0.29	-0.79	0.29
Zconnectivity	0.19	-0.69	0.19
MRsummaryQuality	-0.90	0.44	-0.90
MRsummaryPreser.	-0.54	0.44	-0.54
MRdensity	-0.64	0.46	-0.64
MRconnectivity	-0.33	0.30	-0.33

Table A.6: Preservation of Cholesterol Biosynthesis Process module among 8 tissue/gender combinations in F2 mice (corr method: Spearman)

	Ipsen	Hamming	Mod.Ipsen
Ipsen	1.00	0.69	1.00
Hamming	0.69	1.00	0.73
Mod.Ipsen	1.00	0.73	1.00
ZsummaryQuality	-0.40	-0.44	-0.39
ZsummaryPreser.	0.14	0.10	0.14
Zdensity	0.22	0.13	0.21
Zconnectivity	-0.02	0.11	-0.02
MRsummaryQuality	-0.05	0.09	-0.01
MRsummaryPreser.	-0.02	-0.01	-0.02
MRdensity	-0.04	-0.01	-0.04
MRconnectivity	0.14	0.13	0.14

Table A.7: Correlation between Mod.Ipsen distance and the Network-based module preservation measures for each tissue used as Reference (corr method: Spearman). Missing values are due to zero standard deviation in the considered values

	AdiposeF	AdiposeM	BrainF	BrainM	LiverF	LiverM	MuscleF	MuscleM	Mean
ZsummaryQuality	0.57	0.36	-0.61	-0.32	-0.57	-0.57	0.21	0.11	0.42
ZsummaryPreser.	-0.36	-0.29	-0.61	-0.07	0.79	0.75	-0.46	-0.61	0.49
Zdensity	-0.07	-0.07	-0.46	-0.29	0.82	0.75	-0.64	-0.89	0.50
Zconnectivity	-0.46	-0.18	-0.25	0.71	0.68	0.57	0.39	0.86	0.51
MRsummaryQuality	-0.27						-0.32	-0.87	0.48
MRsummaryPreser.	-0.20	-0.18	-0.41	-0.41	-0.77	-0.79	-0.18	0.09	0.38
MRdensity	-0.20	-0.18	0.61	0.41	-0.91	-0.79	0.40	0.53	0.50
MRconnectivity	-0.20	0.06	-0.60	-0.79	-0.87	-0.87	-0.48	-0.38	0.53
Mean	0.29	0.19	0.51	0.43	0.77	0.73	0.39	0.54	

Table A.8: Correlation between Mod.Ipsen distance and the Network-based module preservation measures for each tissue used as Test (corr method: Spearman). Missing values are due to zero standard deviation in the considered values

	AdiposeF	AdiposeM	BrainF	BrainM	LiverF	LiverM	MuscleF	MuscleM	Mean
ZsummaryQuality	-0.46	-0.25	-0.39	-0.68	0.68	0.50	-0.71	-0.54	0.53
ZsummaryPreser.	-1.00	-0.14	0.75	0.54	0.32	-0.32	0.64	0.32	0.50
Zdensity	-0.71	0.29	0.64	0.18	0.14	-0.82	0.71	0.00	0.44
Zconnectivity	-0.96	-0.21	0.50	0.86	0.64	0.82	0.64	0.46	0.64
MRsummaryQuality	-0.13	-0.13	-0.13	0.30	-0.76	-0.60	0.61	0.61	0.41
MRsummaryPreser.	-0.16	-0.20	-0.41	-0.41	-0.64	-0.41	-0.30	-0.70	0.40
MRdensity	0.69		-0.41				-0.30	-0.56	0.49
MRconnectivity	-0.18	0.12	-0.70	-0.79	-0.87	-0.85	-0.39	-0.38	0.53
Mean	0.54	0.19	0.49	0.54	0.58	0.62	0.54	0.45	0.49