Design of new algorithms for gene network reconstruction applied to *in silico* modeling of biomedical data

Doctoral Thesis

Fernando M. Delgado Chaves

Universidad Pablo de Olavide

Design of new algorithms for gene network reconstruction applied to *in silico* modeling of biomedical data

Fernando Miguel Delgado Chaves

Directors: Dr. Francisco A. Gómez Vela
Dr. Federico Divina

Escuela Politécnica Superior
Universidad Pablo de Olavide

Doctoral thesis

November 2022

D. Francisco A. Gómez Vela y D. Federico Divina, ambos con el cargo de profesor contratado Doctor de Universidad, adscritos al Lenguajes y Sistemas Informáticos, del Departamento de deporte e informática de la Universidad Pablo de Olavide de Sevilla,

CERTIFICAN QUE:

D. Fernando Miguel Delgado Chaves, graduado en Biotecnología por la Universidad Pablo de Olavide de Sevilla, ha realizado bajo su supervisión el trabajo de investigación titulado:

DESIGN OF NEW ALGORITHMS FOR GENE NETWORK RECONSTRUCTION APPLIED TO *IN SILICO* MODELING OF BIOMEDICAL DATA

Una vez revisado, autorizan la presentación del mismo como tesis doctoral en la Universidad Pablo de Olavide de Sevilla y estiman oportuna su presentación al tribunal que habrá de valorarlo. Dicha tesis ha sido realizada dentro del programa de doctorado Biotecnología, Ingeniería y Tecnología Química, de la Universidad Pablo de Olavide de Sevilla.

Noviembre 2022

D. Francisco A. Gómez Vela y D. Federico Divina, profesores contratado Doctor de Universidad, adscritos al Lenguajes y Sistemas Informáticos, del Departamento de deporte e informática de la Universidad Pablo de Olavide de Sevilla, como director y codirector, respectivamente, de la tesis titulada:

DESIGN OF NEW ALGORITHMS FOR GENE NETWORK RECONSTRUCTION APPLIED TO *IN SILICO* MODELING OF BIOMEDICAL DATA

proponen la siguiente composición del tribunal titular, a fin de que la Comisión de Doctorado designe al tribunal encargado de juzgar la tesis doctoral.

PRESIDENTE: Dr. D. Francisco Martínez Álvarez,
Universidad Pablo de Olavide, España.

VOCAL: Dr. D. Jaume Bacardit,
Universidad de Newcastle, Reino Unido.

SECRETARIA: Dra. Dña. Isabel de los Angeles Nepomuceno Chamorro,
Universidad de Sevilla, España.

SUPLENTES:

Dr. D. Mario D.L. Giacobini,
Universidad de Turín, Italia.

Dra. Dña. Beatriz Pontes Balanza,
Universidad de Sevilla, España.

Dr. D. Norberto Díaz Díaz,
Universidad Pablo de Olavide, España.

Noviembre 2022

Declaration

The research that produced the data used in this thesis was conducted by me at Pablo de Olavide University. I am aware of and agree to abide by the university's anti-plagiarism policy. I certify that I authored this thesis and that it was not previously submitted for any other degree or professional certification of equal standing. The research was largely planned and carried out by me, and my analysis and interpretation of the data are wholly original.

The entire dissertation is my own work, with the exception of what is indicated in the text and corresponding acknowledgements sections. Except where work from jointly-authored publications is included, I confirm that the work submitted is my own. My contribution and those of the other authors to this work have been properly indicated within each piece of research. I certify that where references to the work of others have been made in this thesis, proper credit has been given.

Fernando Miguel Delgado Chaves
November 2022

To my father, whom I miss every day.

It doesn't matter if you win or lose. What matters is how you fight between the first bell and the last one. The result of the match is just a piece of news for the public. Who can say you lost if you feel like you've won? Life is like a foot race, Marcus: There will always be people who are faster than you, and there will always be those who are slower than you. What matters, in the end, is how you ran your race.

Joël Dicker,
The Truth About the Harry Quebert Affair.

Acknowledgements

Foremost, I would like to express my special thanks to my directors, Prof. Dr. Francisco A. Gómez Vela and Prof. Dr. Federico Divina, from the Polytechnic School at Pablo de Olavide University, who have offered exhaustive guidance and encouragement throughout the past few years. They were glad to introduce me to the rewarding field of bioinformatics, where I found my real calling and was constantly learning new things. They also gave me my first career opportunities, which were extremely helpful during the difficult pandemic years. Thank you for believing in me and helping me in every way you could.

Second, I would like to thank Pablo de Olavide University as an institution for its support of initiatives like the Grant for Tutored Research, in application of grant modality B.2.-"Predoctoral Bridge Scholarships" of the V Research and Transfer Plan 2018-2020 of the Pablo de Olavide University, according to Official Resolution dated June 7, 2019 (Reference: PPI1903), and the Grants for mobility and improvement of the quality of doctoral theses. Such chances provided me with the means I needed to continue my education over the years.

Third, the completion of this work would not have been possible without the invaluable thoughtful training offered by Prof. Dr. Jan Baumbach and Dr. Olga Zolotareva from the Institute for Computational Systems Biology (CoSy.Bio) at Hamburg University. Not only they have contributed to my education, but also boosted my career by introducing me to fresh ideas and showcasing my value and abilities in ways I could never have imagined. To them and to my other fellow colleagues and friends at CoSy.Bio, my sincerest gratitude.

Then, it would be impossible to miss the talented people I had the opportunity to meet during my time as a doctoral student, who walked this path with me hand in hand, Prof. Dr. Ignacio Luque, Dr. Pedro Martínez García, Dr. Aurelio López Fernández, Dr. Miguel García Torres, Dr. Domingo S. Rodríguez Baena and Prof. Dr. Alicia Troncoso Lora. My undying appreciation to my colleagues at the

Andalusian Centre for Developmental Biology and the IT Department of Pablo de Olavide University.

Finally, the development of this work could not have been possible without the support of my family, who have kept me motivated and encouraged me to pursue my dreams. In particular, I would like to thank my father and mother, Fernando and Rufina, my brother, Diego, my grandmother, Angelita and my uncle, Miguel. Of course, I have to express my gratitude to my friends for traveling with me and supporting me through the worst times. Even though they may not fully grasp what it means to be a predoctoral researcher they are always supportive, always caring and trust me along every road I take. Your unconditional support has meant everything to me.

I sincerely appreciate having you by my side on this journey.

Abstract

The root causes of disease are still poorly understood. The success of current therapies is limited because persistent diseases are frequently treated based on their symptoms rather than the underlying cause of the disease. Therefore, biomedical research is experiencing a technology-driven shift to data-driven holistic approaches to better characterize the molecular mechanisms causing disease. Using omics data as an input, emerging disciplines like network biology attempt to model the relationships between biomolecules. To this effect, gene co-expression networks arise as a promising tool for deciphering the relationships between genes in large transcriptomic datasets. However, because of their low specificity and high false positive rate, they demonstrate a limited capacity to retrieve the disrupted mechanisms that lead to disease onset, progression, and maintenance.

Within the context of statistical modeling, we dove deeper into the reconstruction of gene co-expression networks with the specific goal of discovering disease-specific features directly from expression data. Using ensemble techniques, which combine the results of various metrics, we were able to more precisely capture biologically significant relationships between genes. We were able to find *de novo* potential disease-specific features with the help of prior biological knowledge and the development of new network inference techniques.

Through our different approaches, we analyzed large gene sets across multiple samples and used gene expression as a surrogate marker for the inherent biological processes, reconstructing robust gene co-expression networks that are simple to explore. By mining disease-specific gene co-expression networks we come up with a useful framework for identifying new omics-phenotype associations from conditional expression datasets. In this sense, understanding diseases from the perspective of biological network perturbations will improve personalized medicine, impacting rational biomarker discovery, patient stratification and drug design, and ultimately leading to more targeted therapies.

Resumen

Las causas de las enfermedades siguen siendo poco conocidas. El limitado éxito de las terapias actuales se debe a que als enfermedades persistentes suelen tratarse basándose en sus síntomas y no en sus causas subyacentes. Por ello, la investigación biomédica está experimentando un cambio impulsado por la tecnología hacia enfoques holísticos basados en datos para caracterizar mejor los mecanismos moleculares que causan las enfermedades. Utilizando datos ómicos, disciplinas emergentes como la biología de redes intentan modelar las relaciones entre biomoléculas. En este sentido, las redes génicas de coexpresión surgen como una herramienta prometedora para descifrar las relaciones entre genes en grandes conjuntos de datos transcriptómicos. Sin embargo, debido a su baja especificidad y a la elevada tasa de falsos positivos, demuestran una capacidad limitada para identificar los mecanismos alterados que conducen al inicio, la progresión y el mantenimiento de las enfermedades.

En el contexto del modelado estadístico, nos adentramos en la reconstrucción de las redes de coexpresión de genes con el objetivo específico de descubrir características específicas de las enfermedades directamente a partir de los datos de expresión. Utilizando técnicas *ensemble*, que combinan los resultados de varias métricas, se logra capturar con mayor precisión las relaciones biológicamente significativas entre los genes. Fuimos capaces de encontrar *de novo* potenciales características específicas de la enfermedad incorporando conocimiento biológico previo y desarrollando de nuevas técnicas de inferencia de redes.

A través de nuestros diferentes enfoques, analizamos grandes conjuntos de genes en múltiples muestras y utilizamos la expresión génica como un marcador de los procesos biológicos inherentes, reconstruyendo redes génica de co-expresión robustas y sencillas de explorar. Mediante el minado de redes génicas de coexpresión específicas de una enfermedad, se obtiene un marco útil para identificar nuevas asociaciones ómica-fenotipo a partir de conjuntos de datos de expresión condicional. En este sentido, comprender las enfermedades desde la perspectiva de las perturbaciones de las redes biológicas mejorará la medicina personalizada, repercutiendo en el descubrimiento racional de biomarcadores, la estratificación de pacientes y el diseño de fármacos, y última instancia conduciendo a terapias más específicas.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

A	Adjacency matrix
a	Adjacency matrix position
B_e	Edge betweenness
B_v	Node betweenness
C_i	Clustering coefficient of a node
d	Shortest path between a pair of nodes
E	Set of edges
e	Edge
k	Node degree
l_i	Number of links between node v_i and its k_i neighbors.
M	Number of edges
N	Number of nodes
Q	Modularity
V	Set of nodes
v	Node

Greek Symbols

δ	Number of shortest paths between two nodes

λ Degree exponent

Superscripts

j superscript index

Subscripts

i subscript index

j subscript index

Acronyms / Abbreviations

BN Biological network

CC Correlation Coefficient

CNV Copy number variations

DMR Differentially methylated region

DNA Deoxyribonucleic acid

EnGNet Ensemble and Greedy Networks algorithm

GBA Guilt by association

GCN Gene co-expression network

GEO Gene expression omnibus

GO Gene Ontology

GRN Gene regulatory network

indel Insertion/deletion

KCC Kendall correlation coefficient

MI Mutual Information

MS Mass spectrometry

PCC Pearson correlation coefficient

PC Partial Correlation

PLC Pathway-Level Co-expression

PPI Protein-protein interaction

RNA-Seq RNA sequencing

SB Systems biology

SCC Spearman correlation coefficient

SNP Single nucleotide polymorphism

SRA Sequence read archive

Y2H yeast two-hybrid

Part I

Introduction

Chapter 1

Justification and objectives

BIOMEDICAL research seeks strategies to avert and treat disease. Far from being considered an isolated field, biomedical research involves the cross-talk between many scientific disciplines such as physics, chemistry, biology and, increasingly important, computer science. In fact, in order to provide effective treatments, researchers use new technologies to examine biological processes and disorders, carefully experimenting as part of the scientific progress that is biomedical research.

1.1 A shift in thinking in the field of biomedicine

It takes thorough scientific testing, development, and assessment to find novel drugs and treatments. Identifying the best therapy for a specific disease with any variability in patient features or outcomes is the main scientific goal in the majority of clinical trials. Contrarily, the primary tenet of precision medicine is the individualization of patient care, commonly referred to as "personalized medicine". Precision medicine has so far produced fresh approaches and insights in causal inference, clinical trial design, and machine learning applied to biomedical data. The use of biomarkers to guide treatment decisions together with demographic and physiological data, co-morbid disorders, patient lifestyle, and other factors have all benefited from this growing field [1, 2].

However, precision medicine requires considerable scientific and technological advancements in the areas of infrastructure, engineering, project management, and financial management, in a process that has already started to change the way biomedical research is carried out. For instance, it was thanks to the development

of whole genome sequencing, which determines (nearly) all of the DNA sequence of an organism's genome at once, that the human genome project was successfully completed in 2001 [3]. Such detailed map of the human genome has been used to develop new diagnostic and therapeutic tools. The creation of new methods to investigate novel data space dimensions is a core tenet of modern biology, which states that tackling challenging biological issues needs such synergistic development. Positively, as high-throughput screening methods become more affordable, more and more research organizations are able to take advantage of the latest developments in biomedical research [4].

In the past, reductionist methods were mostly used to study biological systems, in which each experiment usually focused on individual biomolecules or mechanisms. With the aid of modern technology, it is now possible to screen hundreds of molecules at once and estimate how they interact with one another. We can now affirm that a massive shift in biomedical research has occurred, with a new paradigm of data-driven biomedical science complementing hypothesis-driven discoveries. For instance, aside from the aforementioned drop of genome sequencing costs, healthcare organizations have begun embracing the information from digitalized clinical records and imaging data, integrating the data to propose new hypothesis. An inherent issue is that such data is frequently diverse, multi-spectral, incomplete, and inaccurate, which demands advanced data modeling and representation, processing power and algorithmic optimization for data-intensive analytics [5].

Healthcare companies may use data-driven techniques to detect patterns of care, evaluate unstructured data, and develop systems for decision support, prediction, and traceability, among others. Nevertheless, the speed of data collection contrasts with the slower functional characterization of biological data, highlighting the growing disparity between the two types of information. In this context, it is increasingly crucial to integrate biological data, alongside the clinical data of specific individuals from electronic health records. Given the importance of biological data safety due to the general data protection regulation, Big Data analytic solutions for personalized medicine must be distinguished by effective and secure models for data-driven discovery, integration, storage, and interpretation [6].

To go through such massive volumes of data and make sense of them, varied teams are needed due to the complexity of biology in the Big Data era. First, we need to develop new computational approaches because of the great advances in the automated collection of huge volumes of molecular and clinical data [6]. In fact, it is estimated that over the next ten years, the use of genetic data alone would

surpass that of other relevant Big Data sectors like that of astronomy [7]. Second, new data formats may require using new analytical techniques, which are often designed *ad hoc*. Third, the creation of integrative tools for diverse clinical data processing in conjunction with genetic and functional annotations is crucial. Digital health records' organized content and unstructured material (like descriptions of symptoms) can also be included, as these records are an important source for research and model building. Furthermore, owing to new technology that makes data faster and more effective, researchers may re-analyze publicly available datasets, a method that frequently reveals previously overlooked information. As a result of their complementing skill sets, varied groups of researchers can more quickly obtain shared discoveries and improve current understanding of biological challenges.

The fact that tons of biomedical data are readily available has also shifted the way we do research. Hypothesis-driven research, adhering to the classical scientific method, has dominated research for the larger part of the last century [8]. This approach consists on a query and a hypothesis based on observations, followed by a set of experiments to contrast the hypothesis. Contrarily, in discovery-driven research vast volumes of data are gathered and unbiased conclusions are drawn. The creation of conclusions and rigorous examination of the data are essential components of discovery-driven research. Hence, the hypothesis formulation, before or after analyzing the data, is where there is the greatest difference between the two types of research. Because of their complexity, biological systems cannot be fully investigated by hypothesis-driven research; rather, scientific understanding must be founded on experimental facts rather than preconceived ideas [9]. Nevertheless, discovery-driven research frequently yields a lot of hypotheses that may be explored in hypothesis-driven studies due to the large amount of data produced upon research.

The enormous amount of biological data that is currently available is not only stimulating discovery-driven research but also assisting the elucidation of disease mechanisms. In fact, when looking at disease definitions one may realize we barely understand the cause of diseases [11]. For instance, the human body has traditionally been divided in organs, so many diseases are defined by the symptoms and signs observed in each organ, e.g. heart failure, osteoporosis or dermatitis. In other instances, diseases are named after the clinician who originally characterized them, as in Parkinson's, Huntington's, or Alzheimer's. However, such classic definitions are insufficient for understanding diseases, which only

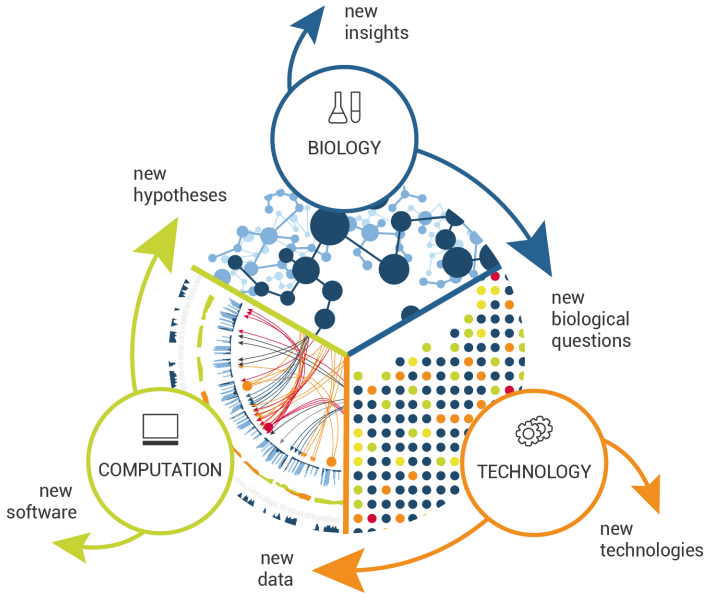


Fig. 1.1 Illustration on the modern view of biomedical research, which often combines hypothesis and data-driven research, together with the design of the appropriate tools for analyzing biological data. Extracted from [10].

become apparent when symptoms appear. Then, we simply give the condition this symptom's name because we do not yet understand what causes it. As a consequence, the molecular interactions that cause disease onset and progression are often not taken into account by current therapeutic practice, which instead focuses on treating symptoms. An illustrative example of this is given in the book *The end of medicine as we know it - and why your health has a future* by H. W. Schmidt [11]: an observable symptom like increased blood pressure is managed with an anti-hypertensive medication to reduce the risk of heart failure and enable the pressure return to a normal range. However, since we are unable to identify a reason for such hypertension in about 90–95% of instances, we cannot be certain if this will treat the problem or even be beneficial to the patient.

The dilemma that follows is whether we should keep treating patients chronically despite the lack of actual evidence for a disease's cause. By "cause", one means the precise molecular mechanism that underlie the symptoms. This means, the precise understanding of which genes, proteins, messengers, hormones, or signaling pathways in our organism are disrupted so that the symptoms appear, and also long-term tragic consequences like a heart attack or stroke. Finding driver genes and molecular markers for complex diseases has become possible thanks to the advancements in technology and the integration of biological data of diverse nature, which have improved our understanding of the molecular etiology

of disease [12]. There is only hope for a treatment if we can identify the disease causes. Because of this, such data-driven research, preferably agnostic of previous conceptions, may be able to offer more accurate prognoses and tailored therapies.

For these reasons, the advancement of personalized medicine is a top priority for both society and governments, requiring broad cooperation, common knowledge and decentralized administration [6]. For instance, the European Commission has allocated more than €2.6 billion to personalized medicine through the FP7 (EU Seventh Framework programme) and Horizon 2020 projects. As an example, the BLUEPRINT initiative [13] for the study of epigenetic processes of hematopoiesis is one of the success stories of the International Consortium for Personalized Medicine [2], which was established by the European Commission. Another example is the REPO4EU initiative [14], which uses point-of-care diagnostics for assessing patients and less animal testing to transition from low precision pharmacological treatment to precision therapy. All significant biomedical projects, including the well-known Human Genome Project, are based on the paradigm of consortium-based research, which was created by community movements and was increasingly accepted by both industries and governments [15].

1.2 Objectives

Keeping the aforementioned ideas in mind, I aimed at contributing to the expanding area of precision medicine by reusing, integrating, and mining biological data of various types in a discovery-driven manner. The major framework of this thesis lies in the use of the substantial quantity of biomedical data made available by next generation high-throughput technologies to approach the *de novo* elucidation of disease mechanisms, directly from data.

In the projects we address in this manuscript, we largely focused on the gene level, using expression data such as that obtained through RNA sequencing or microarrays. As covered in Chapter 2, the expression of genes provides a solid foundation for examining potential disease mechanisms, as they are surrogate indicators of protein levels and alterations at the genome level. Within this framework, reconstructing gene networks, in an approach grounded in statistical modeling that aims to uncover relationships between genes based on their expression level across samples, can be used to integrate and interrogate this data holistically. Gene networks, described in detail in Chapters 3 and 4, are a subset of biological networks,

which depict connections between different types of biomolecules by means of a graph.

As a result, the major goal of this thesis is to investigate gene networks for the *in silico* analysis of biomedical gene expression data, with a focus on developing novel algorithms that may retrieve potential disease-associated genes and mechanisms. To do so, we list the following objectives of this thesis:

Objective 1: Reviewing the methods for the reconstruction of gene networks.

In the past several years, more and more methods have been proposed to infer gene networks from expression data. Many methods use a variety of data sources, such as microarray expression data or previous knowledge stored in databases, to infer gene regulatory networks (GRN) and gene co-expression networks (GCN). We intended to provide a framework for the different approaches for retrieving the associations between genes, evaluating their advantages and disadvantages.

Objective 2: Investigating gene expression datasets and their applicability to gene co-expression approaches.

Publicly accessible repositories would be evaluated as a source of disease-related expression data. Remarkably, control samples from healthy patients would also be considered as a contrast group. Special attention would be paid to the structuring of the information to study the feasibility of its incorporation into network inference algorithms.

Objective 3: Exploring the usage of ensemble approaches for the robust reconstruction of gene co-expression networks.

By examining the pros and cons of different reconstruction methods, we aim to identify the best strategy to retrieve interactions between gene pairs based on their expression profiles. In this, we paid especial attention to ensemble techniques, which evaluate the output of multiple metrics and combine them to obtain a refined result.

Objective 4: Designing a new algorithm for gene network reconstruction based on an ensemble approach.

Gene networks of any size and complexity should be able to be reconstructed by the algorithm. Large volumes of data should be

handled by the algorithm, which should also be scalable and effective in identifying key relationships. Furthermore, the algorithm must be simple and accessible to clinicians and life scientists.

Objective 5: Application of gene networks to the exploration of disease mechanisms and potential biomarkers.

We sought to use our approaches with the explicit objective of identifying disease-specific gene modules that could be used to retrieve potential disease mechanisms, which can provide molecular characterization of diseases in a rational way. We would analyze *in silico* possible novel genetic biomarkers inside this that may be applied in clinical practice.

1.3 Scientific articles contributing to the thesis framework

The publications that comprise the core content of this PhD manuscript are listed below in chronological order. We further discuss the value of our findings and contributions in Chapter 9.

- Delgado-Chaves, F. M. & Gómez-Vela, F. (2019). Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial intelligence in medicine*, 95, 133-145.
- Gómez-Vela, F., Delgado-Chaves, F. M., Rodríguez-Baena, D. S., García-Torres, M. & Divina, F. (2019). Ensemble and Greedy Approach for the Reconstruction of Large Gene Co-Expression Networks. *Entropy*, 21(12), 1139.
- Delgado-Chaves, F. M., Gómez-Vela, F., García-Torres, M., Divina, F. & Vázquez Noguera, J. L. (2019). Computational Inference of Gene Co-Expression Networks for the identification of Lung Carcinoma Biomarkers: An Ensemble Approach. *Genes*, 10(12), 962.
- Delgado-Chaves, F. M., Gómez-Vela, F., Divina, F., García-Torres, M. & Rodriguez-Baena, D. S. (2020). Computational analysis of the global effects of Ly6E in the immune response to coronavirus infection using gene networks. *Genes*, 11(7), 831.

As a starting point, in Delgado-Chaves and Gómez-Vela [16], we conducted a comprehensive review of methods for inferring gene networks from expression data. This work paved the way for subsequent experimental planning, with particular emphasis on information theory and correlation-based methods. A number of key issues in the subject are covered, including: (i) the sort of data required for network reconstruction, (ii) a method-based classification of the main available tools, (iii) model optimization, and (iv) computational methods for result validation. With this, we intend to cover Objectives 1 and 2.

In Delgado-Chaves et al. [17], we combined differential expression and gene co-expression analysis to infer robust GCNs using an ensemble technique with three concomitant similarity metrics. We applied our method to the analysis of a lung cancer dataset. Due to late detection, lung carcinoma, one of the most prevalent cancer types, has a limited life expectancy. Because of this, easy-to-measure lung cancer biomarkers are in great demand in biomedical research. We used microarray data to define disease modules for lung cancer in the reconstructed GCNs, in order to suggest new biomarkers. The genes *NCKAP1L* and *DMD* are emphasized among possible biomarkers because of their relevance to a sizeable part of lung and bronchus primary carcinomas. These results show how our ensemble approach for GCN reconstruction may be used to predict biomarkers in an exploratory way. Consequently, we focused on Objectives 3 and 5.

In a similar spirit, in Gómez-Vela et al. [18], we address two limitations of GCNs: (i) the inability of some reconstruction methods to identify nonlinear dependencies and (ii) sparsity and scale-free topology, properties that many methods fail to achieve. We proposed a novel two-step method, *EnGNet*, which uses an *ensemble* strategy that combines linear and nonlinear measures for GCN generation and then performs topological optimization using a greedy algorithm. Not only is *EnGNet* competitive in terms of the accuracy of the networks when tested on well-characterized datasets, but also it improves their topological characteristics, providing a useful tool for non-specialist end users. A human dataset on post-traumatic stress disorder was applied to demonstrate the method's performance in a real biological context, and the results showed an innate immunity-mediated response to this disease. These outcomes show the method's potential for the identification and characterization of biomarkers. Therefore, we addressed Objectives 3, 4 and 5.

The application of GCNs in the field of biomedicine has provided relevant findings for understanding the molecular mechanisms underlying complex disease.

GCNs have also unveiled the implications of new genes in biological processes of interest and have also assisted the discovery of new biomarkers, following the principle of guilt by association. Many of the predicted interactions have subsequently been experimentally validated. Hence, more recently, we applied *EnGNet* to transcriptomic data corresponding to a murine model of SARS-CoV-2, exploring the role of the *Ly6E* gene in the immune response to coronavirus [19] infection. Understanding host-pathogen pathways and the immune response to them, particularly in viral infections, is seen as a key objective for the logical design of effective therapeutics. The use of gene networks in the SARS-CoV-2 pandemic scenario may possibly boost therapy-related research, organize experimental scrutiny, and lower expenses. In order to examine the time-resolved impact of gene *Ly6E* in the immune response to the SARS-CoV-2-model coronavirus that causes murine hepatitis (MHV), gene co-expression networks were created in this study utilizing RNA-Seq expression data. As a result, we focused on Objective 5.

Part II

Theoretical framework

Chapter 2

Systems biology in biomedical research, a brief introduction

As previously stated, advances in biological data collection are clearly improving biomedical research. New high-throughput technologies, such as next-generation sequencing, have generated an enormous amount of valuable biological information that is very difficult to process using traditional techniques [20, 21]. Complete genomic analysis, transcriptome profiling, and proteomics studies that characterize different types of biomolecules in great detail are examples of studies that link molecular elements with phenotypes at the cell, tissue, or organ level [22].

2.1 The *omizing* complexity of modern biology

Although we focus on biomedical research, technological development has had such a dramatic impact on our understanding of biological systems that the descriptive suffix "omics" was created to distinguish a novel collection of knowledge fields. The major focus of this thesis will be on these so-called omics data, which comprehensively define the varied aspects of biological complexity. In the matter at hand, the thorough analysis of omics data, which can provide mechanistic understanding of biological systems disrupted by disease, is thus one of the trendiest issues in Big Data [23, 24].

In the following sections, we will cover how, in complex diseases like cancer or neurodegenerative disorders, an organism's genotypes and phenotypes participate

in an elaborate network of interactions involving multiple biological pathways, rather than just correlating with one another [25]. Prior to doing so, in this section we hope to provide the essential foundation for understanding the data, a necessary first step before any further computational approach, by determining what and how different features are quantified and what kind of information these could provide. Figure 2.1 provides a summary of the various omics discussed in this section in relation to their main applications and screening platforms.

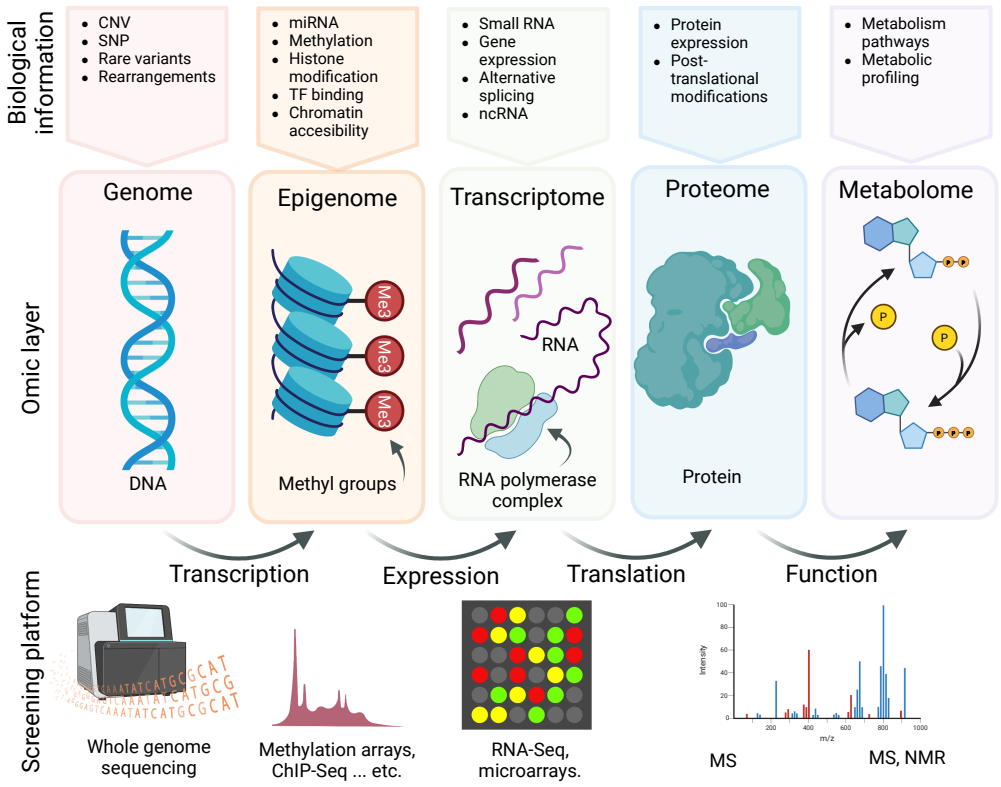


Fig. 2.1 An overview of the relationships between several omic layers, along with the kinds of information they could provide and possible screening platforms. CNV, copy number variation; SNP, single nucleotide polymorphism; miRNA, micro RNA; ncRNA, non-coding RNA; MS, mass spectrometry; NMR, nuclear magnetic resonance.

2.1.1 The genome and epigenome

Starting with the foundations, all of an organism's hereditary features are physically based on the genome of that organism. The genome is the whole genetic composition of an organism and its main attribute is its capacity to communicate information, which is contained in the deoxyribonucleic acid (DNA), composed of chains of four distinct components termed nucleotides. In addition to being

correctly replicated and transmitted, genetic material expresses its information and transforms it into physical or functional characteristics, e.g. from parent to offspring. The development of genetic material changes can be used to describe how new traits evolve.

The process of identifying the nucleotide sequence of a DNA sample is DNA sequencing, which was properly developed as part of the human genome project [26]. DNA sequencing is an effective method for genomics research since sequencing prices are falling rapidly and modern sequencers can generate massive amounts of data. Along these lines, studying complete genomes is possible through whole-genome sequencing. Understandably, clinical medicine has not completely embraced genome sequencing since our ability to sequence human genomes has far outpaced our ability to evaluate and interpret genetic variation.

Whole-genome sequencing can be used to gather and identify genetic variants in an organism, such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs), DNA insertions and deletions (indels), and regulatory sites that impact a certain function or phenotype at the cell or organism level [27]. Some databases with publicly accessible genomic data are *GenBank* [28] and sequence read archive (SRA) [29].

The main objective of genomics in biomedical research is to identify genetic markers linked to disease, drug response, and patient prognosis. For instance, DNA sequencing is progressively changing how we think about cancer, which is fundamentally a disease of the genome. One way DNA sequencing can aid cancer treatment is by suggesting targeted therapies based on mutations detected in a specific tumor, or by sequencing the DNA from circulating tumor cells, allowing for non-invasive diagnosis and/or monitoring [26]. The tremendous genomic heterogeneity of tumors has been shown by large-scale resequencing, which has successfully defined a molecular taxonomy for cancer [30].

There are other means outside the DNA sequence for encoding phenotypic information. Epigenetics is the field that studies inherited modifications of phenotypic characteristics or genomic activity without altering the DNA sequence [31]. Epigenetic marks are able to control DNA transcription and information passing to biological functions. Such epigenetic marking happens in a variety of ways, including nuclear structure reorganization, DNA and histone modifications, histone variations, and nuclear RNA.

DNA methylation is the best-studied epigenetic modification, and it involves the addition of methyl groups to DNA. In mammals, methylation occurs at cytosine-

guanosine dinucleotides, termed CpG sites. Repeated sections in the genome typically experience CpG methylation, which is related to the inhibition of transcriptional activity and the immobilization of transposable elements [32]. High-throughput detection and quantification of the methylation level at CpG sites is possible through three main approaches: DNA differential enzymatic cleavage, affinity capture of methylated DNA, and sodium bisulfite conversion and sequencing [33].

DNA methylation is a well-known factor in cell proliferation and differentiation. Differentially methylated regions (DMRs), which are genomic areas with varying levels of methylation across a set of samples (tissues, cells, people, etc.), are thought to have potential functional roles in the control of gene transcription [34]. In this sense, DMRs show abnormal methylation in tumor samples compared to normal samples. Furthermore, there are also intra-individual DMRs with aging-related alterations in global DNA methylation.

2.1.2 The transcriptome

Inside a single organism, and even inside a single cell, there are a myriad of different forms, functions, and activities that exhibit an incredible level of specialization. The information in the DNA is expressed through the RNA, an intermediary similar carrier that accomplishes such specialization.

Each RNA molecule copies a little portion of the DNA. When a gene is expressed or is in its active state, many copies of RNA, or transcripts corresponding to that gene are created. This is known as transcription, and it is mostly performed by the enzyme RNA polymerase. Since the information coded in RNA may be used to translate DNA into proteins, under some circumstances, RNA can be seen as a transitional molecule between proteins and DNA. In keeping with the original name genome, we refer to the entire collection of RNA transcripts found in a biological system as the transcriptome, and the study of transcriptomics [35].

We can evaluate gene expression by measuring the different RNA levels. Because RNA is easily degraded, it must be turned into the more stable cDNA form. In transcriptomics, RNA levels are assessed across the genome both qualitatively (transcript presence, identification of splice sites), as well as quantitatively (transcript expression value) [36], and the two main techniques are based on hybridization, microarray chips, and sequencing, RNA-sequencing (RNA-Seq).

Originally, microarrays were a key tool for determining RNA levels. These are based on sequence complementarity so the sequences in the sample to be

analyzed will bind a set of known sequences whose presence is to be detected, which are termed probes. Restrictions endonucleases fragment the unidentified DNA molecules, which are then labeled with fluorescent dyes and given a chance to interact with DNA probes arranged on a chip. Non-binding DNA fragments are removed. When a laser beam passes over the target (bound) DNA fragments, its fluorescence emission may be used to identify them. The arrangement of fluorescence intensity and DNA identification are recorded using a computer. This method of using DNA chips is particularly quick, specific, and sensitive for concurrently identifying several DNA fragments [37].

In RNA-Seq, cDNA is fragmented and added adapters that allow their sequencing. The cDNA library is then subjected to amplification, size selection, clean-up, and quality-checking steps before being sent for sequencing analysis. This leads to the generation of short sequences (reads) corresponding to all or a portion of the fragment from which they were derived. By the end of the procedure, there will be millions of reads, which may either be aligned to a reference genome or assembled from scratch to create an RNA sequence map that covers the transcriptome.

There has been a move from microarray chips to sequencing techniques as a consequence of the sharp drop in sequencing costs in recent years. Yet, transcriptional profiling still employs RNA-seq and microarrays [38]. Nevertheless, a benefit of RNA-Seq is that it also records data regarding alternative splicing processes, which result in various transcripts from the same gene sequence. DNA sequencing wouldn't detect these situations. Additionally, it can detect post-transcriptional alterations such as 5′ capping and polyadenylation, which take place during mRNA processing. RNA-seq experiments also do not suffer from cross-hybridization or poor hybridization problems, which can be a problem in microarray research. Additionally, experimental noise may be easily reduced since the cDNA sequences utilized in RNA-seq are mapped to specific locations on the genome. RNA-seq data may be quantified, but microarray data is exclusively presented as relative values to other signals found on the array. The difficulties microarrays have in identifying extremely high or extremely low transcription levels are also avoided by RNA-seq.

If we are to link the information in our DNA with the production of functional proteins, understanding the transcriptome is essential. Transcriptome profiling can determine which genes are active in a sample, their degree of transcription, and the timings at which they are switched on or off [38]. RNA assessment assists in the study of the numerous mechanisms underlying altered gene expression, which are often related to disease. In this sense, differential expression analysis is a widely

used approach that makes use of transcriptional profiling and provides crucial knowledge about the function of genes. Gene Expression Omnibus (GEO) [39] and *ArrayExpress* [40] are two databases that offer transcriptional profiling open to the public.

2.1.3 The proteome

In line with the previous omics, the term "proteome" refers to the whole set of present proteins, which are large, complex molecules that are the active players in countless biological processes.

Numerous thousands of amino acids, which are the smaller building blocks of proteins, are linked together in long chains to form proteins. Generally, to create a protein, 20 distinct kinds of amino acids can be mixed. Sets of three DNA nucleotides, which are defined by the gene DNA sequence, are used to code for amino acids [41]. Each protein's distinctive 3-dimensional structure, or conformation, which defines the protein's function, is determined by the order of the amino acids. Proteins have exceptional properties that underpin the dynamic activities in living cells.

The field of proteomics involves identifying and analyzing a genome's entire protein signature, although in practice we can only identify a growing number of proteins. A popular, high-throughput method for analyzing proteins is mass spectrometry (MS) [42]. In order to identify proteins using this technique, proteins are digested into peptides and then sorted, fragmented, ionized, and collected by the spectrometer. By using mass spectrometry, we can identify proteins and their post-translational modifications, which are added once a protein is synthesized. Indeed, approximately 10,000 proteins may now be identified thanks to developments in computational protein identification using mass spectrometry data [43]. Additionally, traditional, unbiased techniques such as yeast two-hybrid (Y2H) experiments and phage display are used to identify protein interactions [44].

Proteins in body fluids or tissues may be analyzed using MS-based approaches, both for protein quantification and analysis [45]. Consequently, technologies based on proteomics are used in a variety of ways for various research settings, including the identification diagnostic markers, the development of vaccine candidates, the comprehension of pathogenic mechanisms, the alteration of expression profiles in response to various signals, and the understanding of functional protein pathways upon disease [46]. Proteome profiling data is stored in databases like PRIDE [47], *ProteomeXchange* [48], and *ProteomicsDB* [49].

2.1.4 The metabolome

An entire collection of small-molecule types constitute an organism's metabolome. Metabolite levels reflect metabolic function, and abnormal relative ratios and disruptions outside of the normal range could denote disease.

The small molecules from the metabolome may include both endogenous chemicals and exogenous metabolites, which are substances that are not naturally produced by an organism [50]. Amino acids, organic acids, nucleic acids, fatty acids, amines, sugars, vitamins, and co-factors are only a few examples of endogenous metabolites. In contrast, exogenous chemicals include drugs, food additives, toxins, environmental pollutants, and other xenobiotics.

In order to identify new factors that control the relative ratios of small molecules in plasma and other sample types, quantitative measurements of metabolite levels are employed. Liquid chromatography, MS, and nuclear magnetic resonance (NMR) technologies are used in the extensive study of metabolites [51]. To aid compound separation, the majority of MS procedures use a combination of liquid/gas chromatography, or capillary electrophoresis. Based on the instrument or procedure in use, each approach may generally detect or describe 50–5000 distinct metabolites or metabolic "features" at a time. However, no one analytical approach can currently be used to examine the whole spectrum of metabolites.

Target analysis (screening for known molecules), metabolite profiling and fingerprinting are some of the main techniques in metabolomics [52]. A sample's mass profile or metabolic "signature" is created and compared to a large sample set to look for variations across the samples. Since metabolomics may be used for a wide range of purposes, such as phenotyping, determining gene function, and monitoring response to stimuli, may be considered as bridging the gap between genotype and phenotype. Metabolomic data sets and associated metadata may be analyzed and mined to discover novel disease-associated sets of features and targets. Since metabolomics is a young field, there are proportionally less databases on metabolome profiling; some of which are *MetaboLights* [53] and the human metabolome database [54], together with some more general-purpose ones like *Reactome* [55] and the Kyoto encyclopedia of genes and genomes (KEGG) [56].

2.2 Systems biology, towards the holistic understanding of omics

The different omics represent different integral approaches to the study of biological entities (genes, transcripts, proteins, metabolites, etc.), which ultimately are laying the foundation for the progress of Systems Biology, a field that seeks to identify the intricate relationships that occur inside biological systems. Systems biology is an integrative field that uses quantitative reasoning, computational modeling and high-throughput experimental techniques to relate molecular components within a single biological scale (cell, tissue, organ), as well as between different scales, to physiological functions and phenotypes of the organism.

It is often stated that one would get at least 10 different definitions of Systems Biology by asking only five biomedical researchers to define it. Historically, systems biology began to consider cells, tissues and organs as complex biological systems. The fast improvement of genetics and sequencing technologies enabled the creation of massive databases on the underlying ingredients that play a role on these sophisticated systems. It was then demonstrated how interactions between the molecular parts of cells might result in functional behaviors that the individual parts alone are unable to characterize [57]. Systems biology tries then to comprehend the broad picture, whether it be at the level of the organism, tissue, or cell, predicated on the idea that these work together to create a whole that is greater than the sum of its parts. From an epistemological point of view, systems biology is frequently described as a holistic approach because it holds that biological systems and their properties must be studied as a whole, and only with all of its components considered, can we fully understand how the "whole" functions synergistically.

Systems biology can be seen as providing a new and more complete perspective on biomedical research. This is because systems biology integrates molecular biology and biochemistry of molecular components, as well as their interactions and dynamics, to comprehend how organism's functions develop and are controlled, as opposed to classical biology approaches, which primarily use phenomenological approaches to describe the functions of cells, tissues, and organs [58, 22]. Also for this reason, systems biology stands in stark contrast to decades of reductionist biology, which entails separating all the pieces and analyzing each component separately [59].

Still today, Dr. Trey Ideker proposed one of the best, most straightforward descriptions of systems biology nearly 20 years ago. His description may be

summarized as the systematic application of genomic, proteomic, and metabolomic technologies to collect data needed to build models of complicated biological systems and disorders [60]. While Dr. Ideker's concept has inevitably grown to encompass other fields of study, such as microbiomics and epigenetics, it still reflects the fundamental idea of systems biology. The basis for establishing systems biology is to systematically identify the components of biological systems, such as proteins, metabolites, DNA, RNA, etc. [8]. The latter was made possible through the development of screening technologies for these different omics together with their comprehensive analysis to predict how biological systems change over time and under various conditions, which requires the integration of a variety of scientific disciplines, including biology, computer science, engineering, physics, and others.

To decipher the information flow of genes, proteins and other subcellular components of signaling, regulatory and functional pathways that control biological systems, researchers employ a wide range of quantitative experimental and computational methods, which allow understanding the interactions and dynamics within cells, tissues, organs and organisms. Computation is a substantial addition to systems biology when compared to more traditional biological fields like biochemistry and cell biology. As a result from the multiple high-throughput omics screening technologies, large datasets must be analyzed using computational methods to generate sorted lists of molecular entities that may be shown as pathways and networks to deduce their function. To turn the data into knowledge, the information gathered by these system-wide investigations has to be organized and analyzed, often using omics-specific tools that have been developed *ad hoc* [12].

Today, there are various databases that store this data and computational tools to analyze them, including those for genomic characterization, disease-specific SNP profiles, RNA profiles, protein networks, etc. Genome-wide association studies (GWAS), the process of identifying DNA sequence variants, often SNPs, linked with an elevated risk of a particular illness or physiological condition, are an excellent illustration of software-assisted analysis of huge molecular datasets [61].

The systems biology paradigm has recently enabled research in quantitative and systems pharmacology, as well as precision medicine for complex disorders [22]. Along with genome sequencing, experimental platforms for proteome and transcriptome profiling are accelerating the accessibility of biomedical information. Metagenomics, the sequencing of bacterial genomes in human samples, single-cell genome and transcriptome sequencing, liquid biopsies for detecting circulating tumor DNA, and other recent advances in genomics are already having a significant

impact on healthcare and will be used in routine medical procedures. Thus, new Big Data applications include biomarkers and drug discovery, as well as fundamental research in cancer, neurodegenerative and rare diseases, cardiovascular pathologies or diabetes. System biology experts can find novel disease biomarkers thanks to their capacity for designing predictive, multi-scale models, which also enable them to stratify patients according to their particular signature and design targeted therapies. The possibility for completely new ways of research is eventually created by systems biology, which also propels ongoing innovation in biology-based technology and computation.

Chapter 3

Untangling the intricate associations of Life with Network Biology

We have just scratched the surface on the impact that technology development is having on biomedical research, including how it reshaped experimental design and promoted emerging fields like systems biology. We reviewed how the main advantage of systems biology approaches is the holistic analysis of omic layers, which eventually allows characterizing their cross-talk. Especially in the study of complex diseases, the relevance of biological relationships is widely recognized, and systems biology methods have been proven successful.

When we think of biological entities as networks, we can get insights into the intricate relationships between them. This perspective has received a lot of attention recently as a result of advancements in network science and high-throughput technologies. Biological networks (BNs) are effective tools for the comprehensive and integrative study of omics in this context, and their linkage maps between genes, phenotypes, and associated environmental variables are considered crucial to our current understanding of disease [62, 63].

This Chapter outlines the primary biological network methods used to study different omics layers, what these networks represent, and how they are mined. We also cover some of their primary uses in the particular setting of biomedical research.

3.1 From graph theory to biological networks

A network is a collection of nodes that are joined together by edges, and is defined mathematically as a graph. Graph theory is used to study networks' composition and function, since using networks we can organize and combine information at many levels. We may study networks both as computing units and systems, which enables us to understand how they are organized and also what they represent [64]. As classic example, social networks contain nodes that represent individuals and edges that reflect social relationships between them, such as friendship or collaboration.

In a similar spirit, biomolecules can be represented by the network's nodes in systems biology, and their interactions constitute the network's edges. By thinking of cellular systems as a large network of complex relationships, and examining the topology of such network, it is possible to retrieve the functional organization of cells. In this regard, BNs constitute an abstraction of biological relationships by means of a graph composed of nodes and links, where the nodes represent transcripts, proteins, metabolites, etc., and the links represent the relationship between them [65, 66].

A frequently discussed systems biology's idea is the "network of networks", which considers organisms are composed of a variety of networks that interconnect on various scales. [67]. We are, at our core, a network of networks, from our DNA to the chemicals and cells constitute the organs of our body to ourselves in our surroundings. Network biology investigates these networks in order to integrate behaviors, postulate biological functions, and offer spatial and temporal insights into dynamic biological changes. Such models enable a holistic understanding of the dependencies between biological entities, with applications in a wide variety of fields ranging from medicine to nutrition or crop production, among others [68–70].

A key factor in the popularity of BNs is their relatively simple interpretation, which plays a major role when combining different omics layers. Some graph-based multi-omics integration strategies reconstruct composite networks, involving relationships and/or nodes of different omics [71]. In fact, whole-cell interactions and their topologies can theoretically be represented by network-based models, generating a hierarchical structure of a biological system [72]. These topologies are a crucial first step in developing a flexible, multiscale knowledge of the dynamics of cellular systems. They serve as a global map for information transfer throughout cells, tissues, and organs by representing all biological components and their inter-

actions. By locating the main components and network motifs in such networks, it is feasible to look for functional modules, which might help creating new biological hypotheses. Finally, the predictions of the model simulations must be empirically validated in order to meet the assumptions and parameter estimations required to construct dynamic models.

In this setting, networks have quickly emerged as a desirable method for organizing, visualizing, and contextualizing the huge omics data sets in order to get a system- and molecular-level knowledge of biological processes. Emmert-Streib and Dehmer [67] conclude this is due to four main reasons:

1. Networks are able to represent the complex interactions in and across omics levels.

2. Networks provide a mathematical representation that may be used as a model.

3. A network is a type of data structure that may be used in data analysis to derive biological insights using computational and statistical techniques that might not be possible from just looking at the raw data.

4. The process of reconstructing a network itself may help to understand the dynamical processes underlying biological phenomena, such as evolution.

Networks enable us to address issues such as how disease mutations should be interpreted, the impact of mutations on the network model and how these may affect treatment decisions. To illustrate this, below are some examples of biological networks and the information they may supply.

Protein-protein interactions (PPIs) networks: represent the binding or interaction between multiple protein macromolecules. PPIs may be mapped experimentally using a variety of techniques, including Y2H experiments and affinity-purification MS [73]. Typical bioinformatics resources and databases for reconstructing human protein-protein interaction networks include *BioGRID* [74], *HPRD* [75], *STRING* [76] and *Phosphonetworks* [77].

Isoform-isoform networks: represent the relationship between transcript or protein isoforms, generated through alternative transcription, splicing, 3'-end creation, translation, and post-translational modification. Although the most recent human GENCODE project [78] identified over 20,000 protein-coding genes; the human

genome is thought to be capable of producing up to 100,000 different isoform transcripts, which might result in over a million different isoform-isoform interactions. Then, a protein, typically represented in a PPI network as a single node, converts to a sub-network of interactions between the several isoforms produced from a same gene, which directly relates to the network-of-networks idea.

Gene regulatory and gene co-expression networks: Gene regulatory networks (GRNs), are a potent tool for describing the transcriptional control of gene expression [79]. GRNs frequently include the regulatory connections between transcription factors (TF) and their target genes (TGs). On the other hand, gene co-expression (GCNs) networks measure the similarity between genes, so when a strong reliance between two genes' expression profiles is identified, such genes are considered co-expressed. Although conceptually different, many GCN inference methods have been applied to GRN reconstruction and vice versa. We will cover in detail GCNs in Chapter 4, as a central topic of this thesis manuscript.

Sequence Similarity Networks: where the edges refer to the degree of resemblance between amino acid or nucleotide sequences and the nodes respectively depict proteins or genes [80]. The *BLAST* software is the most well-known tool for comparing the sequences of two sequences [81]. Typically, clustering methods are employed on sequence similarity networks to uncover protein families.

Metabolic networks: Metabolic networks, in which nodes are metabolic substrates and products, and edges reflect underlying biochemical reactions, reflect the biochemical and physiological processes that occur within cells. Metabolic networks can incorporate information such as stoichiometric coefficients for biochemical reactions, whether a reaction is reversible, or which enzymes catalyze a given reaction [82]. In order to construct the metabolic network, the KEGG pathway database may be used to collect frequently used metabolic pathways [56].

Drug-target networks: which depict the interaction between pharmaceutical products and their targets, thus containing these two types of nodes. An interaction is formed when a drug binds to a target with a certain binding affinity. There are several free databases with high-quality information for building drug-target interactions, such as *DrugBank* [83]. The human interactome encompasses pharmacological targets, which are linked to both therapeutic and side effect. Each drug

module illustrates how it interacts physiologically and pharmacologically with other drugs [84].

Although this is just the beginning and there is more biomedical data that may be represented in a network, as those in Table 3.1, it is evident how networks are an efficient approach to represent biological data.

Table 3.1 Other examples of networks representing biological data. Adapted from Koutrouli et al. [80].

Network type	**Description**	**Reference works**
Signal Transduction Networks	capture cell-cell signaling, or a sequence of molecular activities inside a cell or from the outside to the inside.	Fabregat et al. [85], Satagopam et al. [86]
Expression Quantitative Trait Loci Network	Useful for condensing the information from genotyping and/or transcriptome analyses that are used to explain the genetic variance of a gene expression phenotype.	Fagny et al. [87]
lncRNA–Protein Interaction Networks	The interactions between lncRNAs and proteins are what give them their activities.	Zhang et al. [88]
Phylogenetic networks	Capture the temporal evolution of the interactions between the organisms.	Huson et al. [89]
Ecological networks	Illustrate the relationships between species in an ecosystem, such as food webs.	Parsana et al. [90]
Epidemiological networks	Networks used to research disease transmission in public health.	Danon et al. [91]

Species interaction networks	Within-species interaction networks quantify associations between individuals and provide information at the species or population level. Between-species interaction networks describe pairwise interactions between species and attempt to determine what factors lead to stability.	Romanuk et al. [92], Croft et al. [93]
Food webs	Feeding interaction, connect organinsms to one another, and networks showing these connections try to address the ecological question regarding the stability of these ties.	Delmas et al. [94]
Disease networks	Created by diseases and the genes that cause them, and the associations between them can be built using databases like the Online Mendelian Inheritance in Man (OMIM).	Goh et al. [95], Goh and Choi [96]
Disease-symptom graphs	Graphs link diseases with symptoms and show how they might progress, making it easier for professionals to quickly follow the more effective medical treatment.	Sonawane et al. [97]
Literature co-occurrence networks	The co-occurrence of bioentities in any text corpus is reflected by these networks.	Pafilis et al. [98]
Knowledge networks	Multi-edge graphs incorporating diverse data and metadata from different sources, e.g. open repositories or biological and literature databases.	Szklarczyk et al. [99], Gioutlakis et al. [100]

3.2 Understanding graph topology

The study of biological network dynamics, function, and their applications are all based on an understanding of network structure, which is the central notion of network science. In this Section, we present a quick overview of the main graph theory terminology that applies to biological networks. These definitions are crucial because they lay the groundwork for the graph-based data analysis techniques in the next sections by clearly defining basic graph classes and graph concepts. In addition, we discuss the key insights about the structural properties of biological networks.

3.2.1 Graph classification

The different kinds of graphs that we can find in biological networks, depend on how we convert real-world biological interactions into a graph problem. Below, we provide a classification of the main types of graphs that are used in network biology, which are summarized in Figure 3.1. A more detailed description on different types of graphs that are commonly found in Systems Biology can be found in Emmert-Streib and Dehmer [67], and, in a broader context, in Kunegis [101]. It should be noted that a graph can include any combination of the following graph types.

Weighted vs. unweighted: We refer to a graph as a weighted graph (Figure 3.1, top left), when we can apply specific weights or attributes to its edges. As an illustration, in a PPI network we may weigh the edges of the graph using affinity or intensity of the protein interaction to create a weighted graph. Unweighted graphs are the ones where the edges can be considered analogous because they have no weight.

Signed vs. unsigned: In a signed graph, each edge is assigned a positive or negative sign 3.1, top right). As an example, gene co-expression networks model the relationship of gene pairs whose expression level is similar (positive-signed edges) or diametrically different (negative-signed edges).

Directed vs. undirected: As shown in Figure 3.1 (center left), the edges of a graph can either be directed or undirected. For instance, in PPI networks, if two nodes (proteins) are connected by an edge, it often indicates that these proteins interact,

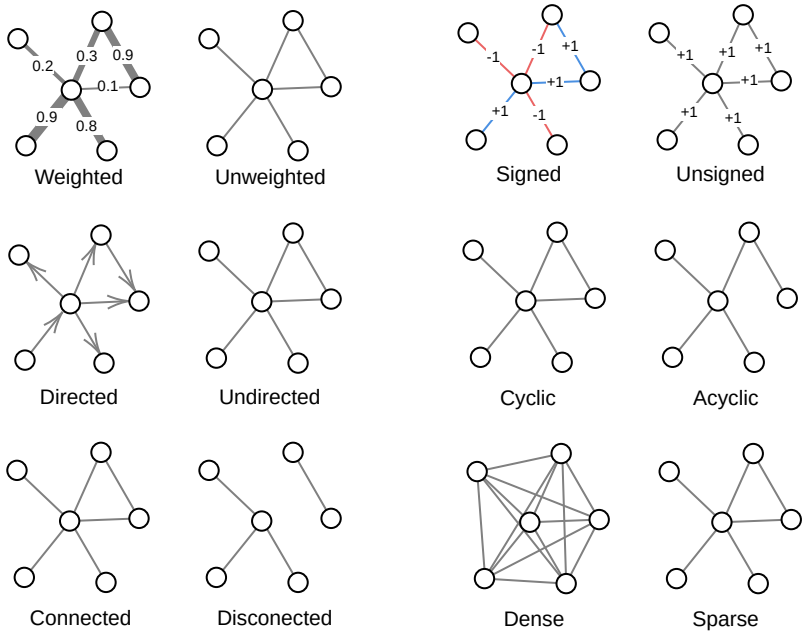


Fig. 3.1 Different types of graphs.

meaning that protein A binds protein B or vice versa. Contrarily in a directed graph, the direction denotes the type of relationship present, e.g. gene regulatory networks are directed graphs where edges may show the regulatory direction from transcription factors to their target genes.

Cyclic vs. acyclic: A graph is referred to as cyclic if it contains cycles (Figure 3.1 center right). If we start at a node and travel through a number of nodes before returning to the original node, we may claim that the graph has a cycle. The term "acyclic graph" refers to a graph in which there are no cycles.

Connected vs. disconnected: In a connected graph (Fig. 3.1 bottom left), there is a path to every node, so all nodes are reachable from every other node. However, one cannot access every node in a disconnected graph if we start at any single node.

Dense vs. sparse: A graph is described as dense or sparse depending on whether the number of edges approaches the total number of potential edges (Fig. 3.1 bottom right). In an undirected graph with N as the total number of nodes, the total number of potential edges M is given by $M = \frac{N \times (N-1)}{2}$. An undirected graph with five nodes, for instance, will have a total of 10 possible edges. Now, the graph

with 5 nodes and 4 edges could be considered sparse, while the graph with 5 nodes and 8 edges could be considered dense.

3.2.2 Graph structural properties

For ease of reference in following subsections, we refer to the biological network $G = (V, E)$, where V is the collection of nodes and E represents the set edges between these nodes. Then, $v \in V$ would represent each node or vertex in the set of nodes and $e \in E$ would be each edge in the set of edges. We denote N and M are the number of nodes and edges, respectively.

Aside from the graph structure, we can represent networks by means of an adjacency matrix or an edge list.

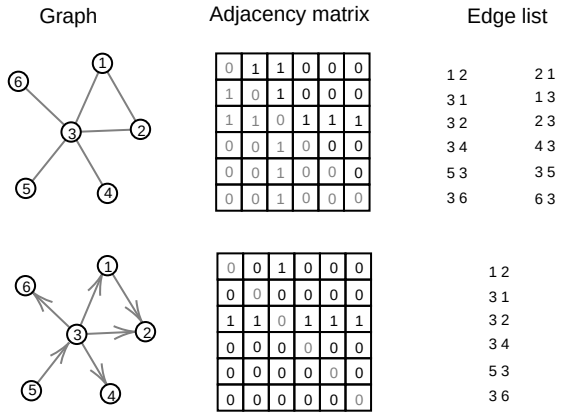


Fig. 3.2 Different ways to represent directed and undirected graphs. In the undirected network, edges could expressed twice in the edge list format to show their bidirectionality.

The network may alternatively be represented as the adjacency matrix A ($N \times N$), where $a_{i,j} = 1$ when the relationship $e_{i,j}$ exists between node v_i and v_j, and $a_{i,j} = 0$ otherwise, as shown in Figure 3.2. For weighted graphs, we may enter the weight of the edge in the corresponding matrix position instead of 0 and 1. Also in this scenario, for undirected graphs the adjacency matrix is symmetrical along the diagonal, which means we may use either the upper or lower triangular submatrix.

If the graph is sparse we might want to use a different data format to represent our graph to save space, since most positions in the adjacency matrix will not be used. For this reason, the edge list is used to tackle the problem of memory waste in the case of a sparse matrix.

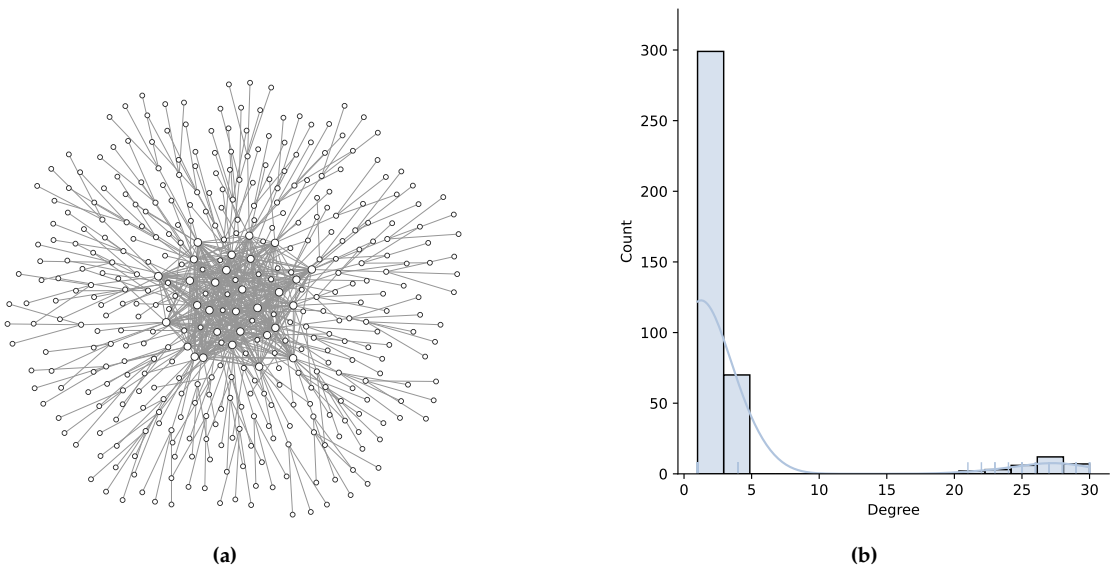
To describe the structural properties of networks, a variety of measurements and definitions have been proposed. These measures assess the fundamental features of networks and are applied, for instance, to characterize network classes, as discussed in Section 3.2.1, or to extract specific network elements, such as nodes or edges, which may be crucial in the processing of molecular information [67].

Node degree and degree distribution

In an undirected network, the number of connections to a node v_i defines its degree k_i. A well-documented network attributes is its degree distribution, i.e. the probability distribution of all node degrees inside a network. In an undirected network, we can calculate the degree distribution of a network as the proportion of nodes having a degree k:

$$P(k) = \frac{\delta_k}{N} \tag{3.1}$$

where, δ_k is the number of nodes in the network with degree k. Hence the probability of choosing a node with degree k is $P(k)$ [67]. The degree distribution of a directed network would be more intricate than that of an undirected network due to the various patterns of in-degree and out-degree. Many biological networks have the scale-free property, which means that the node degree follows a power-law distribution $P(k) \sim k^{-\lambda}$, where λ is the degree exponent [102, 103].



(a) (b)

Fig. 3.3 (a) A network with a topology that is close to scale-free. (b) The network's degree distribution, which resembles a power law.

Clustering coefficient

Many networks have triangular structures; for instance, in social networks, two people are more likely to know one another if they share a friend. A similar case occurs with PPIs, where proteins bind together participating in a certain biological process. The clustering coefficient of a node C_i can be used to characterize this feature:

$$C_i = \frac{2l_i(v_i, n_i)}{n_i(n_i - 1)} \tag{3.2}$$

where l_i is the number of links between node v_i and the n_i neighbors of node v_i, and $n_i(n_i - 1)/2$ is the total number of potential connections between the n_i neighbors of node i, i.e. number of the possible triangles that go through v_i. The majority of biological networks have an average clustering coefficient that is much higher than that of random networks with a comparable degree distribution. Therefore, a strong node clustering coefficient is thought to be a general feature of biological networks [84].

Path-based metrics

The shortest path between two network nodes is critical to the network's structure and behavior [84]. The distance $d(v_i, v_j)$ for a pair of nodes is the length of the shortest path for the pair of nodes $(i, j) \in V$ [67]. Connecting pairs of nodes by shortest pathways is strongly motivated in the setting of biological networks. The shortest path approach is commonly used to assign orientations to edges in protein-protein interaction networks, infer regulatory pathways via matching genes, and discover cancer-related genes or critical components.

In Guney et al. [104] various distance measures are described to to evaluate the proximity between (two) sets of nodes, A and B, including the closest (d_c), shortest (d_s), kernel (d_k), center (d_{cc}), and separation (d_{ss}) measures, whose mathematical formulation is shown below:

$$
\begin{aligned}
d_c(A,B) &= \frac{1}{\|B\|} \sum_{b \in B} \min_{a \in A} d(v_a, v_b) \\
d_s(A,B) &= \frac{1}{\|B\|} \sum_{b \in B} \frac{1}{\|A\|} \sum_{a \in A} d(v_a, v_b) \\
d_k(A,B) &= \frac{-1}{\|B\|} \sum_{b \in B} \ln \sum_{a \in A} \frac{e^{-[d(v_a,v_b)\ +1]}}{\|A\|} \\
d_{cc}(A,B) &= \frac{-1}{\|B\|} d_{center\ a,b} \\
d_{cc}(A,B) &= \frac{\|B\|\ d_c(A,B)\ +\ \|A\|\ d_c(A,B)}{\|B\|\ +\ \|A\|} - \frac{d'_c(A,A)\ +\ d'_c(B,B)}{2}
\end{aligned}
\tag{3.3}
$$

where $a \in A$ is the set of nodes for which to start in calculating the distance, $b \in B$ are the sets of nodes to which the path will arrive, $d_{a,b}$ is the distance between nodes a and b, $\|\|$ the number of nodes in set, *center* denotes the topological center of a set, which for instance for set A is defined as $center_A = argmin_{u \in A} \sum d_{s,u}$, and d'_c a modified d_c in which the shortest path between two nodes is infinite. In the original article by Guney et al. [104], such distances are used to calculate drug-disease proximity, which quantifies the interaction between drug targets and diseases.

Moreover, it is a trait shared by practically all types of complex networks that each node pair does not present a great average distance, i.e. may be connected using only a few links. Such attribute is known as the "small world" property, initially explored in sociological research as the observation of "six degrees of separation". It was discovered that, in comparison to the size of a network, the average route lengths and diameters of several biological networks are rather modest.

3.2.3 Network centrality metrics

Identifying the relevant nodes in a biological network is an important task because of the critical functions these nodes may have. To identify vertices that may behave differently in terms of communication inside these networks, so-called centrality

metrics have been established. Such metrics provide a value to each node in the network so we can rank their importance. Different network-centrality metrics describe different aspects of such node relevance. We provide the key metrics for measuring network centrality below. Except as otherwise stated, we assume undirected networks.

Degree centrality metrics

A node with a high degree has a considerable amount of power since it can influence a lot of other nodes through its connections. Degree centrality is the most basic yet extensively used metric for determining a node's significance. In an undirected network, it is easily computed as $DC(i) = k_i$ and it is assumed that a node with a high degree centrality plays a key role. In the case of directed networks, the degree centrality of a node is split in in-degree, number of edges that reach a node, and out-degree, number of edges that exit a node. Despite its simplicity and low processing complexity, degree centrality works admirably in several areas.

As shown in Subsection 3.2.2, biological networks tend to have a power law-like degree distribution, which means there are few nodes with a high degree centrality [72, 105, 106]. Such nodes are termed hubs, as illustrated in Figures 3.4 and 3.5a. In biological networks, one reason why hubs are originated is gene duplication, a mechanism in the evolutionary divergence processes giving rise to genes that produce similar proteins. For instance, in the context of PPI such proteins may have the same interaction partners than the original protein from which they derive, so the degree of the protein associated with a duplicated protein grows [107]. The importance of hubs in biological systems is highlighted by empirical and modeling evidence; for example, hubs in the PPI develop slowly and are often required for cell survival, as described by the centrality-lethality rule [108]. Furthermore, hub failure has a significant influence on network topology, producing a dramatic rewire that could be responsible for disease.

Degree centrality evaluates just a node's near neighbors while disregarding its network location. However, a node's position is a very effective indicator for depicting the node's relevance. An example of this is that a node in the network's core has a greater effect than one in the network's perimeter. To address this, the k-core decomposition (or k-shell) technique decomposes the network to determine the coreness of each node [109]. A node with a higher coreness is more centrally placed and has a greater influence on network propagation than high-degree nodes with lower coreness. This decomposition is done iteratively based on the remaining

degree of the nodes. The weighted degree of node i is defined in weighted k-shell decomposition as:

$$k_i^w = \alpha k_i + (1-\alpha) \sum_{j \in \Gamma i} w_{i,j} \tag{3.4}$$

where $w_{i,j}$ is the weight of the edge between nodes i and j, $\alpha \in [0,1]$ is a user-defined parameter and Γi is the set of nodes that link to v_i. Another well-known degree-based centrality measure is the H-index, which, based on publications and citations, is a frequently used metric to evaluate the academic contributions of researchers or journals.

Path-based centrality metrics

Betweenness centrality is a path-based metric of node relevance described by the number of shortest paths passing through a node v, and it is defined as:

$$B_v = \sum_{i,j,v \in V, i \neq j} \frac{\delta_{i,j}(v)}{\delta_{i,j}} \tag{3.5}$$

where the number of shortest paths between vertices v_i and v_j is given by $\delta_{i,j}$, and the number of these shortest paths that go through node v is given by $\delta_{i,j}(v)$. Hence, B_v denotes the appearance of node v on all shortest paths of the network, since $\frac{\delta_{i,j}(v)}{\delta_{i,j}}$ is the probability that vertex v lies on a shortest path between nodes v_i and v_j. In scale-free networks, there is usually an abundance of nodes with modest degree and great betweenness. Vertices with high betweenness centrality are termed bottlenecks (Figure 3.5b) and they usually control the information flow in the network.

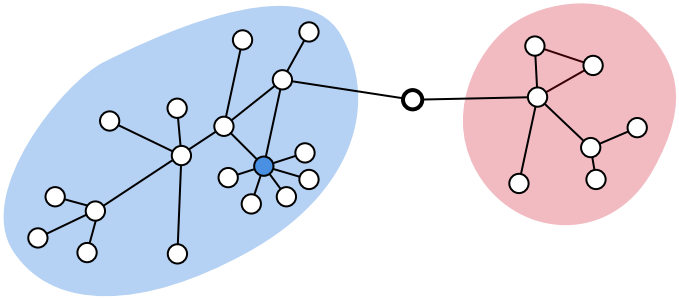


Fig. 3.4 A network having two major communities, circled in blue and red. One node in the blue community has a relative degree that is greater than that of the other nodes. It may be said that the node with the thick border is a bottleneck since it has a high node betweenness.

When the betweenness concept is applied to edges, one can define edge-based betweenness as:

$$B_e = \sum_{i,j \in V, i \neq j, e \in E} \frac{\delta_{i,j}(e)}{\delta_{i,j}} \tag{3.6}$$

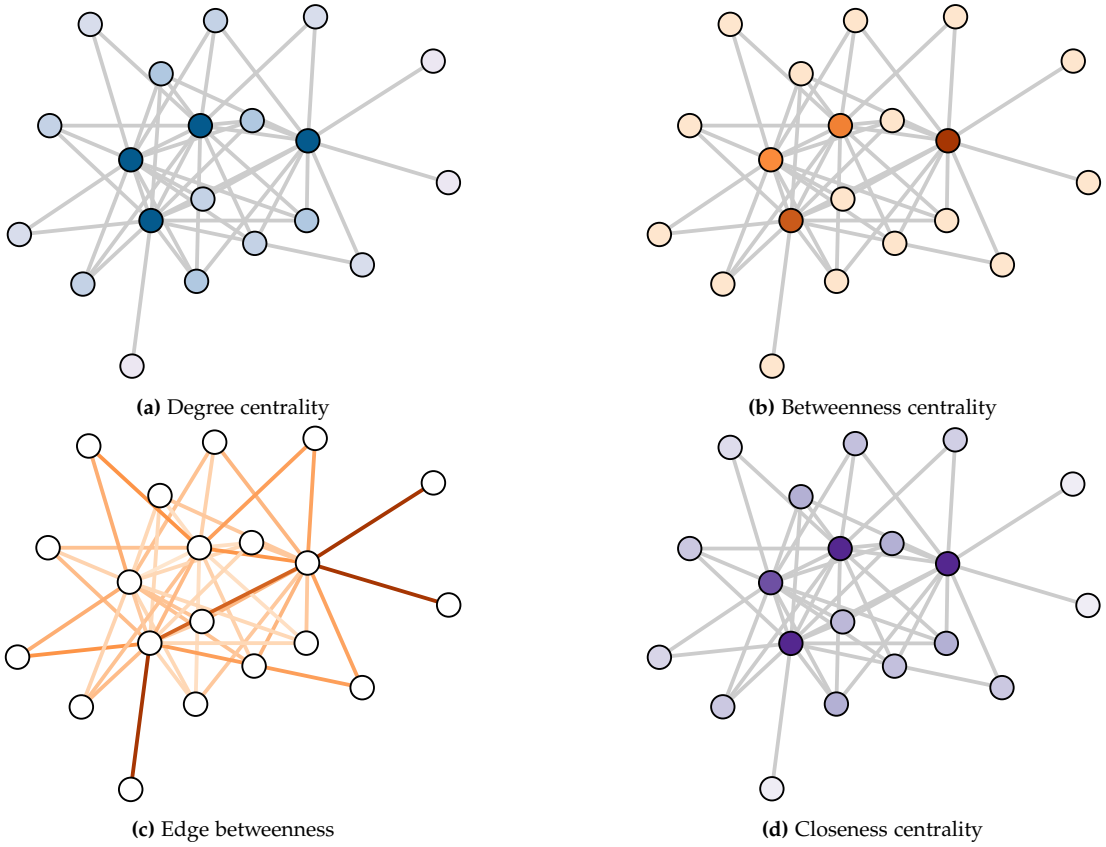
where the number of shortest paths from node v_i to node v_j that go via edge e is given by $\delta_{i,j}(e)$. Edge-betweenness (Figure 3.5c) is an extremely effective clustering method for identifying the community structure.

As previously established, betweenness centrality is the most significant path-based node centrality metric and may be used to assess a node's effect on network information flow. It may be highly useful to interpret the relevance of nodes, particularly bridge nodes that connect distinct communities. However, finding all shortest routes across an entire network is a difficult process, thus obtaining the betweenness centrality for large actual networks is difficult. To deal with this, Closeness centrality (CC), was developed as a path-based centrality that establishes the distance between node v_i and the other nodes in the network. Since being nearer to neighbors generally makes a node more central, closeness centrality may be defined as the average value of the inverse of the distance between v_i and other nodes:

$$CC_i = \frac{1}{\sum_{i=1}^{N} d(v_i, v_j)} \tag{3.7}$$

where $d(v_i, v_j)$ represents the separation between nodes v_i and v_j, and if they are not connected, $\frac{1}{d(i,j)} = 0$. The node with a higher closeness centrality value would be located in a more central location within the network and would have a shorter average information propagation span, reflecting how well it communicates information with other nodes (Figure 3.5d). The closeness centrality for a directed network may be separated into input closeness centrality and output closeness centrality.

There are more complex centrality measures which will not be covered in this Section but for which we refer to Emmert-Streib and Dehmer [67]. Finally, there are two things to think about: (i) each centrality could only be useful in characterizing a certain structural property, and (ii) a centrality's performance across different networks might even differ significantly; e.g. if the networks are random or scale-free.



(a) Degree centrality

(b) Betweenness centrality

(c) Edge betweenness

(d) Closeness centrality

Fig. 3.5 Node rankings according to (a) degree centrality, (b) betweenness centrality and (d) closeness centrality. (c) Edges ranked according to edge betweenness. Color intensity reflects a higher value in each metric.

3.3 Functional organization of biological networks

Biological networks organize in communties or modules. In such a structure, the edges between various communities are scarce and the nodes within the communities are densely linked. It is exciting to learn that communities in network structure are tied to biological processes, so we can identify molecular processes that are functionally relevant to a phenotype of interest. In addition, because biological systems are modular, several tiny, closely related modules can be combined into bigger modules in a hierarchical fashion.

Thus, network motifs, modules and hierarchy are three network ideas that have produced significant results in this area [67]. Even though we conceptually split the categories, we want to highlight that all notions work seamlessly in biology.

3.3.1 Network motifs

In addition to the methods for studying networks globally, and nodes particularly, that were covered in the preceding sections, a thorough examination of local features is essential to understand the functional mechanisms of biological systems.

Motifs are sub-networks that appear more frequently than those seen in random networks. These patterns are believed to be the basic building blocks that enable these networks to conduct biological functions. All 13 potential three-node directed sub-networks are shown in Figure 3.6a and potential three- to five-node undirected sub-networks are shown in Figure 3.6b. These sub-networks are also known as graphlets.

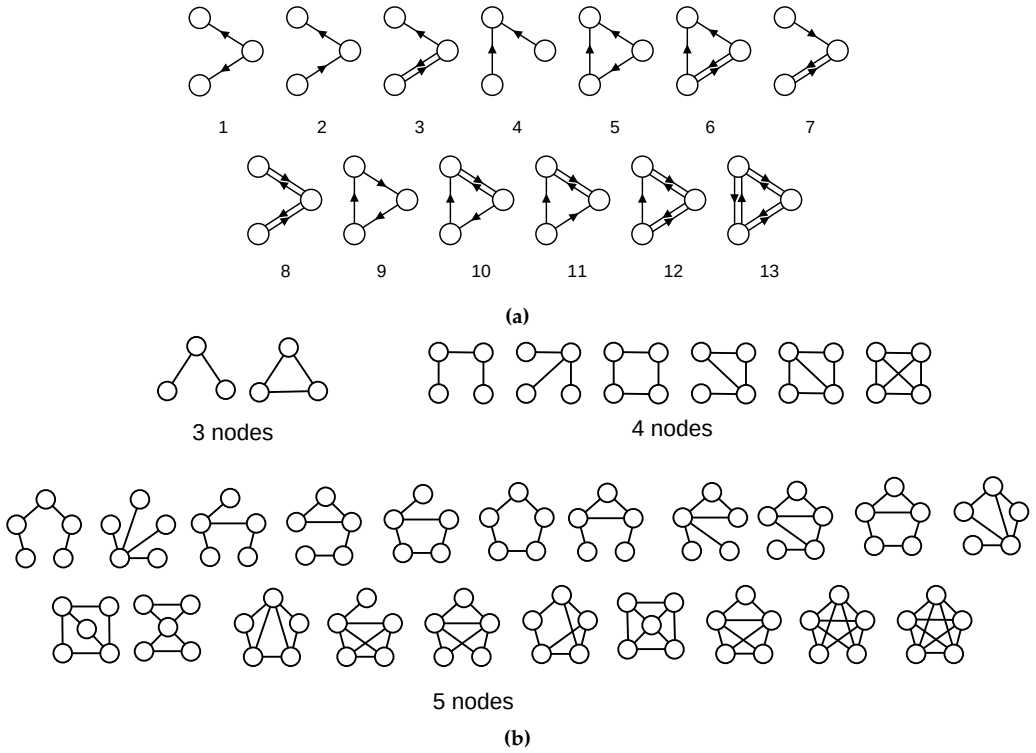


Fig. 3.6 (a) The 13 possible motifs found in 3-node directed networks. Adapted from Schatz et al. [110]. (b) Graphlets variations from 3, 4 and 5 nodes. Adapted from Hulovatyy et al. [111].

The process of defining a motif in a graph G may be described in terms of graph theory as follows: (i) Determine every motif $h \subseteq G$ that is possible. (ii) Create a random distribution for graph motifs by randomly generating a network G' that is the same size as G and has the same degree distribution. (iii) Find all possible motifs $u \subseteq G'$. Within the network G', a subgraph u that is isomorphic to a

motif h is referred to as a match. The frequency of $u \subseteq G'$ is determined by how many times it appears. It is worth noting that determining statistical significance depends on the random network G', which composes the null hypothesis. Random networks are built according to theoretical presumptions because they are not easily obtainable in practice. Hence, the biological context should be taken into account since biological networks may call for different assumptions to produce random networks that are suitable null distributions. This is the rationale of *FANMOD* [112], one of the best-known algorithms for algorithm identification.

The comprehensive study of motifs is particularly relevant in the analysis of gene regulatory networks. For instance, in order to examine cell type transition, Ye et al. [113] investigated gene regulatory networks that may produce a variety of intermediate cellular states using stepwise modulations of transcription factors, and they discovered a motif family that restricts the reversibility of lineage change.

3.3.2 Network communities

A particularly active area of research is identifying the topological community structure (or dense subgraph) of complex networks. Formally, a network module corresponds to a subgraph once more, although modules vary from motifs in that they are larger and include more vertices.

In recent years, a variety of methods for detecting network communities have been developed. Yet, the most often used detection techniques seek to divide the network in a way that maximizes modularity, a parameter defined as follows:

$$Q = \frac{1}{2M} \sum_{i,j \in V} [A_{i,j} - P_{i,j}] \delta_{C_i, C_j} \tag{3.8}$$

where M is the total number of edges in the network, $A_{i,j}$ is the adjacency matrix and $P_{i,j}$ is the expected number of edges between nodes i and j. Then $\delta_{C_i, C_j} = 1$ when nodes i and j are in the same community $C_i = C_j$ and 0 otherwise. The ratio between the proportion of edges found in communities minus the proportion of edges found by random connections is evaluated by the metric Q [67, 84].

We must apply an optimization technique to locate the communities because it is hard to test every partition thoroughly. Many approaches, including the clique percolation-based approaches, the link-community clustering algorithm for the overlapping community structure, and random-walk-based approaches for community detection on weighted and directed networks, have been proposed to address this issue.

Numerous findings suggest that the first neighbors of disease genes in a biological network may also be the genes linked to such disease, or, to put it another way, that genes linked to the same disease are more likely to interact with one another. Although topological communities are frequently capable of representing functional modules, whose detection assists the understanding of biological processes in a systematic way. Functional modules also contribute to the search for new drug targets in medicine, thus promoting drug development. In that direction, instead of revealing the complete community structure of biological networks, we focus on methods for locating disease-associated modules, as described in the main contributions of this thesis. Disease modules are often constituted by a limited set of genes and are indicative of the underlying causes of disease. Although disease modules are inferred in connection to a certain disease, mutliple diseases may contain the same nodes and edges, as diseases often share mechanisms. The topological community discovery techniques are not particularly effective at discovering disease modules, so other approaches need to be applied.

3.3.3 Network hierarchy

Biological processes are typically thought to be hierarchically organized. Nevertheless, in the presence of feedback loops, which are pervasive in gene networks, hierarchies are hard to infer from experimental results.

This idea is most evident in the reconstruction of GRNs, where there is a directed influence from transcription factors to their targets. It should be noted that transcription factors are DNA-binding proteins that can engage in a variety of activities and frequently have many targets. For this reason, in GRNs, transcription factors tend to show high out-degree and behave as regulators above other genes. Allocating nodes without incoming links to the top layer and arranging the remaining nodes as per their Dijkstra distance from such first nodes makes it simple to identify a hierarchy in directed networks [67].

Because there are many shortest routes from nodes in the first layer, the previous definition of hierarchy is more complex for other BN types. For instance, the concept of hierarchy is more nuanced in GCNs, where the role of hub nodes as master regulators should be interpreted carefully. However, there can be modules in a GCN that trigger an effect on other modules.

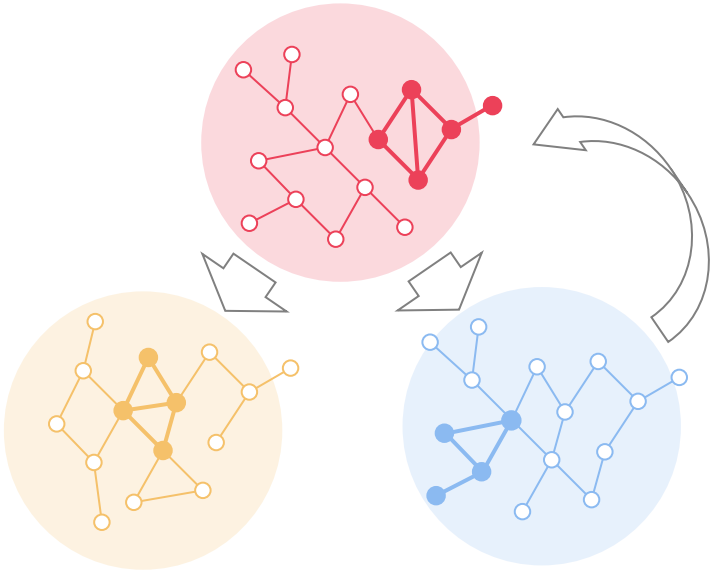


Fig. 3.7 Biological network functional organization illustration. Every module (colored circle) has a corresponding motif (filled). The yellow and blue modules are affected by the red module. The latter reacts to the red module as well.

3.4 Biomedical applications of biological networks

As a result of their intricate and complicated relationships, biological organisms generate very complex networks. Given that a wide range of biological data may be represented as a graph and that only a few instances of biological networks were introduced in Section 3.1, the reconstruction and applications of biological networks could not be more diversified. For clarity, in this section we shall focus on the key uses of biological networks in biomedical research.

The goal of network medicine is to understand human disease from the perspective of how disturbances in the interactome and subcellular systems contribute to disease origin and progression [84]. Network medicine's central premise is that during disease onset, development, and maintenance, sub-cellular networks gradually reorganize, generating progressive modifications in local and global network characteristics and system states, all of which are thought to be responsible for disease-causing variables.

Nowadays, the consequences of mutations on individual biomolecules, known as nodetic effects, are simpler to understand. A typical example of a "nodetic" effect is the direct knockout or knockdown of a gene or protein that removes a node and all of its edges. Such mutations can affect a network directly by altering protein function or indirectly through post-translational modifications and ligand-protein

bindings [114]. In the case of cancer research, the catalogue of somatic mutations in cancer (COSMIC) [115] serves as evidence for this.

On the other hand, recent functional studies, show that disease-associated alleles commonly affect specific PPIs and/or gene regulation. Hence, "edgetic" effects are caused by network-attacking mutations, e.g. affecting protein-protein interactions. PPI-specific mutations can result in the loss or gain of certain PPIs [116]. Since biomolecules operate together in hyper-connected networks and pathways rather than separately, this approach enables a more realistic evaluation of the intricacies of human diseases at the molecular level.

The application of biological networks in the field of biomedicine has provided highly relevant knowledge for the understanding of the molecular mechanisms underlying complex diseases such as cancer [117]. Among others, these networks have been actively involved in the discovery of new disease-associated modules, therapeutic targets and biomarkers that allow detection and monitoring of diseases [118]. Many of the predicted interactions have later been confirmed experimentally [119, 120].

3.4.1 Disease module identification

Biological networks are often too big to be explored in their entirety. The next stage in using network data to get biological insights is frequently module identification, also known as community discovery or graph clustering, in network research. The notion of modularity in cellular networks, which are mainly composed of groups of intricately connected biomolecules in charge of certain cellular functions, motivates module identification. Finding these functional units, also known as modules or pathways, is a significant challenge in the study of diseases [121].

A disease is therefore a disruption of a particular set of molecular interactions that results in aberrant patho-physiological processes. For instance, in a gene network, the majority of disorders are not adequately explained by a single gene, but rather by a set of genes. Indeed, in the human PPI interactome, disease proteins do not constitute a random distribution but rather one or more linked sub-graphs that define the disease module. Such disease gene products, which often have high mutation frequency, has prompted the development of innovative approaches to discover disease-specific modules by connecting individual patient omics data to observed phenotypes.

There is strong evidence that genes involved in complex disorders frequently interact with one another and use the same biological pathways. Given this

overlapping, in Nogales et al. [122], the authors make an argument for changing the way we define disease, from the phenotype we perceive to its real underlying mechanism. Such disease mechanisms will eventually be understood, and they will define new subtypes, endotypes, for diseases. These endotypes are built using related risk, driver genes, proteins, and therapeutic targets to create a disease module or disease signaling network from scratch.

A variety of approaches have been suggested for solving the problem of disease module identification. These are classified in approaches that use pre-existing molecular interaction networks (including PPIs, pathways, and GRNs) and approaches that infer disease-specific networks directly from data in order to build *de novo* disease modules. The latter data-driven network inference methodologies have been proven considerable promising in finding novel features disturbed upon disease, agnostic of preconceived ideas. Using graph clustering techniques, such networks may be decomposed [121]. The computed networks provide insights into aberrant regulations in signal transduction within diseases, assisting hypotheses-making that can be later tested experimentally and identifying potential therapeutic targets (see Section 10.4).

3.4.2 Biomarkers discovery

The key for effective therapy is choosing the appropriate patients, not the correct treatments, i.e. the patients group presenting both the phenotype and the target endotype [122]. As a result, biomarkers turn into a crucial diagnostic tool for identifying diseases, enabling precision treatment. Typically, a scalar feature obtained from biological data is referred to as a biomarker. Either a single element from electronic health records or a composite measurement of several compounds might be considered biomarkers. A biomarker can provide helpful clinical information by (i) prognostically predicting the outcome of a patient, (ii) anticipating the differences in result among potential treatments (moderating), or (iii) choosing the therapy that best optimizes the outcome (prescriptive) [15]. Importantly, a prognostic biomarker predicts the patient's overall disease prognosis independent of therapy, whereas moderating and prescriptive biomarkers anticipate the impact of medical approaches. Prognostic biomarkers could then be used as a starting point for new therapeutic targets [123].

Using complex, rich data as substrate, BN approaches produce a thorough knowledge of how huge numbers of interconnected biomolecules of a system compose networks whose functional qualities emerge as defined functional elements,

rather of focusing on a restricted number of molecular components. Hence, patients may be characterized by the molecular networks underlying them rather than their phenotype alone [124]. In a network context, topology-based approaches allow the identification of molecular or network biomarkers [125].

On the one hand, molecular, or in this case single-molecule, biomarkers refer to quantifiable molecular assessments of biological homeostasis For example, in the case of gene expression, the biomarkers' expression profile should show a clear difference between a disease and a normal state. As explained before, the functional structure of biological networks makes it possible to identify particularly relevant nodes, e.g. *hubs* and *bottlenecks*, which could potentially be used as molecular biomarkers [72, 105]. Hub nodes exhibit relationships with a large number of other nodes, and typically control the biological phenomenon under study [106]. Meanwhile, bottleneck nodes establish functional communication between modules [126].

On the other hand, another strategy in systems biology is to explore a high-level function by picking a few low-level interactions that are thought to be pertinent to a certain phenotype [127]. Hence, network or module biomarkers refer to functional modules or groups of densely connected nodes that are often involved in the same biological process [128]. In contrast to single molecular biomarkers, sets of interacting molecules with similar behavior, known as module biomarkers or network biomarkers, offer a measurable and steady method to typify diseases, which has prompted the progress of network medicine. Researchers suggested studying the interplay of a considerable set of molecules, or network, to more accurately detect a condition [129].

Chapter 4

Gene co-expression networks

In the preceding chapters, we discussed how scientists are becoming increasingly interested in studying biological systems from a network perspective, and the notion that complex diseases are "diseases of the interactome" has received widespread acceptance. We have addressed the biological insights from a variety of data sets, including unipartite networks (such as PPI) and heterogeneous networks, as well as molecular-molecular interactions and cell-cell interactions (e.g., drug-target networks). We have also discussed how molecular profiling approaches make it possible to acquire sizable omics data sets, such as those for transcriptomics, proteomics, and metabolomics, at a cheap cost and with high throughput. Finally, we offered a broad theoretical framework for BN mining.

It is well recognized that platforms must be developed to aggregate biological data and create models that describe biological interactions. Not all omics, though, have profited equally from technological advancement. Transcriptomic technology's fast development, compared to other analytical platforms, has allowed investigations on genetic and environmental modifications at the transcriptome level in many organisms. As a result, many network approaches have been widely used in transcriptomics, which gene expression data to infer gene regulatory networks (GRNs) or gene co-expression networks (GCNs). Nevertheless, although the bulk of the current network inference techniques are employed for GCN and GRN, they are also useful for inferring other types of networks, like those reconstructed from proteomic and metabolomic data [130].

The focus of this thesis will be GCNs, even though we have previously covered a variety of biological networks and outlined fundamental graph analytic concepts

that are generally applicable. GCNs have attracted more interest recently as a result of their ability to analyze enormous transcriptional data sets [16]. A GCN may be used to discover which genes in a dataset are more likely to display a coordinated pattern of expression (co-expressed genes). Through gene co-expression analysis, hundreds of genes with similar expression patterns under various situations may be simultaneously identified, clustered, and explored, which means a systematic approach for determining genes functional state.

Notably, gene expression can be used as a surrogate to interrogate other omics. For instance, mutations at the genome level will produce aberrant expression of transcripts, if any. Numerous techniques also use gene expression as a proxy for protein levels, with the underlying premise that the stronger and longer-lasting the expression, the more likely it is that a protein is present [131, 132]. Then, if two genes are co-expressed and the gene products of those two genes are known to interact, the co-expression of those two genes may also be indicative of the physical interaction of their respective gene products.

Hence, GCNs provide information on genes that are present simultaneously and are frequently engaged in the same biological processes, but they typically cannot provide information regarding causality or make a distinction between genes that are regulated and those that are regulators. Nevertheless, when we combine information of known regulators from which we can infer causation, we may be able to find transcriptional regulatory networks. GCNs enable ranking putative disease genes and relating unknown-function genes to biological processes.

4.1 Data suitable for gene co-expression analysis

As reviewed in Section 1.1, in the post-genomic era, reducing costs for large-scale and high-throughput measurement technologies has been key to obtain an extensive collection of gene expression profiles that capture transcriptional changes during development, between conditions, etc. Furthermore, the sequenced genomes of model organisms (such human, mice, and fruit fly) and the molecular characterization of diseases like cancer, immune disorders, and neurodegenerative disorders significantly improve our understanding of transcriptional dynamics.

As a result, gene expression databases that are accessible to the general public were created using the compendium of produced data. An example of these is shown in Table 4.1. These databases feature intuitive user interfaces that make it

easier to retrieve the data, and the most of them also have built-in data processing tools.

Table 4.1 Repositories for gene expression. GEO, gene expression omnibus; UCSC, University of California Santa Cruz; TCGA, The Cancer Genome Atlas; GTEx, Genotype-Tissue Expression.

Database	Description	Organism	Reference
GEO	Open-access repository for functional genomics data including on arrays and sequencing-based.	Several	Clough and Barrett [133]
Genevestigator	Provides user-friendly visualization tools that enable the interpretation of bulk tissue and single-cell transcriptomic data from public sources.	Several	Hruz et al. [134]
UCSC Xena	Online examination of multi-omic, clinical/phenotype, and public and private data.	Human	Goldman et al. [135]
LinkedOmics	32 TCGA cancer types and 10 CPTAC cancer cohorts, which also contains multi-omics data.	Human	Vasaikar et al. [136]
TCGA	Over 20,000 molecularly described original cancer and matched normal samples from 33 different cancer types.	Human	Weinstein et al. [137]
ArrayExpress	High-throughput functional genomics experiment data made available for reuse.	Several	Brazma et al. [138]
GTEx	Comprehensive public resource to research tissue-specific gene expression and regulation.	Human	Lonsdale et al. [139]

GCNs enable the concurrent analysis of several gene co-expression patterns under a variety of circumstances. Importantly, the approach usually requires an appropriate number of expression samples for the inference process. Therefore, publicly accessible transcriptome data sets are important sources for this kind of analyses. Indeed, about one out of every four studies addresses a biological issue using publicly available data rather than producing fresh raw data. The reuse of such data highlights the need for reliable tools for gene expression analysis [72].

4.1.1 Microarrays vs. RNA-Seq data

As previously mentioned, gene expression data from microarray or RNA-seq technologies can be used to reconstruct GCNs [140]. Although in general, microarray provide more reproducible and robust results, they have a limited number of probes that can be screened in a sample [141]. In this sense, one of the main advantages of RNA-seq is that it measures the expression of the more than 70 000 non-coding RNAs that are often not quantified by microarrays, including the recently identified lincRNAs, long intervening non-coding RNAs, many of which are expected to have regulatory functions that are crucial in disease.

In the specific context of GCN inference, it is relevant to consider the further advantages RNA-seq offers [140]. Firstly, compared to microarray-derived profiles, it improves accuracy for low-abundance transcripts, has a higher precision for detecting tissue-specific expression, and better separates expression profiles of closely related paralogues. As we mentioned before, RNA-seq can also discriminate between the expression of several splice variants, which may have various biological activities and interaction partners. These splice variants, together with non-coding RNAs, may be given potential functions by co-expression analysis on RNA-seq data [142, 143]. Following this method, read distributions that map to a gene's numerous exons are used to infer the expression of distinct isoforms coming from the same gene. By merging all overlapping gene isoforms in the analysis of the RNA-seq data, a typical approach for inferring RNA-seq-based GCNs is to build the network at the gene level. This indicates that if two genes' various splice variants exhibit coordinated expression, only then the two genes are associated. However, this method obscures information on many transcripts encoded by the same gene.

As an alternative, GCNs based on transcripts can be reconstructed. Due to the abundance of gene isoforms and non-coding RNAs, these networks' disadvantage is their rapid growth in size. Co-expression networks are square matrices; therefore, when more genes are added, the size of the network grows quadratically. Given that the human genome has around 200 000 identified transcripts but only about 20 000 protein-coding genes, the resultant network is 100 times larger, significantly increasing the computer resources required for the analysis. Building co-expression network blocks from subsets of the data and combining them later on in the analysis is one way to deal with this issue.

4.1.2 The large p - small n problem in expression datasets

Expression data often contains a large number of genes or predictors (p) and only a small number of instances or samples (n), making its analysis a challenging process. Such data sets are prone to over-fitting, excessive variance, and the generation of non-reproducible results. Additionally, the underlying expression data includes a wide range of noise that contributes to this problem. A well-known method for handling these types of data is the selection of a small subset of features, which facilitates transmitting consistent data throughout the experimental setting. For this reason, there are two primary methods for searching publicly accessible gene expression datasets. These methods are referred to in the literature as targeted (or gene-guided) and non-targeted (or global) techniques [144]. The biological question being addressed and the collection of data are the two main factors that influence the choice for the technique.

The non-targeted methods offer a broad picture of the patterns of co-expression of several genes under a variety of situations. Due to the fact that no *a priori* knowledge is needed to build the network, this method is also considered knowledge-independent or condition-independent. The term "global" networks is typically used to refer to the complete collection of expected connections between gene pairs. Separated modules of genes with similar functions can be found, allowing for further gene prioritization.

However, weak interactions that only take place under particular conditions are frequently overlooked in global networks. This can be avoided by carefully choosing experiment data that is related to the biological inquiry being addressed, resulting in a condition-dependent approach. The resulting condition-dependent networks increase the prediction potential of gene functional annotation by offering insights into certain biological processes. The explanation of system-wide features, such as genes with pleiotropic effects and overlapping biological pathways, may go unnoticed due to this approach's limitations [72].

The heterogeneity of the samples must be taken into account in co-expression analysis. In a GCN inferred from different conditions, it may be difficult to identify condition-specific co-expression modules since the correlation signal of these modules is diluted by the absence of correlation in other situations. For this reason, an alternate strategy makes use of full gene expression data sets clustering the samples before building the network, thus finding heterogeneity in the data [72, 145]. In this instance, the input samples are divided into a specified number of groups based on their overall expression similarity using a clustering method that

is directly applied to the normalized expression matrix. The resulting clusters are then used to construct co-expression networks.

While important elements of certain pathways have been elucidated by experimental data, a targeted technique can aid in more precisely identifying new participants of the same route. In order to reconstruct a seeded GCNs, well-known genes (seeds), or genes that are relevant in the dataset (like differentially-expressed genes), are employed as input. Nevertheless, GCNs may be more predictive when using techniques that combine targeted searches with condition-dependent methods.

4.1.3 Data preprocessing

Preprocessing data is frequently required prior to GCN inference, albeit the procedure differs based on the technique. Gene expression levels are often log2-transformed before the similarity score is computed as a common approach to scaling the data to the same dynamic range.

Recent reviews of various tools and techniques to get accurate expression counts from RNA-seq data may be found in Conesa et al. [146]. In co-expression analysis, different normalization techniques induce bias, often in favor of positive correlations [147]. For the study of single-cell RNA-seq data, specific tools have been developed, and they are discussed in Bacher and Kendziorski [148]. In order to address these normalization challenges, new approaches are constantly being developed.

4.2 Approaches for reconstructing GCNs

Inferring a GCN still requires careful consideration of how to define the edges that connect genes. This Section will focus on the data-driven reconstruction of GCNs, which characterizes the interrelation patterns between gene pairs taking as input expression data sets [149, 144]. By definition, GCNs are undirected graphs in which pairwise dependencies among participating genes are estimated using different co-expression measures [150]. Because they are undirected, GCNs are also considered association networks. Some of the techniques outlined below are also inevitably applied to GRN inference, according to the assumption that if two genes are co-expressed and we recognize one of them as a transcription factor and the other as a target, they are likely to form a directed regulatory link.

There are two main categories of techniques to identify transcriptional connections between genes: supervised and unsupervised approaches [151]. Supervised methods like regression and machine learning-based methods, for example, require a training dataset to retrieve relevant gene interactions and are often used to infer GRNs [152]. In contrast, unsupervised methods mostly capture transcriptional connections that might be of biological significance. Four unsupervised co-expression metrics for estimating pairwise significance that have been widely applied are the following [150]:

1. Mutual information (MI): is based on estimates of density functions and has been demonstrated to perform well enough with nonlinear relationships. The definition of the MI I between the discrete genes X_i and X_j is:

$$I\left(X_i, X_j\right) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p\left(x_i, x_j\right) \log\left(\frac{p\left(x_i, x_j\right)}{p\left(x_i\right) p\left(x_j\right)}\right) \tag{4.1}$$

   where x_i and x_j, are instances of respectively genes i and j, $p\left(x_i\right)$ and $p\left(x_j\right)$ are their marginal probabilities, and $p\left(x_i, x_j\right)$ is their combined probability distribution. In this formulation, it is necessary for X_i and X_j to be discrete variables, e.g. using the equal-width binning discretization method proposed by Meyer et al. [153].

2. Correlation coefficients (CC): either Pearson CC (PCC), Spearman CC (SCC), or Kendall CC, which are the traditional estimators of linear transcriptional association between genes. Because a CC between two genes does not account for the expression of the remaining transcripts in the entire transcriptome, CCs are 2-dimensional distance measurements.

$$PCC_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$SCC = PCC_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}$$

$$KCC = \frac{\text{(number of concordant pairs)} \; - \; \text{(number of discordant pairs)}}{\text{(number of pairs)}} \tag{4.2}$$

where $\text{cov}(X,Y)$ denotes the covariance between the gene expression vectors, σ denotes the mean, R is the ranking operator. In the case of *KCC*, for any gene pair (x_i,y_i) and (x_j,y_j), where $i < j$, are considered concordant if the sorted order of (x_i,x_j) and (y_i,y_j) agrees: i.e., if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. They are described as being discordant if not.

3. Partial Correlation (PC): The following equation can easily be used to calculate first-order PCs [150, 154].

$$R_{xy.z} = \frac{R_{yx} - (R_{yz})(R_{xz})}{\sqrt{(1-(R_{yz}^2)(1-(R_{xz}^2)}} \tag{4.3}$$

where the simple correlations between genes x and y, y and z, and x and z are, respectively, R_{xy}, R_{yz}, and R_{xz}. For large datasets, these calculations could be extremely time-consuming, hence PCs should instead be determined by multiple linear regression with a feature selection step.

4. Ranked CC: the remaining $N-1$ gene, with N representing the total number of genes, are used to calculate all CCs, which are then ranked from 1 to N. Since genes A and B have differential expression profiles and interactions with the other transcripts, rank (A to B) is different from rank (B to A) within a pair of genes. Mutual ranking (MR) and highest reciprocal ranking (HRR) are the two primary approaches [155, 156]. Because they rely on CC values surrounding a gene pair, MR and HRR are expected to be more integrative than unranked CCs. Another possibility is that ranking CCs will only partially account for the "range restriction effect" seen in CCs that provide strong associations for high variance genes [151].

In most cases, unsupervised methods have been used to effectively detect relevant gene interactions. However, they demand specific data characteristics like labels and data size that are unmet by most datasets. Each approach has its own benefits and drawbacks, especially when used in various experimental contexts [157]. Nonetheless, correlation-based approaches are highlighted due to their simplicity, which enables them to deal with enormous volumes of data at a minimal computational cost, being suitable for time-series as well as steady-state scenarios. For this reason, PCC or SCC are frequently employed and are competitive with more complex techniques in terms of identifying gene connections and functionality on massive data sets. On the other hand, PCC, the most often used CC, assumes a

linear correlation and normally distributed data, and it is also sensitive to outliers [72]. In the light of the above, we conclude there is no top strategy, e.g. Song et al. [158] discussed MI does not frequently outperform correlation, contrary to general belief. As a more appealing alternative to PCC, they suggested bi-weight mid-correlation (bicor), which is able to capture non-linearity.

4.2.1 Correlation-based GCN inference

Despite the progress in GCN reconstruction, we are still far from a "one fits all" approach. Instead, the combination of several network inference techniques shows the most effective performance across numerous data sets [157]. This finding implies that various techniques can only adequately capture a subset of network structures when used alone, but fortunately they can perform well when combined. As a result, efforts have been made to develop tools that include networks inferred using different methodologies.

As a rule of thumb, when the data amount is greatly expanded, the majority of algorithms will become constrained, making them more likely to have a large number of indirect connections with a high false-positive rate [90]. In fact, Maetschke et al. [152] found that the accuracy rate is improved for subnetworks and came to the conclusion that genome-scale networks cannot be inferred without first executing a feature selection step to reduce the inference challenge. As discussed in Petralia et al. [159], one straightforward but efficient solution to this problem is to reduce the search space by leveraging known information. For instance, pathway-level co-expression (PLC) is a method for highlighting functional gene sets, such as those in known pathways, by capturing transcriptional relationships in these [151]. These targeted co-expression techniques have been expanded to identify networks that are phenotype-specific, evaluating network inference under different experimental conditions. Hence, many network inference techniques employ seed genes that are related to a particular biological process, such as genes that exhibit differential expression under control and under case scenarios [160].

Other approaches that expand the co-expression analysis by integrating data types such as PPIs, methylation, TF-TF and TF-target gene interactions, and sequence motif analysis of co-expressed genes, might help to further understand the regulatory roles of these genes. This facilitates the discovery of regulatory components that influence expression and compose co-expression modules, such as TFs, expression quantitative trait loci (eQTLs), and methylation patterns.

In brief, we can summarize the process of GCN inference into three main steps:

1. Genes and samples are assembled into expression matrices derived from expression data (microarrays or RNA-Seq), and for each potential gene pair, a distance or similarity estimation is made. Analogously to what is exposed in Section 3.2, a gene-gene similarity matrix that may be utilized in downstream analysis is used to describe GCNs.

2. A network using co-expression associations is reconstructed, where each node corresponds to a gene and each edge denotes the degree and strength of the co-expression link. Such edges can be thresholded, so only edges exceeding the predetermined threshold are then deemed relevant and kept for network reconstruction [151, 72].

3. As an optional third step, using gene prioritizing techniques, modules of closely linked genes may be retrieved from the network using one of the several clustering methods that are available [147].

Following the graph classification detailed in Section 3.2.1, correlation-based GCNs are signed graphs since correlation measurements normally range from -1, a perfect negative correlation, to 1, a perfect positive correlation. However, in an unsigned network, two negatively linked genes will be treated as co-expressed since the absolute correlation values are applied, resulting in the clustering of these genes and interfering with the structure of the network, as they are probably co-expressed with a completely distinct collection of genes. Using a signed network and scaling the correlation values between 0 and 1 is a common strategy to handle this problem, signifying positive correlation with values > 0.5 and negative correlations with values < 0.5. A signed technique builds networks that better distinguish functional modules.

The fact that a similarity measure is estimated for every possible gene pair involves GCNs are inherently weighted graphs. However, by designating gene pairs with a correlation over a particular threshold to be linked and all others disconnected, a weighted network can be converted into an unweighted network, since the interaction between gene pairs is binary, linked (1) or disconnected (0). Also by definition, correlation networks do not discriminate between direct and indirect links, often yielding a large number of indirect connections as false positives. This is addressed by *ARACNE* [161], one of the most well-known GCN-inference algorithms, which prunes edges based on the evaluation of gene triplets. The edge between genes A and C may really be an indirect interaction between A and C via B if genes A, B, and C are fully linked in the network and this edge has the

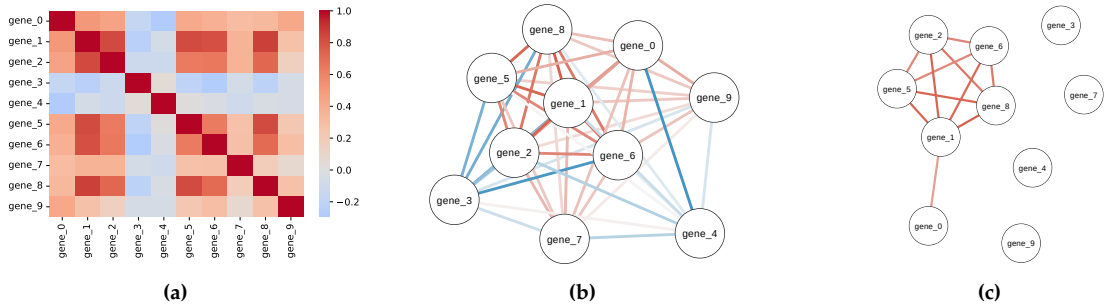


Fig. 4.1 (a) A Pearson's correlation matrix representing gene-gene similarities (b) The corresponding weighted GCN, showing positive and negative associations between gene pairs. (c) the previous network, but only with those relationships whose absolute correlation value is over 0.5. Significantly co-expressed genes are represented in the network as numbered nodes connected by edges.

lowest weight. Also in this regard, the context likelihood of relatedness (CLR) Zhu et al. [162] technique handles indirect correlations by using a score based on the MI distribution as the edge characteristic, enabling the false-positive rate to be managed.

The gene pairs that should be linked in the network are chosen once the similarity scores between all gene pairs are calculated. Although this threshold may seem arbitrary, there are a number of approaches to choose wisely. For instance, only the top 0.5% of positively and negatively linked couples were chosen by Lee et al. [163]. Furthermore, using the Weighted Gene Co-expression Network Analysis (WGCNA) software [164], Bassel et al. [165] selected a cutoff that causes a network to follow a power-law distribution. In another case, using random permutations of the expression data, Butte and Kohane [166] established a limit for significant interactions.

Other methods base their p-value calculations on the null hypothesis that there is no association between any two genes. Instead of using hard cutoffs, Zhang and Horvath [105] suggested using soft thresholds to create weighted gene networks and maintain the correlation's underlying continuous character. However, the neighbors of a node that are directly related to it might be hard to distinguish, making it difficult to visualize these networks.

4.3 Mining GCNs

Along with the general BN applications discussed in Section 3.4, the two primary uses of co-expression network analysis are to identify novel genes engaged in the

biological process being studied and to infer the biological process that a gene may be involved in.

It makes sense that reliable networks are required to infer accurate predictions about gene function. Such networks rely largely on a number of choices made throughout the GCN inference process. Any oversight can lead to unstable networks and subsequent false biological interpretations, depending on the kind, quality, and availability of the input data, the correlation coefficient, the inference method, the previous knowledge, and the experimental and computing resources [72].

GCNs offer a potent paradigm for comprehensively describing gene-gene associations in complex systems. Networks provide not just a wide range of computational and mathematical capabilities but also a very simple visual representation to massive data [167]. However, GCNs can become incredibly complicated when working with huge data sets, which restricts how biologically meaningful they can get. Thus, it is crucial to apply network analysis tools and techniques carefully in order to optimize the information extraction, untangle trustworthy network connections, and infer real biological significance.

4.3.1 GCN modules and the guilt-by-association principle

Module extraction is a common technique for thoroughly examining GCNs and generating controllable subgraphs. Since we already covered the general organization of biological networks in Section 3.3, we will focus on the framework for GCNs in this Section.

Genes having similar expression patterns across several samples are grouped using clustering in co-expression analysis, which results in groups of co-expressed genes rather than just pairings. Graph clustering algorithms aim to cluster the network by taking its topology into account. Several methods try to identify densely connected regions in a network. Others attempt to "break" the links (edges having high betweenness centrality) connecting various communities. Finally, some others are based on node distances or try to look for flow paths [80].

There are many clustering approaches for identifying graph modules, but only a handful of them can handle enormous networks with a huge number of edges. As exhaustively reviewed in D'haeseleer [168], given that it has a significant impact on the analysis' results and significance, the clustering approach should be carefully chosen. In any case, the network, or adjacency matrix, is subjected to a clustering technique in order to infer subnetworks or modules of highly co-expressed genes. The resultant modules might be phenotype-specific and frequently reflect particular

biological processes [70, 169]. These procedures can be categorized using the hierarchical and partitional algorithms' established categories. On the one hand, by repeatedly splitting each cluster into sub-clusters, hierarchical clustering algorithms produce a tree with branches that represent co-expression modules. The network vertices are initially given weights, e.g. using the estimated correlation coefficient. Then, clusters are constructed using vertices with a high weight, and they are gradually increased by adding nearby vertices. Depending on the set threshold, for example, the number of final clusters changes. This is the kind of clustering performed by *WGCNA* [130]. On the other hand, partitional techniques like *k*-means clustering identify a predetermined number of modules based on the input cluster criteria.

The biological information contained within a co-expression module can be evaluated using a variety of techniques. One approach is the functional coherence of the modules. This strategy consists of testing for significant co-expression enrichment between many functionally annotated gene sets defined by curated functional standards (see Section 4.3.2). This can be used as a benchmark to assess how much existing information is possibly reflected by the network, as well as a technique to annotate gene clusters. Other approaches emphasize network structure without using functional annotation.

According to the guilt-by-association (GBA) principle, genes with comparable functions or those implicated in the same regulatory pathway will often have similar expression patterns, forming clusters or modules in the network [151]. Functional modules offer a solid framework for the discovery of novel genes pertinent to biological processes and the functional annotation of those genes in the absence of substantial *a priori* information. As a result, within a module, we can predict the role of co-expressed genes with unknown functions by associating these with other genes whose function is known, using enrichment tools and annotation databases.

4.3.2 Functional analysis

GCNs are a robust method to hasten the understanding of the molecular mechanisms underpinning crucial biological processes. Functional analysis enables us to make use of prior knowledge of networks and pathways to acquire additional insights from a higher-level viewpoint [64].

Using functional enrichment analysis or pathway enrichment analysis, it is possible to find genes or proteins that are over-represented in a large collection of genes or proteins that may be linked to disease symptoms. The strategy applies statistical

tests to identify gene clusters that are substantially enriched or depleted in certain portions of biological knowledge, like the annotations from the gene ontology database. For enrichment analysis, there are three main analytical approaches: over-representation analysis (ORA), rank-based approaches, and topology-based approaches.

Because it is simple to use, ORA, the first-generation technique for enrichment analysis, is still commonly used. Through the use of hypergeometric distribution, chi-square, or Fisher's exact statistics, ORA determines the importance of over-representation. Numerous tools, like DAVID [170], have included this technique. Notably, the input candidates will affect the outcome of the analysis, and the prefiltering requires an arbitrary threshold [64]. All included genes are treated identically in the enrichment stage without extra scrutiny. Another significant flaw in ORA results from this assumption, particularly for large-scale analyses, where some risk indicators may be eliminated if they fall below the cutoff. Additionally, even tiny changes in overlap sizes can have a huge impact on the significance scores.

By adding their significance as an additional input, rank-based techniques, in contrast, take into consideration all of the genes' differences. Gene set enrichment analysis (GSEA) uses the weighted Kolmogorov-Smirnov-like statistic to infer the enrichment score using random walk after ranking all the genes in the list first. In contrast to ORA, the performance of rank-based techniques is not constrained by a predetermined threshold, but depending on the kind of statistics used, it may be significantly impacted by a small number of highly significant genes [171].

The third-generation techniques execute enrichment analysis considering topological information, whereas the previous two approaches only regarded pathways as simple lists of genes/proteins. Given that the importance of pathway structure in biological function has long been recognized, topology-based techniques are thought to offer extra potential for elucidating more precise results. *EnrichNet* [172], as an example, is a tool that calculates enrichment scores for each pathway by calculating its distance from network candidate genes using the random walk with restart method. Another popular topology-based enrichment analysis tool is SPIA [173], which combines evidence from ORA analysis with the assessment of the perturbation on a particular pathway.

The third-generation enrichment techniques have outperformed those that merely use node annotations by taking topological information into consideration. Nevertheless, it is important to keep in mind that these methods significantly

rely on network data recorded in current databases, which can be incomplete, inaccurate, inconsistent, lacking, or not specific. The Kyoto Encyclopedia for Genes and Genomes (KEGG) [56] and the Gene Ontology (GO) [174] database are the two most widely used databases for functional analysis. According to analysis of the disease modules, genes that cause common disorders: (i) have a higher tendency for their products to interact with one another via PPIs; (ii) are more likely to be expressed in tandem in particular tissues; (iii) have higher levels of co-expression; and (iv) share Gene Ontology (GO) terms [84].

4.3.3 Topology analysis and gene prioritization

As exposed in Section 3.3, evidence for understanding the underlying biological organization can be encoded in the network topology, which also offers information about the functional significance of certain nodes. For instance, many different biological networks have demonstrated scale-free behavior, which is represented by a node degree centrality distribution that follows a power law. On a visual level, this kind of network shows a lot of nodes with low connectivity and only a few nodes with high connectivity. On the functional level, we discussed the relevance of nodes with high degree and betweenness centrality in Section 3.2.2.

Gene prioritization may be used to mine the reconstructed GCNs for biologically significant nodes, a pressing matter that has been researched to identify key genes in disease mechanisms. In general, we can distinguish two main approaches: similarity-based and network-based techniques, depending on whether they incorporate previous information or whether they focus on graph properties like the ones described in Section 3.3 [175].

A common strategy is to rank genes according to how closely they resemble known disease genes in the GCN network. Prioritization is achieved via similarity-based approaches that compare candidate and seed genes [176]. This approach combines data sources to rank the network nodes and identify groups of functionally connected genes, down to the important candidate genes involved. Some similarity-based gene prioritization approaches are *ENDEAVOUR* [177], *SUSPECTS* [178] or *MedSim* [179]. Due to their connections to factual information and their ability to include both knowledge databases and raw data, these approaches are thought to produce more promising results [175]. These methods, however, are biased towards a predetermined gene subset and can only identify genes starting from the provided prior information. The latter is especially relevant since multiple

candidate genes related to disease genes could be ignored, as many disease genes are still unknown.

Network-based approaches rank the genes largely based on the network's topology. Such methods use metrics derived from features including clustering coefficient, node degree centrality and betweenness centrality [126]. Finding hubs in a GCN is a common strategy since the network information is typically more dependent on hubs than other nodes. Hubs are also probably connected to crucial genes in the network, as discussed with the above-mentioned centrality-lethality rule. While inter-modular hubs serve as the network's center nodes, intra-modular hubs are more informative since they are representative of the biological process encoded by the individual network modules [180, 84].

Analogously to similarity-based approaches, seed genes can be used to rank nodes. The closeness between nodes and seeds may be determined using methods like the shortest distance, neighborhood associations, or random walk distance. One example is the *DIAMOnD* [181] approach, which, given a set of known disease nodes, produces a rank of possible disease-associated nodes. Network-based methods can help better understand disease processes, as they are based on known genes to find topologically close potential genes, contrary to similarity-based approaches that just compare sets of genes regardless of their connectivity [175]. Nevertheless, they are restricted to genes found in networks, and the goodness of the network to which they are applied greatly affects how well they perform.

4.4 GCNs validation

Compared to PPI network analysis and alignment, using gene co-expression data offers several advantages, such as a significantly bigger supply of data for the study of transcriptomics. Each method has restrictions and drawbacks, much like any analysis, as well as some general limitations that should be noted. First off, a thorough evaluation of the statistical power of many of these methods is challenging [182, 183].

In many cases, the estimates could only be applied to the simulation or data set in question and would therefore be limited in their applicability to other data sets. Technical artifacts due to poor experimental setup, improper data preprocessing, and unsuitable contrast methods can all result in false positive or false negative relationships [90]. Notably, since they cannot discriminate between direct and indirect connections, GCNs might also fail to explain complicated

gene-gene interactions [183]. Besides, due to very long runtimes, the majority of techniques do not conduct a thorough examination of all potential statistical associations, e.g., between every pair of transcripts in the genome. For this reason, true models, which might genuinely describe the underlying biology, might not be readily evaluated.

The purpose of these limitations and warnings is not to dissuade researchers from applying GCNs to analyze expression data but to raise awareness about their usage. A successful gene co-expression study must take into account the model's presumptions, the analysis's constraints, and the need for caution when drawing conclusions and interpreting results [182]. Hence, in this subsection, we focus on the main methods for the *in silico* validation of GCNs. We do not include the experimental validation in wet labs since we consider that this would be the very last step, and many other verifications that can be performed first.

4.4.1 Internal validation

On the basis of the information inherent to the data alone, internal validation procedures estimate the quality of GCN inference.

Simulated data

A straightforward way to validate algorithmic performance is to use simulations to produce data sets with known associations and then test if the GCN inference method can retrieve those relationships accurately.

The development of single-cell RNA-Seq and the increasing number of bioinformatics algorithms created to handle this data have rekindled interest in modeling biological pathways and recreating datasets for benchmarking. For instance, the *ESCO* [184] *R* package employs the copula concept to enforce gene co-expression.

For evaluating GRN inference methods that work with complex expression data, benchmarking datasets that take pathway structure into account are of special importance. This is the case of the *graphsim* [185] *R* package, which generates normally-distributed log-expression data using samples from a multivariate normal distribution, and the correlations between the genes are based on a graph structure. Other methods, like *GeneNetWeaver* [186] use dynamical models based on differential equations.

A thorough evaluation system to rate and contrast various approaches is important to direct the development of GCN inference methods. Three elements

constitute an evaluation framework: (i) a mathematical model to simulate data; (ii) a gold-standard or true network to assess predictions; and (iii) appropriate metrics to assess the effectiveness of each method, e.g., precision, recall, or F1-score, among others [187, 188, 152].

Cross-validation

The possibility of false positives is a crucial factor in large-scale analysis, so it is crucial to find a strategy to distinguish findings that have a greater probability of being actual relationships. In this sense, using a large sample size could decrease the amount of false positive edges in GCNs, yet there is still a large set of edges to consider depending on the thresholds used to determine which edges should be included [183]. Because of this, GCN analysis frequently emphasizes alterations that take place in specific gene sets, like differentially expressed genes.

Another strategy is the use of cross-validation to sparsify GCNs and detect a robust structure. As an example in Pierson et al. [189], for each dataset, samples were randomly separated into five groups. Networks were reconstructed using samples from the first four groups, and the accuracy of each network was assessed using samples from the fifth group based on the kept-out test data. A similar leave-one-out cross-validation (LOOCV) method was used in Panahi et al. [190] to determine the expression levels of hub genes. The initial dataset is divided into a training and a test set for this validation. Consecutively, a sample from the initial dataset is removed for training, and the rest is discarded for testing.

4.4.2 External evaluation

All techniques that assess a GCN result using (partial) previous knowledge of the real co-expression solution are referred to as "external validation measures".

Replication in external datasets

Finding robust prediction models involves looking for the replication of GCNs and reproducing outcomes using independent data. The identical organism and the same experimental setup are required by the classical definition of replication with external data [182].

Independent data sets can be employed when external replication is conceivable under some circumstances. For instance, the validation in Li et al. [191] was based

on the expression profiles of hub genes in the main modules of the reconstructed GCN. GEO [133] was searched using precise inclusion criteria in order to confirm such hub genes in an external dataset. Additionally in Tang et al. [192], hub genes were validated using RNA-seq data and clinical information of breast cancer from TCGA [193]. As an additional step, the Human Protein Atlas [194] was also used to validate the immunohistochemistry of candidate hub genes.

However, and strongly depending on the disease under study, independent data sets are not often readily available due to the cost of various experimental analyses and the limited availability of similar setups.

Comparison with reference databases

We distinguish different evaluations at the graph, gene, and functional levels, depending on the level we consider inside our GCNs, i.e., the set of gene-gene relationships, the genes involved, or their biological implications.

At the graph-level evaluation, an indirect approach to validating GCN outcomes is to compare these to databases of interactions that have already been found [157]. Some of the most used databases are *STRING* [76], *Reactome* [85], *BioGRID* [195] or *GeneMANIA* [196]. Generally, such databases are able to gather and combine information regarding confirmed and predicted interactions from a variety of organisms, for both direct (physical PPI) and indirect (functional) relationships, inside a user-friendly interface [197]. For instance, in the case of *STRING* [76], the following sources are used to derive interaction predictions: (i) detection of shared selective signals across genomes; (ii) comprehensive co-expression analysis; (iii) automated text-mining of the scientific literature; and (iv) computational transfer of interaction knowledge between organisms based on available experimental data on protein-protein interactions.

Gene-level evaluation estimates the ability of the reconstructed GCN to detect relevant known genes that are specific to the case study. For instance, we can assess the overlap between the genes found in the GCN and disease-specific genes using annotations from databases like *DisGeNET* [198], COSMIC [199], OMIM [200] or MONDO [201]. Such overlap can be estimated using the Jaccard index.

Finally, functional and pathway level evaluations rely on estimating the enriched functional terms (e.g. GO) and pathway annotations (e.g., KEGG) that are shared between the results and a gold standard list, respectively. For instance, in Sathyanarayanan et al. [202], methods were assessed based on estimations of gene, GO term, and pathway overlap with the gold standard cancer gene list from the

cBioPortal [203]. A high level view into the processes that are effectively retrieved by a network is established by significant enrichment of co-expression between genes that have previously been annotated for functional terms. This provides a general benchmark of the number of GO terms that are co-expressed as opposed to a random expectation [204, 70].

Part III

Thesis contributions

Chapter 5

Computational methods for Gene Regulatory Networks reconstruction and analysis: A review

Authors	Fernando M. Delgado-Chaves, Francisco Gómez-Vela
Journal	Artificial Intelligence In Medicine
Editorial	Elsevier
eISSN	1873-2860
Published	01/04/2019
DOI	doi.org/10.1016/j.artmed.2018.10.006
Impact factor 2021	6,698
Quartile	Q1

Artificial Intelligence In Medicine 95 (2019) 133–145

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Computational methods for Gene Regulatory Networks reconstruction and analysis: A review

Fernando M. Delgado, Francisco Gómez-Vela*

Division of Computer Science, Pablo de Olavide University, ES-41013 Seville, Spain

ARTICLE INFO

Keywords:
Gene Network
Systems biology
Networks validation
Gene Regulatory Network
Gene Network inference

ABSTRACT

In the recent years, the vast amount of genetic information generated by new-generation approaches, have led to the need of new data handling methods. The integrative analysis of diverse-nature gene information could provide a much-sought overview to study complex biological systems and processes. In this sense, Gene Regulatory Networks (GRN) arise as an increasingly-promising tool for the modelling and analysis of biological processes. This review is an attempt to summarize the state of the art in the field of GRNs. Essential points in the field are addressed, thereof: (a) the type of data used for network generation, (b) machine learning methods and tools used for network generation, (c) model optimization and (d) computational approaches used for network validation. This survey is intended to provide an overview of the subject for readers to improve their knowledge in the field of GRN for future research.

1. Background

The great amount and variety of gene expression information generated in the last few years, have led to the need for processing and interpreting such information. In this sense, Gene Networks (GNs) have become a key tool for the understanding and modelling of complex biological processes. The term Gene Network, also called, Gene Regulatory Network (GRN), is used to describe complicated functional pathways in a given cell or tissue, which represent living processes such as metabolism, gene regulation, transport mechanisms or signal transduction.

GRNs are models used to describe and predict dependencies between molecular entities [1]. These are composed of nodes, representing genes, proteins, metabolites or RNA; and edges, which represent molecular relations, e.g. protein–DNA, protein–protein interactions or other relationships of several kind [2] (see Fig. 1). In Fig. 2, a schematic representation of how the abstraction in the modelling process looks like, is shown.

GRNs are a ground-breaking tool for the discovery of new interactions between biological entities, helping scientists in research and making easier hypotheses formulation. They have been successfully used in diagnostics, as in the case of Liang et al. [3]. Many predicted interactions have been confirmed experimentally, which confirms GRN's reliability [4]. The inference of GRNs has also been proven to be relevant in the study of fundamental processes occurring in living organisms [5], ranging from development to nutrition and metabolic coordination. Multiple applications in fields such as human health or agronomy have been developed thanks to the implementation of GRN models. Moreover, GRNs make easier the manipulation, control and coordination of cell physiologic events related to GRN activity: diseases, biotechnological applications or crop production among others. For example in Yan et al. [6], advances on network reconstruction, analysis and interpretation of GRN reliably allowed the identification of molecular biomarkers for monitoring cancer progression and treatment. Also, GRNs have contributed to the representation of developmental processes, as they can generate developmental patterns [7].

Reverse Engineering deals with the process of network inference or GRN reconstruction out of experimental results. In particular, computational GRN inference process relies on the well-known Knowledge Database Discovery (KDD) workflow. KDD goes from input data preprocessing to the validation of generated models, often performed by data base search and comparison with prior experimental data (see Fig. 3).

The process starts with the input data. This usually consists of gene expression datasets, which can be obtained either experimentally or from databases such as NCBI GEO [8] (Step 1 of Fig. 3). After the input dataset is selected, it may be preprocessed by any computational method in order to improve the quality of the study, either because it does not affect the intended network, or because it shows bad quality (Step 2 of Fig. 3). Then, the preprocessed data is used as input for a

* Corresponding author.
E-mail addresses: fmdelcha@alu.upo.es (F.M. Delgado), fgomez@upo.es (F. Gómez-Vela).

https://doi.org/10.1016/j.artmed.2018.10.006
Received 21 June 2018; Received in revised form 23 October 2018; Accepted 23 October 2018

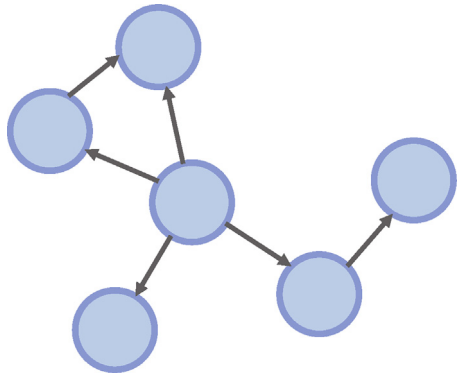


Fig. 1. An example of GRN topology, where nodes represent biological compounds and edges the relations between them.

computational inference algorithm, which provides the resulting network (Step 3 of Fig. 3). Used algorithms may be based on different machine learning approaches. Finally, the obtained model (network) is optimized and validated so true biological insights can be obtained from it, by a comparison with real biological knowledge (Steps 4 and 5 in Fig. 3).

In this work, we present a review of the whole process of GRN reconstruction, based on the KDD workflow (steps in Fig. 3), in an attempt to address the latest and most relevant advances in the field of computational-based GRNs reconstruction. Aiming to offer the most reliable and relevant review, works cited in this paper were classified according to the main taxonomies presented in the field, e.g. the works by Hecker et al. [9] or Dougherty [10] and selected from the main repositories such as Scopus, PubMed or Google Scholar. Besides, most relevant papers, based on citations, presented results and impact of the publication, were given additional priority.

The rest of the paper follows the description of each step of the GRN reconstruction workflow (KDD process). In Section 2, the kind of data and computational methods for data preprocessing used for GRN inference are described (Steps 1 and 2, Fig. 3). Section 3 introduces the main computational approaches for network inference (Step 3). Next, in Section 4, optimisation proposals for the inferred networks are described (Step 4). In Section 5, the different GRNs validation approaches are described (Step 5). Last, final considerations are shown in the Conclusions section.

2. Biological data: basic input for GRN inference

The rapid development of GRNs is linked to the increasing amount of high-throughput technologies, which provide with huge data sets to be managed. These constantly-updated technologies, like Next Generation Sequencing (NGS) [11], which has acquired significant quality, robustness and low noise during the last decade, allow a leader view into RNA and DNA samples. Then, sequencing has become a standard approach, since the Genome is often considered the cornerstone in the study of organisms [12].

Although GRNs have traditionally based on microarray technology [13], depending on the case, NGS may be more efficient than these as a primary source of expression data [14]. Novel techniques like RNA-Seq (RNA sequencing) [15], use NGS to reveal the presence and quantity of RNA in a biological sample at a given moment. RNA-Seq is then used to analyse the continuously-changing cell transcriptome and facilitates the quest for alternatively-spliced transcripts, post-transcriptional modifications, gene fusions, mutations/SNPs and a sort of changes in gene expression. Additionally, RNA-Seq is able to look at different RNA populations including small RNAs (like miRNAs), tRNAs, and ribosomal profiling, thus providing a quite complete overview of the cell state.

Another powerful application of NGS is the study of protein–DNA interactions by looking at protein binding sites in the Genome. This is especially useful in the case of transcription factors (TFs), for which Chromatin Immunoprecipitation (ChIP) techniques are used.

The integration of heterogeneous biological information, e.g. multiple *omics*, may enhance the capabilities of GRN inference. Prior to this points, one needs to know the basics on main *omics*, which are outlined in the following subsection. A general scheme of the process of GRN inference is shown on Fig. 4.

2.1. Omics and related technology

Although genes can be regulated at several levels of integration (transcription factors, co-factors, post-translational modifications of proteins, proteins and transcripts degradation or epigenetics among others), a key step is gene transcription [16]. Then, the choice of expression data for GRN reconstruction is often considered of preference. This is why many GRN approaches, termed *influential* GRNs, only consider transcript levels and try to establish direct or indirect relations between them. Ideally, a robust model, closest to the actual biological system, could be created by the integration of *omics* data sets (whole-genome data sets, transcriptomics, proteomics, interactomics, metabolomics, epigenomics or exomics among others) together with other previous biological knowledge.

In this subsection, the two main data sources for GRN reconstruction, Genome and Transcriptome, are addressed. Nevertheless, GRN inference is shifting towards the integration of heterogeneous data, and models become more complex and closer to reality.

2.1.1. The Genome

The term Genome refers to the collection of genes comprised in a biological system. In the past, these collections were limited to protein-coding genes, but the field has been extended to many other elements such as TF-binding regions, microRNAs, or evolutionarily-conserved regions [17]. In the case of TF, the goal is to detect potential links between these and differential gene expression in the cell/tissue.

At the primary archives level, the most important nucleotide sequence databases are: GenBank (USA) [18], EMBL (Europe) and Data Bank of Japan Center, DDBJ [19]. Also, ENSEMBL database reunites information related to mammals' genomes [20] and seeks for a centralized resource for geneticists, molecular biologists and other genomes researchers studying of our own species and other vertebrates and model organisms.

Furthermore, fields like Epigenetics (i.e. the study of influential factors out of classic genetics) enlarge our understanding of the Genome [21]. As an example, in Ramsey et al. [22], the available public data stored at The Cancer Genome Atlas, an important database for human cancer research, is used to construct the underlying GRN and to identify the key role of the RUNX-1 transcription factor in adenocarcinoma.

2.1.2. The Transcriptome

Functional Genomics or Transcriptomics refers to the analysis of gene expression patterns and tries to find relationships between them and their biological background [23]. Transcription is considered the main control mechanism in gene expression, so GRN reconstruction using expression levels is generally of preference [24].

Interestingly, a lot of RNA transcripts do not code for proteins (non-coding RNA, ncRNA), e.g. tRNA. On the other hand, ncRNA play a key role in multiple processes, including gene regulation. These transcripts are single-stranded RNA folded into structured molecule. Prediction of ncRNA (secondary) structure is possible as well as predicting where such ncRNA genes are located in the genome. Microarray-obtained Gene Expression data can be found at databases like Gene Expression Omnibus (GEO) [25] or ArrayExpress [26].

As an example of this, Kang et al. [27] used genome-wide expression data to generate a model able to predict acute rejection responses in kidney transplantation, using clinical trial data as an input.

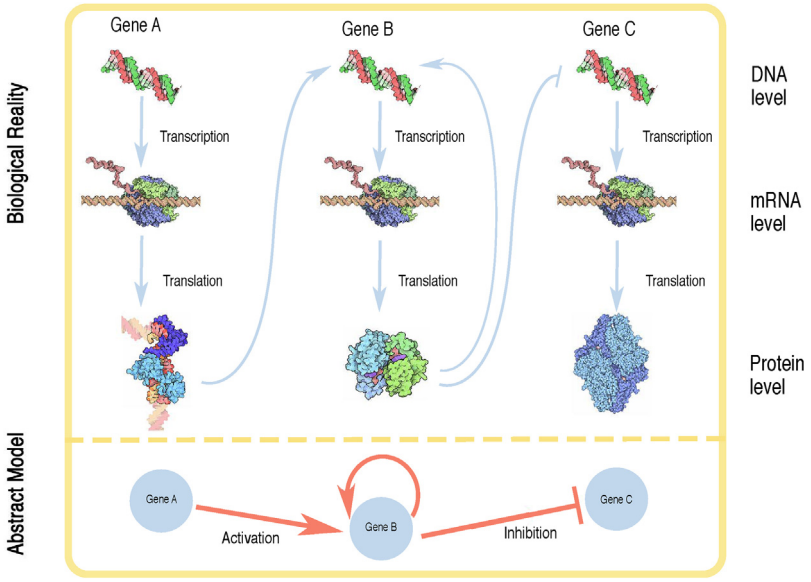


Fig. 2. The main goal in GRNs inference is to generate abstract models for actual biological processes. These models tries to represent complex interactions between molecular entities such as gene activation, inhibition or feedback loops.

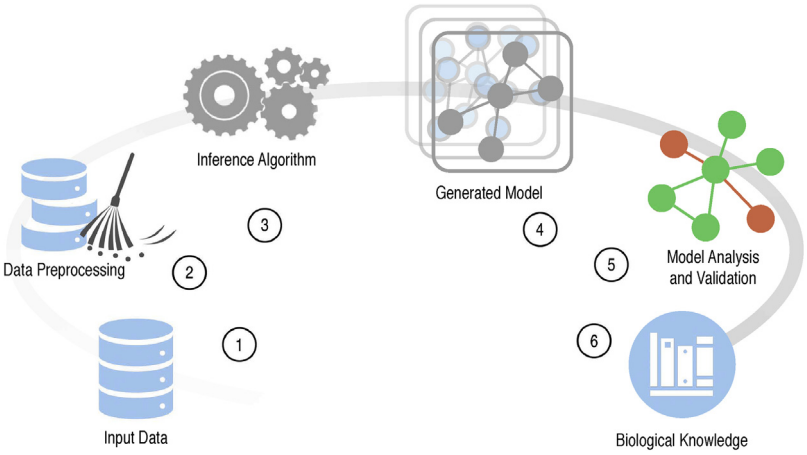


Fig. 3. GRN reconstruction steps based on the KDD workflow.

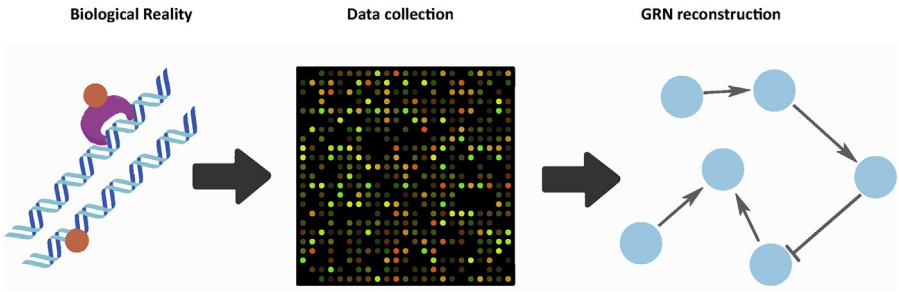


Fig. 4. A global representation of how the GRN are inferred from the biological data to abstract models.

2.1.3. Reading software for the collection of prior available knowledge

Aside from the huge amount of data stored in databases, there is knowledge generated experimentally which can also be integrated into the process of GRN reconstruction. Latest findings in biological data are directly incorporated to databases as mentioned above, some others are described in the literature. For this reason, software has been developed to automatically and reliably extract pieces of information about relationships between molecular elements from the literature [28].

This section does not intend to provide a complete and detailed description of the diversity of available biological data. Additionally, it illustrates the potential benefit as well as the challenge of integrating such diverse and complementary types of biological information to reliably infer GRNs. In the following subsections, some indications on experimental set-up for the obtaining of experimental data are addressed.

2.2. Experimental design to obtain biological data for GRN inference

On a frequent basis, gene expression experiments are performed for GRN reconstruction. Depending on the used approach, quality and quantity of the generated data may vary. Upon the performance of these experiments, two aspects must be taken into account: perturbation and observation of the biological system.

Experimental design includes usually systematic perturbations e.g. shift between different environmental conditions, interventions at the genetic, transcriptomic, proteomic or metabolomic level. As a result, differential expression patterns can be found under imposed conditions. Resultant gene expression patterns can be compared to a non-perturbed profile provided by a presumed model of the network. Thus, estimated *goodness* of the model can be then assessed [29].

Transcriptome level perturbations can also be used for GRN inference, for this, molecular techniques are often applied, e.g. RNA interference (iRNA) can knock-down other RNAs to their degradation in order to study the effect of their loss. Measuring altered gene expression levels provides some insight on the influence of the molecular elements involved, and it is key to model construction. For example, REDuction of UnCertain Edges (REDUCE) algorithm finds optimal gene knock-out experiment for inferring directed graphs of GRNs [30].

Measurements on perturbation experiments can be performed in a static (steady state) or time-course situation, the latter involves the use of dynamic programming. Depending on the knowledge to be achieved, the experimental set-up will vary and so will the choice between a static or a dynamic GRN architecture:

- Generation of static data comes with the assumption of an equilibrium or steady-state situation of the biological system. Depending on the case, the steady-state choice may miss critical dynamic events for reliable GRN construction i.e. dynamic changes occurring with time, as in Kim et al. [31].
- On the other hand, time-series experiments, where samples are taken in a series of time-points after perturbation, constitute the dynamic approach [32]. The experimental set-up determines the number of time-point measurements, thus, the data amount.

2.3. Data requirements

Reliable GRN reconstruction requires a considerable minimum quantity of accurate data. On the other hand, experimental costs and efforts also need to be minimized. Generally, the more nodes involved in the network, the more data will be needed.

In order to provide real biological insights, networks are conventionally based on experimental data. However, experimental data may not be useful for this aim for two main reasons: (i) The data collected from the experiment shows bad quality and (ii) data is unavailable.

Regarding the first case, model quality is proportional to biological data set quality. Alterations on the biological input (measurement noise or inappropriate experimental design) may lead to unreliable GRNs. Models could either estimate high-confidence gene regulatory interactions or just speculative dependencies, but depending on the complexity of the model, they may consist of many parameters and be more data-demanding. However, depending on the information one aims to obtain from the model, the precision required for the data may vary [9]. To cope with this, the inference strategy may use external prior knowledge from databases and literature, so experimental data required will depend on this knowledge and the ability of the used algorithm to integrate this information in the modelling process.

If data is not available, e.g. it is not possible or difficult to obtain, some approaches make use of fuzzy logic to infer missing data. In Bordon et al. [33], fuzzy logic is applied to assess quantitative values to an incomplete kinetic dataset for gene repression. Besides, GRN can still be reconstructed out of in silico generated data. For example, SiGNet (Signal Generator for Networks) is a Cytoscape app that simulates experimental data for a signalling network of known structure [34]. Despite not providing biologically-meaningful networks, this approach is often used for algorithm testing and training, so it can be put to work on an actual database afterwards.

Furthermore, model quality does not only depend on data quality, but also on the inference algorithm, which could vary the efficiency of GRN inference out of the same data set. It is necessary to find precise model parameters using heuristic approaches, which may perform sub-optimally.

As explained before, there is a direct relation between the complexity of the model and the required amount of data, but this certainly has a limit. Dimensionality problems come with difficulty of finding an accurate model (dynamic, large-scale, complex), when the size of the search space increases dramatically [35].

2.4. Data pre-processing

Data pre-processing, prior to GRN inference, is a key step for GRN reconstruction and quality of outcome. Methods for this aim will depend on the type of data and the experimental design.

As mentioned by Hecker et al. [9], there are two main sources of variability in GRN reconstruction: systematic errors (bias) and stochastic effects (noise). To ensure quality of the data and GRN outcome, fundamental analysis is applied, including noise filtering, system affect detection, etc. [36].

Systematic effects can be nearly removed through data normalization, since some genes expression can be very variable in one cell/tissue type. For this, housekeeping genes (relativity constant expression patterns) are often taken into account [37]. On the other hand, replicates performance provides with repeated measurements to reduce stochastic effects.

Finally, further data manipulation may be necessary, whose specificity will depend on the network inference methods e.g. dynamic programming requires the estimation of time derivatives for each measurement point of the time-series. Another example is the case of Boolean networks (see Section 3), which needs the conversion of measured expression levels into binary data prior to network inference. In the following section, the main algorithms for GRN reconstruction are explained.

3. Computational approaches for GRN construction

The term model architecture refers to the logical thinking underneath the GRN reconstruction. In the form of an algorithm, model architecture describes the behaviour of the regulatory dependencies between the biological components involved, basing on multiple other parameters.

The choice of a certain model architecture will shape the resulting GRN. Biological data has to be analysed so network structure (interactions between components) and model parameters (intensity and type of interaction) can be learned from it. Together with the increasing amount of biological data, the needs of analysis have led to the proposition of novel algorithms for GRN reconstruction.

In the following subsections, the main GRN inference methods are exposed, in an attempt to explain the basic thinking underneath each approach. Main GRN inference methods can be summarized in: (a) Information theory, (b) Boolean networks, (c) Differential equations models, (d) Bayesian and (e) Neural models. Also, a comparison between these methods is shown.

3.1. Information theory models

Information theory-based networks are the most common type of networks due to their computational simplicity. They are also called co-expression networks since they establish gene–gene relationships if the

dependence level between both gene is above a threshold set in advance. The higher the threshold, the sparser inferred GRN [38]. The main measures to determine the dependencies between genes are the correlation coefficients like Pearson, Spearman or Kendall coefficients. However, different measures like Euclidean distances or mutual information, were also applied for the inference of GRN. These models are suitable to cover different aspects of cells, which are here understood as time-varying living systems which perform complicated processes inside and between them.

An advantage of information theory models is the discovery of large GRN from low expression data due to their low computational cost. Among the main proposals based on this model for GRN reconstruction we can highlight: REVEAL (The REVerse Engineering ALgorithm) [39]; RELEVANCE [40]; ARACNE (Algorithm for the Reverse engineering of Accurate Cellular Network) [41] or ARACNE-based algorithms [42]; CLR (Context Likelihood of Relatedness) [43]; MRNET (based on the maximum relevance/minimum redundancy) [44] or FyNE [45]. These approaches apply correlation coefficients, Euclidean distances or information theory scores such as mutual information and conditional mutual information for the identification of gene interactions. Also, supervised-DTI and non-supervised-DTI approaches, based on the directed information (DTI) metric, are used [46]. Other algorithms use the minimum description length (best explanation of the data gathered with limited number of observations) to determine a threshold value [47].

Algorithms often undergo an improvement process for a more precise data analysis. This is how novel algorithms work, as: MI3 [48], based on three-way mutual information; SRI-CLR [49] adding the synergy index to CLR algorithms; MRNETB [50] combination of backward selection and sequential search; C3NET [51] which selects the highest mutual information value among the neighbors of each gene; or CMIP [52], which applies conditional mutual information from the lower order to the higher order.

Analogously, graphical Gaussian models (GGMs) are used to represent conditional dependencies between variables and allow distinguishing direct form indirect associations [53]. Also, mutual information models infer regulatory gene interactions basing on pairwise mutual information [54]. Two algorithms commonly used for network inference based on pairwise mutual information are ARACNe [55] and CLR [49].

Overall, main advantages of these models are the simplicity, low computational complexity and low number of required samples [56]. Therefore, this kind of network fits perfectly with the construction of large gene networks. On the other hand, these models are static and do not take into account multiple genes participating in a regulation. As an example, in Wang et al. [57], a total of 14 genes were identified as hubs/nodes in the regulation of postmenopausal osteoporosis disease, by means of an information theory-based approach.

3.2. Boolean networks

Boolean networks are easy to implement and allow capturing the actual dynamical behaviour of GRN. They are able to describe biological phenomena such as oscillations, multi-stationary events, long-range correlations, switch-like behaviour stability and hysteresis [58].

Boolean networks represent genes by variables and their expression level is discretized into Boolean binary values (by clustering and thresholding [59]): '0' for low values (silenced or nearly-silenced genes) or '1' for high values (activated genes) [29]. Thus, a gene i is represented by *on/off* expression values. Operators of logic (and, $\wedge$; or, $\vee$; and not, $\neg$), link Boolean variables in a Boolean function.

Boolean functions reconstructing the network compose directed graphs $G(X, F)$, where X represents a variable, which is associated to other variables by means of the function F [58]. The so-called state of the network (S) for a time t is represented by the values of all different nodes:

$$S(t) = \mathrm{xi}_1(t), \mathrm{xi}_2(t), \ldots \mathrm{xi}_n(t)$$

Although straightforward and simple, Boolean networks find their main limitation in the discretization step. Gene expression is rarely a matter of fully-activation or fully-silencing, since there are often uncountable different gene states in between. Thus, important details of system behaviour might be lost. Boolean networks also need to cope with noisy data problems [60], as the accuracy of thresholding will certainly determine network topology. Nevertheless, Boolean networks are easy to interpret and they offer a simple dynamic approach for GRN.

Considered as the simplest GRN inference approach, Boolean networks have been proven useful in many cases. In Simak et al. [61], regulatory relations and potential candidate biological functions are explored from *Saccharomyces cerevisiae* transcriptomic data. Boolean network function performance was validated using *Saccharomyces cerevisiae* time course data, and its outcome was consistent with all different cell cycle stages. Also, in Claussen et al. [62] a dynamic Boolean network model is used to infer interactions between low-abundance species in human gut microbiome, which are often overlooked. The model uncovered synergistic and competitive interactions between these species. In Polak et al. [63], regulation of immune responses in primary Langerhans cells were analysed using a Boolean GRN model, whose predictions were experimentally validated afterwards. In Moignard et al. [64], a GRN model for blood development was reconstructed and used to predict the role of some transcription factors, later validated experimentally. Finally in Orlando et al. [65], a Boolean network is used to model yeast cell cycle. This model works as an oscillator reflecting a sequenced transcriptional program.

3.3. Differential equations models

Ordinary differential equation (ODE) approaches use continuous instead of discrete variables. This leads to a more accurate model, and enables the dynamic modelling of gene regulation. Differential equations will then represent changes in gene expression as a function of other genes expression and taking into account environmental factors, allowing a quantitative modelling (closer to actual behaviour of the biological system) [58].

Modelling of gene expression dynamics through ODE often can be represented by the equation:

$$\frac{\mathrm{dx}_i}{\mathrm{dt}} = f_i(x_1, x_2, \ldots, x_N, p, u)$$

where x represents the expression level of a certain gene i at a given time t. N represents the number of genes involved, u refers to external perturbations of the system and p to the parameter set of that system. Then, f is the function, which describes the change rate of the state variables x_{1-n} depending on p. In ODE models, continuous-time variables with constraints are used, and negative values are not allowed, i.e. the assumption of protein and mRNA molecules degradation being unregulated [58].

There are multiple solutions to ODE systems if no constraints are assessed. This is why, specifications of the f function and constraints representing prior knowledge (simplification, approximations or educated guesses among others) are required for the identification of model structure and parameters. A disadvantage of many ODE models is that these consider only linear models or just specific types of non-linear functions [66,9], while regulatory processes are often characterized by complex non-linear dynamics. More complex variants of ODE-based models are stochastic differential equations models which take into account the stochastic nature of GRNs [67]. Moreover, ODE models cannot cope with large GRN modelling and value estimation for model parameters results hard in some cases due to their computational complexity.

ODE-based models have provided outstanding results in GRN inference. As an example, in Matsumoto et al. [68], SCODE algorithm

(Single Cell Ordinary Differential Equation) was applied to provide some insight in GRNs related to differentiation processes. Single cell RNA-Seq is performed on individual cells so differences on their expression patterns can be evaluated. Deng et al. [69], proved the efficiency of ODE models in dynamic GRNs reconstruction. Improvement on calculation scheme of the derivatives and data pre-filtration lead to improved scalability to large GRNs. However, combinations between different methods have been proven to be useful depending on the case.

3.4. Bayesian networks

Bayesian networks are one of the most used GRN inference architectures. They make use of the Bayes theorem of probability, then combining probability and graph theory to qualitatively model the properties of GRNs [58].

Bayesian networks are generally directed and acyclic graphs (DAGs) $G = (X, A)$ [70], which characterize the joint distribution of nodes as a series of local probability distributions (P). X represents genes/nodes ($x_1, x_2, \ldots, x_n$), or gene variables and A refers to the directed rod corresponding to probabilistic dependency interactions between these genes. The joint distribution of the variables in a Bayesian network is described in Chai et al. [58] by:

$$P(x_1, x_2, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(x_i))t$$

where x_i is a node, n refers to the total number of involved genes, Parents (x_i) are all *parent* genes regulating the *child node* (gene x_i). Parameter P describes the Conditional Probability Distribution (CPD) or local distribution for the node i. Bayesian networks make use of the Markov assumption [71], which refers to the memoryless property of a stochastic process: given its parents, each node is independent of its non-descendants.

Bayesian networks imply a set of conditional dependencies so the algorithm is able to infer a DAG from them. Since many DAGs can be inferred from the data set (D), the algorithm has to find the best DAG (G) describing the data set and each graph (G) is evaluated through a Bayesian score.

$$S(G: D) = \log P(G|D) + \log P(G) + \text{constant}$$

In Larjo et al. [72], methods for learning Bayesian networks are detailed, having this learning three essential parts: (i) *Model Selection*, DAGs evaluation as candidate graphs of relationships; (ii) *Parameter Fitting*: given graphs and experimental data sets, find the best conditional probabilities for each node; (iii) *Fitness Rating*: scoring of each candidate model so the higher the score, the better the model fits to the data. The model with the highest score represents the GRN resulting from the inference. Learning of Bayesian networks can be based either on discrete (often Boolean) or continuous expression levels. Thus, the probabilistic model underneath could be a multinomial or a Gaussian distribution [73]. Multinomial variables take a finite number of possible values but in the case of continuous variables, in a classical approach, data would be discretized. Bayesian networks are hard to infer from continuous data since it requires large computational power. On the other hand, data does not require discretization.

Flexibility is the main advantage for Bayesian networks, since they combine different types of data as well as prior knowledge for reliable GRN inference [74]. Bayesian approaches are often considered of preference while inferring dynamic GRNs and have shed some light in many fields ranging from evolutionary-development to medicine. In Acerbi et al. [75], continuous time Bayesian networks were used for the successful revelation of well-known regulatory mechanisms in Th17 cells differentiation, also providing some new biological insights. Chekouo et al. [76], proposed a Bayesian model to identify micro RNAs and their target genes obtaining a valid algorithm for kidney cancer biomarkers identification. Also in Chudasama et al. [77], novel cancer biomarkers were identified thanks to Bayesian networks, which may support clinical practice and improve long term outcomes.

3.5. Neural networks

Inspired by animal central nervous systems, these models comprise two main approaches: Artificial Neural Network (ANN) and Recurrent Neural Networks (RNN). The first is purely neural, whereas the second also involves fuzzy logic [78]. RNN is a successful method for GRN inference, since it enables modelling of non-linear and dynamic interaction among genes [79]. Neural models allow continuous variables and their outcome looks similar to the neural connections observed in natural processes.

As described in Chai et al. [58], neural network models can be represented by:

$$\frac{\mathrm{d}e_i}{\mathrm{d}t} = \frac{1}{\tau_i}\left(g\left(\sum_{j}^{N} w_{ij}e_j + \beta_i\right) - \lambda_i e_i\right)$$

where w_{ij} refers to the type and concentration of the relation between genes at the ith and jth position. The reaction decay rate parameter is represented by λ_i, the basal expression level is indicated by β_i and e_i represents the gene expression level for the ith gene. Function g indicates the regulatory effect on each involved gene, defined by a set of weights such as w_{ij}. The computed weighted sum of all potential regulating genes is considered as the regulatory effect on a particular gene. A scoring function is also applied for evaluation of the outcomes, network performance optimization and error minimization.

In Ling et al. [80], a RNN was used to study p53/Mdm2-mediated response to DNA damage. Tong et al. [81], used ANNs to study gene–gene interactions for biomarkers in childhood sarcomas. Siddens et al. [82], used fuzzy neural network models to predict polycyclic aromatic hydrocarbon-mediated perturbations of the Cyp1b1 transcriptional regulatory network in mouse skin. Finally, in Rubiolo et al. [83], a supervised neural model called Extreme Learning Machine (ELM) was successfully used to reconstruct GRNs, even surpassing many commonly-used approaches.

3.6. Additional network architectures

Although GRN reconstruction is usually tackled by means of one of the models above (or a combination of them), there are some other GRN inference approaches that clearly differ from these methods. As an example of this, in Liu et al. [85], fuzzy cognitive maps and a dynamical multi-agent genetic algorithm with the decomposition-based model were used to deal with large-scale GRN inference. Also in Thiagarajan et al. [90], the power of Graphics Processing Units and parallel reverse engineering algorithm was used for the identification and simulation of genome-scale GRN, which pose computer intensive problems. Finally in Ud-Dean et al. [86], an algorithm called TRaCE+ is used for GRN reconstruction, being able to indicate positive or negative regulations, and reduce uncertain edges.

To sum up for this section, a comparison between all 5 exposed inference methods is shown on Table 1.

In Section 4, some important issues on the application of the inference algorithm are explained such as: feature selection, parameter estimation, structure optimization and integration of prior knowledge.

4. Network optimization based on machine learning algorithms

A naïve approach upon the GRN inference process would be the one of enumerating all possible DAGs for a given number of nodes, which is deemed brute-force search. However, the amount of possible DAGs for a given number of nodes grows exponentially, making this search problematic. Therefore, heuristics or/and constraints need to be applied to make the process more efficient.

Table 1
Comparison between different model architectures with some examples.

Computational approach	Strengths	Weaknesses
Information-theory models	• Large GRNs, even out of low expression genes • Mutual and conditional mutual information approach • Not computationally-demanding • Low number of samples	• Regulation by multiple genes is not considered • Static, only suitable for steady-state data
Examples:	REVEAL [36], RELEVANCE [40,84], ARACNE [42], CLR [43], MRNET [44]	
Boolean models	• Capable of inferring large networks • Generally easy to interpret • Simplify underlying complex biological phenomena • Allow supervised learning methods	• Deterministic nature • Discretization bottleneck (only on/off states) • Problems in handling incomplete or inconsistent expression data • High computing time • Most of them use small number of genes
Examples:	RCGA [85], TRaCE+ [86], CABeRNET [87]	
ODE models	• Directed signed graphs • Realistic dynamics • Suitable for both steady-state and time series expression data • Simplification of the system by means of linear functions • Allow prediction of the behaviour of the network under different conditions once parameters are known	• Not suitable for large networks • Linear functions also constrain the dynamic behaviour of cell regulatory functions (e.g. oscillations, multistationarity) • Hard to find appropriate values for model parameters • Noisy data leads to qualitative instead of quantitative GRN inference
Examples:	SCODE [68], HiDi [69]	
Bayesian models	• Noise and uncertainty handling • Do not require a large number of involved variables • Integration of prior knowledge and allowance of enrichment analyses • Statistical inference of gene network	• Feedback loops are not allowed • Fail in the inference using time series expression data • Cannot cope with large GRNs • Inherent combinatorial learning
Examples:	F-MAP [88], MDP [89], POMDP [71], QMR-DT [73]	
Neural models	• Recognize an input pattern • Model any functional relationship inferable from the data • Suitable for both steady-state and time series expression profiles • Noise handling and biologically plausible • Manage non-linear and dynamic behaviour	• Machine training experiments are hard to perform since every situation requires a different learning rate definition • Computational complexity makes them more suitable for very small systems
Examples:	ANN [27], RNN [78], ELM [83]	

Leaving the model algorithm aside, the inference approach needs further adjustment for reliable network reconstruction. First, a dimensionality reduction has to be performed on the large gene expression data in order to reduce computational cost [91,92]. The term *curse of dimensionality* was introduced by Bellman [93] to refer to the exponential growth of required amount of data to provide a reliable analysis. Some strategies to face this issue are:

- Reducing the number of nodes focusing only on features like genes or proteins of interest employing feature selection or feature mapping methods.
- Restricting the number of model parameters: using simple models and network connectivity constraints such as sparseness. This minimizes the number of edges in the network thus reducing the number of model parameters.
- Integrating specific prior knowledge about network structure which increases the model and provides higher quality results.

4.1. Feature selection and feature mapping

Modelling biological system requires assumption-making to focus only on those specific aspects which are important for aim of the study. Feature selection and feature mapping help reducing model complexity by excluding non-relevant features in GRN inference [58].

Upon feature selection, non-responsive or not well measured genes are removed from the data. Machine learning algorithms are often affected by data noise, which should also be reduced to avoid unnecessary model complexity. On the other hand, in feature mapping redundant information is removed. For this, different molecular entities in a functional element are combined when they represent the common behaviour that reflects a particular biological function, thus clustering nodes and reducing complexity [94].

Dimensionality has to be decreased using an appropriate approach so the network is still large enough to provide a significant result on the biological phenomena under study. Some of the approaches consist in the filtering of differentially expressed genes and the clustering those genes that are co-expressed. Differentially expressed genes, are often the only ones considered in some experiments, and are identified thanks to methods like DESeq2 (Love et al. [95]) or Limma (Ritchie et al. [96]). Moreover, these techniques try to reduce the number of model variables.

Feature selection seeks for the minimization of the number of estimated parameters in order to improve performance and generalizability of GRNs, solely using the data to deduce dependencies [94]. A general approach is to consider only genes showing significant expression changes under the studied experimental conditions.

According to Hira and Gillies [91], there are three main categories of feature selection strategies to be considered: (i) *filters*, which extract features from the data (without any involved learning); (ii) *wrappers*, which select useful features through learning techniques; and (iii) *embedded techniques*, combining feature selection and classifier construction. Clustering of co-expressed and/or co-regulated genes or proteins has been conventionally used for the reduction of network components. These are ultimately genes or proteins showing similar expression patterns or genes belonging to the same pathway or biological function.

As stated before, experimental gene expression data is often complemented with prior knowledge obtained from biological databases. In this enrichment analysis, genes are selected and compared with previously-determined functional gene groups. Enrichment analyses normally make use of GO terms, Gene Ontology Annotations (GOA) and databases like KEGG Pathways, OMIM and Gene Prospector [97].

Alternatively, Knowledge Driven Variable Selection (KDVS), performs discovery methods and enrichment analyses at the same time, thus enhancing results interpretability. An example of knowledge-driven enrichment analysis can be found in Sun et al. [98], where an enrichment analysis of a gene/protein network for primary

myelofibrosis is performed, basing on gene expression using KEGG pathways and OMIM database.

4.2. Biological network parameters

GRN structure has been proven to be sparse, and following a scale-free topology [99]. The latter means that nodal degree distribution of the GRN is a power law distribution.

Model structure and model parameters estimation are two main tasks of network inference. The idea behind these tasks is to apply a learning algorithm to fit the output of the mathematical model to the provided experimental data [29]. The choice of a network topology is another important step in GRN reconstruction. Herein, Radcliffe [100] studied the application of form analyses to GRN topology related problems.

Structure optimization is thus the process which finds the network topology (connectivity) that best explains the experimental data, taking into account constraints imposed by the available knowledge. A key point here is network sparseness. On the other hand, parameter optimization seeks for the identification of the best model parameters for a given model structure. As detailed in Valverde et al. [99], several GRN properties have to be taken into account for network reconstruction:

1 *Sparseness*: a gene is usually regulated by a small and limited number of genes. Then, regulatory inputs per node are limited. Exceptionally, the so-called master genes control large parts of the network (high out-degree) [16,101]. Generally, as network connectivity increases, the data will be better fitted by the model. Nevertheless, higher-connected networks bring several difficulties. Thus, there is a compromise between model quality and model complexity. Sparseness helps finding the most likely combination of regulators [102].
2 *Scale-freeness*: the distribution of node degrees in a GRN tends to have the form of a power law. In those scale-free networks most genes are sparsely connected whereas a few are highly connected.
3 *Highly-structured*: networks can be decomposed in modular components that consist of only a few genes, which follow regular hierarchies.
4 *Modularity*: this means, when an ensemble of genes cooperates within a same specific function [103]. This can be observed by clustering co-expressed genes.

4.3. Scoring functions for structure optimization

Explicit structure optimization methods compare different topologies of GRN models by means of a scoring function, which helps achieving network sparseness. Several scoring criteria have been developed for the different inference methods [73]. Gene interactions are added or removed in order to obtain a better-scored topology.

Brute-force search, where all possible combinations of interactions are tested, is only possible for small networks, on which strong restrictions are also applied. The total number of combinations for each node can be estimated as $2^n - 1$, where n is the number of involved nodes. A different optimization strategy has to be applied if this is not the case.

Heuristic methods apply educated guesses to lead the search to the most likely solution [78]. Some examples of search techniques are: breadth first search [104], beam search [105] and hill-climbing [106]. Search techniques add or remove connections in the network. The three main search techniques are: *Forward Selection* (Growing), which starts from a simple model and most important interactions are added first up to a certain limit; *Backward Elimination* (pruning), which starts from a highly connected model and less significant interactions are removed; and *Stepwise Selection*, which combines both previous approaches. As an example, in Gómez-Vela et al. [107], a greedy approach is used for structure optimization of densely-connected networks.

Finally, for a given modelling architecture and a network structure optimization strategy, one can infer the model from the data. Besides, the optimization strategy may be supported by prior knowledge (see Section 4.4).

4.4. Integration of diverse biological information

Modelling of GRNs has to be comprehensive and integrative, thus, as repeated throughout this review, the use of prior available knowledge (other experiments, databases, scientific literature, etc.) is a key step to reduce combinatorial complexity [14].

Knowledge-based modelling approaches are the most robust for realistic GRN inference [14]. Prior biological knowledge hampers biologically plausible assumptions, supporting the reverse engineering process. For example, many inference algorithms make use of the publicly-available data stored at GO and/or KEGG databases, as in the case of Zhu et al. [108].

Biological knowledge can be integrated into mathematical modelling and taken into account for network reconstruction, thus providing accuracy to the process. Methods integrating information from ChIP, microarrays or RNA-Seq experiments are considered of preference. There are ChIP-Seq databases from specific cell types providing with information about transcription factors or epigenetic and transcriptional landscapes. DNA data sets providing with DNA structures, sequence conservation and patterns can also be integrated in GRN inference. Most biological processes involve combinatorial contributions of transcriptional regulation, alternative splicing, post-translational modifications or protein-protein interactions. The Biological General Repository for Interaction Data sets (BioGRID) is a database of reference aiming to annotate protein, genetic and chemical interactions for all model organism species and humans [109].

However, this is a challenging process and the comprehensive integration of the available heterogeneous data may turn into a complex horizon. One of the attempts for this aim is The Cancer Genome Atlas (TCGA), which catalogues and discover major cancer-causing genomic alterations [110]. This platform is based on microarrays and NGS methods such as: RNAseq, MicroRNAseq, DNAseq, SNP-based platforms, array-based DNA methylation sequencing or reverse-phase protein array. As expected, for the integrated-multi-dimensional data analysis, comprehensive exploration is required, for which promising tools are arising. In Pineda et al. [111], several omics data are merged in an integrative analysis for a more comprehensive study of bladder cancer, by using a Global-LASSO. Also in the work by Salehzadeh-Yazdi et al. [112], metabolomic and epigenetic data are integrated in GRN modelling. Finally, in the work by Sinha [113] an improvement on inference accuracy is shown when integrating epigenetic information on a Bayesian model.

Integrative learning approaches start from a template network built out of databases and literature (real network topology) and then, an inference strategy is applied fitting the model to the data and taking into account the template [29]. The template information also known as training data can also be incorporated into the GRN inference process, and it can also be used to constrain explicit search methods [37].

Learning models use training data sets in order to build generalizable models. However, an important parameter to take into account is the data set shift, since most algorithms assume training and test data to be drawn from the same distribution. Data set shift may result in suboptimal fitting of the model. Data normalization and batch correction techniques help to cope with data set shifts. To make sure that two data sets are drawn from the same distribution, quantile normalization is applied, which normalizes target distributions to a reference distribution [37].

5. GRN validation and appraisal of inference methods

Once the final network is obtained, its biological significance has to

be tested. Not all GRN-predicted interactions are biologically meaningful. In general terms, an interaction between a gene and its presumed regulator is considered biologically meaningful when the disruption of one or both of these elements triggers a change on gene expression. The identification of meaningful interactions requires network validation, so further support can be obtained for them [114]. Also the other way round, models should be predictive, meaning this able to generate plausible biological interactions which may be afterwards proved right [10]. However, lack of validation does not necessarily mean 'not-biologically-meaningful' interactions, since the validation methodology plays a crucial role and/or many interactions may have not been described yet.

According to Dougherty and Qian [115], there are two main issues regarding network validation: (i) whether the inferred network provides good predictions on the experimental data (scientific validation) and (ii) whether the applied inference algorithm within a certain network model framework yields networks that are accurate relative to some criterion of goodness (inference validation). The boundaries between both approaches actually blur in practice, since validation of an inference model requires then scientific validation of the inference, and results of the later may be used to improve the inference method [10].

Network validation assesses the quality of the inferred GRN supporting on available knowledge. For this aim, scoring methodologies are often applied to obtain a quantitative evaluation of the model with respect to the information used for its generation (internal validation) and other information (external validation). Also, web-based frameworks have been developed to support GRN validation, as in the case of Genotet [116].

5.1. Quantitative evaluation of inference performance

GRNs can be evaluated using scoring methodologies which allow the comparison between different networks. The process is described in Fig. 5, where the inferred network is compared with a reference network (Gold-Standard, closest to reality to the general knowledge) obtaining a quality measure [117,118]. A Gold Standard enables the estimation of several metrics which would jointly provide an evaluation of model's goodness. This is certainly a key point in GRN inference, since data may provide a massive amount of possible interaction and only a few of them are deemed true [119].

GRN inference relies on the data set handling method, which often cannot depict all interactions. For example, ChIP techniques may cause only the strongest (or most represented) interactions to be considered in the model, leaving many (less represented) others behind. Aside from the experimental techniques limitation, relationships may be missed due to high thresholding, but also due to low expression levels or non-variant behaviour with respect to target genes. Thus, many equally-biologically-meaningful interactions are missed, false negatives (FN) [114].

Conversely, false positives (FP) are also incorporated into GRNs. FPs are deemed technical when the inferred interaction is just sporadic in nature and it is not retrieved even with the same assay and conditions. Technical FPs, depend on assay robustness so the used approach has to be optimized. FPs are deemed biological in case of not biologically meaningful interactions, which are robustly detected. This can be the case of both a TF and its target being regulated by another TF which does not change its expression. Sub-optimal specificity of antibodies in ChIP experiments, which results in binding to lower affinity or non-specific sites, may also result in a FPs gain.

Finally, true positives (TP) and true negatives (TN) are described according to previous knowledge, since these are only considered when a particular interaction has been proved experimentally.

In order to evaluate model quality, it is necessary to analyse if the model correctly predicts the GRN behaviour or if the model represents the true structure of the system. Statistical measurements are used to compare inferred models with the actual behaviour of the network [120]. This is the case of supervised network inference, where part of the actual network is used for model training and optimization.

According to Schrynemackers et al. [121], when the true Gold Standard is known, the inferred networks structure is compared to the first one using several metrics:

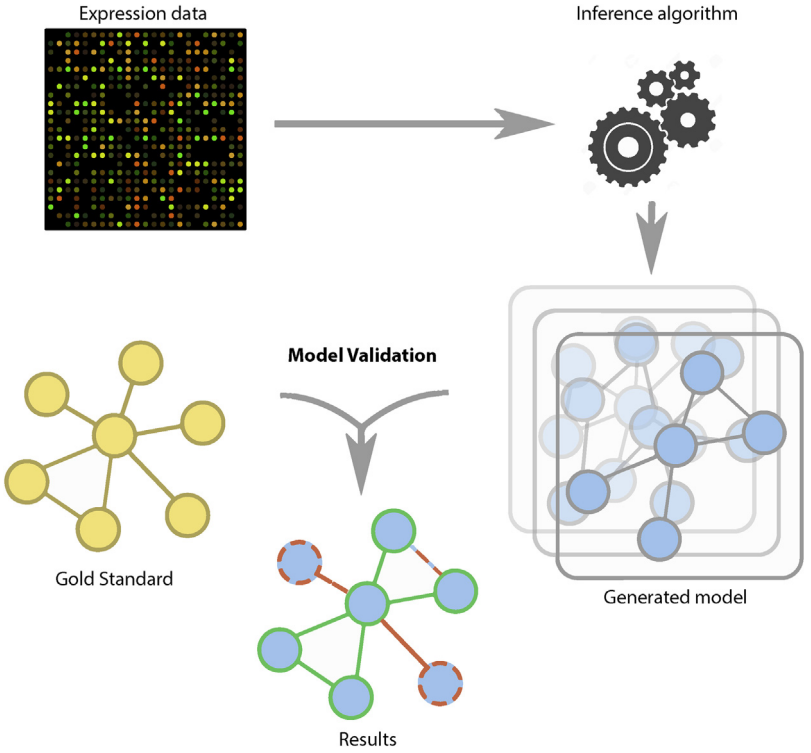


Fig. 5. Validation process scheme. This process usually offer a quality measure of the validated network.

- *True positive rate (TPR)*, *sensitivity* or *recall* is the number of TP divided by the actual number of positives: TP/(TP + FN).
- *True negative rate (TNR)* or *specificity* is the number of true negatives (TNs) divided by the actual number of negatives: TN/(TN + FP).
- *False positive rate (FPR)* also deemed as 1 − *specificity* is the number of FP divided by the number of actual negatives: FP/(FP + TN).
- *False negative rate (FNR)* or *miss* is FN divided by the number of actual negatives: FN/(FN + TP).
- *Precision* is the number of TP divided by the number of predicted positives: TP/(TP + FP).
- *Rate of positive predictions (RPP)* is the number of predicted positives divided by the total number of negatives and positives: (FP + TP)/(N + P).
- *F-score* is a measure of model's accuracy, and it is calculated as the harmonic mean of precision and recall:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

These metrics are combined in the analysis of the inference performance. For this aim, several curves are displayed:

- *Receiver operating characteristic (ROC) curves:* plot TPR by means of the FPR, when the confidence threshold is varied from maximum to minimum confidence score [122]. For network comparison, ROC curves are integrated single real number by measuring the area under the ROC curve (AUROC). AUROC directly compares the inference quality against a random prediction. An AUROC of 1 corresponds to a perfect classifier and an AUROC of 0.5 corresponds to a random classifier. The higher the AUROC, the better the predictions. It is important to achieve low FPR for better precision [120,123].
- *Precision-Recall (PR) curves:* represent precision by means of the recall, when the confidence threshold is varied. An ideal classifier would provide a curve which passing trough (1,1), whereas a random classifier whereas a random classifier would rather provide a plateau [124].
  Analogously to ROC-curves, PR-curves are summarized in a single real number estimated by integrating the area under the curve (AUPR), which represents the average precision of the inference algorithm. The higher the AUPR the better the classifier is [125].

An advantage of ROC curves is that they do not depend on the ration between positives and negatives, whereas PR-curves do. On the other hand, PR curves are better to assess the method performance whilst the P/N ratio is close to the expected ratio when practically applying the model [126].

ROC and PR curves are both used to compare the performance of different inference algorithms, as in the case of De Smet and Marchal [119], Prill et al. [127] or Hase et al. [128].

It is worth mentioning that the presented scoring methodology is used mainly in Synthetic/Biological data-based network validation, which is described in the following subsections.

5.2. GRN inference algorithms performance evaluation

When a new GRN inference method is released, it is normally compared with the main former ones. Thus, GRN inference methods comparison and evaluation are required to find outperforming methods.

Former methods may include well-known, which have been proven useful at many applications. E.g., although outdated, ARACNE is still today often considered a method of reference.

Nevertheless, comparative studies are delicate and their results may be misleading. As an example, input data may vary between them (time series, large microarray or subset of gene expression levels datasets among others), thus impeding a standardized comparison methodology.

However, important contributions have been made in the literature in order to provide a reliable comparison between different inference methods. For example, in Bellot et al. [129], a good part of the above GRN inference methods are compared. In this work the authors present a software (a Bioconductor package called netbenchmark) that is able to carry out comparative between different inference methods. Despite this, they found that no single method is the best across different sources of data as mentioned above.

Also in Chen and Mar [130], a comparison between several GRN inference methods is performed demonstrating these methods, which were originally designed for bulk samples, do not suit biological relating when dealing with single cell expression data. Finally in Marbach et al. [118], the performance of different inference approaches was also analysed, revealing differences between model architectures depending on the case.

There are also worth to mention the DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenges 3–5. In Marbach et al. [131], the authors present a method for the realistic performance assessment of GRN inference methods. Also in Prill et al. [127], an evaluation of GRN inference methods performance is carried out highlighting best strategies. Another interesting result is presented in [132] where the authors present a method and tool for generating in silico benchmark and performance profiling of GRN inference methods called GeneNetWeaver. Finally in Marbach et al. [118], it is shown that single GRN inference methods perform sub-optimally across multiple data sets, when compared to the integration multiple GRN inference methods.

5.3. Synthetic data validation

Some techniques consider the generation of synthetic data, also used to analyse GRN inference algorithm performance like the afore mentioned GeneNetWeaver. Gene expression values are simulated and then used as input for the GRN inference algorithm. The performance of the algorithm is tested by comparing the output network with the actual network inferred from the literature. However, this approach has some drawbacks since it cannot be used to assess biological significance, which requires making use of actual prior knowledge. Novel multimodel approaches integrate experiments based on both real-life biological and synthetic data sets which provide a higher precision for the inference [71]. Some examples of these tools are RegNet [133] or SynTReN [134] are used for this aim.

5.4. Biological data validation

Model predictions are subjected to external information that was not used in the modelling process, which can be found in the literature and databases. External validation often employs text-mining approaches and it is used to compare different network inference methods. Some tools like GeneNetVal [135], GNC [136], GFD-Net [137] or RefNet Builder [138] use the biological information stored in databases like KEGG Biogrid or GO to assess the model goodness through a direct comparison between the computed network with a gold-standard.

However, GRN predictive capability, can be influenced by experimentally-found interactions that are retrieved [114]. Frequently, several interactions would not have been described yet. Thus, experimental approaches for verification should be taken into account, since these have some limitations and scientists need to design a method for a particular interaction to be verified [10].

6. Conclusions

As discussed throughout this manuscript, GRN inference basing on

large-scale data is a major challenge for systems biology which has gained relevance over the recent past years. More complete biological insight will be gained in the upcoming years, when new techniques are developed for the integration of complex omics such as the epigenome or microbiome, which are trends of current research.

There is a huge variety of approaches, inference methods and evaluation metrics for reliable GRN reconstruction. Even if modelling architectures rely on different mathematical formalisms, they all provide similar networks which require some simplifications. Notably, usefulness of GRN reconstruction depends on both its application and the available data. GRN inference methods have their own advantages and disadvantages depending on the available data and the purpose of the inference. Many efforts have been made for their comparison by using these algorithms to infer a GRN from a single data set and then assessing their validity. These comparisons require appropriate evaluation methods to satisfactorily determine algorithm performance.

The curse of dimensionality still makes suffer inference methods dealing with large data sets, thus novel and more efficient algorithms that are highly-scalable are still required. Dimensionality problems usually come with the integration of large prior biological knowledge, and model parameters such as sparseness do little to solve these problems. Together with sparseness, feature selection is important for the inference of GRN from large data sets. These parameters limit the number of regulators per gene and penalize model complexity.

Also, the integration of omics data from single cells is still challenging, so there is a need for standardized methods. Besides this, the integration of multiple-source biological knowledge makes easier data insufficiency problems, and it is a major focus in GRN research.

To sum up, GRN models are certainly a powerful tool for the understanding of biological systems, and their improve is ligated to the advances in biotechnology and bioinformatics, which will enable the characterization of complex relations between biological entities.

Conflict of interest

The authors have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgements

This work was supported in part by the collaboration scholarship of the Ministry of Education, Culture and Sports of the Government of Spain. This scholarship was granted for the collaboration research with DATAi group (Intelligent Data Analysis – TIC200) of the Division of Computer Science of Pablo de Olavide University.

References

[1] McCall MN. Estimation of gene regulatory networks. Postdoc J 2013;1:60.
[2] Emmert-Streib F, Dehmer M, Haibe-Kains B. Untangling statistical and biological models to understand network inference: the need for a genomics network ontology. Front Genet 2014;5:299.
[3] Liang L, Gao L, Zou X-P, Huang M-L, Chen G, Li J-J, et al. Diagnostic significance and potential function of miR-338-5p in hepatocellular carcinoma: a bioinformatics study with microarray and RNA sequencing data. Mol Med Rep 2018;17:2297–312.
[4] Huang R, He Y, Sun B, Liu B. Bioinformatic analysis identifies three potentially key differentially expressed genes in peripheral blood mononuclear cells of patients with Takayasu's arteritis. Cell J (Yakhteh) 2018;19:647.
[5] Ogundijo OE, Elmas A, Wang X. Reverse engineering gene regulatory networks from measurement with missing values. EURASIP J Bioinform Syst Biol 2016;2017:2.
[6] Yan W, Xue W, Chen J, Hu G. Biological networks for cancer candidate biomarkers discovery. Cancer Inform 2016;15:CIN–39458.
[7] Levine M, Davidson EH. Gene regulatory networks for development. Proc Natl Acad Sci U S A 2005;102:4936–42.
[8] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013;41:D991–5.
[9] Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models—a review. Biosystems 2009;96:86–103.
[10] Dougherty ER. Validation of gene regulatory networks: scientific and inferential. Brief Bioinform 2010;12:245–52.
[11] Buermans H, Den Dunnen J. Next generation sequencing technology: advances and applications. Biochim Biophys Acta (BBA)-Mol Basis Dis 2014;1842:1932–41.
[12] Cereb N, Kim HR, Ryu J, Yang SY. Advances in DNA sequencing technologies for high resolution HLA typing. Hum Immunol 2015;76:923–7.
[13] Li Y, Liu L, Bai X, Cai H, Ji W, Guo D, et al. Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. BMC Bioinform 2010;11:520.
[14] Pataskar A, Tiwari VK. Computational challenges in modeling gene regulatory events. Transcription 2016;7:188–95.
[15] Monger C, Kelly PS, Gallagher C, Clynes M, Barron N, Clarke C. Towards next generation CHO cell biology: bioinformatics methods for RNA-Seq-based expression profiling. Biotechnol J 2015;10:950–66.
[16] Larvie JE, Sefidmazgi MG, Homaifar A, Harrison SH, Karimoddini A, Guiseppi-Elie A. Stable gene regulatory network modeling from steady-state data. Bioengineering 2016;3:12.
[17] Chaitankar V, Karakülah G, Ratnapriya R, Giuste FO, Brooks MJ, Swaroop A. Next generation sequencing technology and genomewide data analysis: perspectives for retinal research. Prog Retin Eye Res 2016;55:1–31.
[18] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. Genbank. Nucleic Acids Res 2012;41:D36–42.
[19] Kodama Y, Mashima J, Kosuge T, Katayama T, Fujisawa T, Kaminuma E, et al. The DDBJ Japanese genotype-phenotype archive for genetic and phenotypic human data. Nucleic Acids Res 2014;43:D18–22.
[20] Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res 2014;43:D662–9.
[21] Tammen SA, Friso S, Choi S-W. Epigenetics: the link between nature and nurture. Mol Aspects Med 2013;34:753–64.
[22] Ramsey J, Butnor K, Peng Z, Leclair T, van der Velden J, Stein G, et al. Loss of RUNX1 is associated with aggressive lung adenocarcinomas. J Cell Physiol 2018;233:3487–97.
[23] Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, et al. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. Cell Mol Life Sci 2015;72:3425–39.
[24] Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013;501:506.
[25] Clough E, Barrett T. The gene expression omnibus database. Stat Genom: Methods Protocols 2016:93–110.
[26] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31:68–71.
[27] Kang T, Ding W, Zhang L, Ziemek D, Zarringhalam K. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. BMC Bioinform 2017;18:565.
[28] Fluck J, Hofmann-Apitius M. Text mining for systems biology. Drug Discov Today 2014;19:140–4.
[29] Sverchkov Y, Craven M. A review of active learning approaches to experimental design for uncovering biological networks. PLoS Comput Biol 2017;13:e1005466.
[30] Ud-Dean SM, Gunawan R. Optimal design of gene knockout experiments for gene regulatory network inference. Bioinformatics 2015;32:875–83.
[31] Kim J-W, Yang H-J, Brooks MJ, Zelinger L, Karakülah G, Gotoh N, et al. NRL-regulated transcriptome dynamics of developing rod photoreceptors. Cell Rep 2016;17:2460–73.
[32] Li Y, Varala K, Coruzzi GM. From milliseconds to lifetimes: tracking the dynamic behavior of transcription factors in gene networks. Trends Genet 2015;31:509–15.
[33] Bordon J, Moškon M, Zimic N, Mraz M. Fuzzy logic as a computational tool for quantitative modelling of biological systems with uncertain kinetic data. IEEE/ACM Trans Comput Biol Bioinform 2015;12:1199–205.
[34] Coker EA, Mitsopoulos C, Workman P, Al-Lazikani B. Signet: a signaling network data simulator to enable signaling network inference. PLoS One 2017;12:e0177701.
[35] Wang M, Benedito VA, Zhao PX, Udvardi M. Inferring large-scale gene regulatory networks using a low-order constraint-based algorithm. Mol Biosyst 2010;6:988–98.
[36] Liang Y, Kelemen A. Computational dynamic approaches for temporal omics data with applications to systems medicine. BioData Min 2017;10:20.
[37] Thompson JA, Tan J, Greene CS. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ 2016;4:e1621.
[38] Kourilsky P. The natural defense system and the normative self model. F1000Research 2016;5.
[39] Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific symposium on biocomputing, vol. 3 1998:18–29.
[40] Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Biocomputing 2000. 1999. p. 418–29.
[41] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform 2006;7:S7.
[42] Montes RAC, Coello G, González-Aguilera KL, Marsch-Martínez N, de Folter S, Alvarez-Buylla ER. ARACNE-based inference, using curated microarray data, of *Arabidopsis thaliana* root transcriptional regulatory networks. BMC Plant Biol

2014;14:97.
[43] Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. PLoS One 2010;5:e9803.
[44] Olsen C, Meyer PE, Bontempi G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. EURASIP J Bioinform Syst Biol 2008;2009:308959.
[45] Gómez-Vela F, Barranco CD, Díaz-Díaz N. Incorporating biological knowledge for construction of fuzzy networks of gene associations. Appl Soft Comput 2016;42:144–55.
[46] Rao A, Hero III AO, States DJ, Engel JD. Using directed information to build biologically relevant influence networks. Computational systems bioinformatics: vol. 6). 2007. p. 145–56.
[47] Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol 2007;2007.
[48] Luo W, Woolf PJ. Reconstructing transcriptional regulatory networks using three-way mutual information and Bayesian networks. Computational biology of transcription factor binding. Springer; 2010. p. 401–18.
[49] Watkinson J, Liang K-c, Wang X, Zheng T, Anastassiou D. Inference of regulatory gene interactions from expression data using three-way mutual information. Ann N Y Acad Sci 2009;1158:302–13.
[50] Guo W, Calixto CP, Tzioutziou N, Lin P, Waugh R, Brown JW, et al. Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size. BMC Syst Biol 2017;11:62.
[51] Altay G, Emmert-Streib F. Structural influence of gene networks on their inference: analysis of C3NET. Biol Direct 2011;6:31.
[52] Zheng G, Xu Y, Zhang X, Liu Z-P, Wang Z, Chen L, et al. CMIP: a software package capable of reconstructing genome-wide regulatory networks using gene expression data. BMC Bioinform 2016;17:535.
[53] Xie Y, Liu Y, Valdar W. Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. Biometrika 2016;103:493–511.
[54] Kiani NA, Zenil H, Olczak J, Tegnér J. Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. Semin Cell Dev Biol 2016;51:44–52.
[55] Trescher S, Münchmeyer J, Leser U. Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization. BMC Syst Biol 2017;11:41.
[56] Jakub NAKHZ, Jesper O. Evaluating network inference methods in terms of. networks 2016;19:2271–82.
[57] Wang X-L, Liu Y-M, Zhang Z-D, Wang S-S, Du Y-B, Yin Z-S. Utilizing benchmarked dataset and gene regulatory network to investigate hub genes in postmenopausal osteoporosis. J Cancer Res Ther 2018.
[58] Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. Comput Biol Med 2014;48:55–65.
[59] Melkman AA, Cheng X, Ching W-K, Akutsu T. Identifying a probabilistic Boolean threshold network from samples. IEEE Trans Neural Netw Learn Syst 2018;29:869–81.
[60] Maheshri N, O'Shea EK. Living with noisy genes: how cells function reliably with inherent variability in gene expression. Ann Rev Biophys Biomol Struct 2007;36.
[61] Simak M, Yeang C-H, Lu HH-S. Exploring candidate biological functions by Boolean function networks for *Saccharomyces cerevisiae*. PLoS One 2017;12:e0185475.
[62] Claussen JC, Skiecevičienė J, Wang J, Rausch P, Karlsen TH, Lieb W, et al. Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. PLoS Comput Biol 2017;13:e1005361.
[63] Polak ME, Ung CY, Masapust J, Freeman TC, Ardern-Jones MR. Petri Net computational modelling of Langerhans cell interferon regulatory factor network predicts their role in T cell activation. Sci Rep 2017;7:668.
[64] Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol 2015;33:269.
[65] Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JE, Iversen ES, et al. Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature 2008;453:944.
[66] Voit EO. Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge University Press; 2000.
[67] Rosenfeld S. Mathematical descriptions of biochemical networks: stability, stochasticity, evolution. Prog Biophys Mol Biol 2011;106:400–9.
[68] Matsumoto H, Kiryu H, Furusawa C, Ko MS, Ko SB, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics 2017;33:2314–21.
[69] Deng Y, Zenil H, Tegnér J, Kiani NA. HiDi: an efficient reverse engineering schema for large-scale dynamic regulatory network reconstruction using adaptive differentiation. Bioinformatics 2017;33:3964–72.
[70] Kaderali L, Radde N. Inferring gene regulatory networks from expression data. Computational intelligence in bioinformatics. Springer; 2008. p. 33–74.
[71] Erdogdu U, Polat F, Alhajj R. Employing decomposable partially observable Markov decision processes to control gene regulatory networks. Artif Intell Med 2017;83:14–34.
[72] Larjo A, Shmulevich I, Lähdesmäki H. Structure learning for Bayesian networks as models of biological networks. Data mining for systems biology. Springer; 2013. p. 35–45.
[73] Pineda AL, Gopalakrishnan V. Novel application of junction trees to the interpretation of epigenetic differences among lung cancer subtypes. AMIA Jt Summits Transl Sci Proc 2015;2015:31.
[74] Deeter A, Dalman M, Haddad J, Duan Z-H. Inferring gene and protein interactions using PubMed citations and consensus Bayesian networks. PLoS One 2017;12:e0186004.
[75] Acerbi E, Zelante T, Narang V, Stella F. Gene network inference using continuous time Bayesian networks: a comparative study and application to Th17 cell differentiation. BMC Bioinform 2014;15:387.
[76] Chekouo T, Stingo FC, Doecke JD, Do K-A. miRNA-target gene regulatory networks: a Bayesian integrative approach to biomarker selection with application to kidney cancer. Biometrics 2015;71:428–38.
[77] Chudasama D, Bo V, Hall M, Anikin V, Jeyaneethi J, Gregory J, et al. Identification of novel cancer biomarkers of prognostic value using specific gene regulatory networks (GRN): a novel role of RAD51AP1 for ovarian and lung cancers. Carcinogenesis 2018;39:407–17.
[78] Biswas S, Acharyya S. Neural model of gene regulatory network: a survey on supportive meta-heuristics. Theory Biosci 2016;135:1–19.
[79] Kordmahalleh MM, Sefidmazgi MG, Harrison SH, Homaifar A. Identifying time-delayed gene regulatory networks via an evolvable hierarchical recurrent neural network. BioData Min 2017;10:29.
[80] Ling H, Samarasinghe S, Kulasiri D. Novel recurrent neural network for modelling biological networks: oscillatory p53 interaction dynamics. Biosystems 2013;114:191–205.
[81] Tong DL, Boocock DJ, Dhondalay GKR, Lemetre C, Ball GR. Artificial neural network inference (ANNI): a study on gene–gene interaction for biomarkers in childhood sarcomas. PLoS One 2014;9:e102483.
[82] Siddens LK, Tilton SC, Williams DE, Krueger SK, Larkin A, Waters KM, et al. Application of a fuzzy neural network model in predicting polycyclic aromatic hydrocarbon-mediated perturbations of the Cyp1b1 transcriptional regulatory network in mouse skin. Toxicol Appl Pharmacol 2012.
[83] Rubiolo M, Milone D, Stegmayer G. Extreme learning machines for reverse engineering of gene regulatory networks from expression time series. Bioinformatics 2017;1:8.
[84] Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, et al. DART: Denoising algorithm based on relevance network topology improves molecular pathway activity inference. BMC Bioinform 2011;12:403.
[85] Liu J, Chi Y, Zhu C, Jin Y. A time series driven decomposed evolutionary optimization approach for reconstructing large-scale gene regulatory networks based on fuzzy cognitive maps. BMC Bioinform 2017;18:241.
[86] Ud-Dean SM, Heise S, Klamt S, Gunawan R. Trace+: ensemble inference of gene regulatory networks from transcriptional expression profiles of gene knock-out experiments. BMC Bioinform 2016;17:252.
[87] Paroni A, Graudenzi A, Caravagna G, Damiani C, Mauri G, Antoniotti M. CABeRNET: a Cytoscape app for Augmented Boolean models of gene Regulatory NETworks. BMC Bioinform 2016;17:64.
[88] Shahdoust M, Pezeshk H, Mahjub H, Sadeghi M. F-map: a Bayesian approach to infer the gene regulatory network using external hints. PLoS One 2017;12:e0184795.
[89] Wang YR, Huang H. Review on statistical methods for gene network reconstruction using expression data. J Theor Biol 2014;362:53–61.
[90] Thiagarajan R, Alavi A, Podichetty JT, Bazil JN, Beard DA. The feasibility of genome-scale biological network inference using graphics processing units. Algorithms Mol Biol 2017;12:8.
[91] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform 2015.
[92] Sanchez-Osorio I, Ramos F, Mayorga P, Dantan E. Foundations for modeling the dynamics of gene regulatory networks: a multilevel-perspective review. J Bioinform Comput Biol 2014;12:1330003.
[93] Bellman R. Dynamic programming. Courier Corporation; 2013.
[94] Kerr WT, Douglas PK, Anderson A, Cohen MS. The utility of data-driven feature selection: Re: Chu et al. 2012. NeuroImage 2014;84:1107–10.
[95] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
[96] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
[97] Squillario M, Barbieri M, Verri A, Barla A. Enhancing interpretability of gene signatures with prior biological knowledge. Microarrays 2016;5:15.
[98] Sun C, Cao X, Zhou C, Liu L, Feng F, Liu R, et al. Construction of gene/protein interaction networks for primary myelofibrosis and KEGG pathway-enrichment analysis of molecular compounds. Genet Mol Res 2015;14:16126–32.
[99] Valverde S, Ohse S, Turalska M, West BJ, Garcia-Ojalvo J. Structural determinants of criticality in biological networks. Front Physiol 2015;6:127.
[100] Radcliffe NJ. Genetic set recombination and its application to neural network topology optimisation. Neural Comput Appl 1993;1:67–90.
[101] Jia B, Wang X. Regularized EM algorithm for sparse parameter estimation in nonlinear dynamic systems with application to gene regulatory network inference. EURASIP J Bioinform Syst Biol 2014;2014:5.
[102] Gui S, Rice AP, Chen R, Wu L, Liu J, Miao H. A scalable algorithm for structure identification of complex gene regulatory network from temporal expression data. BMC Bioinform 2017;18:74.
[103] di Bernardo D, Belcastro V. Reverse engineering gene regulatory networks for elucidating transcriptome organisation, gene function and gene regulation in mammalian systems. 2007.
[104] Dechter R, Pearl J. Generalized best-first search strategies and the optimality of A^*. J ACM 1985;32:505–36.
[105] Someren Ev, Wessels L, Backer E, Reinders M. Genetic network modeling.

Pharmacogenomics 2002;3:507–25.
[106] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006;65:31–78.
[107] Gómez-Vela F, Rodriguez-Baena DS, Vázquez-Noguera JL. Structure optimization for large gene networks based on greedy strategy. Comput Math Methods Med 2018;2018.
[108] Zhu X, Chen S, Jiang Y, Xu Y, Zhao Y, Chen L, et al. Analysis of miRNA expression profiles in melatonin-exposed GC-1 spg cell line. Gene 2018;642:513–21.
[109] Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res 2017;45:D369–79.
[110] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 2015;19:A68.
[111] Pineda S, Steen K, Malats N. Integrative eQTL analysis of tumor and host omics data in individuals with bladder cancer. Genet Epidemiol 2017;41:567–73.
[112] Salehzadeh-Yazdi A, Asgari Y, Saboury AA, Masoudi-Nejad A. Computational analysis of reciprocal association of metabolism and epigenetics in the budding yeast: a genome-scale metabolic model (GSMM) approach. PLoS One 2014;9:e111686.
[113] Sinha S. Integration of prior biological knowledge and epigenetic information enhances the prediction accuracy of the Bayesian Wnt pathway. Integr Biol 2014;6:1034–48.
[114] Walhout AJ. What does biologically meaningful mean? A perspective on gene regulatory network validation. Genome Biol 2011;12:109.
[115] Dougherty E, Qian X. Validation of gene regulatory network inference based on controllability. Front Genet 2013;4:272.
[116] Yu B, Doraiswamy H, Chen X, Miraldi E, Arrieta-Ortiz ML, Hafemeister C, et al. Genotet: an interactive web-based visual exploration framework to support validation of gene regulatory networks. IEEE Trans Vis Comput Graphics 2014;20:1903–12.
[117] Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 2004;7:535–45.
[118] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods 2012;9:796.
[119] De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol 2010;8:717.
[120] Haibe-Kains B, Emmert-Streib F. Quantitative assessment and validation of network inference methods in bioinformatics. Front Genet 2014;5:221.
[121] Schrynemackers M, Küffner R, Geurts P. On protocols and measures for the validation of supervised methods for the inference of biological networks. Front Genet 2013;4:262.
[122] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[123] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 1997;30:1145–59.
[124] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. European conference on information retrieval 2005. 2005. p. 45–359.
[125] Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. Joint European conference on machine learning and knowledge discovery in databases. 2013. p. 451–66.
[126] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. 2006. p. 233–40.
[127] Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One 2010;5:e9202.
[128] Hase T, Ghosh S, Yamanaka R, Kitano H. Harnessing diversity towards the reconstructing of large scale gene regulatory networks. PLoS Comput Biol 2013;9:e1003361.
[129] Bellot P, Olsen C, Salembier P, Oliveras-Vergés A, Meyer PE. NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference. BMC Bioinform 2015;16:312.
[130] Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinform 2018;19:232.
[131] Marbach D, Schaffter T, Mattiussi C, Floreano D. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. J Comput Biol 2009;16:229–39.
[132] Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 2011;27:2263–70.
[133] Seifert M, Beyer A. regNet: an R package for network-based propagation of gene expression alterations. Bioinformatics 2017;34:308–11.
[134] Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, et al. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinform 2006;7:43.
[135] Gómez-Vela F, Díaz-Díaz N. Gene network biological validity based on gene–gene interaction relevance. Sci World J 2014;2018.
[136] Gómez-Vela F, Lagares JA, Díaz-Díaz N. Gene network coherence based on prior knowledge using direct and indirect relationships. Comput Biol Chem 2015;56:142–51.
[137] Diaz-Montana JJ, Diaz-Diaz N. Development and use of the Cytoscape app GFD-Net for measuring semantic dissimilarity of gene networks. F1000Research 2014;3.
[138] Li Y, Gong P, Perkins EJ, Zhang C, Wang N. RefNetBuilder: a platform for construction of integrated reference gene regulatory networks from expressed sequence tags. BMC Bioinform 2011;12:S20.

Chapter 6

Computational Inference of Gene Co-Expression Networks for the identification of Lung Carcinoma Biomarkers: An Ensemble Approach

Authors	Fernando M. Delgado-Chaves, Francisco Gómez-Vela, Miguel García-Torres, Federico Divina and José Luis Vázquez Noguera
Journal	Genes
Editorial	MDPI
eISSN	2073-4425
Published	22/11/2019
DOI	doi.org/10.3390/genes10120962
Impact factor 2021	4,141
Quartile	Q2

genes

MDPI

Article

Computational Inference of Gene Co-Expression Networks for the identification of Lung Carcinoma Biomarkers: An Ensemble Approach

Fernando M. Delgado-Chaves [1,†], Francisco Gómez-Vela [1,*,†], Miguel García-Torres [1], Federico Divina [1] and José Luis Vázquez Noguera [2]

1 Division of Computer Science, Pablo de Olavide University, 41013 Seville, Spain; fmdelcha@alu.upo.es (F.M.D.-C.); mgarciat@upo.es (M.G.-T.); fdiv@upo.es (F.D.)
2 Computer Engineer Department, Universidad Americana de Paraguay, Asuncion 1209, Paraguay; jose.vazquez@ua.edu.py
* Correspondence: fgomez@upo.es
† These authors contributed equally to this work.

Received: 14 October 2019; Accepted: 31 October 2019; Published: 22 November 2019

Abstract: Gene Networks (GN), have emerged as an useful tool in recent years for the analysis of different diseases in the field of biomedicine. In particular, GNs have been widely applied for the study and analysis of different types of cancer. In this context, Lung carcinoma is among the most common cancer types and its short life expectancy is partly due to late diagnosis. For this reason, lung cancer biomarkers that can be easily measured are highly demanded in biomedical research. In this work, we present an application of gene co-expression networks in the modelling of lung cancer gene regulatory networks, which ultimately served to the discovery of new biomarkers. For this, a robust GN inference was performed from microarray data concomitantly using three different co-expression measures. Results identified a major cluster of genes involved in SRP-dependent co-translational protein target to membrane, as well as a set of 28 genes that were exclusively found in networks generated from cancer samples. Amongst potential biomarkers, genes *NCKAP1L* and *DMD* are highlighted due to their implications in a considerable portion of lung and bronchus primary carcinomas. These findings demonstrate the potential of GN reconstruction in the rational prediction of biomarkers.

Keywords: co-expression network; lung carcinoma; biomarker discovery; ensemble network; data mining; Bioinformatics

1. Introduction

Over the last two decades, gene networks (GNs) have become an essential tool in the field of biomedicine [1]. Such GNs are usually presented as a graph comprising nodes and rods, where nodes represent genes (or gene products) and rods represent interactions among genes [1,2]. These rods may include a numeric value or *weight* which refers to the strength of these relationships. Therefore, not only are GNs able to identify genes related to biological processes, but also the relationships among these genes, thus providing a comprehensive picture of the studied processes [3]. GNs have been widely applied in various fields such as biology, biomedicine or bioinformatics [4,5] among others.

According to the different works in the literature [1,6], GN inference algorithms lie under four main categories: co-expression, boolean networks, differential equation-based and Bayesian networks. Within this classification, co-expression networks, which are based on information theory algorithms, arise as a significantly relevant approach due to their computational simplicity and extensive use in the literature [1,7]. These kind of networks infers relationships between genes if these show similar

expression patterns, regarding an entropy measure like correlation indices or mutual information approaches among others. To do so, the degree of relationship between each pair of genes is measured, and then considered valid when this degree exceeds a certain threshold. Therefore, this threshold indicates the minimum similarity level for two expression patterns to be considered significant [8].

The main measures to evaluate the degree of co-expression between two genes, according to the literature, are correlation measures such as Pearson, Spearman or Kendall coefficients [9,10]. In addition, other measures for the generation of gene networks have been widely used, such as Mutual Information [11,12]. However, co-expression networks often present a major drawback, as the inference of relationships depends entirely on the chosen measures, which may present some limitations. For instance, the inability of the above mentioned measures to detect non-linear dependencies or their dependence on the input data distribution to obtain reliable results, as in the case of Spearman and Pearson coefficients respectively [13]. In order to overcome these issues, ensemble strategies may well be a solution, as these combine different measures for the evaluation of relationships between genes [14]. Therefore, the obtained networks are more reliable than those obtained by a single measure, also providing more accurate modelling and plausible biological insights.

Ordinarily, GN inference algorithms take gene expression datasets, e.g., microarrays or RNA-Seq, as input for the generation of the gene-gene interactions [6,7,15]. These datasets have been massively generated over the last decade for the study of some type of biological process or specific disease [16], allowing the identification of relationships between DNA, RNA, proteins and other gene products. Researchers may then perform computer analysis on this type of data before checking the results in the laboratory.

In particular, one of the most studied diseases is cancer, due to its high penetrance into the global population [17]. Moreover, cancer expression data have been screened in the quest for cancer biomarkers, which can be defined as substances, structures, or processes that can be quantified in a biological sample or their products and may indicate the prognosis of a disease [18]. In particular, lung carcinoma is among the most common tumor types and it is estimated that around 85% of the cases occur due to tobacco smoking [19,20]. Regrettably, most cases are not curable, partly as a consequence of late diagnosis, which require specific medical tests such as bronchoscopy. For this reason, lung cancer biomarkers are considered of a huge importance in the early diagnosis of the disease, and many approaches have sought for non-invasive methods for their measure. For example, in Peng et al. [21], a method is proposed for the identification of lung carcinoma biomarkers in exhaled air.

In this work we present a study of human lung carcinoma gene expression samples corresponding to smoker patients by means of an ensemble co-expression algorithm. Expression data were computational and comprehensively processed in order to generated a gene co-expression network. The algorithm applied to infer the GNs consists of an ensemble strategy which combines three widely used co-expression measures in order to rate gene-gene relationships. As a result a lung carcinoma network was generated and compared to another network generated from non-cancerous lung samples also corresponding to smoker patients. The cross analysis of these networks yielded meaningful insights on the biological functions affected in both situations, assisting the identification of potentially-novel lung carcinoma biomarkers.

The rest of the paper is organized as follows: In Section 1.1 we introduce some relevant gene networks based works applied to biomedical datasets. Then we describe, in Section 2, the dataset studied and the methods used to perform this work (network inference algorithm and the analysis approaches used). The main results obtained and the discussion are detailed in SectionS 3 and 4. Finally, the main conclusions achieved are presented in Section 5.

1.1. Related Works

Co-expression networks have been extensively used in the literature for the analysis and study of cancer disease. For example, Aggarwal et al. [22] applied a consensus gene co-expression meta-network of gastric cancer, the second most common cause of cancer-related deaths in the world. The results

suggest, at single-gene level, an interaction between the PLA2G2A prognostic marker and the EphB2 receptor. Furthermore, the network analysis also enhances the understanding of gastric cancer at the levels of system topology and functional modules. In another work, Ma et al. [23] adopted weighted co-expression networks to describe the interplay among genes for cancer prognosis. In particular the authors presented six prognosis analyses on breast cancer and lymphoma. The results presented showed that their approach can identify genes that are significantly different from those using different alternatives. Genes that were identified using this approach presented sound biological bases, better prediction performance, and better reproducibility.

In Clarke et al. [24], a weighted version of gene co-expression network is used to analyze breast cancer samples from microarray-based gene expression studies. From the several gene clusters identified, some of them were found to be correlated with clinicopathological variables, survival endpoints for breast cancer as a whole and also its molecular subtypes. Also in 2013, the paper presented by Chang et al. [25], used a weigthed co-expression network in order to identify coexpression modules associated with malignancy menginiomas, one of the most common primary adult brain tumors. The authors identified, at the transcriptome level, 23 coexpression modules from the weighted gene coexpression network. In addition, they were able to identified a module with 356 genes that was highly related to tumorigenesis.

In 2014, the work presented by Yang et al. [26] a prognosis genes analysis based on gene co-expression networks for four cancer types using data from "The Cancer Genome Atlas". The authors performed a systematic analysis of the properties of prognostic genes in the context of biological networks across multiple cancer types. The results of this work suggested that the prognostic mRNA genes tend not to be hub genes (genes with an extremely high connectivity). On the contrary, the prognostic genes are enriched in modules (a group of highly interconnected genes), especially in module genes conserved across different cancer co-expression networks.

In 2015, Liu et al. [27] also uses a weighted co-expression network to investigate how gene interactions influence lung cancer and the roles of gene networks in lung cancer regulation. It was found that the overall expression of one of the modules identified was significantly higher in the normal group than in the lung cancer group.

Recently in 2018, the work presented by Yang et al. [28] weighted gene co-expression network analysis (WGCNA) was applied to investigate intrinsic association between genomic changes and transcriptome profiling in neuroblastoma cancer (a highly complex and heterogeneous cancer in children). The results achieved identified multiple gene coexpression modules in two independent datasets and associated with functional pathways. The results also indicated that modules involved in nervous system development and cell cycle are highly associated with MYCN amplification and 1p deletion.

Finally, in Xu et al. [29] (2019), Xu et al. study Hepatocellular carcinoma, a very common subtype of liver cancer. The authors conducted a WGCNA to identify complex gene interactions that affect prognosis. The final results identified 10 genes that have never been mentioned in hepatocellular carcinoma and that are associated with malignant progression and patient prognosis.

2. Materials and Methods

In this section, the dataset studied and the methods used to perform the analysis are described. To begin with, the used dataset is presented in Section 2.1. Then, the pipeline followed for the analysis of the lung cancer dataset is exposed in the following subsections. First, data preprocessing is specified in Section 2.2. Then, relevant genes were identified in differential expression analyses, as explained in Section 2.3. Afterwards, the GN reconstruction approach is addressed in Section 2.4. Finally the exploration of the inferred networks is described in Section 2.5.

2.1. Lung Cancer Dataset

The dataset presented in this work corresponds to a previous study by Spira et al. [30] and Gustafson et al. [31] carried out in the Boston University Medical Center. In such studies, the gene expression level of epithelial cells coming from the respiratory tract of smoker patients was globally analyzed via microarray.

The dataset in particular retrieves the expression level of 22284 genes, along 192 samples from different smoker patients. Samples were collected from airway tissue during bronchoscopies and total RNA was extracted from these. Patients were divided in three categories: those diagnosed with lung cancer (97), those not diagnosed with lung cancer (90) and those suspected to be under cancer development (5). Although based on a relatively old platform (the Affymetrix U133A array), this dataset in particular was chosen for its suitability to specifically study the underlying genetic impairment in lung carcinoma in smoker patients.

The dataset may be openly-accessed at NCBI's Gene Expression Omnibus (GEO) database [32], dataset record: GDS2771, reference series: GSE4115. The screening platform used to obtain this data was the Affymetrix Human Genome U133A Array [HG-U133A], from which probeset information was retrieved. The available dataset at GEO was already preprocessed in accordance with the original article [30]. In conformity with this paper, the Robust Multichip Average (RMA) algorithm was used to normalize the different datasets and achieving a certain level of similarity between all technical replicates. Also, some samples were removed from the analysis due to their poor quality (Spira et al. [30], Supplementary Methods Revised).

2.2. Data Preprocessing

The original dataset by Spira et al. [30] and Gustafson et al. [31] was imported to RStudio (development environment in R [33]) for data treatment and adaptation to the network inference process. From the original data a subset was selected for the present study, which seeks the comparison between cancer-diagnosed and not diagnosed smokers, thus leaving patients with cancer suspect aside. This decision was made considering the short number of patients with suspected cancer (only 5 patients), as the more analogous samples available, the more robust the GN inference will be.

First, an exploratory multidimensional scaling (MDS) plot or Principal Coordinates Analysis (PCoA) of the subset dataset was performed. This type of analysis helps in the examination of the similarity level between samples, as in the case of Gruvberger et al. [34]. In this case, the classical MDS method was applied, which assumes Euclidean distances. Graphic representation was performed using the *ggplot2* R package [35].

2.3. Differential Expression Analysis

The starting dataset was split in order to generate two different subsets, corresponding to cancerous and non cancerous samples respectively. DEG in cancerous samples vs. non cancerous ones were estimated using the *limma* R package [36]. Basing on linear models, *limma* has been widely used for DEG analysis, yielding prominent results [37,38]. Note those samples corresponding to smoker patients that had not been diagnosed with cancer were used as a control situation.

DEG were filtered using a significance level below 0.05 and a minimum absolute log2 fold change (FC) of 0.25. Note this log2 FC corresponds to ∼20% change in gene expression. Selected *p*-values adjustment method for multiple values was FDR Benjamini Hochberg, as it generally provides a laxer filtering [39], i.e., the larger number of DEG for a same *p*-value. The resulting DEG would be extracted from the starting dataset and would be the only ones to proceed for network inference. *p*-Values were estimated for each gene and corrected with Bonferroni step-down.

DEG information, such as log2 FC, would be additionally imported to the reconstructed networks for biological interpretation purposes. This relatively low threshold was selected in order to filter a reasonable amount of implicated genes to network reconstruction.

2.4. Network Inference

As stated before, co-expression networks have been extensively used in the field of biomedicine. However, they present some limitations that could be overcome by means of an ensemble strategy [40]. Therefore, we applied an ensemble strategy in order to offer a robust GN reconstruction. There are different ensemble strategies in the literature to combine the different results generated such as majority voting or average [41]. For this study, the average strategy was selected due to its good performance in the literature [42].

A schematic representation of the GN inference approach is shown in Figure 1. For this aim, three co-expression measures were used, namely Kendall, Spearman and Blomqvist coefficients, which provide a co-expression index ranging from −1 to 1. The choice for these three measures was made after their extensive use upon GN reconstruction processes [9,13,43]. Definitions for the mentioned co-expression measures are detailed in Appendix A.

The coefficients were estimated for all possible DEG pairs both in for cancer and non cancer samples. In this way, two GNs were generated, respectively corresponding to the cancer situation and the normal situation, which can be used as a control, both under smoking conditions. Then, the average of the values obtained through each of the three coefficients is used as the final weight for the edge between each gene pair. Note that the values resulting from the application of these coefficients were also taken into consideration in the choice of these measures, as the conceived inference approach requires these values to be within a same range for latter averaging.

Finally, a threshold was established in order to keep only significant co-expressions. Thresholds varied from: 0.7, mild co-expression; to 0.8, strong co-expression; and finally 0.9, very strong co-expression. As detailed in Mukaka [44] and Cooke and Clarke [45], a cut-off of 0.5 to 0.7 (or −0.5 to −0.7) provides a moderately positive (or negative) co-expression, a cut-off of 0.70 to 0.9 (or −0.7 to −0.9) yields a high positive (or negative) co-expression and finally, a threshold of 0.9 to 1 (or −0.9 to −1), gives a very high positive (or negative) co-expression. Note that co-expressions between genes may be either positive or negative, so these thresholds are expressed as absolute values. These thresholds were defined in accordance to statistical standards [6,46,47].

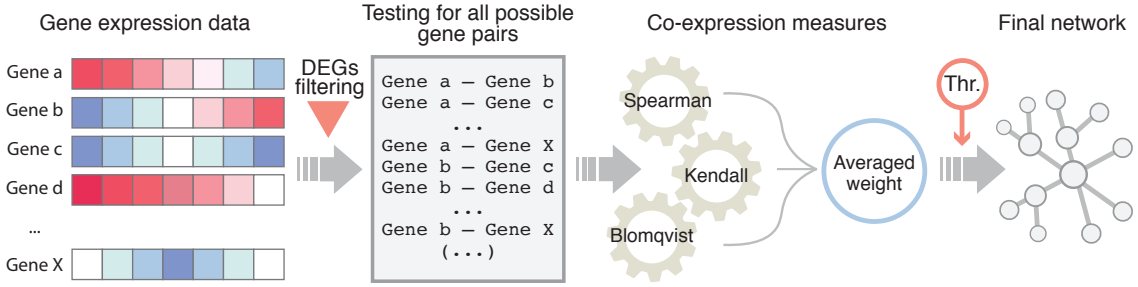


Figure 1. General scheme of the used inference method. For all possible gene pairs, three co-expression coefficients were calculated (Kendall, Spearman and Blomqvist) and averaged for the estimation of the final weight. *Thr*. refers to the thresholding step, using different co-expression indices. DEGs refer to the subset of differentially expressed genes.

Additional parameters for network representation were estimated by means of the *igraph* R package [48]. This package performs adequately with large networks and has been broadly employed in the functional analysis of biological networks [49,50]. In particular, these parameters were node degree, betweenness centrality and rank of the involved nodes. The term degree refers to the number of edges linking a particular node [51]. Those nodes comprising the largest number of relationships in a certain network are termed hubs, which according to the literature, are of a key importance in gene networks [7,52]. On the other hand betweenness centrality is defined as the addition of the fraction of all-pairs shortest paths that go through a specific node [53]. Lastly, node rank is a combination of the two previous measures. Other features such as gene IDs were also added to the nodes information

table, which was imported together with the inferred networks to Cytoscape for network visualization and analysis.

2.5. GN Analysis: Topology and Enrichment Analyses

To perform a comprehensive analysis of the networks, we used the Cytoscape tool [54] and its apps. Cytoscape is a powerful tool to analyze GN and it is commonly used in the literature for such aim [7,55].

As the reconstructed networks were considered to be large and dense, these would be clustered using Cytoscape's clusterMaker app [56] in order to perform an exhaustive analysis of these. The selected clustering algorithm was *GLay*, community clustering [50,57]. Clusterization enables the identification of network modules, i.e., densely-connected regions. According to the GN theory, nodes present in the same cluster are often involved in the same biological function, which will be analyzed in the following steps [58].

With the aim of exploring these functions, a Gene Ontology (GO) terms enrichment analysis was performed over the obtained clusters [59]. For this, *ClueGO* [60] & *CluePedia* [61] Cytoscape apps were used. Additional functional analyses of genes of interest were performed using *DAVID*, the Database for Annotation, Visualization and Integrated Discovery [62,63], an on-line tool for the systematic scrutiny of large lists of genes.

Finally, further infromation on the genetic disruption observed amongst potential biomarkers was revised on the GDC data portal [64] by The Genome Cancer Atlas (TGCA) [65]. The GDC portal is a data-driven platform harboring cancer data, containing information on 3,142,246 mutations registered over 22,872 genes, together with the expression level of these across 37,075 cases of different cancer types.

3. Results

In the following subsections, we report and discuss the main results and biological insights. Noticeably, each step of the GN reconstruction process shapes the final outcome. For this reason, the performed inference and analysis strategies are also addressed along these subsections.

3.1. Data Preprocessing and Exploratory Analyses

MDS plots provided meaningful insights on data distribution and dataset-specific similarity level between samples. According to the performed Euclidean MDS plot, cancerous and non cancerous samples are not clearly differentiated through unsupervised analysis. MDS plot is shown in Figure 2. Although a differential gene expression pattern is suspected between cancerous and non cancerous sample types, differences were found to be fuzzy for a considerable portion of the samples, which could not be classified as part of a delimited group according to the Euclidean method used.

Notwithstanding the fact that slight dissimilarity was found between sample types, presumptive differences in gene expression profiles are thought to be responsible for the cancerous phenotype. Hence, it was assumed that all samples within a same sample type, i.e., cancerous or non cancerous, could be considered homologous. Hence, the original dataset could be split into two portions corresponding to both sample types.

3.2. Obtaining Differentially Expressed Genes

A total of 317 genes were identified as DEG in cancerous samples vs. non cancerous ones, in accordance with the established parameters (log2 FC > 0.25, p-value < 0.05). These genes were filtered from the dataset prior to GN reconstruction, so the generated networks would only comprise these. The identified DEG were considered suitable for GN inference for two reasons: (i) only the relationships between genes of interest will be modeled, and (ii) the number of genes was appropriate for latter network handling in terms of size of the final network.

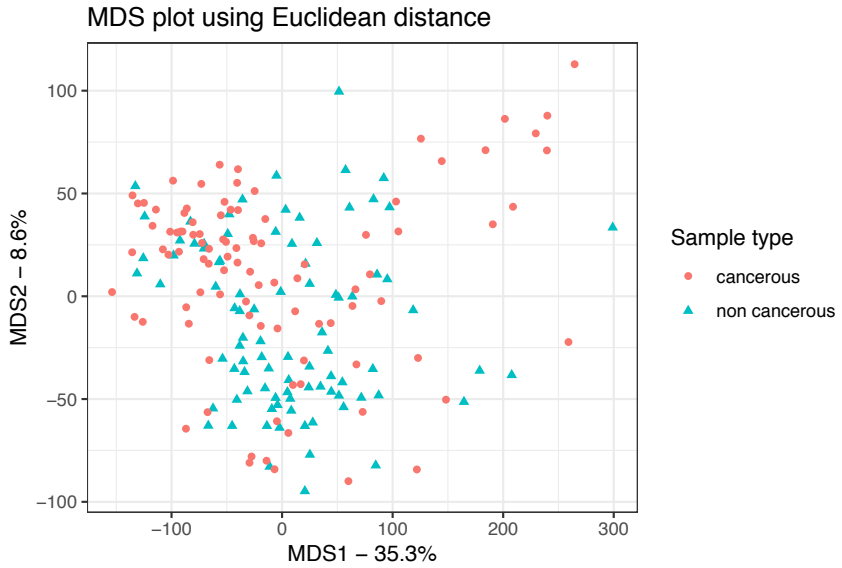
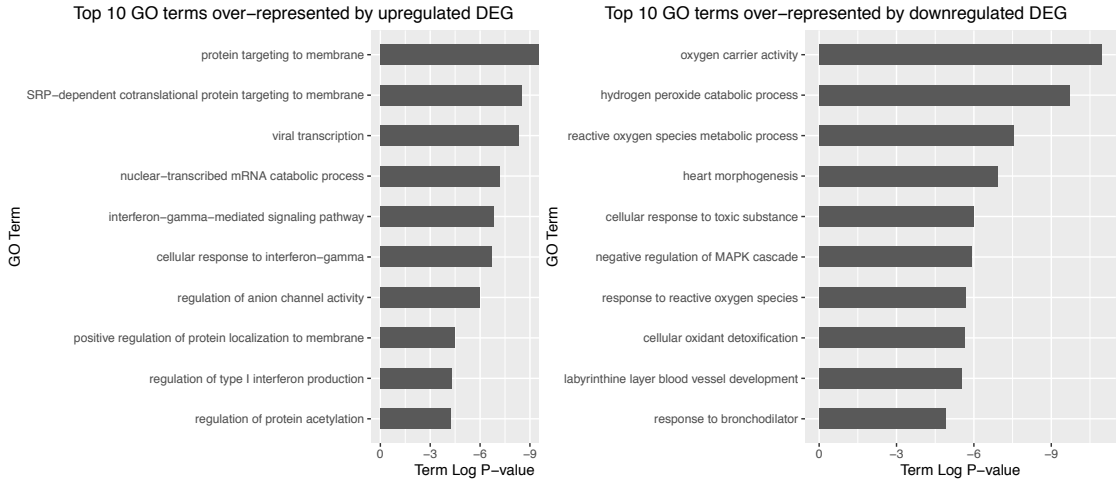


Figure 2. MDS/PCoA plot for the exploratory analysis of the GN inference input data. Since overlapping between sample types is significant, two groups corresponding to cancerous and non cancerous samples cannot be clearly distinguished.

Among these DEG, 165 genes were upregulated in cancer samples when compared to control, whereas the others were found to be downregulated. Log2 FC information was added to the reconstructed networks. Strikingly, only ∼3% of DEG were differentially expressed by a 2 fold factor between sample types. Hence, gene expression levels were not found to change dramatically between cancerous and non cancerous samples. An enrichment analysis was respectively performed over the upregulated and downregulated DEG (Figure 3). As a result, upregulated DEG seemed to be involved in (possibly SRP-dependent) protein targeting to membrane (p-value: 1.180907×10^{-5}), whereas downregulated genes appeared related to oxygen carrier activity (p-value: 1.744030×10^{-5}). Further details on which genes are involved in the impaired biological processes upon the development of lung carcinoma will be addressed in Section 3.4.



(**a**) (**b**)

Figure 3. (**a**) Top 10 GO terms over-represented by the upregulated DEG. (**b**) Top 10 GO terms over-represented by the downregulated DEG. Term p-value was corrected with Bonferroni step-down. Note the lower the p-value, the more the over-represented the GO term is.

3.3. GN Reconstruction and Topology Analysis

As mentioned above, two networks were inferred, corresponding to cancerous and non cancerous samples. These networks will be respectively referred as cancer and non cancer from now on for the sake of simplicity. The comparison between these networks provided meaningful biological insights on the genetic routes that were disrupted in lung carcinoma samples, as well as the impaired biological processes.

Among the three different thresholds that were established, the one corresponding to mild co-expression (0.7) was chosen. Other thresholds provided considerably smaller networks, which were not as informative and less suitable for latter enrichment analyses. However, the results obtained with other thresholds are addressed in the Appendix Section B. The cancer network comprised 197 genes and 2738 interactions, whereas the non cancer network comprised 183 genes and 2499 interactions (Appendix B, Figure A1). Networks corresponding to the strong and very strong co-expression thresholds are also shown in the Appendix B, Figures A2 and A3.

Clustering analysis revealed a major cluster in both inferred networks, respectively comprising around the 70% of the nodes present in both cancer and non cancer networks. This is indicative of a main biological process being affected by DEG in cancerous vs. non cancerous samples. With this assumption, the rest of the cluster will not henceforth be considered for this study, as proposed by previous work like the one by Nepomuceno-Chamorro et al. [55].

In order to detect samples-specific genes, both networks were merged and reclustered in the so-called merged network. Although most genes are present in both cancer and non cancer networks, 28 cancer-exclusive genes were identified, as these were present in the main cluster of the cancer network, but not at its non cancer counterpart (Appendix D, Table A1). Among these, 25 showed genetic downregulation in cancer compared to non cancerous samples, whereas the three resting genes were upregulated in cancerous samples. On the other hand, 7 genes were identified as exclusively belonging to the main cluster of the non cancer network.

3.4. Enrichment Analysis over the Identified Network Clusters

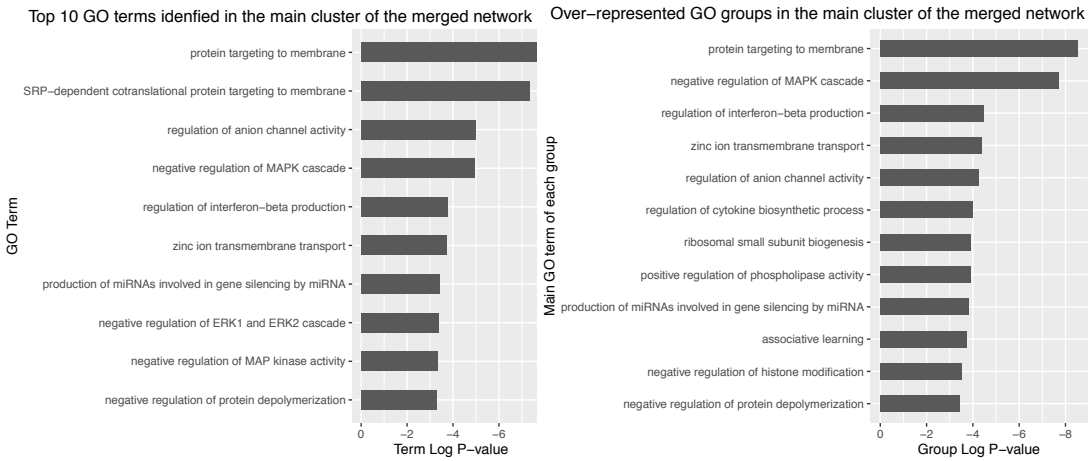
Attending to the merged network, enrichment analysis of these clusters revealed that the major cluster might be implied in protein targeting to membrane (p-value < 0.0005, Figure 4a). The most over-represented GO terms group is also related to this biological process (p-value < 0.0005, Figure 4b). Given that most genes are common between cancer and non cancer networks, and the fact that the main cluster of the merged network comprises most of these common genes, the genes involved in the reconstructed networks would be involved in the above mentioned biological functions. These analyses were also performed separately over the cancer and non cancer networks (Appendix C, Figures A4 and A5).

Gene information of the 28 cancer-exclusive genes was retrieved using *DAVID* (Appendix D, Table A1). Functional analyses revealed the implication of three genes of this list in type 2 diabetes mellitus (T2DM), p-value: 5.6×10^{-3}. These genes are VAMP3, HMGCR and KLF4. Interestingly enough, HMGCR is also related to lung cancer, which suggests an interplay between T2DM and lung cancer. Besides, 4/28 genes were found to be involved in enzyme regulation: HMGCR, PRPS1, PTP4A1 and SLC4A4. These processes are suggested to occur in the cytoplasm according to the functional analysis. GO enrichment analysis showed that 14/28 genes were involved in developmental processes (Appendix D, Table A2). Finally, regarding the tissue-specific genes, genes were associated with brain neoplasia (p-value: 4.9×10^{-4}) and lung tissue (p-value: 1.0×10^{-3}).

On the other hand, there are 7 nodes that are exclusively present at the main cluster of the non cancer network (Appendix D, Table A3). Unfortunately, some of the Affymetrix IDs could not be mapped by *DAVID*, which precluded functional analyses with this tool.

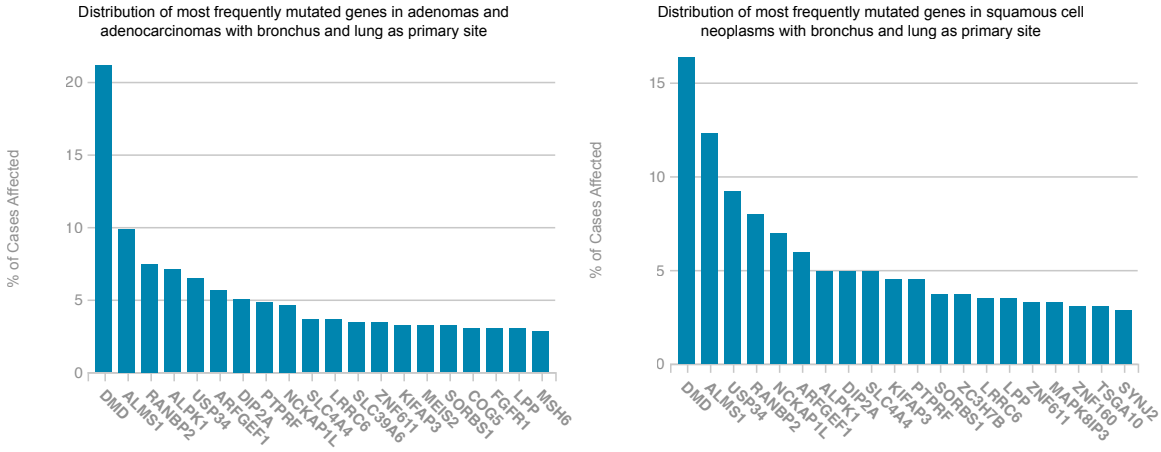
Finally, the observed genetic disruption was explored in the GDC portal. The 28 genes identified as cancer-exclusive were found to be affected in 7081 registered cancer cases, from which 2495 corresponded to adenomas and adenocarcinomas and 1045 corresponded to squamous cell

neoplasms. Both neoplasms lie under the context of lung or bronchus carcinoma. Amongst the 28 cancer-exclusive genes, the gene *NCKAP1L* (NCK associated protein 1 like) was found to be affected in the 8.19% of the mentioned cases (N = 415) of lung and bronchus squamous cell neoplasms. It was also affected in the 6.15% of these cases (N = 374) of lung and bronchus adenomas and adenocarcinomas. On the other hand, when taking into consideration all genes from the main cluster of the cancer network (165), results significantly improve, as the identified gene *DMD* (dystrophin) is disrupted in the 21.13% of the registered cases of adenomas and adenocarcinomas with bronchus and lung as primary site, and also in the 16.35% squamous cell neoplasm cases at this same primary site, as it is shown in Figure 5. This genetic disruption was quantified in terms of simple somatic mutations (SSM), as this data was available for most cases at the GDC portal.



(**a**) (**b**)

Figure 4. (**a**) Top 10 GO terms over-represented by the genes comprised in the main cluster of the merged network. (**b**) GO groups over-represented by the genes in the main cluster of the merged network. The main GO term of each identified group is presented as group label. Term and group *p*-value was corrected with Bonferroni step-down. Note the slower the *p*-value, the more the over-represented the GO term is.



(**a**) (**b**)

Figure 5. Distribution of the most frequently mutated genes in the cases of adenomas and adenocarcinomas (**a**) and squamous cell neoplasms (**b**) registered at the GDC portal [64] presenting bronchus and lung as primary site. These genes belong to the main cluster of the reconstructed cancer network. The number of cases for adenomas and adenocarcinomas was of 497, and 489 for squamous cell neoplasms .

4. Discussion

Firstly, the reconstruction approach used demonstrated its efficacy in the generation of informative GNs for biomedical research. As stated in Section 2.4, these methods have been widely used for GN reconstruction and their ensemble application yielded robust inferences. The present approach was conceived as a rational biomarker discovery tool, which enables the comprehensive analysis of complex expression data to infer data that can be tested experimentally.

The utilization of DEG for GN reconstruction allowed the reconstruction of two networks, namely cancer and no cancer, which assist the modeling of the differences between sample types, thus helping in the identification of network-exclusive elements. An initial enrichment analysis was performed over DEG, in order to identify the main biological networks affected, which corresponded to the ones identified in the major clusters of the reconstructed networks.

Topology analyses revealed a major cluster for each of the two reconstructed networks. According to the literature, clustered co-expressed genes usually take part in a same biological process [15]. Taking into consideration the reconstruction approach, and the fact that DEGs were filtered prior to GN reconstruction, it can be stated that DEGs are involved in a biological process that changes between cancer and non cancer samples. The GO enrichment analysis of the cancer network's major cluster indicated, with high significance, the involvement of these genes in SRP-dependent cotranslational protein targeting to membrane. SRP refers to signal recognition particle, which is added to nascent peptides in the endoplasmic reticulum for their latter targeting to a specific cell component. The connection between SRP and cancer histology has been previously suggested in multiple works [66,67]. For instance, in Zhong et al. [68], this GO term was found to be significantly represented by a set of DEGs which were downregulated in HER2-positive breast cancer compared to normal tissue. Also in Fahrmann et al. [69], samples non-small cell lung cancer adenocarcinoma samples were integratively analysed from metabolomic and proteomic approaches. In this work, SRP-dependent cotranslational protein to membrane was one of the top 10 most significantly disrupted pathways in cancer samples when compared to normal tissue. Taking the above into consideration, the underlying connection between SRPs and lung cancer development is yet to be clarified, but the presented approach was capable of providing a starting point for hypotheses making.

The independent reconstruction of GNs for each sample type allowed the identification of cancer and non cancer-exclusive genes. These sample type-exclusive genes could be responsible for tumor growth, potentially serving as biomarkers. Furthermore, the fact that 25/28 cancer-exclusive genes were downregulated in cancer samples compared to control normal tissue suggests the strong genetic inhibition upon cancer development. What is more, some of these cancer-exclusive genes were found to be associated with T2DM, whose implications in cancer have long been addressed [70–72]. It is known that cancer cells show impaired glucose metabolism, which promotes their uncontrolled proliferation and the preservation of tumor microenvironment [73]. For this reason, many newly-engineered, but also old drugs designed for other diseases such as T2DM, are used to target tumor metabolism as part of anticancer therapies [74,75]. Hence, disruptions at the genetic level can be considered either the effect or the cause of the aberrant cancer metabolism, and their deeper understanding could provide the rational design of new antitumoral drugs.

Notably, half of cancer-exclusive genes were involved in developmental processes, which could be indicative of tumor progression (Appendix D, Table A2). This GO term has also been found in previous studies, as in the case of Heller et al. [76], in which "developmental processes" was represented by tumor-specifically methylated genes in non-small cell lung cancer. Besides, 4/28 genes were found to be involved in enzyme regulation: HMGCR, PRPS1, PTP4A1 and SLC4A4. Only some of the genes in the cluster are found to be associated with the mentioned biological functions, which leads to believe that other genes within the cancer-exclusive gene list might also be involved in these processes, either directly or indirectly, but their implications might have not been discovered yet.

Furthermore, 7 genes were exclusively-found in the non cancer network (Appendix D, Table A3), which means that these genes are taking part in the processes represented in the major cluster of both networks but only in the normal situation. Besides, although these genes probably take part in the same biological process than most DEGs, the co-expressions between them were not so evident in the reconstruction process, which classified them as non cancer-exclusive genes. These genes would require further exploration as their lack in the cancer situation could also be part of cancer onset. Nevertheless, the sequences corresponding to some of these genes could not be mapped from their Affymetrix IDs using *DAVID*.

Regarding the information retrieved from the GDC portal on the potential biomarkers, the role of gene *NCKAP1L* in proliferation and invasion has previously been described breast and hepatocellular carcinoma [77,78]. However, poor has been described within the context of lung carcinoma, hereby suggesting potentially shared mechanisms between the three mentioned cancer types. On the other hand, the role of gene *DMD*, long known for its intrinsic relationship with muscular dystrophies, has previously been addressed in lung and breast cancer. In the work by Luce et al. [79], 1765 samples corresponding to 16 different non-myogenic tumors were analyzed, finding a downregulation of *DMD* the majority of the samples. Besides, a mutated version of *DMD* were observed to shorten the overall survival of patients.

Note these two identified genes were further studied because they were found to be affected in most cases of the cohort at the GDC portal. Ideally, a biomarker should be indicative and present for all cases from a same cancer type. This situation rarely occurs, being necessary to check multiple biomarkers for early cancer detection. Nevertheless, the GDC portal presents some limitations as not every gene has been tested in every sample and cancer type for SSM, so the actual affection of other identified potential biomarkers cannot be verified using this database. But even so, this leaves a door open for further experimental research, delving deeper into the implications of the suggested biomarkers, since GN are considered a powerful predictive tool.

5. Conclusions

In this work we presented a case of study of lung cancer by means of GN approach. To do so, the algorithm applied for inferring the GNs consists of an ensemble of three widely used co-expression measures in order to rate gene-gene relationships. As a result, two networks were generated, a lung carcinoma network and a non-cancerous lung network, both corresponding to smoker patients.

The analyses performed reveal that most DEGs between cancer and non-cancer samples were found to be associated to SRP-dependent cotranslational protein targeting to membrane. Moreover, 28 DEGs were only found in the cancer network, indicating their cancer exclusiveness. Some of these genes were associated with T2DM, developmental processes and enzyme regulation. In addition, 7 DEGs were exclusively found in the non cancer network, and their further analysis could provide further insights on their lack in the cancer situation. Finally, it is worth to mention that among DEGs present in the analyzed clusters, biomarkers exploration is possible and considered a subsequent step in this research.

Genes *NCKAP1L* and *DMD*, identified in the main cluster of the cancer network, were identified as mutated in a considerable percentage of the cases of adenomas, adenocarcinomas and squamous cell neoplasms whose primary site was bronchus and lung, and which were registered at the GDC portal by TCGA.

As future works, we will attempt to refine the process of generating the networks. To this end, we will study new measures that take into account not only linear relations of gene expression, but also non-linear relations. This is due to the fact that non linearity is a grounded assumption when it comes to gene expression [80,81]. Nevertheless, the reconstruction method provided meaningful biological insights even obviating non-linear dependencies.

Author Contributions: conceptualization, F.G.V. and F.D.; methodology, F.G.V.; software, F.G.V. and F.D.; validation, F.G.V. and F.D.C.; Visualization, F.G.V., F.D.C., M.G.T., F.D. and J.L.V ;data curation, F.D.C. and M.G.T.; writing—original draft preparation, F.G.V.,F.D.C. and M.G.T.; writing–review and editing, F.G.V., F.D.C., M.G.T., F.D. and J.L.V; supervision, F.G.V. and F.D.; project administration, F.G.V.; Funding acquisition J.L.V

Funding: This research was funded by Universidad Americana de Paraguay.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DEG	Differentially Expressed Genes
FC	Fold Change
GO	Gene Ontology
GN	Gene Networks
PCoA	Principal Coordinates Analysis
MDS	Multidimensional scaling
T2DM	Type 2 diabetes mellitus

Appendix A. Definitions of the Used Co-Expression Measures

Co-expression measures are used in a bivariate analysis measuring the association strength between two genes, and whether this relationship is positive or negative. The presented co-expression measures take values ranking from -1 to +1. Hence, a value of +1 indicates a perfect positive association degree between two genes, whereas a value of -1, does likewise in perfect negative correlations. On the other hand, a value towards 0, is indicative of no/weak relationship.

The three chosen co-expression measures accomplish the above-mentioned features, which makes them suitable for a straight-forward ensemble strategy implementation. In the following subsections, these measures are described in mathematical detail.

Appendix A.1. Kendall Co-Expression Measure

Kendall co-expression measure is a non-parametric hypothesis test which assess the weight of a relationship between two genes, e.g., a and b, whose expression level has been measured n times. Hence, the total number of pairings between a and b is $n(n-1)/2$.

The dataset containing all n expression level observations corresponding to genes a and b will look like (a_1, b_1), (a_2, b_2), ..., (a_n, b_n). Thus, for every pair of observations (a_i, b_i) and (a_j, b_j), given $j > i$, are considered concordant if $a_i > a_j$ and $b_j > b_j$, or if, $a_i < a_j$ and $b_j < b_j$. In the contrary case, if $a_i > a_j$ and $b_j < b_j$, or if, $a_i < a_j$ and $b_j > b_j$, the pair of observations is considered discordant [46]. Hence, Kendall co-expression measure can be estimated using the following equation:

$$\tau = \frac{Nc - Nd}{\frac{1}{2}n\,(n-1)}$$

Where Nc refers to the number of concordant pairs of observations and Nd to the number of discordant pairs of observations. Finally τ refers to Kendall co-expression value.

Appendix A.2. Spearman Co-Expression Measure

Spearman co-expression measure is also a non-parametric hypothesis test which assess the degree of relationship between two genes a and b, which have been observed at their expression level n times. The Spearman co-expression measure does not consider any prior assumption on the data distribution and it is useful in the analysis of monotonic relationships (linear or not).

Again, datasets for each gene pair looks like $a = (a_1, ..., a_n)$ & $b = (b_1, ..., b_n)$. In this case, the Spearman co-expression measure acts on the ranks of the data rather than the raw data. This way, the respective ranks for both distributions, of the form $(R_1, ..., R_n)$ and $(S_1, ..., S_n)$, are calculated [82]. Thus, the Spearman co-expression measure can be calculated using the following formula:

$$\rho = 1 - \frac{\sum_{i=1}^{n}(R_i - S_i)^2}{n\,(n^2 - 1)}$$

Where ρ refers to Spearman co-expression value and n is the number of observations.

Appendix A.3. Blomqvist Co-Expression Measure

Finally, Blomqvist co-expression measure is also a non-parametric hypothesis test for the association of two genes. This measure places the emphasis on the difference of observed values among the first ranks in the orderings induced by the variables.

Again if $(a_1, b_1), ..., (a_n, b_n)$ represent the expression level of genes a and b across n measurements, a cumulative distribution function (cdf) can be defined as cdf F (a, b). Provided $\bar{a}$ and $\bar{b}$ denote the average expression level for genes a and b, let the a,b plane be divided in four areas by the lines x = $\bar{a}$ and b = $\bar{b}$. Thus, information on the co-expression of these genes can be obtained from the number of samples belonging to any of the quadrants 1 or 3 (n_1), when compared with the number of samples belonging to either the second or fourth quadrant (n_1) [83]. Blomqvist co-expression measure is then defined as:

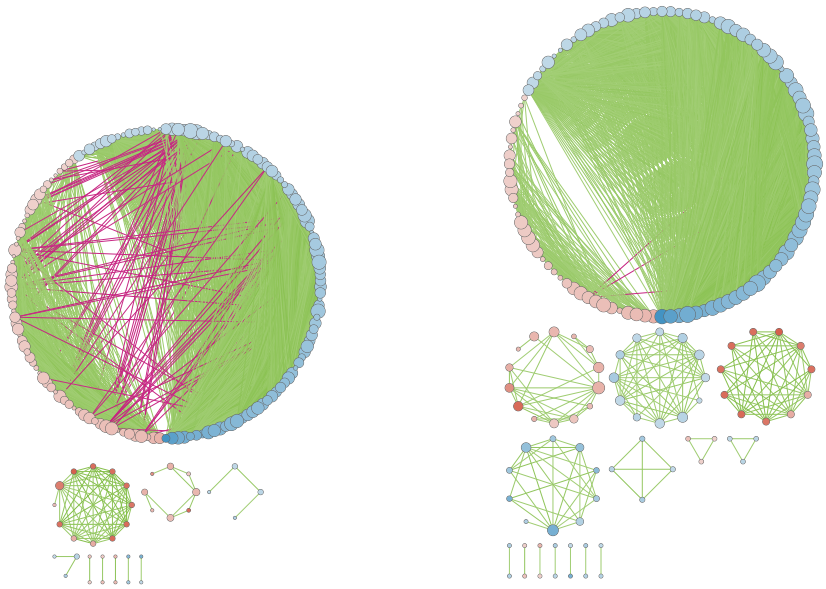
$$B = \frac{2n_1}{n_1 + n_2} - 1 - 1 \leq B \leq 1$$

Appendix B. Reconstructed Networks with High Thresholding

The cancer and no cancer networks corresponding to mild co-expression (0.7) are shown in Figure A1. These networks would proceed for latter topology and enrichment analysis as preliminary analyses revealed their suitability for the goal of our study.

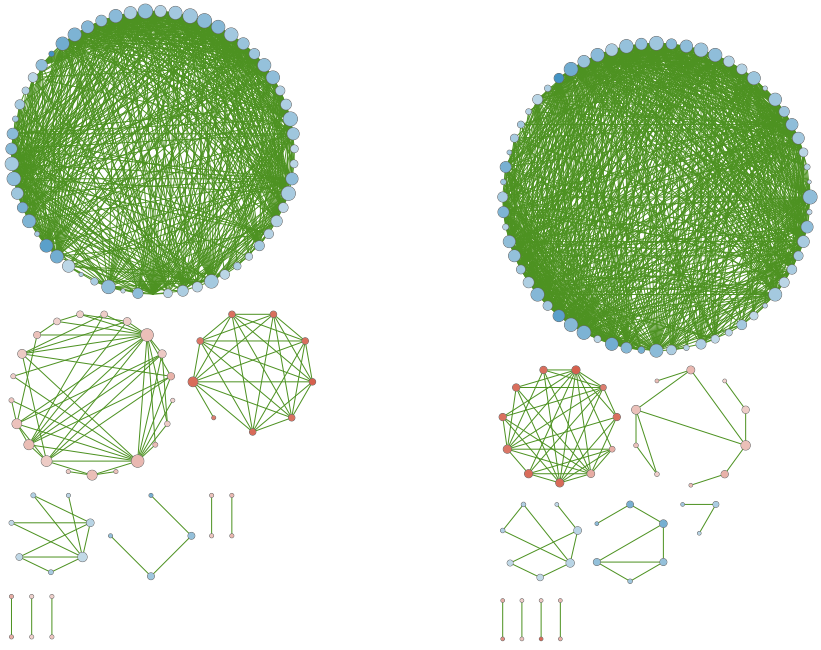
As mentioned in the main text, strong and very strong co-expression thresholds, respectively 0.8 and 0.9, were also used for the GN inference process. The cancer network for the strong co-expression threshold (weight cutoff: 0.8) comprised 110 nodes and 740 rods, whereas its non cancer equivalent comprised 109 nodes and 888 rods. On the other hand, the cancer network for the very strong co-expression threshold (weight cutoff: 0.9) comprised 15 nodes and 17 rods, whereas its non cancer counterpart comprised 21 nodes and 38 rods.

Notably, all co-expressions in these networks are positive. Clustering also revealed genetic interactions in the case of the 0.8 network (Figure A2). Nodes within these clusters represent around the 50% of the total number of nodes in these networks. After conducting similar analyses to the one presented with the 0.7 network, no new biological results were found for these networks compared to those already exposed.



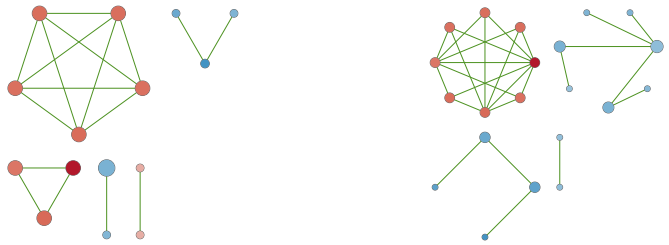
(**a**) (**b**)

Figure A1. Inferred networks corresponding to (**a**) cancerous samples and (**b**) non cancerous samples, using the mild co-expression threshold (0.7). Log2 FC is represented by node color, so blue and red intensity is related to gene up or down regulation respectively. Positive co-expressions are represented in green and negative co-expressions are shown in magenta. Node size is represented according to their rank. Edge transparency is represented according to edge weight. Circle layout is represented for independent clusters. Note both networks are clustered, showing a major connected module.



(**a**) (**b**)

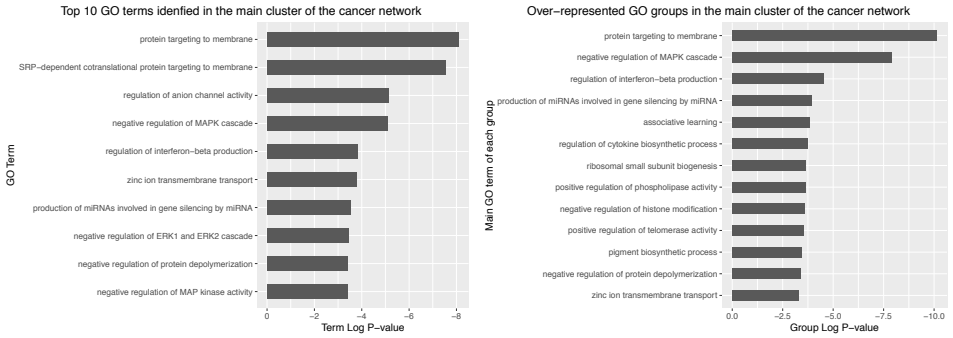
Figure A2. Inferred networks corresponding to (**a**) cancerous samples and (**b**) non cancerous samples, using the strong co-expression threshold (0.8). Log2 FC is represented by node color, so blue and red intensities are related to gene up or down regulation respectively. Node size is represented according to their rank. Positive co-expressions are represented in green and negative co-expressions are shown in magenta. Edge transparency is represented according to edge weight. Circle layout is represented for independent clusters. Note both networks show a major connected module.



(**a**) (**b**)

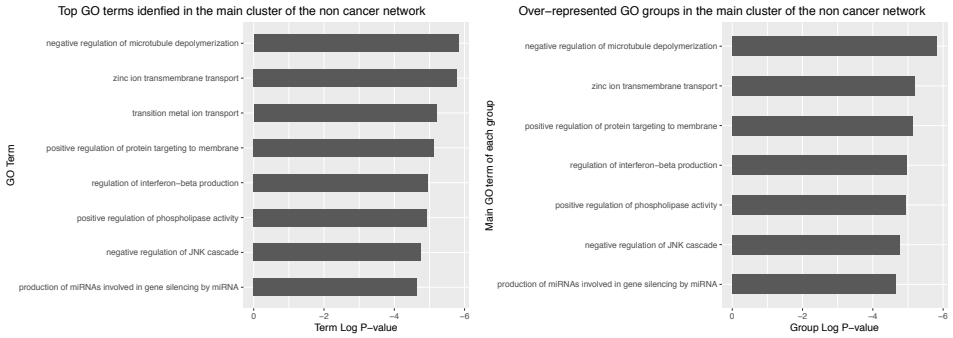
Figure A3. Inferred networks corresponding to (**a**) cancerous samples and (**b**) non cancerous samples, using the very strong co-expression threshold (0.9). Log2 FC is represented by node color, so blue and red intensity is related to gene up or down regulation respectively. Node size is represented according to their rank. Positive co-expressions are represented in green and negative co-expressions are shown in magenta. Edge transparency is represented according to edge weight. Circle layout is represented for independent clusters.

Appendix C. Main Over-Represented GO Terms and GO Groups for the Cancer and Non Cancer Networks



(**a**) (**b**)

Figure A4. (**a**) Top 10 GO terms over-represented by the genes comprised in the main cluster of the cancer network. (**b**) GO groups over-represented by the genes comprised in the main cluster of the cancer network. The main GO term of each identified group is presented as group label. Term and group *p*-value was corrected with Bonferroni step-down. Note the slower the *p*-value, the more the over-represented the GO term is.



(**a**) (**b**)

Figure A5. (**a**) Top 10 GO terms over-represented by the genes comprised in the main cluster of the non cancer network. (**b**) GO groups over-represented by the genes comprised in the main cluster of the non cancer network. The main GO term of each identified group is presented as group label. Term and group *p*-value was corrected with Bonferroni step-down. Note the slower the *p*-value, the more the over-represented the GO term is.

Appendix D. Detailed Lists of Sample Type-Exclusive Genes

Comparison of the cancer vs. the non cancer network yielded a list of 28 cancer-exclusive genes. These were submitted to the *DAVID* database for information retrieval, which is shown in Table A1. According to *DAVID* functional analysis genes are VAMP3, HMGCR and KLF4 are related to T2DM. Besides, HMGCR is also related to lung cancer.

In Table A2, the 14 cancer-exclussive genes that were found to share the GO term 'developmental process' (GO:0032502) are listed. This GO term is related to processes resulting in the progression of subcellular structures, cells, tissues or organs from a starting situation to a final situation. This could be related to tumor progression airway epithelial cells.

Table A1. The 28 cancer-exclusive genes, found in the main cluster of the cancer network which were not found at its non-cancer counterpart. Regulation refers to the increase (up) or decrease (down) of the gene expression levels.

Affymetrix ID	Gene Name	Gene Description	Regulation
202539_s_at	HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase	Down
211672_s_at	ARPC4-TTLL3	ARPC4-TTLL3 readthrough	Down
209288_s_at	CDC42EP3	CDC42 effector protein 3	Down
213826_s_at	H3F3A	H3 histone family member 3A	Up
220266_s_at	KLF4	Kruppel like factor 4	Down
212327_at	LIMCH1	LIM and calponin homology domains 1	Down
207480_s_at	MEIS2	Meis homeobox 2	Down
217549_at	NCKAP1L	NCK associated protein 1 like	Up
203582_s_at	SPHAR	S-phase response (cyclin related)	Down
216064_s_at	AGA	aspartylglucosaminidase	Down
201942_s_at	CPD	carboxypeptidase D	Down
203492_x_at	CEP57	centrosomal protein 57	Down
213753_x_at	EIF5A	eukaryotic translation initiation factor 5A	Up
218343_s_at	GTF3C3	general transcription factor IIIC subunit 3	Down
206483_at	LRRC6	leucine rich repeat containing 6	Down
218212_s_at	MOCS2	molybdenum cofactor synthesis 2	Down
206302_s_at	NUDT4	nudix hydrolase 4	Down
208447_s_at	PRPS1	phosphoribosyl pyrophosphate synthetase 1	Down
200730_s_at	PTP4A1	protein tyrosine phosphatase type IVA, member 1	Down
218276_s_at	SAV1	salvador family WW domain containing protein 1	Down
203908_at	SLC4A4	solute carrier family 4 member 4	Down
217975_at	TCEAL9	transcription elongation factor A like 9	Down
209149_s_at	TM9SF1	transmembrane 9 superfamily member 1	Down
204426_at	TMED2	transmembrane p24 trafficking protein 2	Down
211689_s_at	TMPRSS2	transmembrane protease, serine 2	Down
214007_s_at	TWF1	twinfilin actin binding protein 1	Down
211763_s_at	UBE2B	ubiquitin conjugating enzyme E2 B	Down
201337_s_at	VAMP3	vesicle associated membrane protein 3	Down

Table A2. List of 14/28 cancer-exclusive genes associated with the GO term developmental process (GO:0032502).

Affymetrix ID	Gene Name	Gene Description
202539_s_at	HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase
209288_s_at	CDC42EP3	CDC42 effector protein 3
213826_s_at	H3F3A	H3 histone family member 3A
220266_s_at	KLF4	Kruppel like factor 4
207480_s_at	MEIS2	Meis homeobox 2
217549_at	NCKAP1L	NCK associated protein 1 like
203492_x_at	CEP57	centrosomal protein 57
206483_at	LRRC6	leucine rich repeat containing 6
208447_s_at	PRPS1	phosphoribosyl pyrophosphate synthetase 1
200730_s_at	PTP4A1	protein tyrosine phosphatase type IVA, member 1
218276_s_at	SAV1	salvador family WW domain containing protein 1
204426_at	TMED2	transmembrane p24 trafficking protein 2
211763_s_at	UBE2B	ubiquitin conjugating enzyme E2 B
201337_s_at	VAMP3	vesicle associated membrane protein 3

Table A3. The 7 non cancer-exclusive genes identified at the main cluster of the non cancer network which were not found at its cancer counterpart. Regulation refers to the increase (up) or decrease (down) of the gene expression levels. Note some Affymetrix IDs could not be mapped.

Affymetrix ID	Gene Name	Gene Description	Regulation
212206_s_at	H2AFV	H2A histone family member V	Down
209703_x_at	METTL7A	methyltransferase like 7A	Up
217734_s_at	WDR6	WD repeat domain 6	Up
215359_x_at	LOC101060181	zinc finger protein ZnFP12	Up
222339_x_at	-	-	Up
220856_x_at	-	-	Up
208082_x_at	-	-	Up

References

1. Delgado, F.M.; Gómez-Vela, F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. In *Artificial Intelligence in Medicine*; Elsevier: Amsterdam, The Netherlands, 2018.
2. Díaz-Montaña, J.J.; Gómez-Vela, F.; Díaz-Díaz, N. GNC–app: A new Cytoscape app to rate gene networks biological coherence using gene–gene indirect relationships. *Biosystems* **2018**, *166*, 61–65. [CrossRef] [PubMed]
3. Gómez-Vela, F.; Lagares, J.A.; Díaz-Díaz, N. Gene network coherence based on prior knowledge using direct and indirect relationships. *Comput. Biol. Chem.* **2015**, *56*, 142–151. [CrossRef] [PubMed]
4. Darrason, M. Mechanistic and topological explanations in medicine: The case of medical genetics and network medicine. *Synthese* **2018**, *195*, 147–173. [CrossRef]
5. Liang, L.; Gao, L.; Zou, X.P.; Huang, M.L.; Chen, G.; Li, J.J.; Cai, X.Y. Diagnostic significance and potential function of miR-338-5p in hepatocellular carcinoma: A bioinformatics study with microarray and RNA sequencing data. *Mol. Med. Rep.* **2018**, *17*, 2297–2312. [CrossRef] [PubMed]
6. Gómez-Vela, F.; Barranco, C.D.; Díaz-Díaz, N. Incorporating biological knowledge for construction of fuzzy networks of gene associations. *Appl. Soft Comput.* **2016**, *42*, 144–155. [CrossRef]
7. Gómez-Vela, F.; Rodriguez-Baena, D.S.; Vázquez-Noguera, J.L. Structure Optimization for Large Gene Networks Based on Greedy Strategy. *Comput. Math. Methods Med.* **2018**, *2018*. [CrossRef]
8. Zhao, W.; Langfelder, P.; Fuller, T.; Dong, J.; Li, A.; Hovarth, S. Weighted gene coexpression network analysis: State of the art. *J. Biopharm. Stat.* **2010**, *20*, 281–300. [CrossRef]
9. Kumari, S.; Nie, J.; Chen, H.S.; Ma, H.; Stewart, R.; Li, X.; Lu, M.Z.; Taylor, W.M.; Wei, H. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS ONE* **2012**, *7*, e50411. [CrossRef]
10. De Siqueira Santos, S.; Takahashi, D.Y.; Nakata, A.; Fujita, A. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings Bioinform.* **2013**, *15*, 906–918. [CrossRef]

11. D'haeseleer, P.; Liang, S.; Somogyi, R. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics* **2000**, *16*, 707–726. [CrossRef]
12. Emamjomeh, A.; Robat, E.S.; Zahiri, J.; Solouki, M.; Khosravi, P. Gene co-expression network reconstruction: A review on computational methods for inferring functional information from plant-based expression data. *Plant Biotechnol. Rep.* **2017**, *11*, 71–86. [CrossRef]
13. Liu, Z.P. Quantifying gene regulatory relationships with association measures: A comparative study. *Front. Genet.* **2017**, *8*, 96. [CrossRef] [PubMed]
14. Zhong, R.; Allen, J.D.; Xiao, G.; Xie, Y. Ensemble-based network aggregation improves the accuracy of gene network reconstruction. *PLoS ONE* **2014**, *9*, e106319. [CrossRef] [PubMed]
15. Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* **2009**, *96*, 86–103. [CrossRef]
16. Wang, Z.; Xu, P.; Chen, B.; Zhang, Z.; Zhang, C.; Zhan, Q.; Huang, S.; Xia, Z.; Peng, W. Identifying circRNA-associated-ceRNA networks in the hippocampus of Aβ1-42-induced Alzheimer's disease-like rats using microarray analysis. *Aging (Albany NY)* **2018**, *10*, 775. [CrossRef]
17. Jemal, A.; Bray, F.; Center, M.M.; Ferlay, J.; Ward, E.; Forman, D. Global cancer statistics. *CA Cancer J. Clin.* **2011**, *61*, 69–90. [CrossRef]
18. WHO. *International Programme on Chemical Safety: Biomarkers in Risk Assessment: Validity and Validation, 2001*; WHO: Genewa, Switzerland, 2015.
19. Murray, J.F.; Nadel, J.A. *Murray & Nadel's Textbook of Respiratory Medicine*; Elsevier Saunders: Amsterdam, The Netherlands, 2016.
20. Stewart, B.; Wild, C.P. *World Cancer Report 2014*; WHO: Genewa, Switzerland, 2014.
21. Peng, G.; Tisch, U.; Adams, O.; Hakim, M.; Shehada, N.; Broza, Y.Y.; Billan, S.; Abdah-Bortnyak, R.; Kuten, A.; Haick, H. Diagnosing lung cancer in exhaled breath using gold nanoparticles. *Nat. Nanotechnol.* **2009**, *4*, 669. [CrossRef]
22. Aggarwal, A.; Guo, D.L.; Hoshida, Y.; Yuen, S.T.; Chu, K.M.; So, S.; Boussioutas, A.; Chen, X.; Bowtell, D.; Aburatani, H.; et al. Topological and Functional Discovery in a Gene Coexpression Meta-Network of Gastric Cancer. *Cell Tumor Stem Cell Biol.* **2006**, *66*, 232–241. [CrossRef]
23. Ma, S.; Shi, M.; Li, Y.; Yi, D.; Shia, B.C. Incorporating gene co-expression network in identification of cancer prognosis markers. *BMC Bioinform.* **2010**, *11*, 271. [CrossRef]
24. Clarke, C.; Madden, S.F.; Doolan, P.; Aherne, S.T.; Joyce, H.; O'Driscoll, L.; Gallagher, W.M.; Hennessy, B.T.; Moriarty, M.; Crown, J.; et al. Correlating transcriptional networks to breast cancer survival: A large-scale coexpression analysis. *Carcinogenesis* **2013**, *34*, 2300–2308. [CrossRef]
25. Chang, X.; Shi, L.; Gao, F.; Russin, J.; Zeng, L.; He, S.; Chen, T.C.; Giannotta, S.L.; Weisenberger, D.J.; Zada, G.; et al. Genomic and transcriptome analysis revealing an oncogenic functional module in meningiomas. *Neurosurg. Focus* **2013**, *35*, E3. [CrossRef] [PubMed]
26. Yang, Y.; Han, L.; Yuan, Y.; Li, J.; Hei, N.; Liang, H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **2014**, *5*, 3231. [CrossRef] [PubMed]
27. Liu, R.; Cheng, Y.; Yu, J.; Lv, Q.L.; Zhou, H.H. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene* **2015**, *563*, 56–62. [CrossRef] [PubMed]
28. Yang, L.; Li, Y.; Wei, Z.; Chang, X. Coexpression network analysis identifies transcriptional modules associated with genomic alterations in neuroblastoma. *Biochim. Biophys. Acta (BBA) Mol. Basis Dis.* **2018**, *1864*, 2341–2348. [CrossRef] [PubMed]
29. Xu, B.; Lv, W.; Li, X.; Zhang, L.; Lin, J. Prognostic genes of hepatocellular carcinoma based on gene coexpression network analysis. *J. Cell. Biochem.* **2019**, *120*, 11616–11623. [CrossRef] [PubMed]
30. Spira, A.; Beane, J.E.; Shah, V.; Steiling, K.; Liu, G.; Schembri, F.; Gilman, S.; Dumas, Y.M.; Calner, P.; Sebastiani, P.; et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.* **2007**, *13*, 361. [CrossRef]
31. Gustafson, A.M.; Soldi, R.; Anderlind, C.; Scholand, M.B.; Qian, J.; Zhang, X.; Cooper, K.; Walker, D.; McWilliams, A.; Liu, G.; et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci. Transl. Med.* **2010**, *2*, 26ra25–26ra25. [CrossRef]
32. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [CrossRef]

33. Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **1996**, *5*, 299–314.
34. Gruvberger, S.; Ringnér, M.; Chen, Y.; Panavally, S.; Saal, L.H.; Borg, Å.; Fernö, M.; Peterson, C.; Meltzer, P.S. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **2001**, *61*, 5979–5984.
35. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: Berlin, Germany, 2016.
36. Smyth, G.K. Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Springer: Berlin, Germany, 2005; pp. 397–420.
37. Diboun, I.; Wernisch, L.; Orengo, C.A.; Koltzenburg, M. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genom.* **2006**, *7*, 252. [CrossRef] [PubMed]
38. Ben-Dov, I.Z.; Whalen, V.M.; Goilav, B.; Max, K.E.; Tuschl, T. Cell and microvesicle urine microRNA deep sequencing profiles from healthy individuals: Observations with potential impact on biomarker studies. *PLoS ONE* **2016**, *11*, e0147249. [CrossRef] [PubMed]
39. Thissen, D.; Steinberg, L.; Kuang, D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *J. Educ. Behav. Stat.* **2002**, *27*, 77–83. [CrossRef]
40. Marbach, D.; Mattiussi, C.; Floreano, D. Combining multiple results of a reverse engineering algorithm: Application to the DREAM five gene network challenge. *Ann. N. Y. Acad. Sci.* **2009**, *1158*, 102–113. [CrossRef] [PubMed]
41. Zainal, A.; Maarof, M.A.; Shamsuddin, S.M. Ensemble classifiers for network intrusion detection system. *J. Inf. Assur. Secur.* **2009**, *4*, 217–225.
42. Altay, G.; Emmert-Streib, F. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics* **2010**, *26*, 1738–1744. [CrossRef] [PubMed]
43. Martin, A.J.; Dominguez, C.; Contreras-Riquelme, S.; Holmes, D.S.; Perez-Acle, T. Graphlet Based Metrics for the comparison of gene regulatory networks. *PLoS ONE* **2016**, *11*, e0163497. [CrossRef]
44. Mukaka, M.M. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **2012**, *24*, 69–71.
45. Cooke, D.; Clarke, G. *A Basic Course in Statistics*; Arnold: London, UK, 1989.
46. Kendall, M.G. *Rank Correlation Methods*; American Psychological Association: Washington, DC, USA, 1948.
47. Mindrila, D.; Balentyne, P. Scatterplots and correlation. *Retrieved From* **2017**. Available online: https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf (accessed on 15 October 2019).
48. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **2006**, *1695*, 1–9.
49. Kolaczyk, E.D.; Csárdi, G. *Statistical Analysis of Network Data with R*; Springer: Berlin, Germany, 2014; Volume 65.
50. Contreras-Lopez, O.; Moyano, T.C.; Soto, D.C.; Gutiérrez, R.A. Step-by-step construction of gene co-expression networks from high-throughput arabidopsis RNA sequencing data. In *Root Development*; Springer: Berlin, Germany, 2018; pp. 275–301.
51. Godsil, C.; Royle, G.F. *Algebraic Graph Theory*; Springer Science & Business Media: Berlin, Germany, 2013; Volume 207.
52. Parikshak, N.N.; Gandal, M.J.; Geschwind, D.H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **2015**, *16*, 441. [CrossRef] [PubMed]
53. Rizzolo, K.; Kumar, A.; Kakihara, Y.; Phanse, S.; Minic, Z.; Snider, J.; Stagljar, I.; Zilles, S.; Babu, M.; Houry, W.A. Systems analysis of the genetic interaction network of yeast molecular chaperones. *Mol. Omics* **2018**, *14*, 82–94. [CrossRef] [PubMed]
54. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
55. Nepomuceno-Chamorro, I.A.; Aguilar-Ruiz, J.S.; Riquelme, J.C. Inferring gene regression networks with model trees. *BMC Bioinform.* **2010**, *11*, 517. [CrossRef] [PubMed]
56. Morris, J.H.; Apeltsin, L.; Newman, A.M.; Baumbach, J.; Wittkop, T.; Su, G.; Bader, G.D.; Ferrin, T.E. clusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinform.* **2011**, *12*, 436. [CrossRef] [PubMed]

57. Su, G.; Kuchinsky, A.; Morris, J.H.; States, D.J.; Meng, F. GLay: Community structure analysis of biological networks. *Bioinformatics* **2010**, *26*, 3135–3137. [CrossRef] [PubMed]
58. Milenković, T.; Pržulj, N. Uncovering biological network function via graphlet degree signatures. *Cancer Inform.* **2008**, *6*, S680. [CrossRef]
59. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2008**, *37*, 1–13. [CrossRef]
60. Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **2009**, *25*, 1091–1093. [CrossRef]
61. Bindea, G.; Galon, J.; Mlecnik, B. CluePedia Cytoscape plugin: Pathway insights using integrated experimental and in silico data. *Bioinformatics* **2013**, *29*, 661–663. [CrossRef]
62. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44. [CrossRef] [PubMed]
63. Jiao, X.; Sherman, B.T.; Huang, D.W.; Stephens, R.; Baseler, M.W.; Lane, H.C.; Lempicki, R.A. DAVID-WS: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **2012**, *28*, 1805–1806. [CrossRef]
64. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a shared vision for cancer genomic data. *New Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef] [PubMed]
65. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M.; Network, C.G.A.R.; et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113. [CrossRef] [PubMed]
66. Apiwattanakul, M.; Milone, M.; Pittock, S.J.; Kryzer, T.J.; Fryer, J.P.; O'toole, O.; Mckeon, A.; Lennon, V.A. Signal recognition particle immunoglobulin g detected incidentally associates with autoimmune myopathy. *Muscle Nerve* **2016**, *53*, 925–932. [CrossRef]
67. Ren, Y.G.; Wagner, K.W.; Knee, D.A.; Aza-Blanc, P.; Nasoff, M.; Deveraux, Q.L. Differential regulation of the TRAIL death receptors DR4 and DR5 by the signal recognition particle. *Mol. Biol. Cell* **2004**, *15*, 5064–5074. [CrossRef]
68. Zhong, B.; Bian, L.; Wang, G.; Zhou, Y.; Chen, Y.; Peng, F. Identification of key genes involved in HER2-positive breast cancer. *Eur. Rev. Med. Pharmacol. Sci.* **2016**, *20*, 664–672.
69. Fahrmann, J.F.; Grapov, D.; Wanichthanarak, K.; DeFelice, B.C.; Salemi, M.R.; Rom, W.N.; Gandara, D.R.; Phinney, B.S.; Fiehn, O.; Pass, H.; et al. Integrated metabolomics and proteomics highlight altered nicotinamide and polyamine pathways in lung adenocarcinoma. *Carcinogenesis* **2017**, *38*, 271–280. [CrossRef]
70. Currie, C.J.; Poole, C.D.; Jenkins-Jones, S.; Gale, E.A.; Johnson, J.A.; Morgan, C.L. Mortality after incident cancer in people with and without type 2 diabetes: Impact of metformin on survival. *Diabetes Care* **2012**, *35*, 299–304. [CrossRef]
71. Currie, C.; Poole, C.; Gale, E. The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia* **2009**, *52*, 1766–1777. [CrossRef]
72. Evans, J.M.; Donnelly, L.A.; Emslie-Smith, A.M.; Alessi, D.R.; Morris, A.D. Metformin and reduced risk of cancer in diabetic patients. *Bmj* **2005**, *330*, 1304–1305. [CrossRef] [PubMed]
73. Hsu, P.P.; Sabatini, D.M. Cancer cell metabolism: Warburg and beyond. *Cell* **2008**, *134*, 703–707. [CrossRef] [PubMed]
74. Hamanaka, R.B.; Chandel, N.S. Targeting glucose metabolism for cancer therapy. *J. Exp. Med.* **2012**, *209*, 211–215. [CrossRef] [PubMed]
75. Hirsch, H.A.; Iliopoulos, D.; Tsichlis, P.N.; Struhl, K. Metformin selectively targets cancer stem cells, and acts together with chemotherapy to block tumor growth and prolong remission. *Cancer Res.* **2009**, *69*, 7507–7511. [CrossRef]
76. Heller, G.; Babinsky, V.N.; Ziegler, B.; Weinzierl, M.; Noll, C.; Altenberger, C.; Müllauer, L.; Dekan, G.; Grin, Y.; Lang, G.; et al. Genome-wide CpG island methylation analyses in non-small cell lung cancer patients. *Carcinogenesis* **2012**, *34*, 513–521. [CrossRef]
77. Teng, Y.; Qin, H.; Bahassan, A.; Bendzunas, N.G.; Kennedy, E.J.; Cowell, J.K. The WASF3–NCKAP1–CYFIP1 complex is essential for breast cancer metastasis. *Cancer Res.* **2016**, *76*, 5133–5142. [CrossRef]
78. Xiao, C.Z.; Wei, W.; Guo, Z.X.; Zhang, M.Y.; Zhang, Y.F.; Wang, J.H.; Shi, M.; Wang, H.Y.; Guo, R.P. MicroRNA-34c-3p promotes cell proliferation and invasion in hepatocellular carcinoma by regulation of NCKAP1 expression. *J. Cancer Res. Clin. Oncol.* **2017**, *143*, 263–273. [CrossRef]

79. Luce, L.N.; Abbate, M.; Cotignola, J.; Giliberto, F. Non-myogenic tumors display altered expression of dystrophin (DMD) and a high frequency of genetic alterations. *Oncotarget* **2017**, *8*, 145. [CrossRef]
80. Duggan, D.J.; Bittner, M.; Chen, Y.; Meltzer, P.; Trent, J.M. Expression profiling using cDNA microarrays. *Nat. Genet.* **1999**, *21*, 10. [CrossRef]
81. Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **1999**, *6*, 281–297. [CrossRef]
82. Majd, H.A.; Talebi, A.; Gilany, K.; Khayyer, N. Two-Way Gene Interaction From Microarray Data Based on Correlation Methods. *Iran. Red Crescent Med. J.* **2016**, *18*.
83. Blomqvist, N. On a measure of dependence between two random variables. *Ann. Math. Stat.* **1950**, *21*, 593–600. [CrossRef]

Chapter 7

Ensemble and Greedy Approach for the Reconstruction of Large Gene Co-Expression Networks

Authors	Francisco Gómez-Vela, Fernando M. Delgado-Chaves, Domingo S. Rodríguez-Baena, Miguel García-Torres and Federico Divina
Journal	Entropy
Editorial	MDPI
eISSN	1099-4300
Published	21/11/2019
DOI	doi.org/10.3390/e21121139
Impact factor 2021	2,738
Quartile	Q2

entropy

MDPI

Article

Ensemble and Greedy Approach for the Reconstruction of Large Gene Co-Expression Networks

Francisco Gómez-Vela [1,*], Fernando M. Delgado-Chaves [2], Domingo S. Rodríguez-Baena [1], Miguel García-Torres [1] and Federico Divina [1]

[1] Computer Science Division, Pablo de Olavide University, ES-41013 Seville, Spain; dsrodbae@upo.es (D.S.R.-B.); mgarciat@upo.es (M.G.-T.); fdiv@upo.es (F.D.)

[2] Faculty of Experimental Sciences, Pablo de Olavide University, ES-41013 Seville, Spain; fmdelcha@alu.upo.es

* Correspondence: fgomez@upo.es

Received: 1 November 2019; Accepted: 13 November 2019; Published: 21 November 2019

Abstract: Gene networks have become a powerful tool in the comprehensive analysis of gene expression. Due to the increasing amount of available data, computational methods for networks generation must deal with the so-called curse of dimensionality in the quest for the reliability of the obtained results. In this context, ensemble strategies have significantly improved the precision of results by combining different measures or methods. On the other hand, structure optimization techniques are also important in the reduction of the size of the networks, not only improving their topology but also keeping a positive prediction ratio. In this work, we present Ensemble and Greedy networks (EnGNet), a novel two-step method for gene networks inference. First, EnGNet uses an ensemble strategy for co-expression networks generation. Second, a greedy algorithm optimizes both the size and the topological features of the network. Not only do achieved results show that this method is able to obtain reliable networks, but also that it significantly improves topological features. Moreover, the usefulness of the method is proven by an application to a human dataset on post-traumatic stress disorder, revealing an innate immunity-mediated response to this pathology. These results are indicative of the method's potential in the field of biomarkers discovery and characterization.

Keywords: gene networks; scale-free networks; ensemble networks; graph theory; computational biology; mutual information networks; biomarkers discovery

1. Introduction

Arising at the beginning of the century, Gene Networks (GN) have become a breakthrough in the analysis of biological processes in most gene expression studies. Such networks represent relationships between genes (or gene products) by means of a graph composed of nodes and edges, where nodes represent genes and edges the relationships among them. GNs have been widely used in both basic and applied research, such as biology [1], medicine [2], and diagnostics [3], among others.

GNs models also pave the way for hypotheses-making, which can be empirically validated afterwards. The results show significant reliability of GNs in this sense, since many predicted interactions have been experimentally confirmed later [4]. Therefore, algorithms and computational methods for GNs reconstruction have gained relevance among the Bioinformatics community [5]. These methods usually take gene expression datasets as inputs, e.g., microarrays or RNA-Seq data, for the inference of gene–gene relationships. To a greater extent, the vast amount of genetic information generated in the last decade has allowed the inference of relationships among DNAs, RNAs, proteins and other cellular components [6,7].

In this context, it is possible to classify GNs according to the inference approach used, including Bayesian, information theory, Boolean, or differential equations models, among others [8]. Consistently with this classification, co-expression networks, which are based on information theory, appear as a remarkably relevant approach due to their computational simplicity and low computational demands [9]. These networks infer relationships between genes that show similar patterns of expression. This is achieved by measuring the degree of relationship between each pair of genes, so the relationship is only approved when this degree exceeds a certain threshold. This threshold value indicates the minimum level of similarity between two expression patterns for the relationship to be considered significant. Therefore, the higher this threshold is, the sparser inferred GNs will be [10]. According to the published literature, the main measures to evaluate the co-expression degree between two genes are correlation measures such as Pearson, Spearman or Kendall coefficients [11,12]. Additionally, other measures have been widely used for the generation of GNs such as Mutual information [13].

Nevertheless, co-expression networks usually present two main drawbacks: (a) the above-mentioned measures present some limitations [14], for example, their inability to detect non-linear dependencies or their dependence on the distribution of the data, as in the case of Spearman and Pearson coefficients, respectively [15]; and (b) inferred networks are often too densely-connected to perform comprehensive analyses, being actual GNs known to be sparse [16].

As far as the topology of the networks inferred is concerned, GNs should generally meet a series of requirements. First, GNs should follow a scale-free topology, as they have been proven to be sparse [17,18]. Thus far, scale-free GNs reconstruction entails a major challenge as algorithms themselves show limitations in distinguishing truly-significant interactions, thus providing densely-connected networks. Second, it is to be highlighted that biological networks contain hubs, which are genes influenced by a significant number of relationships. Hubs are then key elements in the control and regulation of the genes comprised in the network, and have proven their importance in the modeling and analysis of genetic interactions [19–21]. It follows that inferred GNs should contain hubs. As consequence of these two requirements, GNs topology optimization arises as a major issue to be faced.

In this work, we propose a novel approach for the reconstruction of large gene co-expression networks. In particular, we propose a two steps strategy to induce gene networks. In a first phase, an ensemble approach is used in order to generate co-expression networks. The so-obtained network is then optimized in a second stage, where a greedy strategy optimizes both the size and the topological features of the network.

Not only is this method able to overcome the limitations of using a single measure to assess gene co-expression thanks to an ensemble strategy, it also carries out a greedy heuristic topological optimization of the inferred GNs. Therefore, we can summarize our contributions as follows:

- The method is able to overcome the limitations of a single information theory measure thanks to an ensemble strategy.
- The method is also able to perform a topology optimization.
- The experiments carried out show that our approach achieved good results against other state of the art methods.
- The usefulness of the proposed method becomes evident in an application to a study of a post traumatic stress disorder on human dataset.
- The method's results show its potential in the field of biomarkers discovery and characterization.

1.1. Related Work

Co-expression analysis assumes that genes whose mRNAs show similar level of variation upon perturbations are involved in the same, or closely related, biological processes. Approaches based on such assumption haven been considered as promising for the discovery of genes implicated in

biological processes of interest [22]. Particularly, co-expression networks have provided valuable insights on diseases' underlying molecular mechanisms, as in the case of cancer [23].

In Reference [24], weighted gene co-expression networks were analyzed to investigate the role of gene regulation in lung cancer. Using Pearson correlation coefficients for gene pairs, the authors detected a lung cancer-specific module of co-expressed genes with clear functional interpretations. Pearson's measure, and the Weighted Gene Co-expression Network Analysis (WGCNA) methodology [25], were also used by Ivliev et al. [26] to identify gene co-expression modules covering a range of known tumour features. The WGCNA methodology implies not only taking into account the correlation between a gene pair, but also whether these genes are correlated with similar sets of genes across the entire transcriptome. Other works use different co-expression measures. For example, Yuejie et al. [27] assumed that two genes that use the same dictionary to represent their original expression values must share similar co-expression patterns. In this case, the authors used a sparse coding and dictionary learning algorithms.

Despite the good results achieved in previous approaches, the measures used present some limitations, as mentioned in the previous section. Thus, recent works have been focused on the possibility of combining different inference methods and co-expression measures. For example, in [28], an Ensemble-based Network Aggregation method (ENA) is proposed to integrate gene networks derived from different methods and datasets, in order to improve the accuracy of network inference. Other works try to combine different pre-processing methods (see, e.g., [29]). In this work, the network inference problem between g genes is decomposed as g separate regression problems. Thus, an ensemble of several feature selection algorithms are used to find those genes most suitable in modeling the expression values of every target gene. Besides looking for the best co-expression measure, other studies try to use different inference methods. In [30], three normalization methods and 10 inference methods, including six correlation and four mutual information methods, were tested. Liue et al. [31] presented a novel inference algorithm, namely Local Bayesian Network (LBN). This algorithm applies an iterative methodology, in which, firstly, conditional mutual information is used to generate an initial network. Then, it uses a k-nearest neighbor approach to decompose the network into smaller sub-networks. Finally, the algorithm identifies and removes redundant relationships between genes using a Bayesian method. These new sub-network are integrated into a new gene network and the process restarts until the topological structure of the network remains unchanged.

In addition, the optimization of gene co-expression networks represents a challenge due to the size and complexity of the data from which the networks are obtained. Hence, the goal is to reduce both size and complexity of the final network while maintaining biological relevance. Network structure optimization is a NP-hard problem, so some works use heuristic algorithms to explore the possible combinations of all interactions in order to simplify the network structure [32]. However, these approaches usually present computational limitations due to the high dimensionality of the networks [33]. Other works use evolutionary techniques to reduce the large search spaces. For example, in [34], a genetic algorithm is used to reconstruct gene networks from time-series expression profiles based on fuzzy cognitive maps. Some research works based their optimization efforts on objective functions and scores (see, e.g., [35]). In this work, an undirected confidence-weighted likelihood matrix is created using pairwise confidence scores from functional association databases. Using this matrix, GNs are inferred with a high accuracy level. Other researchers, e.g., Lopes et al. [36], use a scale-free topology information to prune search space during inference problem. Finally, in the research presented by Yang et al. [37], a bayesian-based inference process is used to evaluate the relative importance of nodes.

2. Materials and Methods

In this section, we present the different methods and datasets used in this paper. In particular, Section 2.1 describes the proposed method for large GNs reconstruction, while, in Section 2.2,

we describe the datasets used in the experiments. Finally, Section 2.3 introduces the measures used to assess the performance of the method.

2.1. EnGNet: Gene Network Reconstruction Based on Ensemble Strategy and Greedy Optimization

In this section, we introduce the proposed method for large co-expression networks generation, which we name Ensemble and Greedy networks (EnGNet). A EnGNet JAVA-based implementation is available at: https://github.com/fgomezvela/EnGNet (accessed on 15 November 2019). As previously introduced, EnGNet comprises two main steps, described in Figure 1: (a) an ensemble-based method to infer gene–gene co-expression relationships; and (b) a greedy strategy for the topological optimization of the network. As a result, the final network exhibits not only reliable interactions but also lower topological complexity and sparseness than other techniques that adopt single co-expression measurements. As stated in Section 1, the spareness in a GN is a desirable feature, involving a significant improvement over other methodologies.

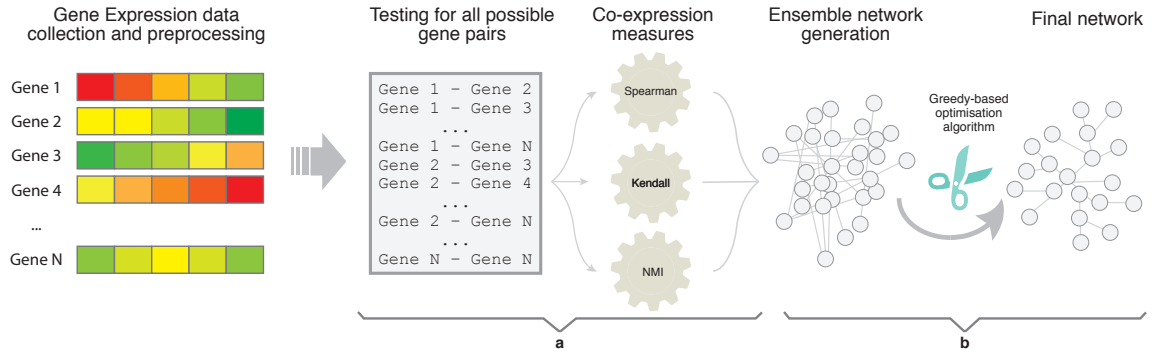


Figure 1. Global workflow of EnGNet for GNs reconstruction. As shown, the method is based on two different steps: (**a**) an ensemble strategy for network inference; and (**b**) a greedy-based approach for the final optimization (maximum spanning tree algorithm).

2.1.1. Ensemble Strategy for Network Generation

In the first phase, EnGNet induces a single co-expression network, using three different evaluation measures. In this case, three widely-used co-expression measures were selected for assessing the significance of gene–gene interactions. In particular, we used the Spearman, Kendall coefficients and Normalized Mutual Information (NMI) measures. Our choice is motivated by the following observations. The Spearman coefficient is able to detect linear dependencies between two genes unaffected by data distribution. Kendall's measure is also able to detect linear dependencies but has advantages over Spearman's in approaching distribution normality more rapidly [15]. Finally, the NMI is able to detect linear and also non-linear dependencies between genes [38].

The three measures used provide a value v_i, $0 \leq v_i \leq 1$, where 0 represents no dependency and 1 a total dependency between the genes.

The reconstruction process is based on the evaluation of all possible gene pairs. As shown in Figure 2, the three measures are used for evaluating every gene pair relationship. For each measure, a significance threshold ($Th_i, 1 \leq i \leq 3$) is used in order to determine whether or not the relationship is considered valid by a specific measure.

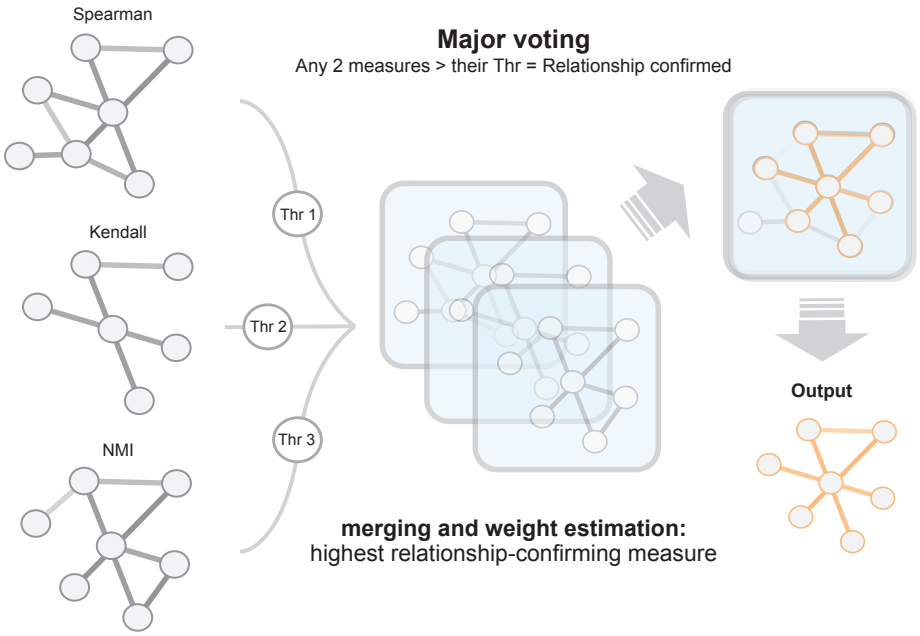


Figure 2. Schematic representation of the Ensemble step of EnGNet. Three well-known measures for the generation of co-expression networks are combined, here by Spearman, Kendall and NMI, by means of an ensemble strategy.

The final significance assessment is carried out through a voting system. Thus, a relationship is confirmed if it is considered significant by at least two measures (see Table 1).

Table 1. Example representation of the major voting strategy to evaluate gene pairs.

Gene Pair	Spearman	Kendall	NMI	Final
g_1, g_2	Correct	Correct	Correct	Correct
g_2, g_3	Incorrect	Correct	Correct	Correct
g_4, g_5	Incorrect	Correct	Incorrect	Incorrect
g_5, g_2	Correct	Incorrect	Incorrect	Incorrect

Hence, a relationship is added to the final network if it is considered correct, and its final weight, denoted as w_{en}, is set to the average value v_i of the three measures. Doing so, we subsume the information of the three measures in a single value. The so-created network represent the input to the second step of the proposed strategy.

2.1.2. Topological Optimization Based on Greedy MST Algorithm

In this step, the topological features of the network obtained in the first step described in previous section are optimized by means of two phases: pruning and adding relevant edges (see Figure 3). In the first phase, the ensemble network is pruned using a greedy-based heuristic algorithm, which takes into account the most relevant interactions, i.e., those showing the highest co-expression weight according to ensemble step. In particular, we used the modification of the Kruskal's minimum spanning tree (MST) algorithm presented in [7] to obtain the longest path between each pair of genes. This modification consists of selecting the most significant edges, instead the less relevant ones, until all nodes are connected with no cycles. Thanks to this, the method obtains the most significant path between each pair of nodes that comprising the network [39].

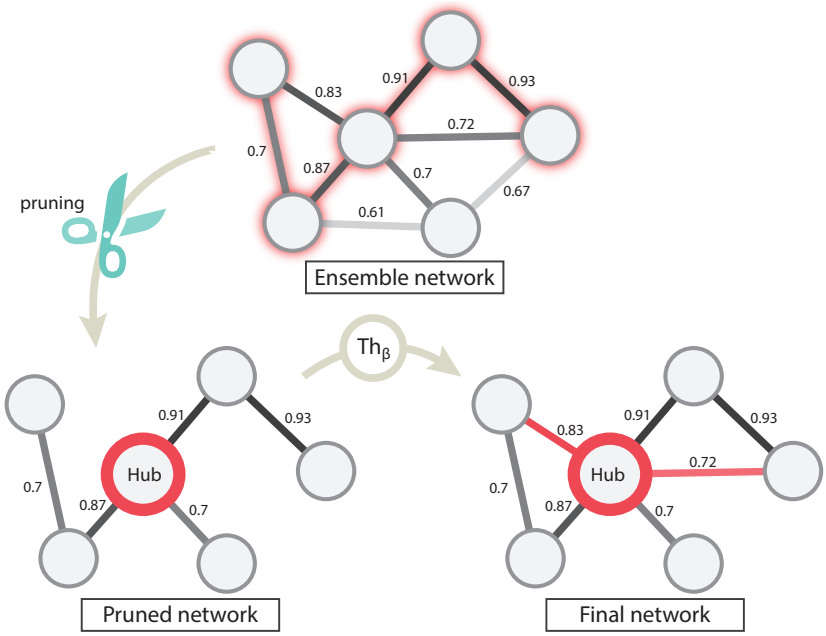


Figure 3. Graphical description of the second step of EnGNet. First the previously-obtained ensemble network is pruned by a MST (minimum spanning tree)greedy algorithm. In a second phase, the most relevant edges, which were initially pruned, are evaluated with a threshold (Th_β), and added to the final network again.

As a result, the method computes a pruned network (see "Pruned network" in Figure 3), which contains the same number of nodes as the original network, albeit keeping only most relevant relationships. This reduction in edges significantly improves the sparseness of the network.

However, not all removed relationships are necessarily irrelevant to the network. For this reason, in the second phase, a topological analysis of the pruned network is performed in order to identify network's hubs. As stated in Section 1, hubs play a crucial role in how the information is distributed through the network and usually these are key regulators of the genes involved. For this reason, hubs are selected as those nodes whose connection degree exceeds the average network connectivity (see "Pruned network" in Figure 3 where the hub is highlighted as the node showing the greatest number of relationships).

Once the hubs have been identified upon the pruning process, they are independently processed. For each hub, its linking edges that were removed in the ensemble network are again evaluated using a threshold Th_β. This threshold is a user set parameter, which is employed to determine the biological relevance level of the removed edges. Each individual edge will be added to the network if its weight w_{en} (calculated in the ensemble step) exceeds Th_β. Note edges are not recalculated as they are preserved from the first step (Section 2.1.1).

Note that, after pruning, those nodes exceeding the average node's degree are selected as potential hubs. In addition, the pruning step drastically reduces the average node degree. After the second step, where edges are added using the threshold Th_β, hubs are enriched in edges so these greatly exceed the average network connectivity. The final network generated by EnGNet is obtained after this step (see Figure 3).

A complete pseudo-code of EnGNet is described in Algorithm 1 and 2.

Algorithm 1: A general pseudocode of EnGNet. The method is divided into two different steps: (a) the ensemble network generation; and (b) a structure optimization by means of MTS algorithm. The pseudocode of the function *ensembleEdge* is given in Algorithm 2

input : Input Dataset, D
$g_i \in$ D (g_i:Gene i)
input : Relevant measures Thresholds:Th_1, Th_2,Th_3
input : Relevance level Threshold: Th_β
output: Final network, $G_\beta := < V, E_\epsilon >$
, where $E_\epsilon \in E$

/*Step 1: Ensemble generation of the network */
Let $G \leftarrow EmptyGraph$
for $g_i \in D$ **do**
  for $g_j \in D \wedge g_i \neq g_j$ **do**
    if $ensembleEdge(g_i, g_j, Th_1, Th_2, Th_3)$ **then**
      $e_i \leftarrow newEdge(g_i, g_j)$;
      $G \leftarrow addEdge(e_i)$;
    end
  end
end
/*Step 2: Topological optimization based on MST algorithm*/
$G_\beta \leftarrow$ ***MTSKruskal***(G);
$i \leftarrow 0$;
for $v_i \in V$ **do**
  if $isHub(v_i)$ **then**
    $j \leftarrow 0$;
    for $e_j \in E$ **do**
      if $contains(e_j, v_i) \wedge e_j.w_{en} \geq Th_\beta$ **then**
        $G_\beta \leftarrow addEdge(e_i)$;
      end
      $j \leftarrow j + 1$
    end
  end
  $i \leftarrow i + 1$
end
Return G_β;

Algorithm 2: A general pseudocode of *ensembleEdge* function.

input : g_i, g_j
input : Input data for g_i and g_j
input : Thresholds: Th_1, Th_2, Th_3
input : Weight of *ensembleEdge*: w_i
output: Weight of *ensembleEdge*: w_i
output: Boolean value: *true* or $false$ depending on whether the edge is labeled as correct or not.

$v_1 \leftarrow Spearman(g_i, g_j)$;
$v_2 \leftarrow Kendall(g_i, g_j)$;
$v_3 \leftarrow NMI(g_i, g_j)$;
$w_i \leftarrow average(v_1, v_2, v_3)$;
$vote \leftarrow 0$;
if $v_1 \geq Th_1$ **then**
  $vote \leftarrow vote + 1$;
end
if $v_2 \geq Th_2$ **then**
  $vote \leftarrow vote + 1$;
end
if $v_3 \geq Th_3$ **then**
  $vote \leftarrow vote + 1$;
end
$isCorrect \leftarrow false$;
if $vote \geq 2$ **then**
  $isCorrect \leftarrow true$;
end
Return *isCorrect*;

2.2. Datasets

In this section, the datasets used to test the usefulness of the proposed method are described. To this aim, we selected three datasets related to two different organisms that present different features: *Saccharomyces cerevisiae* and *Homo sapiens*. These organisms represent evolutionary-distant species, showing increasing complexity.

***Saccharomyces cerevisiae* cell cycle dataset** The dataset presented by Spellman et al. [40] was selected, which has been widely used for gene networks inference. This dataset contains the information about yeast cell cycle-related genes through a microarray analysis of the expression level of 5521 genes. RNA samples were collected from yeast cultures, which were synchronized by means of three different methods: α factor arrest, elutriation, and cdc15 thermosensible mutant.

***Homo sapiens* SNP dataset** The first selected human dataset, which was presented by Hodo et al. [41], was used in a study of the associations between interleukin 28B SNPs and recurrence of hepatocellular carcinoma (HCC) in patients with chronic hepatitis C (CHC). For the original purpose, the effects of a certain IL-28B genotype were tested by comparison of microarray data of 20 HCC patients vs. 91 CHC patients. The mentioned dataset stores expression levels of 54,616 human genes.

***Homo sapiens* Post-Traumatic Stress Disorder (PTSD)** Finally, a dataset testing PTSD, presented in the work by Breen et al. [42], was selected. This dataset was obtained to compare lymphocytic gene expression levels between PTSD-diagnosed US marines and control cases. Samples were collected from 94 marines (47 cases and 47 controls) both previously and subsequently to battlefield deployment. Thus, the dataset is divided into pre-deployment samples (controls) and post-deployment samples (cases). For the sake of simplicity, they are named "Pre" and "Post" for the rest of the paper. The dataset, harboring 27974 genes, were normalized as they comprise microarray (pre-deployment samples) and RNA-Seq (post-deployment samples) expression data. Additionally, this dataset was comprehensively analyzed to test the biological utility of the EnGNet tool in the experiment section.

2.3. Performance Evaluation of Gene Association Network

To assess the quality of our proposal, we present a comparison of the results obtained by EnGNet with those obtained from different methods from the literature on the datasets described in the previous section. To do so, we selected GeneMANIA [43] as the gold-standard to obtain different quality measures of the evaluated networks.

GeneMANIA is a gene interactions web-repository, which stores information presented in the form of web application for generating hypotheses about gene functions. It is possible to access online and freely the information stored in GeneMANIA. The genetic relationships identified in this database range from curated relationships that have been experimentally demonstrated to others that have been predicted in silico. A gene–gene relation is maintained in the database if at least one piece of evidence of such relationship exists in the literature. We selected GeneMANIA since it is a reliable source to test the correctness of gene–gene interactions [7,44,45], and it has demonstrated its suitability for this purpose in multiple previous works.

In this paper, the information stored for the two used organisms, i.e., *S. cerevisiae* and Human, was selected. The final networks obtained from GeneMANIA database are composed by 6462 nodes and 4,833,480 edges for yeast, and 19,551 nodes with 6,979,630 relationships for Human network.In particular, we based the comparison on two well known measures, namely precision and recall [9,15], which are defined as in the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where

- True positives (TP) is the number of edges contained in both the network obtained by EnGNet and in GeneMANIA.
- False positives (FP) is the number of edges that are contained in the network obtained by EnGNet but not in GeneMANIA.
- False negatives (FN) is the number of edges appearing in GeneMANIA but not in the network obtained by EnGNet.

2.4. Topological Features of Biological Networks

With the aim of evaluating the biological attributes of the networks that are presented (from a topological point of view), various criteria can be used. In the following, we present the most commonly used topological feature criteria of scale-free networks [6,46]:

Average Clustering Coefficient : Calculated as the number of edges linking nodes within its neighborhood divided by the number of links that are possible among them. A low clustering coefficient for a network is an indicator of the existence of biological relationships, as the lower this parameter, the sparser the network. Sparseness is also considered a main feature of GRNs.

Characteristic Path Length (CPL): Indicates the average length of the shortest paths between each pair of nodes comprising the network. A high path length indicates that the network is in a linear chain, while a lower value means that it is more compact. Scale-free networks usually have larger CPLs.

Diameter : Indicates the maximum distance between two nodes. As in the case of CPL, a high value indicates that the network follows a biological pattern.

Graph Density: Defines the ratio between the number of edges of a network and the number of all possible edges. Gene networks are generally sparsely connected so a low density is indicative of a biologically-meaningful pattern.

Node Degree Distribution: Defined as the number of edges linking a node. The larger is the degree. the more relevant is the node in a certain network. A distribution function $P(k)$ defines the spread of node degrees over a network. This function represents the probability of finding a degree of k in a randomly-selected node. The degree distribution usually follows a power law of the form $P(k) \sim k^{-\gamma}$, where γ is a constant (≥ 0). A high γ is indicative of a scale-free topology [47].

3. Results and Discussion

In this section, we present the results of the experimentation carried out in order to assess the reliability and usefulness of EnGNet. We first compared EnGNet with three standard information theory approaches commonly used in the literature to infer large GNs (based on NMI, Spearman and Kendall measures). Moreover, we compared our proposal with the ensemble strategy of these methods (i.e., only the first step of EnGNet). The aim of these experiments was to test the performance of EnGNet against other classical methods from the literature to infer large co-expression networks, and also to test the relevance of the prune step in the final results obtained. Thus, we not only tested the reliability of the inferred networks, but also the ability of EnGNet to reduce the size of final networks and their topological features.

In the second experiment, we also tested the performance of EnGNet against different algorithms from the literature for generating small gene networks. In particular, we present a study on 20 yeast genes that encode the Cell Cycle G1 phase.

Finally, with the aim of proving the effectiveness of our proposal in a biomedical study, we applied EnGNet to a human dataset regarding post traumatic stress disorder (PTSD).

3.1. Comparative Analysis Of EnGNet For Large Gene Networks

In the experiments, we used five approaches to generate networks from each dataset. In particular, we used EnGNet, the first phase of EnGNet, i.e., only the ensemble strategy without the pruning phase, and three information theory based methods. These last three methods are based on the NMI, Kendall and Spearman measures, in a similar way as the experiments presented in [7,15]. These approaches have been widely used in the biomedical literature for studying with gene co-expression networks (e.g., Xu et al. [48], Johnson et al. [49] and Liu et al. [50]).

For each information theory method used, we needed to set a validity threshold, and in the case of EnGNet, we needed four thresholds (see Section 2.1). For this experiment, we selected three different thresholds for all methods: $0.7, 0.8, 0.9$. For a fair comparison, EnGNet and the ensemble approach also used the same thresholds for $Th_{1,2,3}$ and Th_{β}. These thresholds represent a complete full spectrum from a mid correlation (0.7) to a very strong one (0.9). Thus, 60 networks were generated and analyzed (5 methods $\times$ 3 thresholds $\times$ 4 datasets).

3.1.1. Networks Performance Against GeneMANIA

As mentioned above, we first tested the biological significance of the obtained networks in a direct comparison with GeneMANIA database. The results obtained, in terms of nodes, edges, precision and recall, are presented in Tables 2–5, respectively.

Table 2 shows how EnGNet achieved the second best results of the experiment (only behind Kendall's) in terms of average precision. However, it is important to notice that EnGNet is the method that presents the most stable precision and size values for the different thresholds, obtaining the sparser networks for all methods considered (almost half the average size compared to Kendall's). This result confirms the overall stability of EnGNet.

The experiment carried out on the Human SNP dataset shows that EnGNet obtains the best results in terms of average precision (see last row of Table 4). We can also notice that the NMI approach infers smaller networks than EnGNet. However, the precision is so low that these networks do not appear to be biologically significant.

For the experiments with "Pre" and "Post" PTSD datasets (Tables 3 and 5, respectively), the results present the same pattern: EnGNet obtains the best results in term of precision and size of the networks.

Table 2. The results obtained by different gene networks on the yeast dataset using different thresholds. The precision and recall results were obtained using GeneMANIA database as gold standard. The last row presents the average results in terms of precision and size of the network for the experiment.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Nodes**	3123	2684	2581	5371	3123
	Rods	7129	26,633	14771	455,776	33,715
	Precision	0.480	0.365	0.541	0.334	0.43
	Recall	0.002	0.01	0.009	0.041	0.01
0.8	**Nodes**	1057	1070	544	4180	620
	Rods	1296	4518	599	88,508	781
	Precision	0.555	0.416	0.773	0.412	0.514
	Recall	0.005	0.012	0.011	0.016	0.001
0.9	**Nodes**	258	1032	8	1375	258
	Rods	176	4398	4	3471	245
	Precision	0.657	0.409	1	0.639	0.651
	Recall	0.012	0.013	0.04	0.008	0.015
	Avg. Precision	0.56	0.39	**0.77**	0.46	0.53
	Avg. Size	2808.51	10,383.8	5123.59	181,428.13	11,498.83

Table 3. The results obtained by different gene networks on the Pre-deployment samples of the PTSD dataset using different thresholds. The precision and recall results were obtained using GeneMANIA database as gold standard. The last row presents the average results in terms of precision and size of the network for the experiment.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Nodes**	1104	1026	941	5407	1098
	Rods	1222	9299	10,055	605,409	10,274
	Precision	0.407	0.112	0.294	0.138	0.294
	Recall	0.009	0.023	0.068	0.08	0.053
0.8	**Nodes**	131	823	98	2716	131
	Rods	110	8971	110	108,861	142
	Precision	0.635	0.112	0.611	0.195	0.633
	Recall	0.06	0.034	0.1	0.073	0.081
0.9	**Nodes**	5	775	0	624	5
	Rods	3	8943	0	4177	4
	Precision	1	0.112	0	0.301	1
	Recall	0.333	0.037	0	0.059	0.444
	Avg. Precision	**0.67**	0.11	0.30	0.21	0.64
	Avg. Size	462.66	9071	3388.33	239,482.33	3473.33

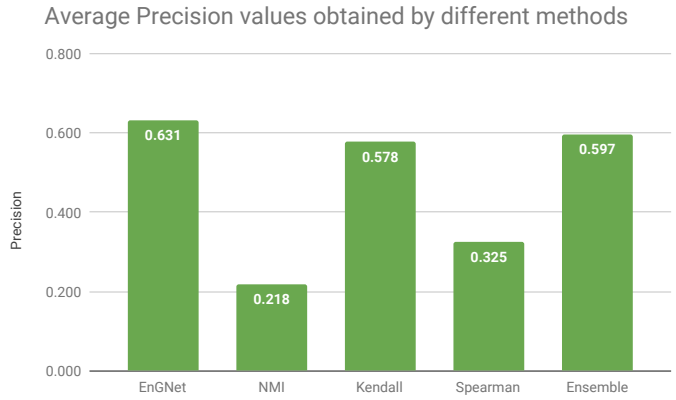
Table 4. The results obtained by different gene networks on the Human SNP dataset using different thresholds. The precision and recall results were obtained using GeneMANIA database as gold standard. The last row presents the average results in terms of precision and size of the network for the experiment.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Nodes**	1553	259	1595	20,668	1544
	Rods	1963	202	5314	725,553	5049
	Precision	0.653	0.380	0.675	0.200	0.684
	Recall	0.020	0.023	0.043	0.022	0.044
0.8	**Nodes**	280	59	251	6853	241
	Rods	467	39	403	50309	381
	Precision	0.840	0.190	0.7607	0.398	0.771
	Recall	0.074	0.032	0.101	0.020	0.1120
0.9	**Nodes**	30	37	32	813	30
	Rods	16	26	25	2023	21
	Precision	0.6	0.15	0.5	0.727	0.428
	Recall	0.1875	0.0338	0.1818	0.0610	0.1875
	Avg. Precision	**0.69**	0.24	0.64	0.44	0.62
	Avg. Size	815.33	89	1914	259,295	1817

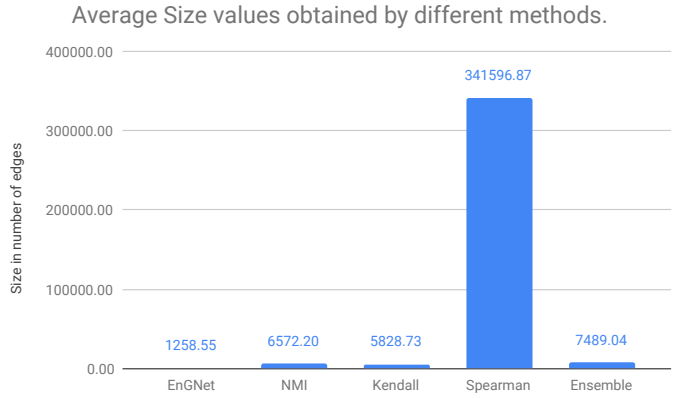
Finally, Figure 4 shows the average values of precision and size of the networks for all experiments presented above. Considering the precision results presented in Figure 4a, we can observe that our algorithm is the one that obtains the best values, followed by the Ensemble approximation and Kendall's. Regarding the size of the networks, it can be verified in Figure 4bthat EnGNet obtains the smallest networks (approximately 271 times smaller than Spearman's network or six times smaller than Ensemble's network, which is the second approximation in precision values) with the highest precision values.

Table 5. The results obtained by different gene networks on the Post-deployment samples of the PTSD dataset using different thresholds. The precision and recall results were obtained using GeneMANIA database as gold standard. The last row presents the average results in terms of precision and size of the network for the experiment.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Nodes**	1723	1303	1508	5958	1715
	Rods	2491	7381	37912	1718641	38641
	Precision	0.318	0.125	0.253	0.104	0.252
	Recall	0.006	0.012	0.091	0.147	0.075
0.8	**Nodes**	352	882	273	3516	351
	Rods	347	6479	753	325270	855
	Precision	0.456	0.119	0.522	0.155	0.503
	Recall	0.02	0.02	0.079	0.109	0.057
0.9	**Nodes**	9	750	4	982	9
	Rods	5	6375	2	14635	5
	Precision	1	0.116	1	0.294	1
	Recall	0.71	0.028	0.667	0.086	0.714
	Avg. Precision	**0.59**	0.12	0.59	0.18	0.58
	Avg. Size	947.66	6745	12889	686182	13167



(**a**)



(**b**)

Figure 4. Visual comparison of the average results presented in the tables for all datasets. As it is possible to see in the chart, EnGNet obtains smaller networks with the best results in the precision experiments and the sparsest networks. As discussed above, these are desirable features for any method that infer large gene networks: (**a**) average precision values; and (**b**) average size of the networks.

In summary, we can conclude that EnGNet is successful in reducing the size of the networks while keeping competitive results in terms of precision and recall (against other methods studied). In fact, networks generated by EnGNet are significantly sparser than those obtained by other methods (see Figure 4). As stated above, this is a significant result, since sparseness is a desirable feature in GNs reconstruction from a large dataset. In fact, the smaller is the networks, the easier is their analysis [51]. Additionally, although networks are sparser in terms of the number of edges, precision and recall values do not suffer a relevant loss. This observation is confirmed from the results presented, since EnGNet obtains average precision values above 0.5 in all the cases studied (presented in the tables).

Finally, Figure 4 shows that EnGNet obtains the best average precision value, whilst the size of the network is significantly reduced (especially against the Spearman's approach). This result indicates that EnGNet networks do not lose biological significance upon pruning. As a conclusion, we can affirm that EnGNet is a competitive and reliable method for the generation of large gene networks.

3.1.2. Topological Features Analysis

In addition to network sparseness, the topological properties of gene networks should be considered in order to estimate the performance of EnGNet upon network reconstruction [7,9,16]. As discussed in Section 1, biological networks tend to be sparse and to follow a scale-free topology. Therefore, it is desirable for the reconstruction methods to provide networks that present such topological features.

With the aim of performing a topological analysis of the generated networks, we extracted the topological features presented in Section 2.4 for all networks discussed in Section 3.1. The results are shown in Tables 6–9.

Table 6. Yeast feature.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Clust. Coef**	0.114	0.282	0.262	0.416	0.272
	CPL	7.201	6.947	5.406	2.978	4.358
	Diameter	29	28	22	10	20
	Density	0.001	0.007	0.004	0.032	0.007
	Gamma	1.413	0.958	1.529	0.915	1.286
0.8	**Clust. Coef**	0.283	0.524	0.162	0.342	0.163
	CPL	4.567	2.011	5.56	4.083	6.984
	Diameter	18	10	19	13	23
	Density	0.004	0.008	0.004	0.01	0.004
	Gamma	1.203	0.823	2.223	1.202	1.825
0.9	**Clust. Coef**	0.409	0.549	-	0.239	0.167
	CPL	2.401	1.007	1	6.726	2.57
	Diameter	6	2	1	24	7
	Density	0.007	0.008	0.143	0.004	0.007
	Gamma	0.934	0.66	-	1.782	1.981
Average	**Clust. Coef**	0.27	0.45	0.21	0.33	**0.20**
	CPL	**4.72**	3.32	3.99	4.60	4.64
	Diameter	**17.67**	13.33	14.00	15.67	16.67
	Density	**0.004**	0.008	0.050	0.015	0.01
	Gamma	1.18	0.81	**1.88**	1.30	1.70

Table 7. HUMANSNP.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Clust. Coef**	0.055	0.119	0.235	0.219	0.239
	CPL	9.469	1.719	6.605	3.685	6.687
	Diameter	24	6	18	13	18
	Density	0.001	0.006	0.003	0.03	0.003
	Gamma	1.415	2.124	1.31	1.272	1.305
0.8	**Clust. Coef**	0.145	0.169	0.25	0.224	0.24
	CPL	2.543	1.026	2.551	5.231	2.573
	Diameter	8	2	8	23	7
	Density	0.007	0.022	0.009	0.02	0.009
	Gamma	1.01	1.98	1.447	1.374	1.486
0.9	**Clust. Coef**	0	0.27	0.073	0.238	0.1
	CPL	1.111	1.037	1.174	4.429	1.056
	Diameter	2	2	2	14	2
	Density	0.037	0.039	0.038	0.004	0.039
	Gamma	3.807	1.72	1.712	1.407	2.221
Average	**Clust. Coef**	**0.07**	0.19	0.19	0.23	0.19
	CPL	4.37	1.26	3.44	**4.45**	3.44
	Diameter	11.33	3.33	9.33	**16.67**	9.00
	Density	**0.015**	0.022	0.017	0.018	0.02
	Gamma	**2.08**	1.94	1.49	1.35	1.67

Table 8. Pre-deployment samples of the PTSD dataset.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Clust. Coef**	0.031	0.689	0.499	0.615	0.444
	CPL	6.315	1.464	3.346	3.191	3.72
	Diameter	28	11	16	14	18
	Density	0.002	0.018	0.023	0.041	0.017
	Gamma	1.589	0.252	1.043	0.748	1.081
0.8	**Clust. Coef**	0.13	0.797	0.171	0.581	0.195
	CPL	2.859	1.001	2.721	3.485	2.785
	Diameter	9	3	6	15	8
	Density	0.012	0.027	0.023	0.03	0.017
	Gamma	1.426	0.142	1.571	0.847	1.681
0.9	**Clust. Coef**	0	0.843	0	0.449	0.6
	CPL	1.25	1	0	3.041	1
	Diameter	2	1	0	10	1
	Density	0.03	0.03	0	0.021	0.4
	Gamma	2	0.107	0	1.125	0.585
Average	**Clust. Coef**	**0.05**	0.78	0.22	0.55	0.41
	CPL	**3.47**	1.16	2.02	3.24	2.50
	Diameter	**13.00**	5.00	7.33	**13.00**	9.00
	Density	**0.01**	0.03	0.02	0.03	0.14
	Gamma	**1.67**	0.17	0.87	0.91	1.12

Table 9. Post-deployment samples of the PTSD dataset.

Thr		EnGNet	NMI	Kendall	Spearman	Ensemble
0.7	**Clust. Coef**	0.069	0.572	0.556	0.694	0.51
	CPL	5.207	7.142	3.17	2.592	3.309
	Diameter	17	22	12	13	12
	Density	0.002	0.009	0.033	0.097	0.026
	Gamma	1.301	0.862	0.859	0.507	0.763
0.8	**Clust. Coef**	0.245	0.706	0.344	0.632	0.295
	CPL	4.085	1.16	3.325	2.998	3.418
	Diameter	13	10	8	20	8
	Density	0.008	0.017	0.02	0.053	0.014
	Gamma	1.206	0.321	1.266	0.707	1.399
0.9	**Clust. Coef**	0	0.813	-	0.515	0
	CPL	1.167	1	1	3.523	1.167
	Diameter	2	1	1	13	2
	Density	0.139	0.023	0.333	0.03	0.139
	Gamma	3	0.218	-	0.954	3
Average	**Clust. Coef**	**0.10**	0.70	0.45	0.61	0.27
	CPL	**3.49**	3.10	2.50	3.04	2.63
	Diameter	10.67	11.00	7.00	**15.33**	7.33
	Density	**0.05**	0.02	0.13	0.06	0.06
	Gamma	**1.84**	0.47	1.06	0.72	1.72

From these results, we can observe that EnGNet obtains the most stable results over the experiments carried with respect to the majority of the topological features studied (see "Average" rows in the tables). To clarify these results, we also calculated the average values for all datasets and thresholds presented. These results are reported in Table 10. In the table, it is possible to observe that, for all topological features studied, EnGNet is the algorithm achieving the best results, except for the network diameter. For the network's diameter, only the Spearman's method obtains better results. This is a logical result since Spearman's method generates the biggest networks (271 times bigger than EnGNet). It is remarkable, from a topological point of view, that our method reaches a diameter in a similar range with a significantly smaller size than Spearman's network.

Table 10. Average topological feature results for all methods in all datasets.

	EnGNet	NMI	Kendall	Spearman	Ensemble
Clust. Coef	**0.123**	0.528	0.268	0.430	0.269
CPL	**4.015**	2.210	2.988	3.830	3.302
Diameter	13.167	8.167	9.417	**15.167**	10.500
Density	**0.021**	0.018	0.053	0.031	0.057
Gamma	**1.692**	0.847	1.325	1.070	1.551

In summary, EnGNet obtains the best results on all topological features, for all the networks, indicating that EnGNet networks follow a biological pattern (scale-free topology). Furthermore, EnGNet-generated networks improve the results obtained by information theory methods and ensemble networks. Bearing this in mind and the results presented in the comparison with the network contained in GeneMANIA, we can affirm that EnGNet is a suitable tool for large co-expression GNs reconstruction in biomedical research.

3.2. Comparative Analysis Of EnGNet For Small Networks

The ability of our approach to infer small gene networks was also tested. To do so, we performed a similar experiment to the one presented by Gallo et al. [52]. In this experiment, precision was used as quality measure to rate the reliability of the input GNs. The main objective of the experiment

is to compare the precision values of different gene networks algorithms from the literature on the same dataset.

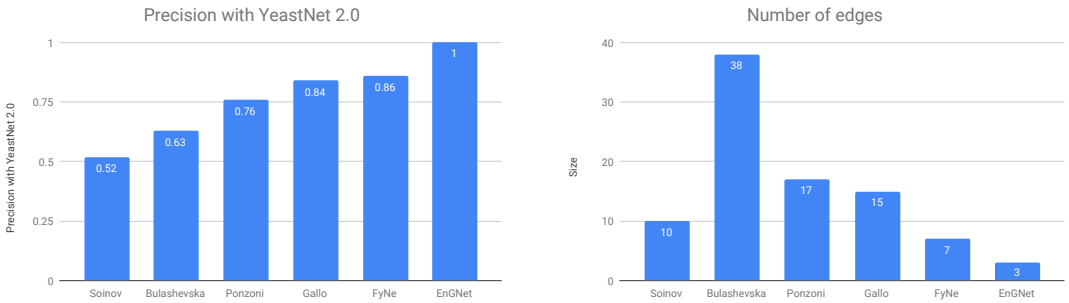
To obtain the input networks, we used different methods from the literature, which are described in the works by:

- Soinov et al. [53], a C4.5-based method;
- Bulashevska et al. [54], a Bayesian-based method;
- Ponzoni et al. [55], a combinatorial optimization algorithm (GRNCOP);
- Gallo et al. [52], an upgraded version of the previous algorithm named GRNCOP2; and
- Gomez-Vela et al. [15], a fuzzy method to infer gene co-expression networks named FyNe.

These methods were applied to the same dataset from the Yeast Cell Cycle—more specifically, to a subset of 20 well-described genes. These genes code for key proteins in cell-cycle regulation, as presented by Martinez-Ballesteros et al. [56].

As in the experiment performed by Gallo et al. [52], the quality of the networks was assessed regarding the precision values obtained against the data stored in YeastNet [57]. YeastNet is a repository that comprises a probabilistic functional GN generated from verified protein-coding open reading frames (ORFs) of the yeast genome. This repository combines protein–protein interactions, protein–DNA interactions, co-expression, phylogenetic conservation and literature information, in total covering more than 102,803 linkages among 5483 yeast proteins (95% of the validated proteome).

The results of the experiment are presented in Figure 5a,b, where it can be verified that EnGNet yields the best results amongst all studied methods, and again with the smaller network. Note that the inference of small gene networks usually provides higher precision results than in the case of large ones, as detailed in Hecker et al. [16]. The results show that not only is EnGNet suitable for large gene networks studies, but also obtains competitive results for studies with small datasets.



(**a**) Precision values (**b**) Size of the networks

Figure 5. (**a**) Results from different methods on the 20 genes from the yeast cell cycle dataset. The results show that EnGNet is also a reliable method for inference of small co-expression networks with a high precision. (**b**) Size in terms of number of relationships. Note that EnGNet is again the method that obtain the smaller network.

3.3. Application to the Study of Human Post Traumatic Stress Disorder

The second objective of this study was to prove the usefulness of EnGNet in actual life sciences research. To do so, EnGNet was applied to a human PTSD dataset obtained by Breen et al. [42], so as to shed some light over the genes involved in this pathology.

In this case-control study, expression data were obtained from US marines peripheral blood leukocytes both before and after deployment to conflict zones (that called "Pre" and "Post"). As stated above, 94 marines (47 cases and 47 controls) were analyzed. According to the original article by Breen et al. [42], controls refer to selected marines who did not show signs of PTSD. These are used as a reference for cases, which are marines who show a broad spectrum of signs that classify them

as under PTSD after battlefield deployment. PTSD was scored through a diagnostic interview and annotated in the Clinician Administered PTSD Scale (CAPS) [58]. In the experimental design, cases are analogous to controls prior to battlefield deployment, i.e., none are under PTSD symptomatology. On the other hand, after battlefield deployment cases significantly differ from controls in terms of the CAPS score (see the original article by Breen et al. [42] for further details).

Overall, PTSD signs may be observed in the second group when compared to the first one. An exploratory multidimensional scaling (MDS) plot or Principal Coordinates Analysis (PCoA) was performed in order to roughly examine these differences. MDS assisted the examination of sample similarity. On this occasion, the classical MDS method was applied, assuming Euclidean distances. An illustrative distribution of this dataset is shown in Figure 6, in which differences can be observed between post- and pre-deployment marines. However, these differences are fuzzy and there is a spectrum of sample states between pre- and post-deployment situations.

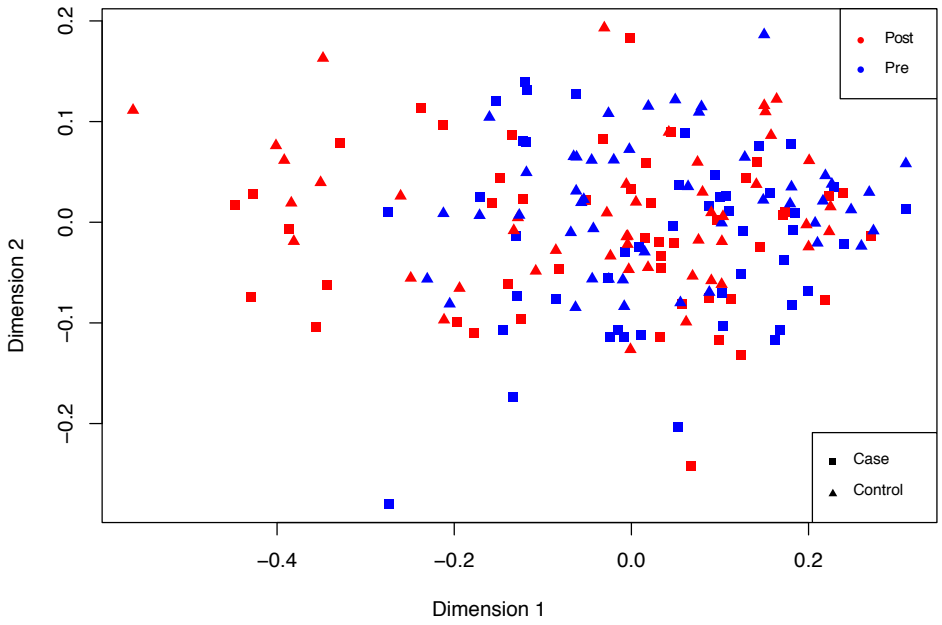


Figure 6. Non-supervised exploratory MDS plot showing differences between the input samples. RNA-seq (cases, squares in the figure) and microarray data (controls, triangles in the figure) were normalized and joined in a single dataset. Thus, no significant differences were expected between them. However, two groups for pre-deployment (red) and post-deployment (blue) are modestly differentiated, although cases in between are also appreciated.

First, a differential gene expression analysis was carried out to verify the mentioned differences using the *DESeq*2 [59] R package, a tool for the estimation of differentially-expressed genes (DEGs). The information on gene up- or down-regulation was of especial interest in the analysis of the biological processes underlying PTSD development. Hence, data provided by *DESeq*2 were latter imported into Cytoscape for network interpretation purposes.

EnGNet was used to reconstruct two different networks corresponding to pre-deployment and post-deployment samples, respectively. To this aim, the EnGNet $Th_{1,2,3}$ thresholds were set to the values that yield the best results in the experimentation presented in Section 2.3, namely $Th_1 = 0.7$, $Th_2 = 0.8$ and $Th_3 = 0.9$. As far as the Th_β threshold is concerned, a new analysis was carried out to determine the optimal threshold for each sample. The results of this study are presented in Table 11 and show the values of the precision and recall measure obtained by different networks against GeneMANIA.

Table 11. Analysis to determine the Th_β optimal value.

		Th_β Values		
		0.7	**0.8**	**0.9**
Pre	**Nodes**	116	**105**	105
	Rods	119	**90**	87
	Precision	0.59	**0.63**	0.61
	Recall	0.07	**0.07**	0.06
Post	**Nodes**	437	298	**298**
	Rods	945	295	**272**
	Precision	0.313	0.481	**0.5**
	Recall	0.002	0.002	**0.002**

Therefore, considering the results presented in the table, candidate networks for this study correspond to Th_β = 0.8 in the pre-deployment case and Th_β = 0.9 in the post-deployment situation.

Once the networks were generated, a significant increase in the number of genes was found in the post-deployment network compared to its pre-deployment counterpart, which is indicative of gene up-regulation in lymphocytes upon PTSD development. Pre- and post-deployment networks are shown in Appendix A (see Figure A1). Remarkably, the reconstructed networks for pre-deployment and post-deployment samples were significantly different, which is indicative for the discrimination power of the GN reconstruction approach over other unsupervised techniques such as PCoA.

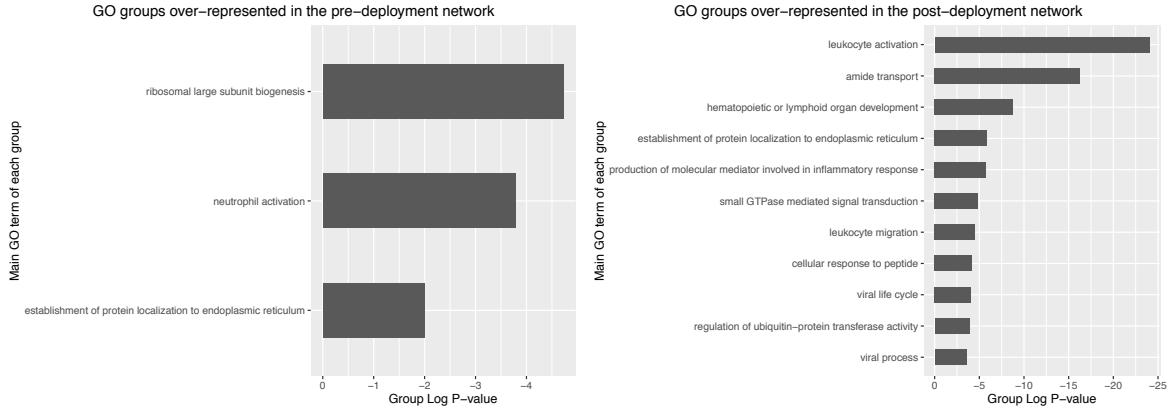
Pre- and post-deployment networks were merged in order to graphically observe the differences in gene expression upon PTSD development. Overall, 73.8% of the nodes in this merged network were found to be upregulated in the post-deployment situation compared to pre-deployment, which suggest the importance of gene activation upon PTSD development. Genes up/down-regulation in the merged network is shown in Appendix A (see Figure A2).

Enrichment analysis was performed by means of Cytoscape's plugins ClueGO [60] and CluePedia [61], which shows over-represented GO-terms in a ensemble of genes. ClueGO + CluePedia analyses provided useful information about the biological processes in which the genes comprised at the pre-deployment and post-deployment networks were involved.

Regarding the pre-deployment network (105 nodes), three different GO groups were identified, corresponding to ribosomal biogenesis, neutrophil activation and establishment of protein localization to endoplasmic reticulum (Figure 7a). Group p-values observed were of the order of 10^{-6}.

In the case of the post-deployment network (298 nodes), 10 GO groups were identified, mostly corresponding to leukocyte activation, amide transport and hematopoietic or lymphoid organ development (Figure 7b). Observed group P-values were of the order of 10^{-25}, thus representing a dramatic increase in significancy compared to the pre-deployment GO groups. Further exploration of the main GO group in the post-deployment revealed GO terms such as leukocyte activation involved in immune response, myeloid cell activation involved in immune response, myeloid leukocyte activation, and leukocyte degranulation. Main GO terms comprised in the main GO group of the post-deployment network are shown in Figure 8.

Enrichment analyses thus revealed a dramatically different situation in the post-deployment network compared to pre-deployment one, in terms of the biological processes these represent. Whereas the pre-deployment network shows biological processes more related to an unexcited steady-state immune system, the post-deployment network displays several GO groups and GO terms which lie under the context of immunoenhancement. Reconstructed GNs thereby model two different situations in terms of the biological context. This also suggests the potential use of GNs for diagnostic purposes.



(**a**) (**b**)

Figure 7. Bar plots showing the different groups of analogous GO terms that were identified in: (**a**) the pre-deployment network; and (**b**) the post-deployment network. The main GO term of each identified group, i.e., the one with lowest term P-value, is presented as group label. Group P-value was corrected with Bonferroni step-down. Note the lower is the P-value, the more the over-represented is the GO term.

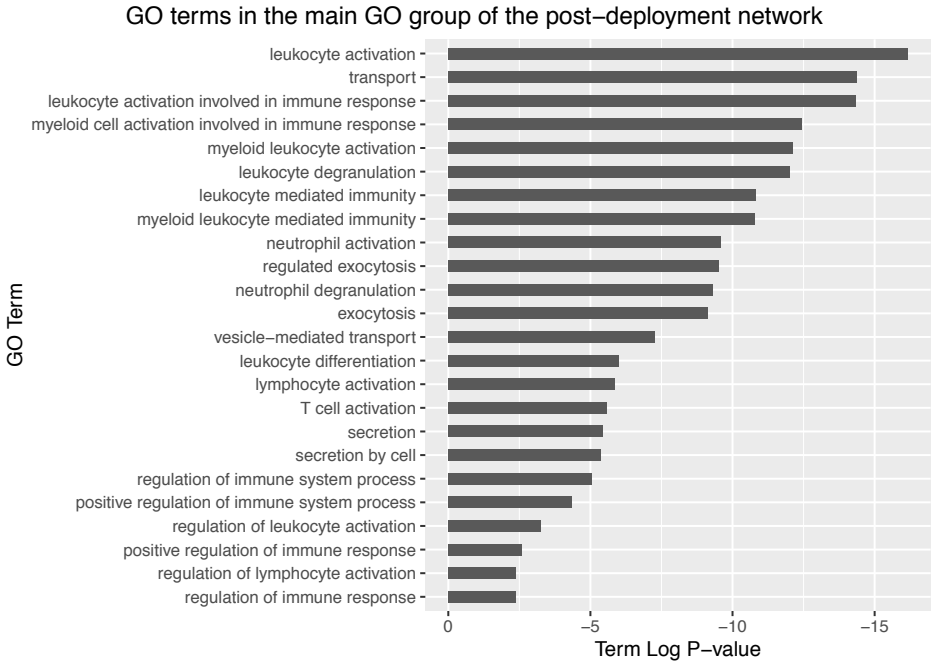


Figure 8. Top GO terms in the main GO group of the post-deployment network. Term P-value was corrected with Bonferroni step-down. Note the lower is the P-value, the more over-represented is the GO term.

With regard to differential expression, a considerable gene up-regulation is observed, which correlates to immunoenhancement upon PTSD development. In general, the above mentioned GO terms are indicative of a nonspecific immune response, characteristic of innate immunity, suggesting the potential role of myeloid leukocytes in PTSD. Quite significant is also the GO group "hematopoietic or lymphoid organ development", as the immune system is generated from multipotent hematopoietic stem cells, which branch in myeloid and lymphoid progenitors. This myeloid cell line comprises cells such as basophils, neutrophils, eosinophils and macrophages, which through immunosurveillance are responsible for the so-called unspecific or innate immunity. This is consistent with the results found by Breen et al. [42], who predicted the intrinsic role of innate immunity upon PTSD. These findings were

also highlighted in a previous study by Watson et al. [62], who observed enhanced immunological features in PTSD-diagnosed Vietnam combat veterans in comparison with civilians.

4. Conclusions

In this paper, we introduce EnGNet, an ensemble-based novel method for the inference of large gene co-expression networks. First, EnGNet applies an ensemble approach for large co-expression networks reconstruction. Second, a greedy strategy optimizes both the size and topological features of the final network.

When compared with other standard approaches from the literature, EnGNet-inferred networks were smaller in size than those of other approaches, regarding the number of edges. In addition to achieving competitive results in terms of the presented biological information, EnGNet-inferred networks showed better performance in respect of networks topological, and thus biological, features. Among these features, sparseness and scale-free topology are to be highlighted as a major convenience of EnGNet networks, in concordance with actual GRN. In addition, EnGNet was demonstrated to be a competitive solution for studies on small datasets, by means of the experiments carried out. Moreover, topological features of EnGNet networks enable friendlier interpretation and hypothesis-making by life scientists.

Finally, the biological relevance of EnGNet was successfully tested in the application to human PTSD dataset. EnGNet inferred gene association networks from the gene expression dataset, revealing an innate immunity-mediated response in PTSD cases, which was accompanied by considerable gene upregulation. In particular, myeloid cells activation was detected in PTSD cases when compared to non-PTSD ones. Such PTSD-associated genes could then be considered as potential biomarkers, which can be used as pathology indicators. Besides, the GN inference approach distinguished between two different biological situations basing on gene expression, whereas analyses such as PCoA did not. These results demonstrate the usefulness of EnGNet in the field of biomarkers discovery, a field that has become one of the most relevant in personalized medicine.

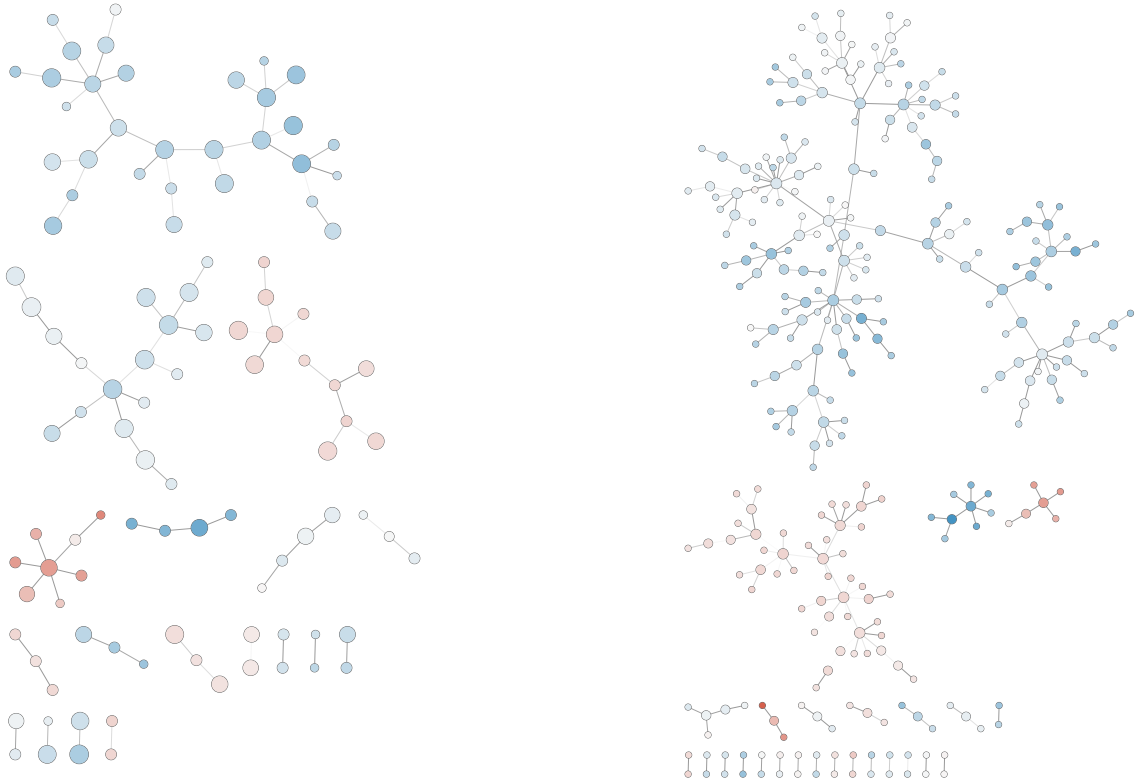
Author Contributions: Conceptualization, F.G.-V. and F.D.; methodology, F.G.-V.; software, F.G.-V.; validation, F.G.-V. and F.M.D.-C.; data curation, F.G.-V., F.D.C. and M.G.-T.; writing-original draft preparation, F.G.-V., F.M.D.-C., D.S.R.-B., M.G.-T. and F.D.; writing-review and editing, F.G.-V., F.M.D.-C., D.S.R.-B, M.G.-T. and F.D.; supervision, F.G.-V. and F.D.; and project administration, F.G.-V.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. PTSD Application Reconstructed Networks

Pre- and post-deployment networks, respectively, comprising 105 and 298 nodes, are shown in Figure A1. An increase is observed in the number of genes involved in post-deployment samples compared to pre-deployment ones. Such increase may well be the result of the genetic regulation upon PTSD that is addressed along Section 3.3. Gene FC is also represented in Figure A2, which revealed an overall genetic upregulation.



(**a**) (**b**)

Figure A1. Inferred networks corresponding to: (**a**) pre-deployment samples; and (**b**) post-deployment samples. Log2 FC is represented by node color, so blue and red intensities are related to gene up- or down-regulation, respectively. Node size is represented according to their rank. Edge transparency is represented according to edge weight. Note both networks show a major connected module and exhibit a scale-free topology.

The union of the reconstructed networks is shown in Figure A2. Among the 310 genes comprised in this merged network, 229 showed an upregulation in the post-deployment situation compared to the pre-deployment samples.

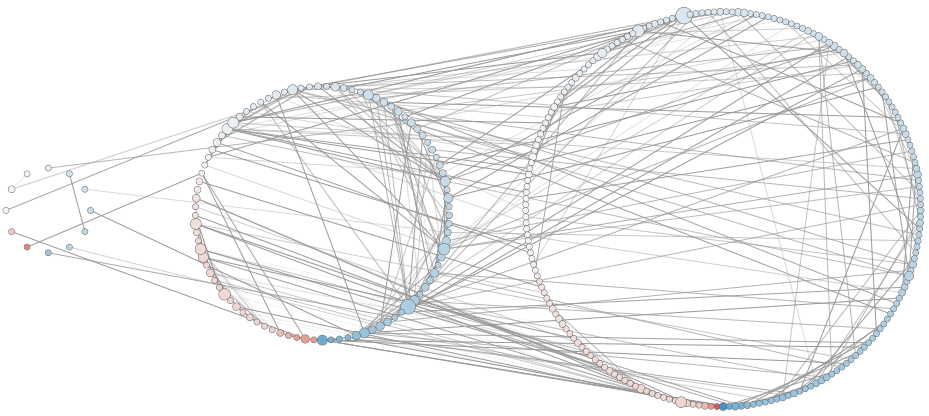


Figure A2. Union of pre- and post-deployment reconstructed networks. Nodes are sorted depending on whether they are exclusively present at the pre-deployment network (far left), exclusively present at the post-deployment network (far right) or present at both networks (center). Node size is represented according to their degree. Edges transparency is represented according to their weight. Upregulated and downregulated genes in post-deployment samples compared to pre-deployment samples are, respectively, shown in shades of blue and red.

References

1. Parikshak, N.N.; Gandal, M.J.; Geschwind, D.H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **2015**, *16*, 441. [CrossRef] [PubMed]
2. Darrason, M. Mechanistic and topological explanations in medicine: The case of medical genetics and network medicine. *Synthese* **2018**, *195*, 147–173. [CrossRef]
3. Liang, L.; Gao, L.; Zou, X.P.; Huang, M.L.; Chen, G.; Li, J.J.; Cai, X.Y. Diagnostic significance and potential function of miR-338-5p in hepatocellular carcinoma: A bioinformatics study with microarray and RNA sequencing data. *Mol. Med. Rep.* **2018**, *17*, 2297–2312. [CrossRef] [PubMed]
4. Huang, R.; He, Y.; Sun, B.; Liu, B. Bioinformatic Analysis Identifies Three Potentially Key Differentially Expressed Genes in Peripheral Blood Mononuclear Cells of Patients with Takayasu's Arteritis. *Cell J.* **2018**, *19*, 647. [PubMed]
5. Brugere, I.; Gallagher, B.; Berger-Wolf, T.Y. Network structure inference, a survey: Motivations, methods, and applications. *ACM Comput. Surv.* **2018**, *51*, 24. [CrossRef]
6. Pavlopoulos, G.A.; Secrier, M.; Moschopoulos, C.N.; Soldatos, T.G.; Kossida, S.; Aerts, J.; Schneider, R.; Bagos, P.G. Using graph theory to analyze biological networks. *BioData Min.* **2011**, *4*, 10. [CrossRef]
7. Gómez-Vela, F.; Rodriguez-Baena, D.S.; Vázquez-Noguera, J.L. Structure Optimization for Large Gene Networks Based on Greedy Strategy. *Comput. Math. Method Med.* **2018**, *2018*. [CrossRef]
8. Barbosa, S.; Niebel, B.; Wolf, S.; Mauch, K.; Takors, R. A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. *Biosystems* **2018**, *174*, 37–48. [CrossRef]
9. Delgado, F.M.; Gómez-Vela, F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artif. Intell. Med.* **2019**, *95*, 133–145. [CrossRef]
10. Kourilsky, P. The natural defense system and the normative self model. *F1000Res* **2016**, *5*, 797. [CrossRef]
11. Kumari, S.; Nie, J.; Chen, H.S.; Ma, H.; Stewart, R.; Li, X.; Lu, M.Z.; Taylor, W.M.; Wei, H. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS ONE* **2012**, *7*, e50411. [CrossRef] [PubMed]
12. de Siqueira Santos, S.; Takahashi, D.Y.; Nakata, A.; Fujita, A. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief. Bioinform.* **2013**, *15*, 906–918. [CrossRef] [PubMed]
13. Butte, A.J.; Kohane, I.S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*; World Scientific: Singapore, 1999; pp. 418–429.
14. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Aderhold, A.; Bonneau, R.; Chen, Y.; et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804. [CrossRef] [PubMed]
15. Gómez-Vela, F.; Barranco, C.D.; Díaz-Díaz, N. Incorporating biological knowledge for construction of fuzzy networks of gene associations. *Appl. Soft Comput.* **2016**, *42*, 144–155. [CrossRef]
16. Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: Data integration in dynamic models: A review. *Biosystems* **2009**, *96*, 86–103. [CrossRef]
17. Dougherty, E.R. Validation of inference procedures for gene regulatory networks. *Curr. Genom.* **2007**, *8*, 351–359. [CrossRef]
18. Winterbach, W.; Van Mieghem, P.; Reinders, M.; Wang, H.; de Ridder, D. Topology of molecular interaction networks. *BMC Syst. Biol.* **2013**, *7*, 90. [CrossRef]
19. Yip, A.M.; Horvath, S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinform.* **2007**, *8*, 22. [CrossRef]
20. Goh, K.I.; Cusick, M.E.; Valle, D.; Childs, B.; Vidal, M.; Barabási, A.L. The human disease network. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8685–8690. [CrossRef]
21. Barabási, A.L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef]
22. Ruprecht, C.; Vaid, N.; Proost, S.; Persson, S.; Mutwil, M. Beyond Genomics: Studying Evolution with Gene Coexpression Networks. *Trends Plant Sci.* **2017**, *22*. [CrossRef] [PubMed]

23. Fehrmann, R.S.; Karjalainen, J.M.; Krajewska, M.; Westra, H.J.; Maloney, D.; Simeonov, A.; Pers, T.H.; Hirschhorn, J.N.; Jansen, R.C.; Schultes, E.A.; et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genet.* **2015**, *47*, 115–125. [CrossRef] [PubMed]
24. Liu, R.; Cheng, Y.; Yu, J.; Lv, Q.L.; Zhou, H.H. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene* **2015**, *563*, 56–62. [CrossRef] [PubMed]
25. Horvath, S.; Langfelder, P. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 1–13.
26. Ivliev, A.; Hoen, P.; Borisevich, D.; Yuri, N.; Marina, S. Drug Repositioning through Systematic Mining of Gene Coexpression Networks in Cancer. *PLoS ONE* **2016**, *11*, 1–19. [CrossRef] [PubMed]
27. Yujie, L.; Hanbo, C.; Xi, J.; Li, X.; Lv, J.; Peng, H.; Tsien, J.; Liu, T. Discover mouse gene coexpression landscapes using dictionary learning and sparse coding. *Brain Struct. Funct.* **2017**, *222*, 4253–4270.
28. Zhong, R.; Allen, J.; Xiao, G.; Xie, Y. Ensemble-Based Network Aggregation Improves the Accuracy of Gene Network Reconstruction. *PLoS ONE* **2014**, *9*, e106319. [CrossRef]
29. Ruyssinck, J.; Huynh-Thu, V.; Geurts, P.; Dhaene, T.; Demeester, P.; Saeys, Y. NIMEFI: Gene Regulatory Network Inference using Multiple Ensemble Feature Importance Algorithms. *PLoS ONE* **2014**, *9*, e92709. [CrossRef]
30. Huang, J.; Vendramin, S.; Shi, L.; McGinnis, K. Construction and Optimization of a Large Gene Coexpression Network in Maize Using RNA-Seq Data. *Plant Physiol.* **2017**, *175*, 568–583. [CrossRef]
31. Liue, F.; Shang, S.; Shi, L.; Guo, W.; Wei, Z.; Chen, L. Inference of Gene Regulatory Network Based on Local Bayesian Networks. *PLoS Comput. Biol.* **2016**, *12*, e1005024. [CrossRef]
32. Wang, Y.; Zhang, X.S.; Chen, L. Optimization meets systems biology. *BMC Syst. Biol.* **2010**, *4*, 1–4. [CrossRef] [PubMed]
33. Angus Thomas, S.; Jin, Y. Reconstructing biological gene regulatory networks: Where optimization meets big data. In *Evolutionary Intelligence*; IEEE: Piscataway, NJ, USA, 2014; Volume 7, pp. 29–47.
34. Liu, J.; Chi, Y.; Zhu, C. A Dynamic Multiagent Genetic Algorithm for Gene Regulatory Network Reconstruction Based on Fuzzy Cognitive Maps. *IEEE Trans. Fuzzy Syst.* **2016**, *24*, 419–431. [CrossRef]
35. Studham, M.; Tjärnberg, A.; Nordling, T.; Nelander, S.; Sonnhammer, E. Functional association networks as priors for gene regulatory network inference. *Bioinformatics* **2014**, *30*, 130–138. [CrossRef] [PubMed]
36. Lopes, F.; Martins, D.; Barrera, j.; Cesar, R. A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. *Inf. Sci.* **2014**, *272*, 1–15. [CrossRef]
37. Yang, B.; Xu, J.; Liu, B.; Wu, Z. Inferring Gene Regulatory Networks with a ScaleFree Property Based Informative Prior. In Proceedings of the 8th International Conference on BioMedical Engineering and Informatics, Shenyang, China, 14–16 October 2015; pp. 542–547.
38. Mousavian, Z.; Díaz, J.; Masoudi-Nejad, A. Information theory in systems biology. Part II: Protein–protein interaction and signaling networks. In *Seminars in Cell & Developmental Biology*; Elsevier: Amsterdam, The Netherlands, 2016; Volume 51, pp. 14–23.
39. Kruskal, J.J. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math Soc.* **1956**, *7*, 48–50. [CrossRef]
40. Spellman, P.T.; Sherlock, G.; Zhang, M.Q.; Iyer, V.R.; Anders, K.; Eisen, M.B.; Brown, P.O.; Botstein, D.; Futcher, B. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **1998**, *9*, 3273–3297. [CrossRef]
41. Hodo, Y.; Honda, M.; Tanaka, A.; Nomura, Y.; Arai, K.; Yamashita, T.; Sakai, Y.; Yamashita, T.; Mizukoshi, E.; Sakai, A.; et al. Association of interleukin-28B genotype and hepatocellular carcinoma recurrence in patients with chronic hepatitis C. *Clin. Cancer Res.* **2013**, *19*, 1827–1837. [CrossRef]
42. Breen, M.S.; Maihofer, A.X.; Glatt, S.J.; Tylee, D.S.; Chandler, S.D.; Tsuang, M.T.; Risbrough, V.B.; Baker, D.G.; O'Connor, D.T.; Nievergelt, C.M.; et al. Gene networks specific for innate immunity define post-traumatic stress disorder. *Mol. Psychiatr.* **2015**, *20*, 1538–1545. [CrossRef]
43. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Franz, M.; Grouios, C.; Kazi, F.; Lopes, C.T.; et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **2010**, *38*, 214–220. [CrossRef]

44. Montojo, J.; Zuberi, K.; Shao, Q.; Bader, G.D.; Morris, Q. Network Assessor: An automated method for quantitative assessment of a network's potential for gene function prediction. *Front. Genet.* **2014**, *5*, 123. [CrossRef]
45. Huang, J.K.; Carlin, D.E.; Yu, M.K.; Zhang, W.; Kreisberg, J.F.; Tamayo, P.; Ideker, T. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **2018**, *6*, 484–495. [CrossRef] [PubMed]
46. Assenov, Y.; Ramírez, F.; Schelhorn, S.E.; Lengauer, T.; Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **2007**, *24*, 282–284. [CrossRef] [PubMed]
47. Wang, X.F.; Chen, G. Complex networks: Small-world, scale-free and beyond. *IEEE Circuits Syst. Mag.* **2003**, *3*, 6–20. [CrossRef]
48. Xu, B.; Lv, W.; Li, X.; Zhang, L.; Lin, J. Prognostic genes of hepatocellular carcinoma based on gene coexpression network analysis. *J. Cell. Biochem.* **2019**, *120*, 11616–11623. [CrossRef]
49. Johnson, M.R.; Shkura, K.; Langley, S.R.; Delahaye-Duriez, A.; Srivastava, P.; Hill, W.D.; Rackham, O.J.L.; Davies, G.; Harris, S.E.; Moreno-Moral, A.; et al. Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease. *Nat. Neurosci.* **2016**, *19*, 223–232. [CrossRef]
50. Liu, F.F.; Tu, T.T.; Zhang, H.F.; Hu, F.; Huang, L.; Deng, L.F.; Guo, M.; Wei, Q.; Li, K. Coexpression network analysis of platelet genes in sickle cell disease. *Platelets* **2019**, *30*, 1–8. [CrossRef]
51. Espinosa-Soto, C. On the role of sparseness in the evolution of modularity in gene regulatory networks. *PLoS Comput. Biol.* **2018**, *14*, e1006172. [CrossRef]
52. Gallo, C.A.; Carballido, J.A.; Ponzoni, I. Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinform.* **2011**, *12*, 123. [CrossRef]
53. Soinov, L.; Krestyaninova, M.; Brazma, A. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol.* **2003**, *4*, R6. [CrossRef]
54. Bulashevska, S.; Eils, R. Inferring genetic regulatory logic from expression data. *Bioinformatics* **2005**, *21*, 2706–2713. [CrossRef]
55. Ponzoni, I.; Azuaje, F.A.; Juan Glass, D. Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning. *IEEE-ACM Trans. Comput. Biol. Bioinform.* **2007**, *4*, 624–634. [CrossRef] [PubMed]
56. Martínez-Ballesteros, M.; Nepomuceno-Chamorro, I.A.; Riquelme, J.C. Discovering gene association networks by multi-objective evolutionary quantitative association rules. *J. Comput. Syst. Sci.* **2014**, *80*, 118–136. [CrossRef]
57. Kim, H.; Shin, J.; Kim, E.; Kim, H.; Hwang, S.; Shim, J.E.; Lee, I. YeastNet v3: A public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Res.* **2013**, *42*, D731–D736. [CrossRef] [PubMed]
58. Blake, D.D.; Weathers, F.W.; Nagy, L.M.; Kaloupek, D.G.; Gusman, F.D.; Charney, D.S.; Keane, T.M. The development of a clinician-administered PTSD scale. *J. Trauma Stress* **1995**, *8*, 75–90. [CrossRef] [PubMed]
59. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
60. Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **2009**, *25*, 1091–1093. [CrossRef]
61. Bindea, G.; Galon, J.; Mlecnik, B. CluePedia Cytoscape plugin: Pathway insights using integrated experimental and in silico data. *Bioinformatics* **2013**, *29*, 661–663. [CrossRef]
62. Watson, I.P.B.; Muller, H.K.; Jones, I.H.; Bradley, A.J. Cell-mediated immunity in combat veterans with post-traumatic stress disorder. *Med. J. Aust.* **1993**, *159*, 513–517. [CrossRef]

Chapter 8

Computational analysis of the global effects of *Ly6E* in the immune response to coronavirus infection using gene networks

Authors	Fernando M. Delgado-Chaves, Francisco Gómez-Vela, Federico Divina, Miguel García-Torres and Domingo S. Rodriguez-Baena
Journal	Genes
Editorial	MDPI
eISSN	2073-4425
Published	21/07/2020
DOI	doi.org/10.3390/genes11070831
Impact factor 2021	4,141
Quartile	Q2

genes

MDPI

Article

Computational Analysis of the Global Effects of *Ly6E* in the Immune Response to Coronavirus Infection Using Gene Networks

Fernando M. Delgado-Chaves *,†, **Francisco Gómez-Vela** †, **Federico Divina**, **Miguel García-Torres** and **Domingo S. Rodriguez-Baena**

Pablo de Olavide University, Carretera de Utrera km 1, ES-41013 Seville, Spain; fgomez@upo.es (F.G.-V.); fdiv@upo.es (F.D.); mgarciat@upo.es (M.G.-T.); dsrodbae@upo.es (D.S.R.-B.)

* Correspondence: fmdelcha@upo.es

† These authors contributed equally to this work.

Received: 23 May 2020; Accepted: 13 July 2020 ; Published: 21 July 2020

Abstract: Gene networks have arisen as a promising tool in the comprehensive modeling and analysis of complex diseases. Particularly in viral infections, the understanding of the host-pathogen mechanisms, and the immune response to these, is considered a major goal for the rational design of appropriate therapies. For this reason, the use of gene networks may well encourage therapy-associated research in the context of the coronavirus pandemic, orchestrating experimental scrutiny and reducing costs. In this work, gene co-expression networks were reconstructed from RNA-Seq expression data with the aim of analyzing the time-resolved effects of gene *Ly6E* in the immune response against the coronavirus responsible for murine hepatitis (MHV). Through the integration of differential expression analyses and reconstructed networks exploration, significant differences in the immune response to virus were observed in $Ly6E^{\Delta HSC}$ compared to *wild type* animals. Results show that *Ly6E* ablation at hematopoietic stem cells (HSCs) leads to a progressive impaired immune response in both liver and spleen. Specifically, depletion of the normal leukocyte mediated immunity and chemokine signaling is observed in the liver of $Ly6E^{\Delta HSC}$ mice. On the other hand, the immune response in the spleen, which seemed to be mediated by an intense chromatin activity in the normal situation, is replaced by ECM remodeling in $Ly6E^{\Delta HSC}$ mice. These findings, which require further experimental characterization, could be extrapolated to other coronaviruses and motivate the efforts towards novel antiviral approaches.

Keywords: gene co-expression network; murine coronavirus; viral infection; immune response; data mining; systems biology

1. Introduction

The recent SARS-CoV-2 pandemic has exerted an unprecedented pressure on the scientific community in the quest for novel antiviral approaches. A major concern regarding SARS-CoV-2 is the capability of the *coronaviridae* family to cross the species barrier and infect humans [1]. This, along with the tendency of coronaviruses to mutate and recombine, represents a significant threat to global health, which ultimately has put interdisciplinary research on the warpath towards the development of a vaccine or antiviral treatments.

Given the similarities found amongst the members of the *coronaviridae* family [2,3], analyzing the global immune response to coronaviruses may shed some light on the natural control of viral infection, and inspire prospective treatments. This may well be achieved from the perspective of systems biology, in which the interactions between the biological entities involved in a certain process are represented

by means of a mathematical system [4]. Within this framework, gene networks (GN) have become an important tool in the modeling and analysis of biological processes from gene expression data [5]. GNs constitute an abstraction of a given biological reality by means of a graph composed by nodes and edges. In such a graph, nodes represent the biological elements involved (i.e., genes, proteins or RNAs) and edges represent the relationships between the nodes. In addition, GNs are also useful to identify genes of interest in biological processes, as well as to discover relationships among these. Thus, they provide a comprehensive picture of the studied processes [6,7].

Among the different types of GNs, gene co-expression networks (GCNs) are widely used in the literature due to their computational simplicity and good performance in order to study biological processes or diseases [8–10]. GCNs usually compute pairwise co-expression indices for all genes. Then, the level of interaction between two genes is considered significant if its score is higher than a certain threshold, which is set *ad hoc*. Traditionally, statistical-based co-expression indices have been used to calculate the dependencies between genes [5,7]. Some of the most popular correlation coefficients are Pearson, Kendall or Spearman [11–13]. Despite their popularity, statistical-based measures present some limitations [14]. For instance, they are not capable of identifying non-linear interactions and the dependence on the data distribution in the case of parametric correlation coefficients. In order to overcome some of these limitations, new approaches, e.g., the use of information theory-based measures or ensemble approaches, are receiving much attention [15–17].

Gene Co-expression Networks (GCNs) have already been applied to the study of dramatic impact diseases, such as cancer [18], diabetes [19] or viral infections (e.g., HIV) in order to study the role of immune response to these illnesses [20,21]. Genetic approaches are expected to be the best strategy to understand viral infection and the immune response to it, potentially identifying the mechanisms of infection and assisting the design of strategies to combat infection [22,23]. The current gene expression profiling platforms, in combination with high-throughput sequencing, can provide time-resolved transcriptomic data, which can be related to the infection process. The main objective of this approach is to generate knowledge on the immune functioning upon viral entry into the organism, which means mean a perturbation to the system.

In the context of viral infection, a first defense line is the innate response mediated by interferons, a type of cytokines which eventually leads to the activation of several genes of antiviral function [24]. Globally, these genes are termed interferon-stimulated genes (ISGs), and regulate processes like inflammation, chemotaxis or macrophage activation among others. Furthermore, ISGs are also involved in the subsequent acquired immune response, specific for the viral pathogen detected [25]. Gene *Ly6E* (lymphocyte antigen 6 family member e), which has been related to T cell maturation and tumorogenesis, is amongst the ISGs [26]. This gene is transcriptionally active in a variety of tissues, including liver, spleen, lung, brain, uterus and ovary. Its role in viral infection has been elusive due to contradictory findings [27]. For example, in Liu et al. [28], *Ly6E* was associated with the resistance to Marek's disease virus (MDV) in chickens. Moreover, differences in the immune response to mouse adenovirus type 1 (MAV-1) have been attributed to *Ly6E* variants [29]. Conversely, *Ly6E* has also been related to an enhancement of human immunodeficiency viruses (HIV-1) pathogenesis, by promoting HIV-1 entry through virus–cell fusion processes [30]. Also in the work by Mar et al. [31], the loss of function of *Ly6E* due to gene *knockout* reduced the infectivity of Influenza A virus (IAV) and yellow fever virus (YFV). This enhancing effect of *Ly6E* on viral infection has also been observed in other enveloped RNA viruses such as in West Nile virus (WNV), dengue virus (DEN), Zika virus (ZIKV), O'nyong nyong virus (ONNV) and Chikungunya virus (CHIKV) among others [32]. Nevertheless, the exact mechanisms through which *Ly6E* modulates viral infection virus-wise, and sometimes even cell type-dependently, require further characterization.

In this work we present a time-resolved study of the immune response of mice to a coronavirus, the murine hepatitis virus (MHV), in order to analyze the implications of gene *Ly6E*. To do so, we have applied a GCN reconstruction method called *EnGNet* [33], which is able to perform an ensemble strategy to combine three different co-expression measures, and a topology optimization of the final

network. *EnGNet* has outscored other methods in terms of network precision and reduced network size, and has been proven useful in the modeling of disease, as in the case of Human post-traumatic stress disorder.

The rest of the paper is organized as follows. In the next section, we propose a description of related works. In Section 3, we first describe the dataset used in this paper, and then we introduce the *EnGNet* algorithm and the different methods used to infer and analyze the generated networks. The results obtained are detailed in Section 4, while, in Section 5, we propose a discussion of the results presented in the previous section. Finally, in Section 6, we draw the main conclusions of our work.

2. Related Works

As already mentioned, gene co-expression networks have been extensively applied in the literature for the understanding of the mechanisms underlying complex diseases like cancer, diabetes or Alzheimer [34–36]. Globally, GCN serve as an *in silico* genetic model of these pathologies, highlighting the main genes involved in these at the same time [37]. Besides, the identification of modules in the inferred GCNs, may lead to the discovery of novel biomarkers for the disease under study, following the 'guilt by association' principle. Along these lines, GCNs are also considered suitable for the study of infectious diseases, as those caused by viruses to the matter at hand [38]. To do so, multiple studies have analyzed the effects of viral infection over the organism, focusing on immune response or tissue damage [39,40].

For instance, the analysis of gene expression using co-expression networks is shown in the work by Pedragosa et al. [41], where the infection caused by Lymphocytic Choriomeningitis Virus (LCMV) is studied over time in mice spleen using GCNs. In Ray et al. [42], GCNs are reconstructed from different microarray expression data in order to study HIV-1 progression, revealing important changes across the different infection stages. Similarly, in the work presented by McDermott et al. [43], the over- and under-stimulation of the innate immune response to severe acute respiratory syndrome coronavirus (SARS-CoV) infection is studied. Using several network-based approaches on multiple *knockout* mouse strains, authors found that ranking genes based on their network topology made accurate predictions of the pathogenic state, thus solving a classification problem. In [39], co-expression networks were generated by microarray analysis of pediatric influenza-infected samples. Thanks to this study, genes involved in the innate immune system and defense to virus were revealed. Finally, in the work by Pan et al. [44], a co-expression network is constructed based on differentially-expressed microRNAs and genes identified in liver tissues from patients with hepatitis B virus (HBV). This study provides new insights on how microRNAs take part in the molecular mechanism underlying HBV-associated acute liver failure.

The alarm posed by the COVID-19 pandemic has fueled the development of effective prevention and treatment protocols for 2019-nCoV/SARS-CoV-2 outbreak [45]. Due to the novelty of SARS-CoV-2, recent research takes similar viruses, such as SARS-CoV and Middle East Respiratory Syndrome coronavirus (MERS-CoV), as a starting point. Other coronaviruses, like Mouse Hepatitis Virus (MHV), are also considered appropriate for comparative studies in animal models, as demonstrated in the work by De Albuquerque et al. [46] and Ding et al. [47]. MHV is a murine coronavirus (M-CoV) that causes an epidemic illness with high mortality, and has been widely used for experimentation purposes. Works like the ones by Case et al. [48] and Gorman et al. [49], study the innate immune response against MHV arbitrated by interferons, and those interferon-stimulated genes with potential antiviral function. This is the case of gene *Ly6E*, which has been shown to play an important role in viral infection, as well as various orthologs of the same gene [50,51]. Mechanistic approaches often involved the ablation of the gene under study, like in the work by Mar et al. [31], where gene *knockout* was used to characterize the implications of *Ly6E* in Influenza A infection. As it is the case of Giotis et al. [52], these studies often involve global transcriptome analyses, via RNA-seq or microarrays, together with computational efforts, which intend to screen the key elements of the immune system that are required

for the appropriate response. This approach ultimately leads experimental research through predictive analyses, as in the case of co-expression gene networks [53].

3. Materials and Methods

In the following subsections, the main methods and GCN reconstruction steps are addressed. First, in Section 3.1, the original dataset used in the present work is described, together with the experimental design. Then, in Section 4.1, the data preprocessing steps are described. Subsequently in Section 3.3, key genes controlling the infection progression are extracted through differential expression analyses. Finally, the inference of GCNs and their analysis are detailed in Sections 3.4 and 3.5, respectively.

3.1. Original Dataset Description

The original experimental design can be described as follows. The progression of the MHV infection at genetic level was evaluated in two genetic backgrounds: wild type (*wt*, Ly6Efl/fl) and Ly6E *knockout* mutants (*ko*, $Ly6E^{\Delta HSC}$). The ablation of gene *Ly6E* in all cell types is lethal, hence the $Ly6E^{\Delta HSC}$ strain contains a disrupted version of gene Ly6E only in hematopoietic stem cells (HSC), which give rise to myeloid and lymphoid progenitors of all blood cells. *Wild type* and $Ly6E^{\Delta HSC}$ mice were injected intraperitoneally with 5000 PFU MHV-A59. At 3 and 5 days post-injection (d p.i.), mice were euthanized and biological samples for RNA-Seq were extracted. The overall effects of MHV infection in both *wt* and *ko* strains was assessed in liver and spleen.

In total 36 samples were analyzed, half of these corresponding to liver and spleen, respectively. From the 18 organ-specific samples, 6 samples correspond to mock infection (negative control), 6 to MHV-infected samples at 3 d p.i. and 6 to MHV-infected samples at 5 d p.i. For each sample, two technical replicates were obtained. Libraries of cDNA generated from the samples were sequenced using Illumina NovaSeq 6000. Further details on sample preparation can be found in the original article by Pfaender et al. [54]. For the sake of simplicity, MHV-infected samples at 3 and 5 d p.i. will be termed 'cases', whereas mock-infection samples will be termed 'controls'.

The original dataset consists of 72 files, one per sample replicate, obtained upon the mapping of the transcript reads to the reference genome. Reads were recorded in three different ways, considering whether these mapped introns, exons or total genes. Then, a count table was retrieved from these files by selecting only the total gene counts of each sample replicate file.

3.2. Data Pre-Processing

Pre-processing was performed using the *EdgeR* [55] R package. The original dataset by Pfaender et al. [54] was retrieved from GEO (accession ID: GSE146074) using the *GEOquery* [56] package. Additional files on sample information and treatment were also used to assist the modeling process.

By convention, a sequencing depth per gene below 10 is considered neglectable [57,58]. Genes meeting this criterion are known as low expression genes, and are often removed since they add noise and computational burden to the following analyses [59]. In order to remove genes showing less than 10 reads across all conditions, counts per million (CPM) normalization was performed, so possible differences between library sizes for both replicates would not affect the result.

Afterwards, Principal Components Analyses (PCA) were performed over the data in order to detect the main sources of variability across samples. PCA were accompanied by unsupervised k-medoid clustering analyses, in order to identify different groups of samples. In addition, multidimensional scaling plots (MDS) were applied to further separate samples according to their features. Last, between-sample similarities were assessed through hierarchical clustering.

3.3. Differential Expression Analyses

The analyses of differential expression served a two-way purpose, (i) the exploration of the directionality in the gene expression changes upon viral infection, and (ii) the identification of key regulatory elements for the subsequent network reconstruction. In the present application,

differentially-expressed genes (DEG) were filtered from the original dataset and proceeded to the reconstruction process. This approximation enabled the modeling of the genetic relationships that are considered of relevance in the presented comparison [60–62]. In the present work mice samples were compared organ-wise depending on whether these corresponded to control, 3 d p.i. and 5 d p.i.

The identification of DEG was performed using the *Limma* [63] R package, which provides non-parametric robust estimation of the gene expression variance. This package includes *Voom*, a method that incorporates RNA-Seq count data into the *Limma* workbench, originally designed for microarrays [64]. In this case, a minimum log2-fold-change (log2FC) of 2 was chosen, which corresponds to four fold changes in the gene expression level. P-value was adjusted by Benjamini-Hochberg [65] and the selected adjusted p-value cutoff was 0.05.

3.4. Inference of the Gene Networks: EnGNet

In order to generate gene networks the *EnGNet* algorithm was used. This technique, presented in Gómez-Vela et al. [33], is able to compute gene co-expression networks with a competitive performance compared other approaches from the literature. *EnGNet* performs a two-step process to infer gene networks: (a) an ensemble strategy for a reliable co-expression networks generation, and (b) a greedy algorithm that optimizes both the size and the topological features of the network. These two features of *EnGNet* offer a reliable solution for generating gene networks. In fact, *EnGNet* relies on three statistical measures in order to obtain networks. In particular, the measures used are the Spearman, Kendall and normalized mutual information (NMI), which are widely used in the literature for inferring gene networks. *EnGNet* uses these measures simultaneously by applying an ensemble strategy based on major voting, i.e., a relationship will be considered correct if at least 2 of the 3 measures evaluate the relationship as correct. The evaluation is based on different independent thresholds. In this work, the different thresholds were set to the values originally used in [33]: 0.9, 0.8 and 0.7 for Spearman, Kendall and NMI, respectively.

In addition, as mentioned above, *EnGNet* performs an optimization of the topological structure of the networks obtained. This reduction is based on two steps: (i) the pruning of the relations considered of least interest in the initial network, and (ii) the analysis of the hubs present in the network. For this second step of the final network reconstruction, we have selected the same threshold that was used in [33], i.e., 0.7. Through this optimization, the final network produced by *EnGNet* results easier to analyze computationally, due to its reduced size.

3.5. Networks Analyses

Networks were imported to R for the estimation of topology parameters and the addition of network features that are of interest for the latter network analysis and interpretation. These attributes were added to the reconstructed networks to enrich the modeling using the *igraph* [66] R package. The networks were then imported into *Cytoscape* [67] through RCy3 [68] for examination and analyses purposes. In this case, two kind of analyses were performed: (i) a topological analysis and (ii) an enrichment analysis.

Regarding the topological analysis, clustering evaluation was performed in order to identify densely connected nodes, which, according to the literature, are often involved in a same biological process [69]. The chosen clustering method was community clustering (GLay) [70], implemented via *Cytoscape*'s *ClusterMaker* app [71], which has yielded significant results in the identification of densely connected modules [72,73]. Among the topology parameters, *degree* and *edge betweenness* were estimated. The *degree* of a node refers to the number of its linking nodes. On the other hand, the *betweenness* of an edge refers to the number of shortest paths which go through that edge. Both parameters are considered as a measure of the implications of respectively nodes and edges in a certain network. Particularly, nodes whose *degree* exceeds the average network node *degree*, the so called *hubs*, are considered key elements of the biological processes modeled by the network. In this particular case, the distribution of nodes' degree network was analyzed so those nodes whose degree

exceeded a threshold were selected as hubs. This threshold is defined as $Q3 + 1.5 \times IQR$, where $Q3$ is the third quartile and IQR the interquartile range of the degree distribution. This method has been widely used for the detection of upper outliers in non-parametric distributions [74,75], as it is the case. However, the outlier definition does not apply to this distribution since those nodes whose degree are far above the median degree are considered hubs.

On the other hand, Gene Ontology (GO) Enrichment Analysis provides valuable insights on the biological reality modeled by the reconstructed networks. The Gene Ontology Consortium [76] is a data base that seeks for a unified nomenclature for biological entities. GO has developed three different ontologies, which describe gene products in terms of the biological processes, cell components or molecular functions in which these are involved. Ontologies are built out of GO terms or annotations, which provide biological information of gene products. In this case, the *ClusterProfiler* [77] R package, allowed the identification of the statistically over-represented GO terms in the gene sets of interest. Additional enrichment analyses were performed using *DAVID* [78]. For both analyses, the complete genome of *Mus musculus* was selected as background. Finally, further details on the interplay of the genes under study was examined using the *STRING* database [79].

4. Results

The reconstruction of gene networks that adequately model viral infection involves multiple steps, which ultimately shape the final outcome. First, in Section 4.1, exploratory analyses and data preprocessing are detailed, which prompted the modeling rationale. Then, in Section 4.2, differential expression is evaluated for the samples of interest. Finally, networks reconstruction and analysis are addressed in Section 4.3. At the end, four networks were generated, both in an organ- and genotype-wise manner. A schematic representation of the GCN reconstruction approach is shown in Figure 1.

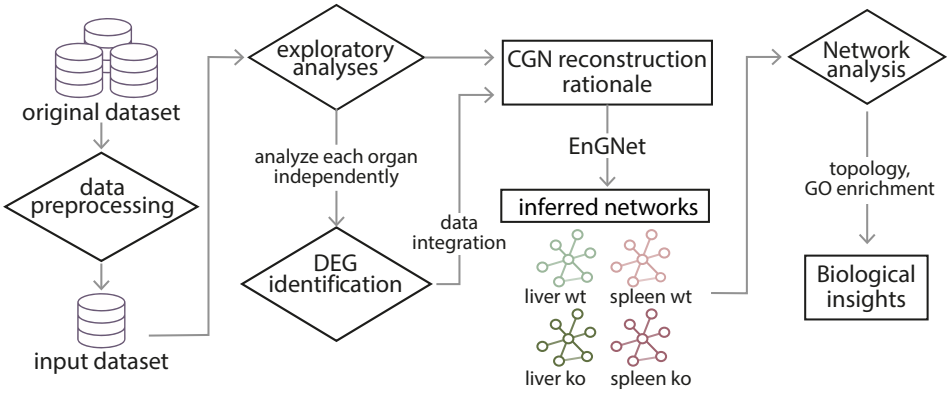


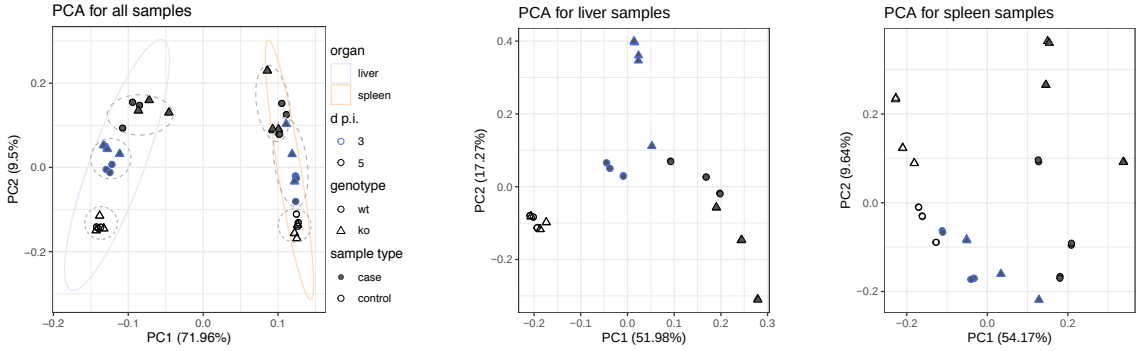
Figure 1. General scheme for the reconstruction method. The preprocessed data was subjected to exploratory and differential expression analyses, which imposed the reconstruction rationale. Four groups of samples were used to generate four independent networks, respectively modeling the immune response in the liver, both in the *wt* and the *ko* situations; and in the spleen, also in the *wt* and the *ko* scenarios.

4.1. Data Pre-Processing and Exploratory Analyses

In order to remove low expression genes, a sequencing depth of 10 was found to correspond to an average CPM of 0.5, which was selected as threshold. Hence, genes whose expression was found over 0.5 CPM in at least two samples of the dataset were maintained, ensuring that only genes which are truly being expressed in the tissue will be studied. The dataset was Log2-normalized with priority to the following analyses, in accordance to the recommendations posed in Law et al. [64].

The results of both PCA and k-medoid clustering are shown in Figure 2a. Clustering of the Log2-normalized samples revealed clear differences between liver and spleen samples. Also, for each organ, three subgroups of analogous samples that cluster together are identified. These groups

correspond to mock infection, MHV-infected mice at 3 d p.i. and MHV-infected mice at 5 d p.i. (dashed lines in Figure 2a). Finally, subtle differences were observed in homologous samples of different genotypes (Figure A1).



(**a**) (**b**) (**c**)

Figure 2. (**a**) PCA plot of the Log2-normalized counts for the exploratory analysis of all samples under study. The metric used for k-medoid partitioning was the Euclidean distance. Both replicates are included. Two groups, respectively corresponding to liver and spleen samples, are clearly differentiated. Dashed lines were added for improved visualization of the different groups that are distinguished within each organ. Organ-specific PCA for (**b**) liver and (**c**) spleen samples. Both replicates are included. PCA suggests the progressive nature of the MHV infection, where groups corresponding to mock infections, 3 d p.i. and 5 d p.i. are distinguished in varying degrees. Differences between controls and cases are more evident in liver samples. Figure 2a legend is the same for Figure 2b,c.

Organ-specific PCA revealed major differences between MHV-infected samples for $Ly6E^{\Delta HSC}$ and *wt* genotypes, at both 3 and 5 d p.i. These differences were not observed in the mock infection (control situation). Organ-wise PCA are shown in Figure 2b,c. The distances between same-genotype samples illustrate the infection-prompted genetic perturbation from the uninfected status (control) to 5 d p.i., where clear signs of hepatitis were observed according to the original physiopathology studies [54]. On the other hand, the differences observed between both genotypes are indicative of the role of gene *Ly6E* in the appropriate response to viral infection. These differences are subtle in control samples, but in case samples, some composition biass is observed depending on whether these are *ko* or *wt*, especially in spleen samples. The comparative analysis of the top 500 most variable genes confirmed the differences observed in the PCA, as shown in Figure A2. Among the four different features of the samples under study: organ, genotype, sample type (case or control) and days post injection; the dissimilarities in terms of genotype were the subtlest.

In the light of these exploratory findings, the network reconstruction approach was performed as follows. Networks were reconstructed organ-wise, as these exhibit notable differences in gene expression. Additionally, a main objective of the present work is to evaluate the differences in the genetic response in the *wt* situation compared to the $Ly6E^{\Delta HSC}$ *ko* background, upon the viral infection onset in the two mentioned tissues.

For each organ, Log2-normalized samples were coerced to generate time-series-like data, i.e., for each genotype, 9 samples will be considered as a set, namely 3 control samples, 3 case samples at 3 d p.i. and 3 case samples at 5 d p.i. Both technical replicates were included. This rational design seeks for a gene expression span representative of the infection progress. Thereby, control samples may well be considered as a time zero for the viral infection, followed by the corresponding samples at 3 and 5 d p.i. The proposed rationale is supported by the exploratory findings, which position 3 d p.i. samples between control and 5 d p.i. samples. At the same time, the reconstruction of gene expression becomes robuster with increasing number of samples. In this particular case, 18 measuring points are attained for the reconstruction of each one of the four intended networks, since two technical replicates were obtained per sample [80].

4.2. Identification of Differentially-Expressed Genes Between Wild Type and Ly6E$^{\Delta HSC}$ Samples

The differential expression analyses were performed over the four groups of 9 samples explained above, with the aim of examining the differences in the immune response between *Ly6E*$^{\Delta HSC}$ and *wt* samples. *Limma - Voom* differential expression analyses were performed over the Log2-normalized counts, in order to evaluate the different genotypes whilst contrasting the three infection stages: control vs. cases at 3 d p.i., control vs. cases at 5 d p.i. and cases at 3 vs. 5 d p.i. The choice of a minimum absolute log2FC ≥ 2, enabled considering only those genes that truly effect changes between *wt* and *Ly6E*$^{\Delta HSC}$ samples, whilst maintaining a relatively computer-manageable number of DEG for network reconstruction. The latter is essential for the yield of accurate network sparseness values, as this is a main feature of gene networks [5].

For both genotypes and organs, the results of the differential expression analyses reveal that MHV injection triggers a progressive genetic program from the control situation to the MHV-infected scenario at 5 d p.i., as shown in Figure 3a. The absolute number of DEG between control vs. cases at 5 d p.i. was considerably larger than in the comparison between control vs. cases at 3 d p.i. Furthermore, in all cases, most of the DEG in control vs. cases at 3 d p.i. are also differentially-expressed in the control vs. cases at 5 d p.i. comparison, as shown in Figure 4.

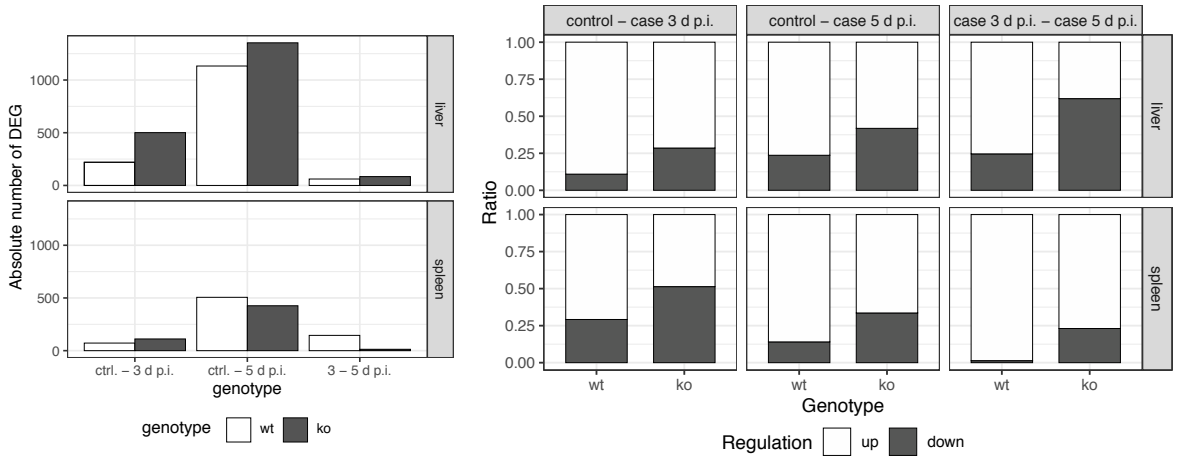
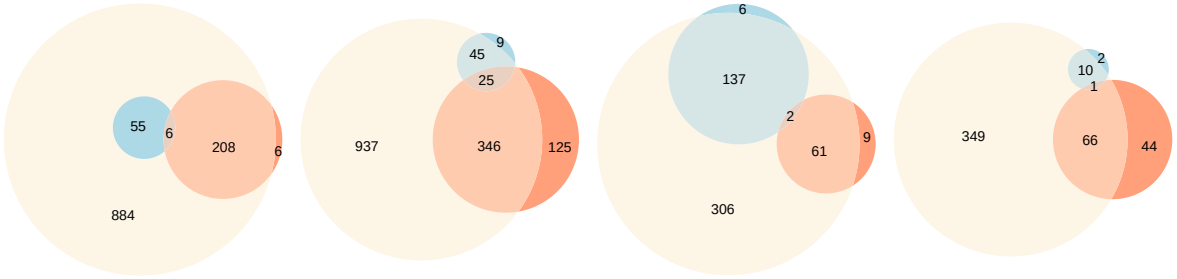


Figure 3. (**a**) Absolute numbers of DEG in the different comparisons (**b**) Ratio of up- and downregulated DEG in the different performed comparisons. Three comparisons were performed: control vs. case samples at 3 d p.i., control vs. case samples at 5 d p.i. and case samples at 3 vs. 5 d p.i. *ko* refers to *Ly6E*$^{\Delta HSC}$ samples.

Regarding genes fold change, an overall genetic up-regulation is observed upon infection. Around 70% of DEG are upregulated for all the comparisons performed for *wt* samples, as shown in Figure 3b. Nonetheless, a dramatic reduce in this genetic up-regulation is observed, by contrast, in *knockout* samples, even limiting upregulated genes to nearly 50% in the control vs. cases at 3 d p.i. comparison of liver *Ly6E*$^{\Delta HSC}$ samples. The largest differences are observed in the comparison of controls vs. cases at 5 d p.i (Figures A3 and A4). These DEG are of great interest for the understanding of the immune response of both *wt* and *ko* mice to viral infection. These genes were selected to filter the original dataset for latter network reconstruction.

The commonalities between *wt* and *ko* control samples for both organs were also verified through differential expression analysis following the same criteria (Log2FC > 2, p value < 0.05). The number of DEG between *wt* and *ko* liver control samples (2) and between *wt* and *ko* spleen control samples (20) were not considered significant, so samples were taken as analogous starting points for infection.



(**a**) DEG liver *wt* (**b**) DEG liver *ko* (**c**) DEG spleen *wt* (**d**) DEG spleen *ko*

Figure 4. Euler diagrams showing the overlapping of DEG between the three possible contrast situations: control vs. cases at 3 d p.i. (red), control vs. cases at 5 d p.i. (yellow) and cases at 3 d p.i. vs. cases at 5 d p.i. (blue) *ko* refers to $Ly6E^{\Delta HSC}$ samples. These comparisons were performed both organ and genotype-wise considering four groups of samples: (**a**) liver *wt*, (**b**) liver $Ly6E^{\Delta HSC}$, (**c**) spleen *wt*, (**d**) spleen $Ly6E^{\Delta HSC}$.

4.3. Reconstruction and Analysis of Gene Networks

As stated above, the samples were arranged both organ and genotype-wise in order to generate networks which would model the progress of the disease in each scenario. GCNs were inferred from Log2-normalized expression datasets. A count of 1 was added at log2 normalization so the problem with remaining zero values was avoided. Each network was generated exclusively taking into consideration their corresponding DEG at control vs. cases at 5 d p.i., where larger differences were observed. Four networks were then reconstructed from these previously-identified DEG for liver *wt* samples (1133 genes), liver *ko* samples (1153 genes), spleen *wt* samples (506 genes) and spleen *ko* samples (426 genes). This approach results in the modeling of only those relationships that are related to the viral infection. Each sample set was then fed to *EnGNet* for the reconstruction of the subsequent network. Genes that remained unconnected due to weak relationships, which do not overcome the set threshold, were removed from the networks. Furthermore, the goodness of *EnGNet*-generated models outperformed other well-known inference approaches, as detailed in Appendix B.

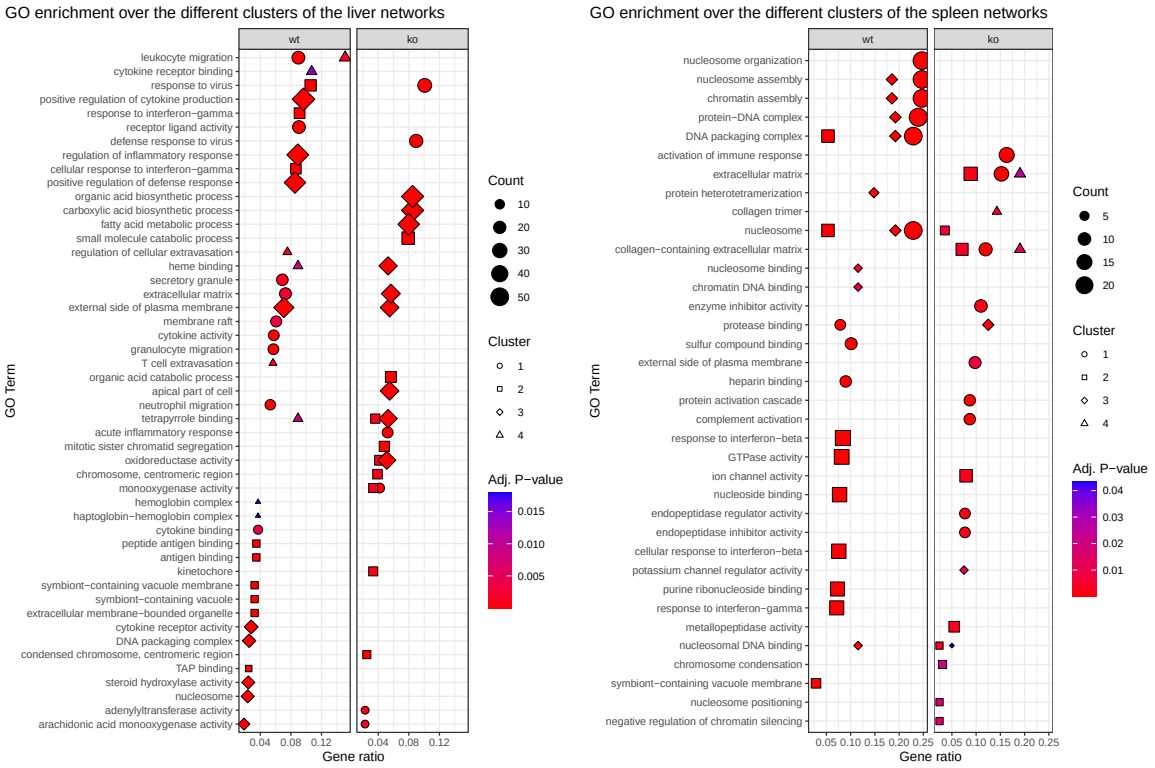
Topological parameters were estimated and added as node attributes using *igraph*, together with Log2FC, prior to Cytoscape import. Specifically, networks were simplified by removing potential loops and multiple edges. The clustering topological scrutiny of the reconstructed networks revealed neat modules in all cases, as shown in Figure A5. The number of clusters identified in each network, as well as the number of genes harbored in the clusters is shown in Table A1.

As already mentioned, according to gene networks theory, nodes contained within the same cluster are often involved in the same biological process [5,81]. In this context, the GO-based enrichment analyses over the identified clusters may well provide an idea of the affected functions. Only clusters containing more than 10 genes were considered, since this is the minimum number of elements required by the enrichment tool *ClusterProfiler*. The results of the enrichment analyses revealed that most GO terms were not shared between *wt* and *ko* homologous samples, as shown in Figure 5.

In order to further explore the reconstructed networks, the intersection of *ko* and *wt* networks of a same organ was computed. This refers to the genes and relationships that are shared between both genotypes for a specific organ. Additionally, the genes and relationships that were exclusively present at the *wt* and *ko* samples were also estimated, as shown in Figure A6. The enrichment analyses over the nodes, separated using this criterion, would reveal the biological processes that make the difference between in $Ly6E^{\Delta HSC}$ mice compared to *wt* ones. The results of such analyses are shown in Figure A7.

Finally, the exploration of nodes' *degree* distribution would reveal those genes that can be considered hubs. Those nodes comprised within the top genes with highest degree (degree $>$ Q3 + 1.5 $\times$ IQ), also known as upper outliers in the nodes distribution, were considered hubs. A representation of nodes' degree distribution throughout the four reconstructed networks is shown in Figure 6.

These distributions are detailed in Figure A8. This method provided four cutoff values for the degree, 24, 39, 21 and 21, respectively for liver *wt* and *ko*, spleen *wt* and *ko* networks. Above these thresholds, nodes would be considered as hubs in each network. These hubs are shown in Tables A2–A5.



(**a**) (**b**)

Figure 5. Enrichment analyses performed over the main clusters identified in *wt* and *ko* networks of (**a**) liver and (**b**) spleen networks. Gene ratio is defined by the number of genes used as input for the ernichment analyses associated with a particular GO term divided by the total number of input genes.

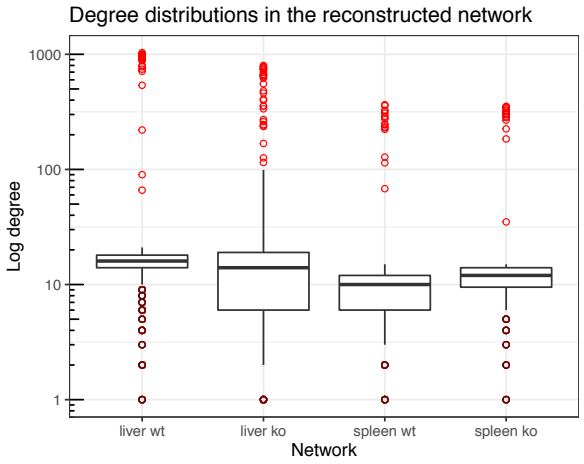


Figure 6. Boxplots representative of the degree distributions for each one of the four reconstructed networks. Identified hubs, according to the $Q3 + 1.5 \times IQR$ criterion, are highlighted in red. The degree cutoffs, above which nodes would be considered as hubs, were 24, 39, 21 and 21, respectively for liver *wt*, liver *ko*, spleen *wt* and spleen *ko* networks. Note degree is represented in a log scale given that the reconstructed networks present a scale-free topology.

5. Discussion

In this work four gene networks were reconstructed to model the genetic response MHV infection in two tissues, liver and spleen, and in two different genetic backgrounds, *wild type* and $Ly6E^{\Delta HSC}$. Samples were initially explored in order to design an inference rationale. Not only did the designed approach reveal major differences between the genetic programs in each organ, but also, between different subgroups of samples, in a time-series-like manner. Noticeably, disparities between *wt* and $Ly6E^{\Delta HSC}$ samples were observed in both tissues, and differential expression analyses revealed relevant differences in terms of the immune response generated. Hereby, our results predict the impact of *Ly6E ko* on HSC, which resulted in an impaired immune response compared to the *wt* situation.

5.1. Exploratory Analyses Revealed a Time-Series Llike Behaviour on Raw Data, Assisting Network Reconstruction

Overall, results indicate that the reconstruction rationale, elucidated from exploratory findings, is suitable for the modeling of the viral progression. Regarding the variance in gene expression in response to virus, PCA and K-medoid clustering revealed strong differences between samples corresponding to liver spleen, respectively (Figure 2a). These differences set the starting point for the modeling approach, in which samples corresponding to each organ were analyzed independently. This *modus operandi* is strongly supported by the tropism that viruses exhibit for certain tissues, which ultimately results in a differential viral incidence and charge depending on the organ [82]. In particular, the liver is the target organ of MHV, identified as the main disease site [83]. On the other hand, the role of the spleen in innate and adaptive immunity against MHV has been widely addressed [84,85]. The organization of this organ allows blood filtration for the presentation of antigens to cognate lymphocytes by the antigen presenting cells (APCs), which mediate the immune response exerted by T and B cells [86].

As stated before, PCA revealed differences between the three sample groups on each organ: control and MHV-infected at 3 and 5 d p.i. Interestingly, between-groups differences are specially clear for liver samples (Figure 2b), whereas spleen samples are displayed in a continuum-like way. This becomes more evident in organ-wise PCA (Figure 2), and was latter confirmed by the exploration of the top 500 most variable genes and differential expression analyses (Figure A2). Furthermore, clear differences between *wt* and $Ly6E^{\Delta HSC}$ samples are observed in none of these analyses, although the examination of the differential expression and network reconstruction did exposed divergent immune responses for both genotypes.

5.2. Differential Expression Analyses Revealed Significant Changes between Wild Type and Knockout Samples

The differential expression analyses revealed the progressive genetic response to virus for both organs and genotypes (Figures 3a and 4). In a *wt* genetic background, MHV infection causes an overall rise in the expression level of certain genes, as most DEG in cases vs. control samples are upregulated. However, in a $Ly6E^{\Delta HSC}$ genetic background, this upregulation is not as prominent as in a *wt* background, significantly reducing the number of upregulated genes (Figure 3b). Besides, the number of DEG in each comparison varies from *wt* to $Ly6E^{\Delta HSC}$ samples.

Attending at the DEG in the performed comparisons, for both the *wt* and *ko* genotypes, liver cases at 3 d p.i. are more similar to liver cases at 5 d p.i. than to liver controls, since the number of DEG between the first two measuring points is significantly lower than the number of DEG between control and case samples at 3 d p.i. (Figure 4a,b). A different situation occurs in the spleen, where *wt* cases at 3 d p.i. are closer to control samples (Figure 4c), whereas *ko* cases at 3 d p.i. seem to be more related to cases at 5 d p.i. (Figure 4d). This was already suggested by hierarchical clustering in the analysis of the top 500 most variable genes, and could be indicative of a different progression of the infection impact on both organs, which could be modulated by gene *Ly6E*, at least for the spleen samples.

Moreover, the results of the DEG analyses indicate that the sole *knockout* of gene *Ly6E* in HSC considerably affects the upregulating genetic program normally triggered by viral infection

in *wild type* individuals (in both liver and spleen). Interestingly, there are some genes in each organ and genotype that are differentially expressed in every comparison between the possible three sample types, controls, cases at 3 d p.i. and cases at 5 d p.i. These genes, which we termed highly DEG, could be linked to the progression of the infection, as changes in their expression level occur with days post injection, according to the data. The rest of the DEG, show an uprise or fall when comparing two sample types, which does not change significantly in the third sample type. Alternatively, highly DEG, shown in Table A6, exhibited three different expression patterns: (i) Their expression level, initially low, rises from control to cases at 3 d p.i. and then rises again in cases at 5 d p.i. (ii) Their expression level, initially high in control samples, falls at 3 d p.i. and falls even more at 5 d p.i cases. (iii) Their expression level, initially low, rises from control to cases at 3 d p.i. but then falls at cases at 5 d p.i., when it is still higher than the initial expression level. These expression patterns, which are shown in Figure A9, might be used to keep track of the disease progression, differentiating early from late infection stages.

In some cases, these genes exhibited inconsistent expression levels, specially at 5 d p.i. cases, which indicates the need for further experimental designs targeting these genes. Highly DEG could be correlated with the progression of the disease, as in regulation types (i) and (ii) or by contrast, be required exclusively at initial stages, as in regulation type (iii). Notably, genes *Gm10800* and *Gm4756* are predicted genes which, to date, have been poorly described. According to the *STRING* database [79], *Gm10800* is associated with gene *Lst1* (Leukocyte-specific transcript 1 protein), which has a possible role in modulating immune responses. In fact, *Gm10800* is homologous to human gene PIRO (Progranulin-Induced-Receptor-like gene during Osteoclastogenesis), related to bone homeostasis [87,88]. Thus, we hypothesize that bone marrow-derived cell lines, including erythrocytes and leukocytes (immunity effectors), could also be regulated by *Gm10800*. On the other hand, *Gm4756* is not associated to any other gene according to *STRING*. Protein *Gm4756* is homologous to Human protein DHRS7 (dehydrogenase/reductase SDR family member 7) isoform 1 precursor. Nonetheless and to the best of our knowledge, these genes have not been previously related to *Ly6E*, and could play a role in the immune processes mediated by this gene.

Finally, highly DEG were not found exclusively present in *wt* nor *ko* networks, instead, these were common nodes of these networks for each organ. This suggests that highly DEG might be of core relevance upon MHV infection, with a role in those processes independent on $Ly6E^{\Delta HSC}$. Besides, genes *Hykk*, *Ifit3* and *Ifit3b*; identified as highly DEG throughout liver $Ly6E^{\Delta HSC}$ samples were also identified as hubs in the liver *ko* network. Also gene *Saa3*, highly DEG across spleen $Ly6E^{\Delta HSC}$ samples was considered a hub in the spleen *ko* network. Nevertheless, these highly DEG require further experimental validation.

5.3. The Ablation of Ly6E in HSC Results in Impaired Immune Response as Predicted by Enrichment Analyses

The enrichment analyses of the identified clusters at each network revealed that most GO terms are not shared between the two genotypes (Figure 5), despite the considerable amount of shared genes between the two genotypes for a same organ. The network reconstructed from liver *wt* samples reflects a strong response to viral infection, involving leukocyte migration or cytokine and interferon signaling among others. These processes, much related to immune processes, are not observed in its *ko* counterpart.

The liver *wt* network presented four clusters (Figure A5a). Its cluster 1 regulates processes related to leukocyte migration, showing the implication of receptor ligand activity and cytokine signaling, which possibly mediates the migration of the involved cells. Cluster 2 is related to interferon-gamma for the response to MHV, whereas cluster 3 is probably involved in the inflammatory response mediated by pro-inflammatory cytokines. Last, cluster 4 is related to cell extravasation, or the leave of blood cells from blood vessels, with the participation of gene *Nipal1*. The positive regulation observed across all clusters suggests the activation of these processes. Overall, hub genes in this network have been related to the immune response to viral infection, as the innate immune response to the virus is the

mediated by interferons. Meanwhile, the liver *ko* network showed three main clusters (Figure A5b). Its cluster 1 would also be involved in defense response to virus, but other processes observed in the liver *wt* network, like leukocyte migration or cytokine activity, are not observed in this cluster nor the others. Cluster 2 is then related to the catabolism of small molecules and cluster 3 is involved in acids biosynthesis. These processes are certainly ambiguous and do not correspond the immune response observed in the *wt* situation, which suggests a decrease in the immune response to MHV as a result of *Ly6E* ablation in HSC.

On the other hand, spleen *wt* samples revealed high nuclear activity potentially involving nucleosome remodeling complexes and changes in DNA accessibility. Histone modification is a type of epigenetic modulation which regulates gene expression. Taking into account the central role of the spleen in the development of immune responses, the manifested relevance of chromatin organization could be accompanied by changes in the accessibility of certain DNA regions with implications in the spleen-dependent immune response. This is supported by the reduced reaction capacity in the first days post-infection of $Ly6E^{\Delta HSC}$ samples compared to *wt*, as indicated by the number of DEG between control and cases at 3 d p.i for these genotypes. The spleen *wt* network displayed three clusters (Figure A5c). Cluster 1, whose genes were all upregulated in $Ly6E^{\Delta HSC}$ samples at 5 d p.i. compared to mock infection, is mostly involved in nucleosome organization and chromatin remodelling, together with cluster 3. Cluster 2 would also be related to DNA packaging complexes, possibly in response to interferon, similarly to liver networks. Instead, in spleen *ko* most genes take part in processes related to the extracellular matrix. In the spleen *ko* network, four clusters were identified (Figure A5d). Cluster 1 is related to the activation of an immune response, but also, alongside with clusters 2 and 4, to the extracellular matrix, possibly in relation with collagen, highlighting its role in the response to MHV. Cluster 3 is implied in protease binding. The dramatic shut down in the *ko* network of the nuclear activity observed in the spleen *wt* network, leads to the hypothesis that the chromatin remodeling activity observed could be related to the activation of certain immunoenhancer genes, modulated by gene *Ly6E*. In any case, further experimental validation of these results would provide meaningful insights in the face of potential therapeutic approaches (See Appendix A for more details).

The exploration of nodes memebership, depending on whether these exclusively belonged to *wt* or *ko* networks or, by contrast, were present in both networks, helped to understand the impairment caused by $Ly6E^{\Delta HSC}$. In this sense, GO enrichment analyses over these three defined categories of the nodes in the liver networks revealed that genes at their intersection are mainly related to cytokine production, leukocyte migration and inflammatory response regulation, in accordance to the phenotype described for MHV-infection [89]. However, a differential response to virus is observed in *wt* mice compared to *Ly6E*-ablated. The nodes exclusively present at the *wt* liver network are related to processes like regulation of immune effector process, leukocyte mediated immunity or adaptive immune response. These processes, which are found at a relatively high gene ratio, are not represented by nodes exclusively present in the liver *ko* network. Additionally, genes exclusively present at the *wt* network and the intersection network are upregulated in case samples with respect to controls (Figure A6a), which suggests the activation of the previously mentioned biological processes. On the other hand, genes exclusively-present at the liver *ko* networks, mostly down-regulated, were found to be associated with catabolism.

As for the spleen networks, genotype-wise GO enrichment results revealed that the previously-mentioned intense nuclear activity involving protein-DNA complexes and nucleosome assembly is mostly due to *wt*-exclusive genes. Actually, these biological processes could be pinpointing cell replication events. Analogously to the liver case, genes that were found exclusively present in the *wt* network and the intersection network are mostly upregulated, whereas in the case of *ko*-exclusive genes the upregulation is not that extensive. Interestingly, the latter are mostly related to extracellular matrix (ECM) organization, which suggest the relevance of *Ly6E* on these. Other lymphocyte antigen-6 (LY-6) superfamily members have been related to ECM remodelling processes such as the Urokinase

receptor (*uPAR*), which participates in the proteolysis of ECM proteins [90]. However and to the best of our knowledge, the implications of *Ly6E* in ECM have not been reported.

The results presented are in the main consistent with those by Pfaender et al. [54], who observed a loss of genes associated with the type I IFN response, inflammation, antigen presentation, and B cells in infected $Ly6E^{\Delta HSC}$ mice. Genes *Stat1* and *Ifit3*, selected in their work for their high variation in absence of *Ly6e*, were identified as hub genes in the networks reconstructed from liver *wild type* and *knockout* samples, respectively. It is to be noticed that our approach significantly differs to the one carried out in the original study. In this particular case, we consider that the reconstruction of GCN enables a more comprehensive analysis of the data, potentially finding the key genes involved in the immune response onset and their relationships with other genes. For instance, the transcriptomic differences between liver and spleen upon *Ly6E* ablation become more evident using GCN.

Altogether, the presented results show the relevance of gene *Ly6E* in the immune response against the infection caused by MHV. The disruption of *Ly6E* significantly reduced the immunogenic response, affecting signaling and cell effectors. These results, combining *in vivo* and *in silico* approaches, deepen in our understanding of the immune response to viruses at the gene level, which could ultimately assist the development of new therapeutics. For example, basing on these results, prospective studies on *Ly6E* agonist therapies could be inspired, with the purpose of enhancing the gene expression level via gene delivery. Given the relevance of *Ly6E* in SARS-CoV-2 according to previous studies [54,91], the overall effects of *Ly6E* ablation in HSCs upon SARS-CoV-2 infection, putting special interest in lung tissue, might show similarities with the deficient immune response observed in the present work.

6. Conclusions

In this work we have presented an application of co-expression gene networks to analyze the global effects of *Ly6E* ablation in the immune response to MHV coronavirus infection. To do so, the progression of the MHV infection on the genetic level was evaluated in two genetic backgrounds: wild type mice (*wt*, Ly6Efl/fl) and Ly6E *knockout* mutants (*ko*, $Ly6E^{\Delta HSC}$) mice. For these, viral progression was assessed in two different organs, liver and spleen.

The proposed reconstruction rationale revealed significant differences between MHV-infected *wt* and $Ly6E^{\Delta HSC}$ mice for both organs. In addition we observed that MHV infection triggers a progressive genetic response of upregulating nature in both liver and spleen. In addition, the results suggest that the ablation of gene *Ly6E* at HSC caused an impaired genetic response in both organs compared to *wt* mice. The impact of such ablation is more evident in the liver, consistently with the disease site. At the same time, the immune response in the spleen, which seemed to be mediated by an intense chromatin activity in the normal situation, is replaced by ECM remodeling in $Ly6E^{\Delta HSC}$ mice.

We infer that the presence of *Ly6E* limits the damage in the above mentioned target sites. We believe that the characterization of these processes could motivate the efforts towards novel antiviral approaches. Finally, in the light of previous works, we hypothesize that *Ly6E* ablation might show analogous detrimental effects on immunity upon the infection caused by other viruses including SARS-CoV, MERS and SARS-CoV-2. In future works, we plan to investigate whether the over-expression of *Ly6E* in *wt* mice has an enhancement effect in immunity. In this direction, *Ly6E* gene mimicking (agonist) therapies could represent a promising approach in the development of new antivirals.

Author Contributions: Conceptualization, F.M.D.-C. and F.G.-V.; methodology, F.M.D.-C. and F.G.-V.; software, F.M.D.-C. and F.G.-V.; validation, F.M.D.-C. and F.G.-V.; Visualization, F.M.D.-C., F.G.-V., M.G.-T., F.D.; data curation, F.M.D.-C. and M.G.-T.; writing-original draft preparation, F.M.D.-C., D.S.R.-B., F.G.-V. and M.G.-T.; writing-review and editing, F.M.D.-C., F.G.-V., M.G.-T., D.S.R.-B. and F.D.; supervision, F.G.-V. and F.D.; project administration, F.G.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Pablo de Olavide University: Scholarships for Tutored Research, V Pablo de Olavide University's Research and Transfer Plan 2018-2020 (Grant No. PPI1903).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Figures and Tables

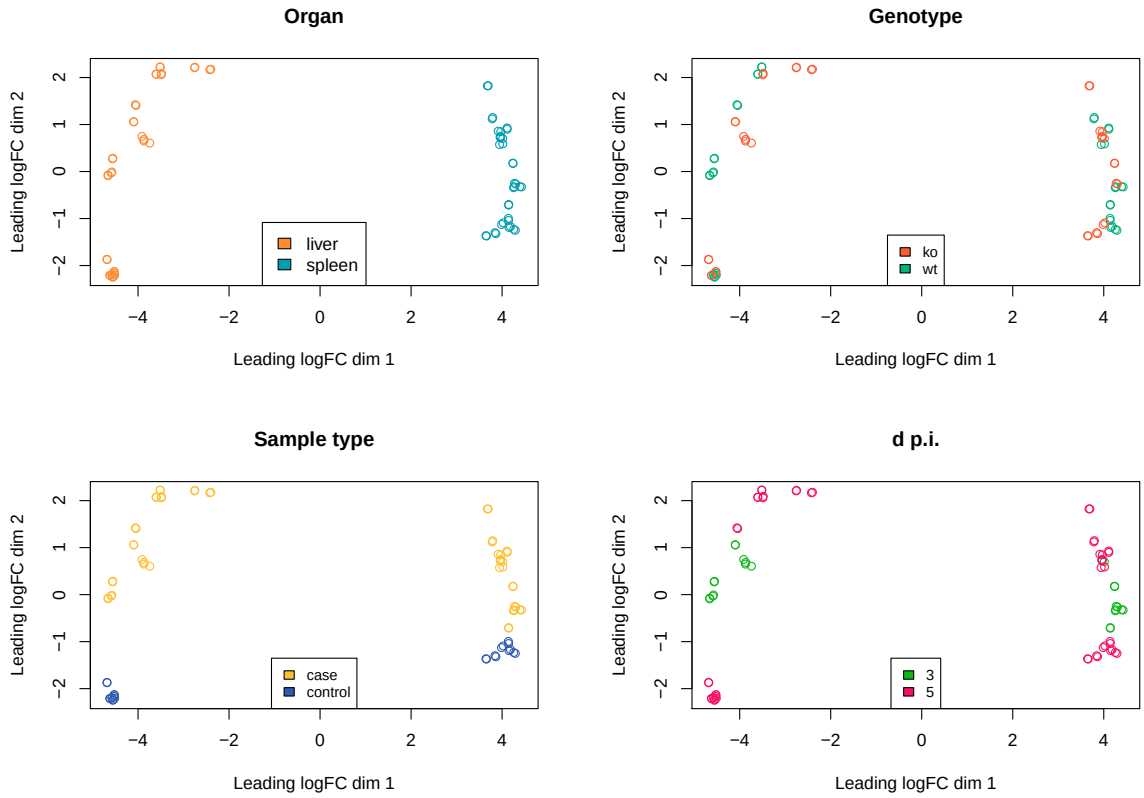
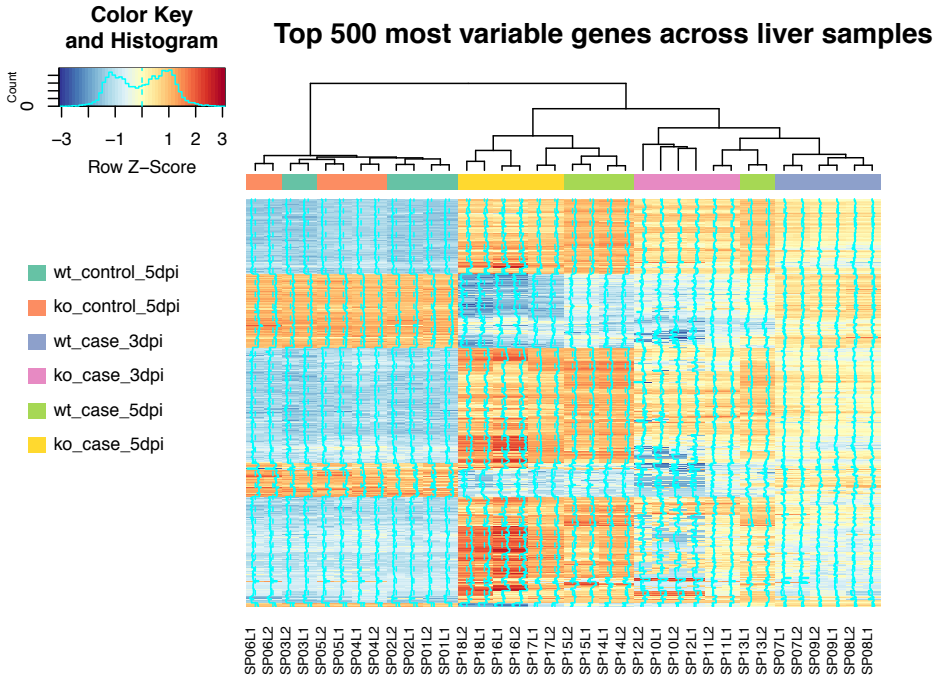
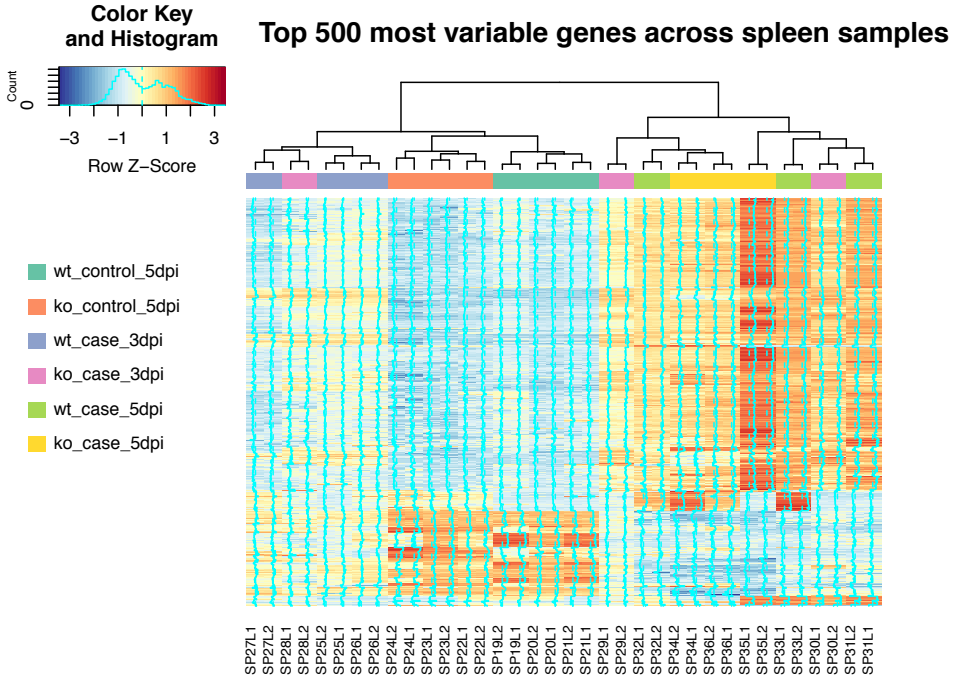


Figure A1. Multidimensional Scaling (MDS) plots showing main differences between individual samples according to the four features these present: organ procedence, genotype, sample type (mock infection or MHV-infected) and days post injection.



(**a**)



(**b**)

Figure A2. Top 500 most variable genes in (**a**) liver and (**b**) spleen samples. Log2-normalization was applied over the Counts per Million (CPMs) in order to properly compare distributions. Variance estimation reaffirms the homogenity of control vs. case samples. Overall, differences are also observed between 3 and 5 d p.i. case samples.

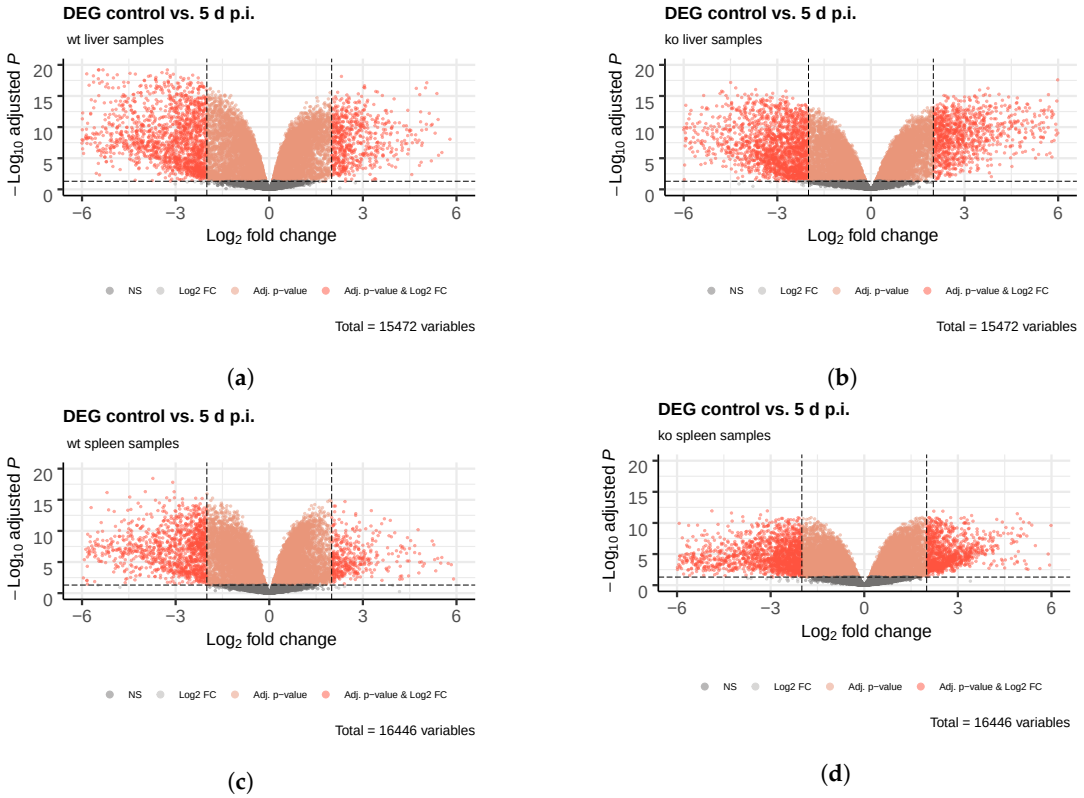


Figure A3. Volcano plots showing the differentially-expressed genes (DEG) that proceeded to the analyses. DEG were filtered by log2FC ≥ 2 and adjusted p value ≤ 0.05. These comparisons were performed both organ and genotype-wise: (**a**) liver *wt*, (**b**) liver *ko*, (**c**) spleen *wt*, (**d**) spleen *ko*. *ko*, $Ly6E^{\Delta HSC}$.

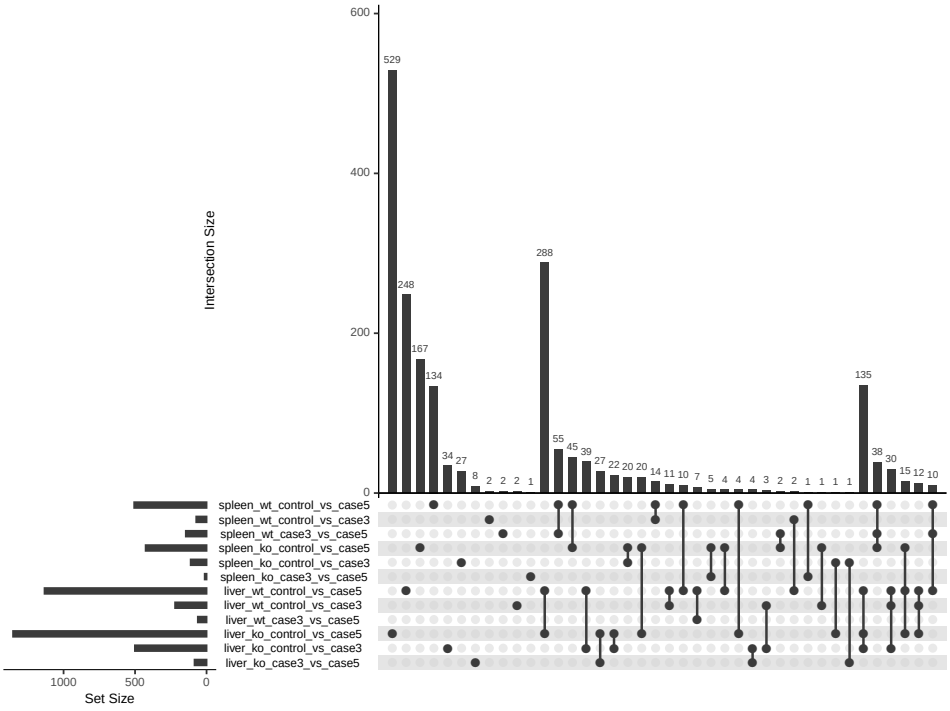


Figure A4. UpSet plot representing the commonalities between the 12 differentially-expressed genes (DEG) groups identified in differential expression analyses. The comparison of controls vs. samples at 5 d p.i. comprised the greatest number of genes for all sample types.

Table A1. Number of DEG used as input to EnGNet for network reconstruction and their latter distribution in inferred networks. Genes that were not assigned to a cluster (or were comprised in minoritary clusters) were not taken into consideration for enrichment analyses.

	Liver *wt*	**Liver *ko***	**Spleen *wt***	**Spleen *ko***
Input genes	1133	1153	506	426
Network genes	1118	1300	485	403
Cluster 1	262	284	180	109
Cluster 2	218	379	255	190
Cluster 3	579	624	36	77
Cluster 4	59			25
Unconnected/minor clustered	0	13	14	2

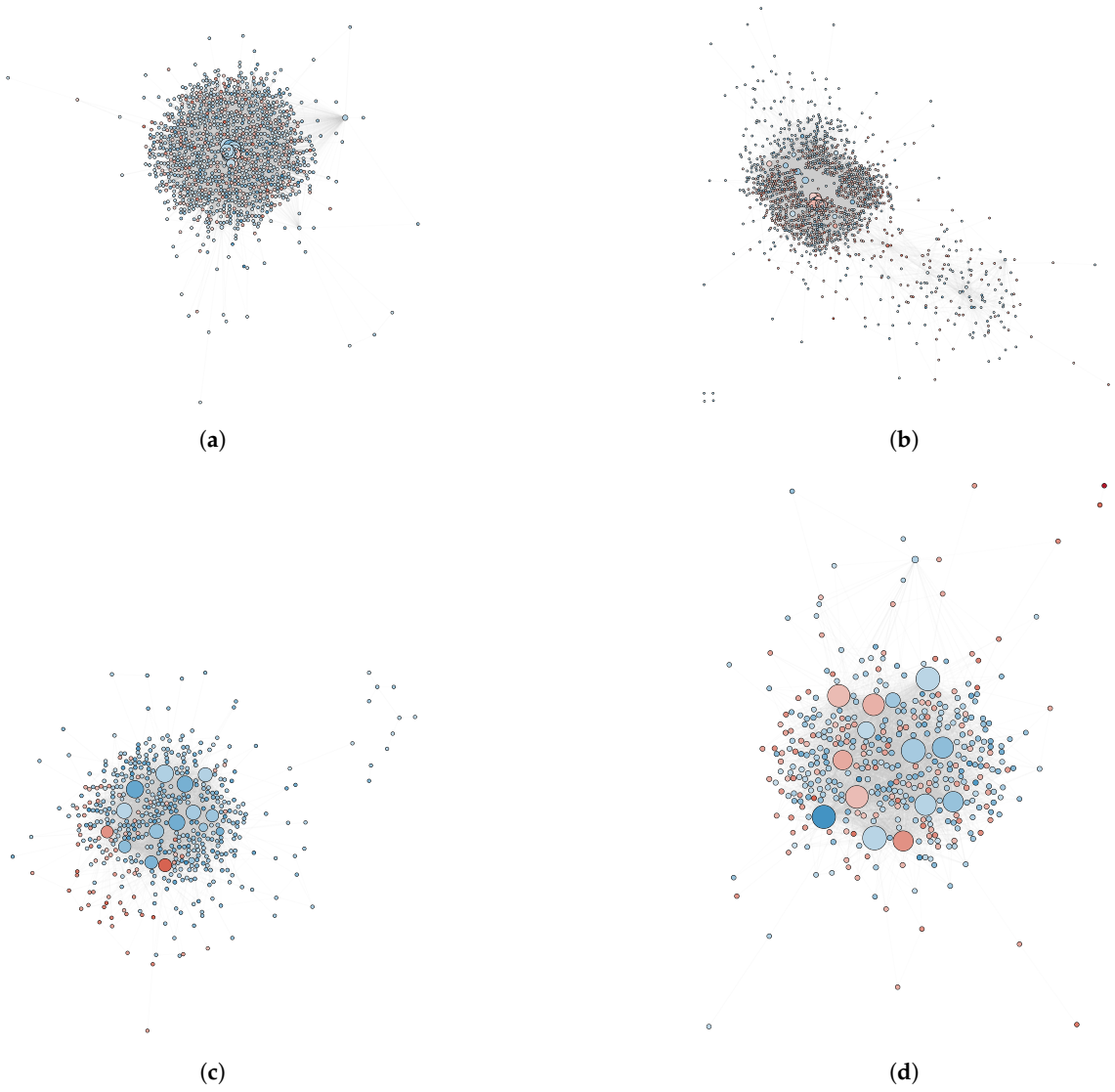
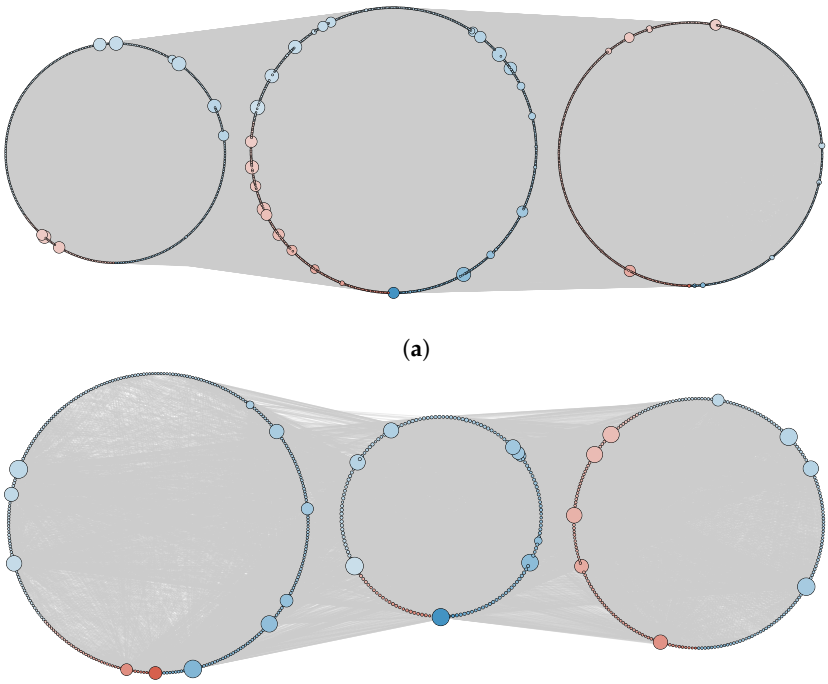


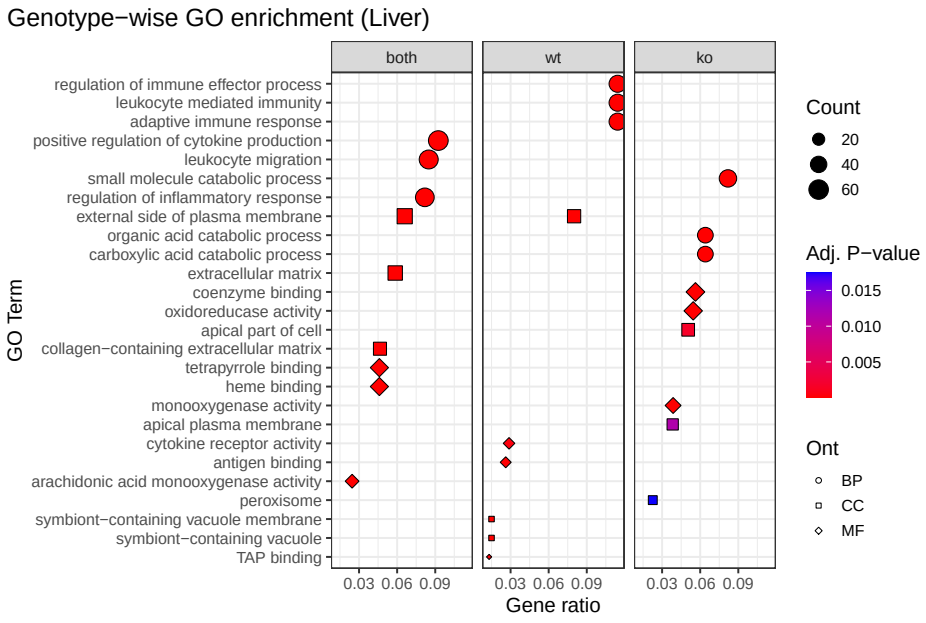
Figure A5. Inferred networks for (**a**) liver *wt* (1118 nodes, 16,281 edges, 4 clusters), (**b**) liver *ko* (1300 nodes, 15,727 edges, 3 clusters), (**c**) spleen *wt* (485 nodes, 4042 edges, 3 clusters), (**d**) spleen *ko* (403 nodes, 4220 edges, 4 clusters). Nodes are colored according to log2FC, upregulated genes in blue, downregulated genes in red. Clusters are numbered from left to right. Node size is represented according to node's degree. Edge transparency is represented according to edge weight. Networks are displayed using the yfiles organic layout [92].



(**a**)

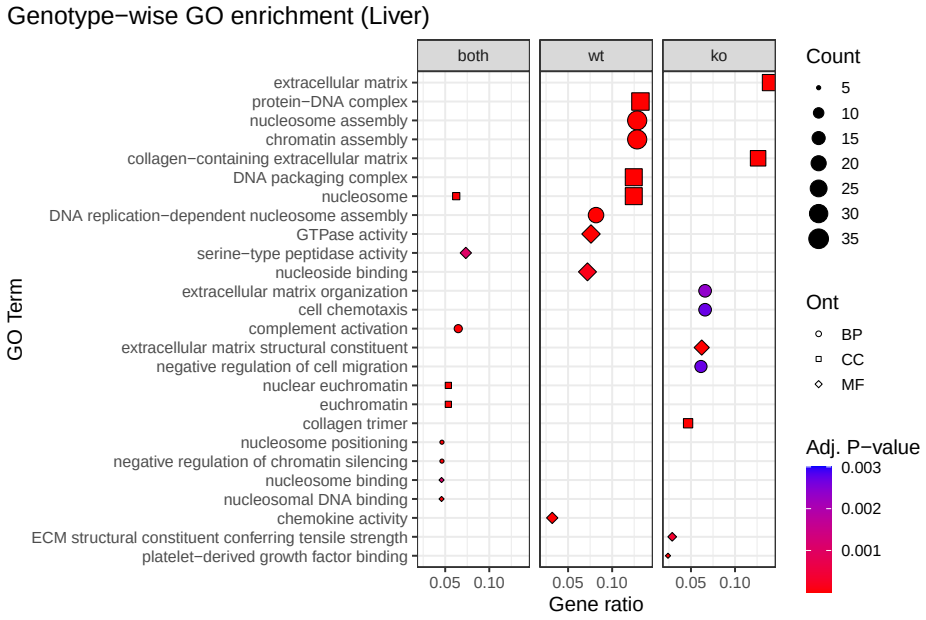
(**b**)

Figure A6. Networks resulting from the organ-wise merging of (**a**) *wt* and (**b**) *ko* samples. From left to right, nodes are displayed in circles depending on whether genes are contained exclusively at the *wt*, in the intersection between the *ko* and *wt* networks and in the *ko* network exclusively. Nodes are sorted and colored according to log2FC, upregulated genes in blue, downregulated genes in red. Node size is represented according to node's degree.



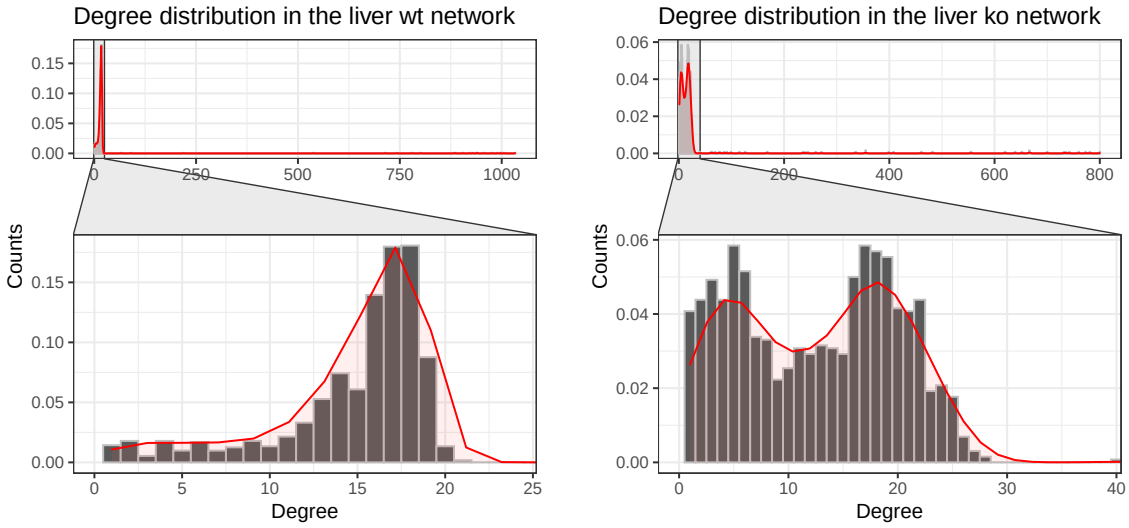
(**a**)

Figure A7. *Cont.*



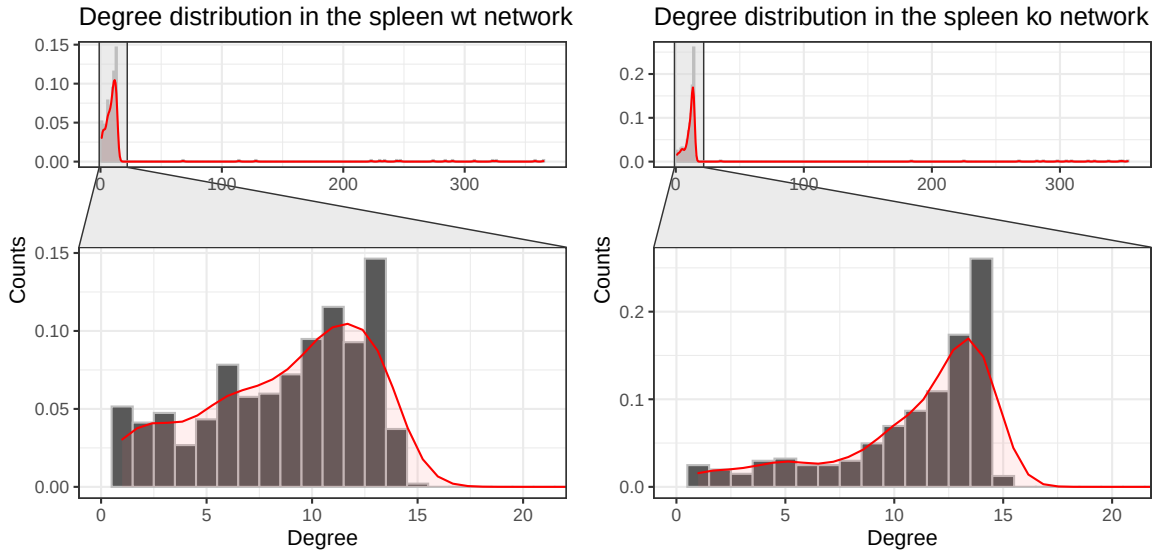
(**b**)

Figure A7. Enrichment analyses based on node exclusiveness of (**a**) liver and (**b**) spleen networks. *wt* refers to nodes exclusively present at those networks reconstructed from *wt* samples; *ko* refers to nodes exclusively present at networks reconstructed from $Ly6E^{\Delta HSC}$ samples; *both* addresses shared nodes between *wt* and *ko* networks. Gene ratio is defined by the number of genes used as input for the ernichment analyses associated with a particular GO term divided by the total number of input genes.



(**a**) (**b**)

Figure A8. *Cont*.



(**c**) (**d**)

Figure A8. Distribution of node's degree throughout the networks reconstructed from (**a**) liver *wt* samples, (**b**) liver *ko* samples, (**c**) spleen *wt* samples and (**d**) spleen *ko* samples. The distribution trendline is shown in red. Nodes that are not present in the zoomed area are considered hubs. Note degree distributions do not fit a normal distribution (Shapiro–Wilk normality test, p-value < 0.05).

Table A2. Hubs identified in the network reconstructed from liver *wt* samples. Degree cutoff: 24. Reg. regulation.

Ensembl ID	Cluster	Degree	Reg.	Symbol	Description
ENSMUSG00000034593	1	1033	up	Myo5a	myosin VA
ENSMUSG00000000982	3	1006	up	Ccl3	chemokine (C-C motif) ligand 3
ENSMUSG00000030745	2	997	up	Il21r	interleukin 21 receptor
ENSMUSG00000032322	3	989	up	Pstpip1	proline-serine-threonine phosphatase-interacting protein 1
ENSMUSG00000079227	3	975	up	Ccr5	chemokine (C-C motif) receptor 5
ENSMUSG00000031304	3	957	up	Il2rg	interleukin 2 receptor, gamma chain
ENSMUSG00000069268	3	940	up	Hist1h2bf	histone cluster 1, H2bf
ENSMUSG00000027071	1	938	down	P2rx3	purinergic receptor P2X, ligand-gated ion channel, 3
ENSMUSG00000019232	3	929	down	Etnppl	ethanolamine phosphate phospholyase
ENSMUSG00000032643	3	921	up	Fhl3	four and a half LIM domains 3
ENSMUSG00000033763	3	904	down	Mtss2	MTSS I-BAR domain containing 2
ENSMUSG00000032094	1	887	up	Cd3d	CD3 antigen, delta polypeptide
ENSMUSG00000050896	3	883	up	Rtn4rl2	reticulon 4 receptor-like 2
ENSMUSG00000067219	4	801	down	Nipal1	NIPA-like domain containing 1
ENSMUSG00000110439	3	780	down	Mup22	major urinary protein 22
ENSMUSG00000004105	2	743	down	Angptl2	angiopoietin-like 2
ENSMUSG00000081650	1	713	up	Gm16181	-
ENSMUSG00000050395	2	538	up	Tnfsf15	tumor necrosis factor (ligand) superfamily, member 15

Table A2. *Cont.*

Ensembl ID	Cluster	Degree	Reg.	Symbol	Description
ENSMUSG00000038067	1	220	up	Csf3	colony stimulating factor 3 (granulocyte)
ENSMUSG00000026104	2	90	up	Stat1	signal transducer and activator of transcription 1
ENSMUSG00000037965	2	66	up	Zc3h7a	zinc finger CCCH type containing 7 A

Table A3. Hubs identified in the network reconstructed from liver *Ly6E*$^{\Delta HSC}$ samples. Degree cutoff: 39. Reg. regulation.

Ensembl ID	Cluster	Degree	Reg.	Symbol	Description
ENSMUSG00000029445	2	800	down	Hpd	4-hydroxyphenylpyruvic acid dioxygenase
ENSMUSG00000037071	3	781	down	Scd1	stearoyl-Coenzyme A desaturase 1
ENSMUSG00000041773	3	773	up	Enc1	ectodermal-neural cortex 1
ENSMUSG00000075015	3	760	up	Gm10801	-
ENSMUSG00000021250	3	742	up	Fos	FBJ osteosarcoma oncogene
ENSMUSG00000031618	3	735	down	Nr3c2	nuclear receptor subfamily 3, group C, member 2
ENSMUSG00000022419	1	732	down	Deptor	DEP domain containing MTOR-interacting protein
ENSMUSG00000033610	3	700	down	Pank1	pantothenate kinase 1
ENSMUSG00000024349	3	667	up	Tmem173	transmembrane protein 173
ENSMUSG00000006519	3	666	up	Cyba	cytochrome b-245, alpha polypeptide
ENSMUSG00000035878	3	666	down	Hykk	hydroxylysine kinase 1
ENSMUSG00000054630	2	652	down	Ugt2b5	UDP glucuronosyltransferase 2 family, polypeptide B5
ENSMUSG00000041757	3	639	down	Plekha6	pleckstrin homology domain containing, family A member 6
ENSMUSG00000053398	3	620	up	Phgdh	3-phosphoglycerate dehydrogenase
ENSMUSG00000022025	3	555	down	Cnmd	chondromodulin
ENSMUSG00000029659	2	482	up	Medag	mesenteric estrogen dependent adipogenesis
ENSMUSG00000062380	2	461	up	Tubb3	tubulin, beta 3 class III
ENSMUSG00000069309	3	408	up	Hist1h2an	histone cluster 1, H2an
ENSMUSG00000034285	3	399	down	Nipsnap1	nipsnap homolog 1
ENSMUSG00000027654	3	355	up	Fam83d	family with sequence similarity 83, member D
ENSMUSG00000073435	2	355	down	Nme3	NME/NM23 nucleoside diphosphate kinase 3
ENSMUSG00000021062	2	336	up	Rab15	RAB15, member RAS oncogene family
ENSMUSG00000037852	3	271	up	Cpe	carboxypeptidase E
ENSMUSG00000096201	2	260	up	Gm10715	-
ENSMUSG00000022754	2	245	up	Tmem45a	transmembrane protein 45a
ENSMUSG00000038233	1	239	down	Gask1a	golgi associated kinase 1A

Table A3. *Cont.*

Ensembl ID	Cluster	Degree	Reg.	Symbol	Description
ENSMUSG00000043456	2	236	up	Zfp536	zinc finger protein 536
ENSMUSG00000095891	2	168	up	Gm10717	-
ENSMUSG00000096688	1	126	down	Mup17	major urinary protein 17
ENSMUSG00000099398	2	115	up	Ms4a14	membrane-spanning 4-domains, subfamily A, member 14
ENSMUSG00000025002	1	99	down	Cyp2c55	cytochrome P450, family 2, subfamily c, polypeptide 55
ENSMUSG00000074896	1	91	up	Ifit3	interferon-induced protein with tetratricopeptide repeats 3
ENSMUSG00000062488	1	86	up	Ifit3b	interferon-induced protein with tetratricopeptide repeats 3B
ENSMUSG00000029417	1	78	up	Cxcl9	chemokine (C-X-C motif) ligand 9
ENSMUSG00000057465	1	77	up	Saa2	serum amyloid A 2
ENSMUSG00000050908	2	69	up	Tvp23a	trans-golgi network vesicle protein 23A
ENSMUSG00000030142	1	63	up	Clec4e	C-type lectin domain family 4, member e
ENSMUSG00000038751	1	61	down	Ptk6	PTK6 protein tyrosine kinase 6
ENSMUSG00000068606	1	40	up	Gm4841	predicted gene 4841

Table A4. Hubs identified in the network reconstructed from spleen *wt* samples. Degree cutoff: 21. Reg. regulation.

Ensembl ID	Cluster	Degree	Reg.	Symbol	Description
ENSMUSG00000019505	2	365	up	Ubb	ubiquitin B
ENSMUSG00000094777	2	358	up	Hist1h2ap	histone cluster 1, H2ap
ENSMUSG00000057729	3	326	up	Prtn3	proteinase 3
ENSMUSG00000056071	1	323	up	S100a9	S100 calcium binding protein A9 (calgranulin B)
ENSMUSG00000025403	2	308	up	Shmt2	serine hydroxymethyltransferase 2 (mitochondrial)
ENSMUSG00000023132	2	290	up	Gzma	granzyme A
ENSMUSG00000078920	2	284	up	Ifi47	interferon gamma inducible protein 47
ENSMUSG00000037894	1	274	up	H2afz	H2A histone family, member Z
ENSMUSG00000035472	2	247	down	Slc25a21	solute carrier family 25 (mitochondrial oxodicarboxylate carrier), member 21
ENSMUSG00000009350	1	244	up	Mpo	myeloperoxidase
ENSMUSG00000103254	1	234	up	Ighv1-15	-
ENSMUSG00000069274	1	230	up	Hist1h4f	histone cluster 1, H4f
ENSMUSG00000028328	2	223	down	Tmod1	tropomodulin 1
ENSMUSG00000094322	1	128	up	Ighv9-4	-
ENSMUSG00000094124	1	114	up	Ighv1-74	-
ENSMUSG00000094546	1	68	up	Ighv1-26	-

Table A5. Hubs identified in the network reconstructed from spleen $Ly6E^{\Delta HSC}$ samples. Degree cutoff: 21. Reg. regulation

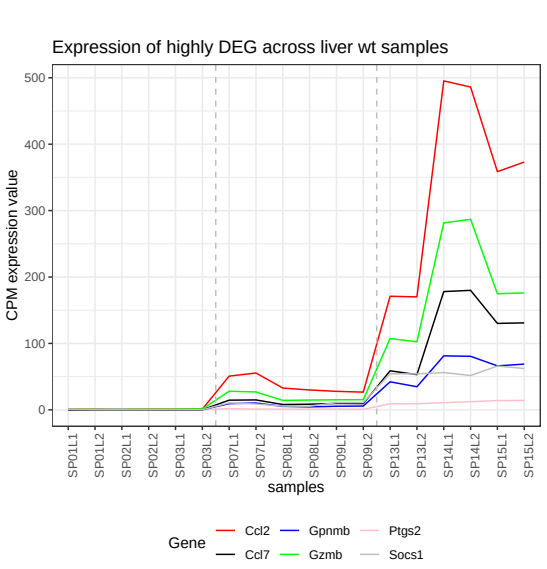
Ensembl ID	Cluster	Degree	Reg.	Symbol	Description
ENSMUSG00000027715	2	353	up	Ccna2	cyclin A2
ENSMUSG00000024742	3	349	up	Fen1	flap structure specific endonuclease 1
ENSMUSG00000024640	2	347	up	Psat1	phosphoserine aminotransferase 1
ENSMUSG00000040026	2	338	up	Saa3	serum amyloid A 3
ENSMUSG00000039713	2	327	down	Plekhg5	pleckstrin homology domain containing, family G (with RhoGef domain) member 5
ENSMUSG00000075289	4	322	down	Carns1	carnosine synthase 1
ENSMUSG00000067610	2	309	down	Klri1	killer cell lectin-like receptor family I member 1
ENSMUSG00000031503	1	305	up	Col4a2	collagen, type IV, alpha 2
ENSMUSG00000095700	3	298	up	Ighv10-3	-
ENSMUSG00000076613	3	287	up	Ighg2b	-
ENSMUSG00000051079	2	282	down	Rgs13	regulator of G-protein signaling 13
ENSMUSG00000036027	2	268	down	1810046K07Rik	RIKEN cDNA 1810046K07 gene
ENSMUSG00000027962	1	225	up	Vcam1	vascular cell adhesion molecule 1
ENSMUSG00000049130	1	184	up	C5ar1	complement component 5a receptor 1
ENSMUSG00000066861	1	35	up	Oas1g	2′-5′ oligoadenylate synthetase 1G

Table A6. Highly DEG. List of DEG that are differentially-expressed for every of the comparisons performed: control vs. cases at 3 d p.i., control vs. cases at 5 d p.i. and cases at 3 vs. 5 d p.i. Memb, membership to the group of samples genes belong; *ko*, $Ly6E^{\Delta HSC}$ samples. Reg. Type refers to the three expression patterns observed, described in Section 5.

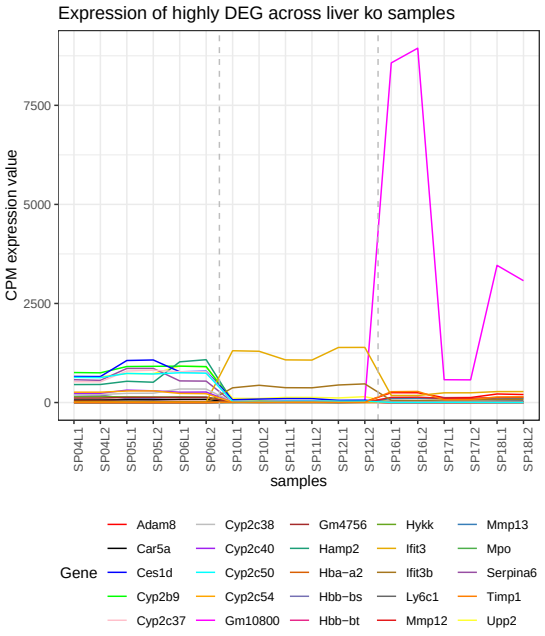
Ensembl ID	Symbol	Description	Memb.	Reg. Type
ENSMUSG00000032487	Ptgs2	prostaglandin-endoperoxide synthase 2	liver *wt*	1
ENSMUSG00000029816	Gpnmb	glycoprotein (transmembrane) nmb	liver *wt*	1
ENSMUSG00000035385	Ccl2	chemokine (C-C motif) ligand 2	liver *wt*	1
ENSMUSG00000035373	Ccl7	chemokine (C-C motif) ligand 7	liver *wt*	1
ENSMUSG00000015437	Gzmb	granzyme B	liver *wt*	1
ENSMUSG00000038037	Socs1	suppressor of cytokine signaling 1	liver *wt*	1
ENSMUSG00000026839	Upp2	uridine phosphorylase 2	liver *ko*	2
ENSMUSG00000075014	Gm10800	-	liver *ko*	1
ENSMUSG00000040660	Cyp2b9	cytochrome P450, family 2, subfamily b, polypeptide 9	liver *ko*	2
ENSMUSG00000056978	Hamp2	hepcidin antimicrobial peptide 2	liver *ko*	2
ENSMUSG00000073940	Hbb-bt	hemoglobin, beta adult t chain	liver *ko*	2
ENSMUSG00000052305	Hbb-bs	hemoglobin, beta adult major chain	liver *ko*	2
ENSMUSG00000025473	Adam8	a disintegrin and metallopeptidase domain 8	liver *ko*	1
ENSMUSG00000056973	Ces1d	carboxylesterase 1D	liver *ko*	2
ENSMUSG00000025317	Car5a	carbonic anhydrase 5a, mitochondrial	liver *ko*	2
ENSMUSG00000050578	Mmp13	matrix metallopeptidase 13	liver *ko*	1
ENSMUSG00000049723	Mmp12	matrix metallopeptidase 12	liver *ko*	1

Table A6. *Cont.*

Ensembl ID	Symbol	Description	Memb.	Reg. Type
ENSMUSG00000035878	Hykk	hydroxylysine kinase 1	liver *ko*	2
ENSMUSG00000069917	Hba-a2	hemoglobin alpha, adult chain 2	liver *ko*	2
ENSMUSG00000009350	Mpo	myeloperoxidase	liver *ko*	1
ENSMUSG00000109482	Gm4756	-	liver *ko*	2
ENSMUSG00000060807	Serpina6	serine (or cysteine) peptidase inhibitor, clade A, member 6	liver *ko*	2
ENSMUSG00000079018	Ly6c1	lymphocyte antigen 6 complex, locus C1	liver *ko*	1
ENSMUSG00000074896	Ifit3	interferon-induced protein with tetratricopeptide repeats 3	liver *ko*	3
ENSMUSG00000062488	Ifit3b	interferon-induced protein with tetratricopeptide repeats 3B	liver *ko*	3
ENSMUSG00000032808	Cyp2c38	cytochrome P450, family 2, subfamily c, polypeptide 38	liver *ko*	2
ENSMUSG00000025004	Cyp2c40	cytochrome P450, family 2, subfamily c, polypeptide 40	liver *ko*	2
ENSMUSG00000042248	Cyp2c37	cytochrome P450, family 2, subfamily c, polypeptide 37	liver *ko*	2
ENSMUSG00000067225	Cyp2c54	cytochrome P450, family 2, subfamily c, polypeptide 54	liver *ko*	2
ENSMUSG00000054827	Cyp2c50	cytochrome P450, family 2, subfamily c, polypeptide 50	liver *ko*	2
ENSMUSG00000001131	Timp1	tissue inhibitor of metalloproteinase 1	liver *ko*	1
ENSMUSG00000015437	Gzmb	granzyme B	spleen *wt*	1
ENSMUSG00000022584	Ly6c2	lymphocyte antigen 6 complex, locus C2	spleen *wt*	1
ENSMUSG00000040026	Saa3	serum amyloid A 3	spleen *ko*	1

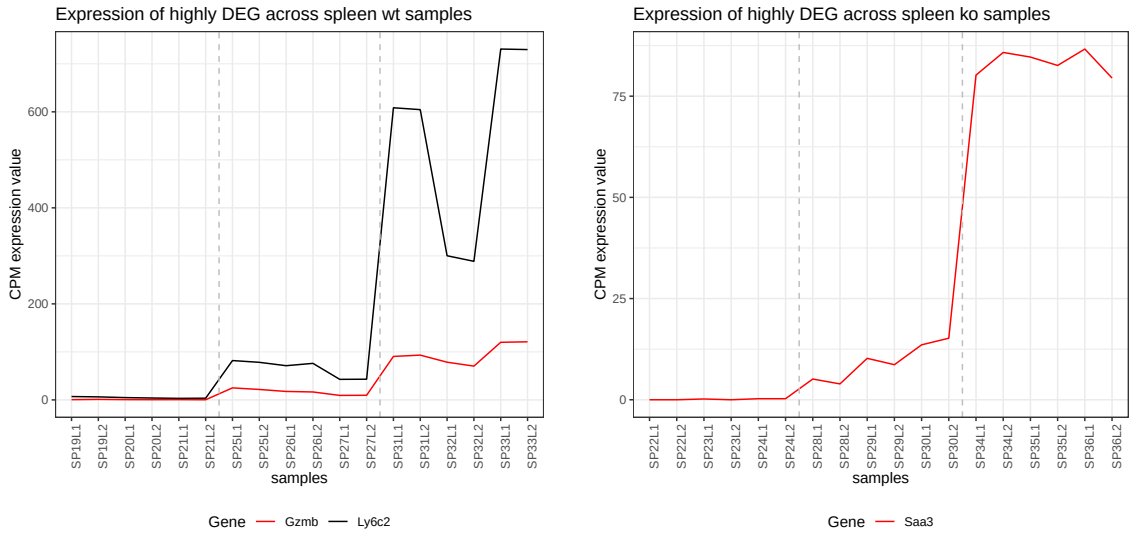


(**a**)



(**b**)

Figure A9. *Cont.*



(**c**) (**d**)

Figure A9. CPM-normalized expression values of highly DEG identified across (**a**) liver *wt* samples, (**b**) liver *ko* samples, (**c**) spleen *wt* samples and (**d**) spleen *ko* samples. Dashed lines separate samples from the three groups under study: controls, cases at 3 d p.i. and cases at 5 d p.i. Note sample order within same group is exchangeable.

Appendix B. Validation of the Reconstruction Method

The reconstruction method employed in this case study was validated against other thee well-known inference methods: *ARACNe* [93], *WGCNA* [94] and *wTO* [95]. The output of each reconstruction method, using default values (including *EnGNet*) was compared to a gold standard (GS), retrieved from the *STRING* database.

Four different GSs were taken into consideration, since these were reconstructed from the DEG that were identified in the comparison of control vs. case samples at 5 d p.i., as shown in Section 4.2. These DEG were mapped to the *STRING* database gene identifiers selecting *Mus musculus* as model organism (taxid: 10090). A variable percentage of DEG (6–20%) could not be assigned to a STRING identifier, and were thus removed from the analysis. The interactions exclusively concerning the resulting DEG in each case were retrieved from the STRING database. These interaction networks would serve as GSs. The mentioned DEG (without unmapped identifiers) would also serve as input for the four reconstruction methods to be compared.

The *ARACNe* networks were inferred using the Spearman correlation coefficient following the implementations in the *minet* [96] *R* package. In this case, mutual information values were normalized and scaled in the range 0–1. On the other hand, the *WGCNA* networks were reconstructed following the original tutorial provided by the authors [97]. The power was defined as 5. Additionally, the *wTO* networks were built using Pearson correlation in accordance to the documentation. Absolute values were taken as relationship weights. Finally, *EnGNet* networks were inferred using the default parameters described in the original article by Gómez-Vela et al. [33]. For the comparison, the Receiver operating characteristic (ROC)-curve was estimated using the *pROC* [98] *R* package. ROC curves are shown in Figure A10.

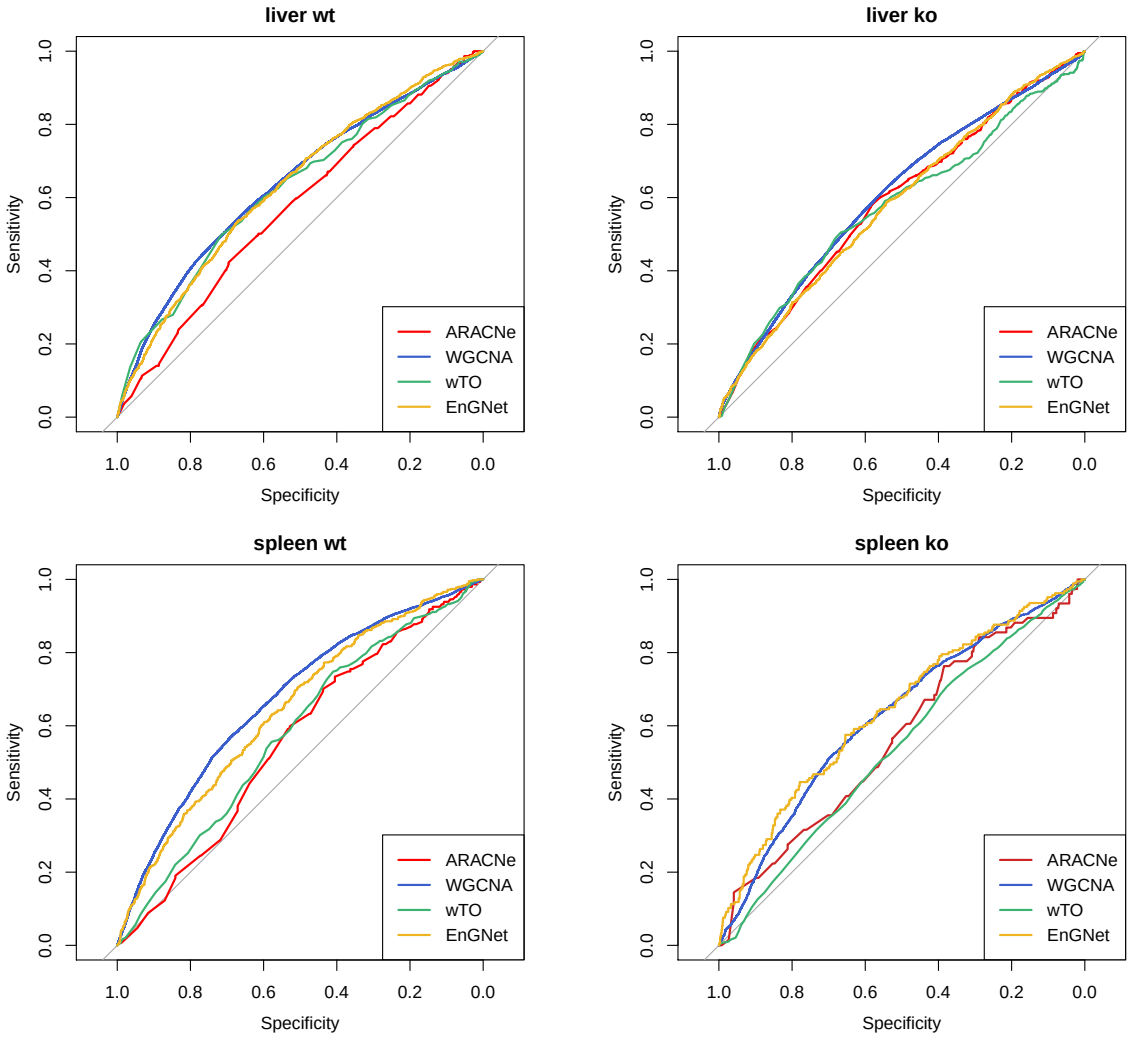
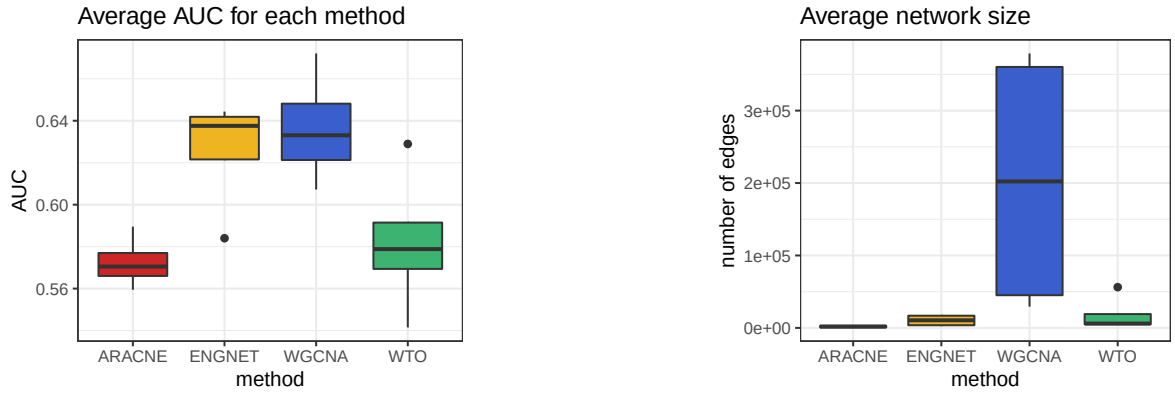


Figure A10. Receiver operating characteristic (ROC) curves for the four datasets obtained in our study using different reconstruction methods. Sensitivity is the true positive rate: $TP/(TP+FN)$. Specificity is the true negative rate: $TN/(TN+FP)$. TP, true positive; TN, true negative; FN, false negative; FP, false positive.

The area under the ROC curve (AUC) was also computed in each case for the quantitative comparison of the methods, as shown in Figure A11a. The AUC compares the reconstruction quality of each method against random prediction. An AUC ≈ 1 corresponds to the perfect classifier whereas am AUC ≈ 0.5 approximates to a random classifier. Thus, the higher the AUC, the better the predictions. On average, *EnGNet* provided the best AUC results, whilst maintaining a good discovery rate. In addition, *EnGNet* provided relatively scarce networks compared to *WGCNA*, as shown in Figure A11b. This is considered of relevance given that sparseness is a main feature of gene networks [7].



(**a**) (**b**)

Figure A11. (**a**) Comparison of the average area under the ROC curve (AUC) for the four reconstruction methods under comparison across the four used datasets. On average, EnGNet outperformed the other three methods in terms of AUC. (**b**) Size comparison of the inferred networks. EnGNet exhibited competitive results in terms of network size, providing considerably sparser networks than WGCNA's.

References

1. Corman, V.M.; Muth, D.; Niemeyer, D.; Drosten, C. Hosts and Sources of Endemic Human Coronaviruses. *Adv. Virus Res.* **2018**, *100*, 163–188. [PubMed]
2. Prentice, E.; McAuliffe, J.; Lu, X.; Subbarao, K.; Denison, M.R. Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. *J. Virol.* **2004**, *78*, 9977–9986. [CrossRef] [PubMed]
3. Sheahan, T.P.; Sims, A.C.; Zhou, S.; Graham, R.L.; Pruijssers, A.J.; Agostini, M.L.; Leist, S.R.; Schäfer, A.; Dinnon, K.H.; Stevens, L.J.; et al. An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Sci. Transl. Med.* **2020**, *12*, eabb5883. [CrossRef]
4. Voit, E. *A First Course in Systems Biology*; Garland Science: New York, NY, USA, 2017.
5. Delgado, F.M.; Gómez-Vela, F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artif. Intell. Med.* **2019**, *95*, 133–145. [CrossRef] [PubMed]
6. Gómez-Vela, F.; Lagares, J.A.; Díaz-Díaz, N. Gene network coherence based on prior knowledge using direct and indirect relationships. *Comput. Biol. Chem.* **2015**, *56*, 142–151. [CrossRef]
7. Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: data integration in dynamic models—A review. *Biosystems* **2009**, *96*, 86–103. [CrossRef] [PubMed]
8. Gómez-Vela, F.; Rodriguez-Baena, D.S.; Vázquez-Noguera, J.L. Structure Optimization for Large Gene Networks Based on Greedy Strategy. *Comput. Math. Methods Med.* **2018**, *2018*, 9674108. [CrossRef] [PubMed]
9. Zhang, Q.; Ding, Z.; Wan, L.; Tong, W.; Mao, J.; Li, L.; Hu, J.; Yang, M.; Liu, B.; Qian, X. Comprehensive analysis of the long noncoding RNA expression profile and construction of the lncRNA-mRNA co-expression network in colorectal cancer. *Cancer Biol. Ther.* **2020**, *21*, 157–169. [CrossRef]
10. Díaz-Montaña, J.J.; Gómez-Vela, F.; Díaz-Díaz, N. GNC–app: A new Cytoscape app to rate gene networks biological coherence using gene–gene indirect relationships. *Biosystems* **2018**, *166*, 61–65. [CrossRef]
11. Kumari, S.; Nie, J.; Chen, H.S.; Ma, H.; Stewart, R.; Li, X.; Lu, M.Z.; Taylor, W.M.; Wei, H. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS ONE* **2012**, *7*, e50411. [CrossRef]
12. de Siqueira Santos, S.; Takahashi, D.Y.; Nakata, A.; Fujita, A. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief. Bioinform.* **2013**, *15*, 906–918. [CrossRef]
13. Liesecke, F.; Daudu, D.; Dugé de Bernonville, R.; Besseau, S.; Clastre, M.; Courdavault, V.; de Craene, J.O.; Crèche, J.; Giglioli-Guivarc'h, N.; Glévarec, G.; et al. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci. Rep.* **2018**, *8*, 10885. [CrossRef] [PubMed]

14. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Aderhold, A.; Bonneau, R.; Chen, Y.; et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804. [CrossRef] [PubMed]
15. Song, L.; Langfelder, P.; Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinform.* **2012**, *13*, 328. [CrossRef] [PubMed]
16. Villaverde, A.F.; Ross, J.; Morán, F.; Banga, J.R. MIDER: Network inference with mutual information distance and entropy reduction. *PLoS ONE* **2014**, *9*, e96732. [CrossRef]
17. Zhang, X.; Bai, J.; Yuan, C.; Long, L.; Zheng, Z.; Wang, Q.; Chenand, F.; Zhou, Y. Bioinformatics analysis and identification of potential genes related to pathogenesis of cervical intraepithelial neoplasia. *J. Cancer* **2020**, *11*, 2150–2157. [CrossRef]
18. Sehrawat, A.; Gao, L.; Wang, Y.; Bankhead, A.; McWeeney, S.K.; King, C.J.; Schwartzman, J.; Urrutia, J.; Bisson, W.H.; Coleman, D.J.; et al. LSD1 activates a lethal prostate cancer gene network independently of its demethylase function. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E4179–E4188. [CrossRef]
19. Sandor, C.; Beer, N.L.; Webber, C. Diverse type 2 diabetes genetic risk factors functionally converge in a phenotype-focused gene network. *PLoS Comput. Biol.* **2017**, *13*, e1005816. [CrossRef]
20. Wang, L.; Huang, J.; Jiang, M.; Sun, L. Survivin (BIRC5) cell cycle computational network in human no-tumor hepatitis/cirrhosis and hepatocellular carcinoma transformation. *J. Cell. Biochem.* **2011**, *112*, 1286–1294. [CrossRef]
21. He, D.; Liu, Z.P.; Honda, M.; Kaneko, S.; Chen, L. Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. *J. Mol. Cell Biol.* **2012**, *4*, 140–152. [CrossRef]
22. Nogales, A.; Martínez-Sobrido, L. Reverse genetics approaches for the development of influenza vaccines. *Int. J. Mol. Sci.* **2017**, *18*, 20. [CrossRef] [PubMed]
23. Rajoriya, N.; Combet, C.; Zoulim, F.; Janssen, H.L. How viral genetic variants and genotypes influence disease and treatment outcome of chronic hepatitis B. Time for an individualised approach? *J. Hepatol.* **2017**, *67*, 1281–1297. [CrossRef] [PubMed]
24. Wong, H.H.; Fung, T.S.; Fang, S.; Huang, M.; Le, M.T.; Liu, D.X. Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. *Virology* **2018**, *515*, 165–175. [CrossRef] [PubMed]
25. Schneider, W.M.; Chevillotte, M.D.; Rice, C.M. Interferon-stimulated genes: a complex web of host defenses. *Annu. Rev. Immunol.* **2014**, *32*, 513–545. [CrossRef] [PubMed]
26. Luo, L.; McGarvey, P.; Madhavan, S.; Kumar, R.; Gusev, Y.; Upadhyay, G. Distinct lymphocyte antigens 6 (Ly6) family members Ly6D, Ly6E, Ly6K and Ly6H drive tumorigenesis and clinical outcome. *Oncotarget* **2016**, *7*, 11165. [CrossRef] [PubMed]
27. Yu, J.; Liu, S.L. Emerging Role of LY6E in Virus–Host Interactions. *Viruses* **2019**, *11*, 1020. [CrossRef]
28. Liu, H.C.; Niikura, M.; Fulton, J.; Cheng, H. Identification of chicken lymphocyte antigen 6 complex, locus E (LY6E, alias SCA2) as a putative Marek's disease resistance gene via a virus-host protein interaction screen. *Cytogenet. Genome Res.* **2003**, *102*, 304–308. [CrossRef]
29. Stier, M.T.; Spindler, K.R. Polymorphisms in Ly6 genes in Msq1 encoding susceptibility to mouse adenovirus type 1. *Mamm. Genome* **2012**, *23*, 250–258. [CrossRef]
30. Yu, J.; Liang, C.; Liu, S.L. Interferon-inducible LY6E protein promotes HIV-1 infection. *J. Biol. Chem.* **2017**, *292*, 4674–4685. [CrossRef]
31. Mar, K.B.; Rinkenberger, N.R.; Boys, I.N.; Eitson, J.L.; McDougal, M.B.; Richardson, R.B.; Schoggins, J.W. LY6E mediates an evolutionarily conserved enhancement of virus infection by targeting a late entry step. *Nat. Commun.* **2018**, *9*, 1–14. [CrossRef]
32. Hackett, B.A.; Cherry, S. Flavivirus internalization is regulated by a size-dependent endocytic pathway. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4246–4251. [CrossRef] [PubMed]
33. Gómez-Vela, F.; Delgado-Chaves, F.M.; Rodríguez-Baena, D.S.; García-Torres, M.; Divina, F. Ensemble and Greedy Approach for the Reconstruction of Large Gene Co-Expression Networks. *Entropy* **2019**, *21*, 1139. [CrossRef]

34. Giulietti, M.; Occhipinti, G.; Principato, G.; Piva, F. Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene co-expression network analysis. *Cell. Oncol.* **2017**, *40*, 181–192. [CrossRef]
35. Ray, S.; Hossain, S.M.M.; Khatun, L.; Mukhopadhyay, A. A comprehensive analysis on preservation patterns of gene co-expression networks during Alzheimer's disease progression. *BMC Bioinform.* **2017**, *18*, 579. [CrossRef]
36. Medina, I.R.; Lubovac-Pilav, Z. Gene co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes. *PLoS ONE* **2016**, *11*, e0156006.
37. van Dam, S.; Vosa, U.; van der Graaf, A.; Franke, L.; de Magalhaes, J.P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* **2018**, *19*, 575–592. [CrossRef] [PubMed]
38. Argilaguet Marqués, J.; Pedragosa Marín, M.; Esteve-Codina, A.; Riera Domínguez, M.G.; Vidal, E.; Peligero Cruz, C.; Casella, V.; Andreu Martínez, D.; Kaisho, T.; Bocharov, G.A.; et al. Systems analysis reveals complex biological processes during virus infection fate decisions. *Genome Res.* **2019**, *29*, 907–919. [CrossRef] [PubMed]
39. Ghobadi, M.Z.; Mozhgani, S.H.; Farzanehpour, M.; Behzadian, F. Identifying novel biomarkers of the pediatric influenza infection by weighted co-expression network analysis. *Virol. J.* **2019**, *16*, 124. [CrossRef] [PubMed]
40. Michlmayr, D.; Pak, T.R.; Rahman, A.H.; Amir, E.A.D.; Kim, E.Y.; Kim-Schulze, S.; Suprun, M.; Stewart, M.G.; Thomas, G.P.; Balmaseda, A.; et al. Comprehensive innate immune profiling of chikungunya virus infection in pediatric cases. *Mol. Syst. Biol.* **2018**, *14*, e7862. [CrossRef]
41. Pedragosa, M.; Riera, G.; Casella, V.; Esteve-Codina, A.; Steuerman, Y.; Seth, C.; Bocharov, G.; Heath, S.C.; Gat-Viks, I.; Argilaguet, J.; et al. Linking cell dynamics with gene coexpression networks to characterize key events in chronic virus infections. *Front. Immunol.* **2019**, *10*, 1002. [CrossRef] [PubMed]
42. Ray, S.; Hossain, S.M.M.; Khatun, L. Discovering preservation pattern from co-expression modules in progression of HIV-1 disease: An eigengene based approach. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21–24 September 2016; pp. 814–820.
43. McDermott, J.; Mitchell, H.; Gralinski, L.; Eisfeld, A.J.; Josset, L.; Bankhead, A., 3rd; Neumann, G.; Tilton, S.C.; Schäfer, A.; Li, C.; et al. The effect of inhibition of PP1 and TNFα signaling on pathogenesis of SARS coronavirus. *BMC Syst. Biol.* **2016**, *10*, 93. [CrossRef] [PubMed]
44. Pan, K.; Wang, Y.; Pan, P.; Xu, G.; Mo, L.; Cao, L.; Wu, C.; Shen, X. The regulatory role of microRNA-mRNA co-expression in hepatitis B virus-associated acute liver failure. *Ann. Hepatol.* **2019**, *18*, 883–892. [CrossRef]
45. Sungnak, W.; Huang, N.; Bécavin, C.; Berg, M.; Queen, R.; Litvinukova, M.; Talavera-López, C.; Maatz, H.; Reichart, D.; Sampaziotis, F.; et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* **2020**, *26*, 681–687. [CrossRef] [PubMed]
46. De Albuquerque, N.; Baig, E.; Ma, X.; Zhang, J.; He, W.; Rowe, A.; Habal, M.; Liu, M.; Shalev, I.; Downey, G.P.; et al. Murine hepatitis virus strain 1 produces a clinically relevant model of severe acute respiratory syndrome in A/J mice. *J. Virol.* **2006**, *80*, 10382–10394. [CrossRef] [PubMed]
47. Ding, Z.; Fang, L.; Yuan, S.; Zhao, L.; Wang, X.; Long, S.; Wang, M.; Wang, D.; Foda, M.F.; Xiao, S. The nucleocapsid proteins of mouse hepatitis virus and severe acute respiratory syndrome coronavirus share the same IFN-β antagonizing mechanism: attenuation of PACT-mediated RIG-I/MDA5 activation. *Oncotarget* **2017**, *8*, 49655. [CrossRef]
48. Case, J.B.; Li, Y.; Elliott, R.; Lu, X.; Graepel, K.W.; Sexton, N.R.; Smith, E.C.; Weiss, S.R.; Denison, M.R. Murine hepatitis virus nsp14 exoribonuclease activity is required for resistance to innate immunity. *J. Virol.* **2018**, *92*, e01531-17. [CrossRef]
49. Gorman, M.J.; Poddar, S.; Farzan, M.; Diamond, M.S. The interferon-stimulated gene Ifitm3 restricts West Nile virus infection and pathogenesis. *J. Virol.* **2016**, *90*, 8212–8225. [CrossRef] [PubMed]
50. Loughner, C.; Bruford, E.; McAndrews, M.; Delp, E.E.; Swamynathan, S.; Swamynathan, S.K. Organization, evolution and functions of the human and mouse Ly6/uPAR family genes. *Hum. Genom.* **2016**, *10*, 10. [CrossRef] [PubMed]
51. Mar, K.B.; Eitson, J.; Schoggins, J. Interferon-stimulated gene LY6E enhances entry of diverse RNA viruses. *J. Immunol.* **2016**, *196*, 217.7.

52. Giotis, E.S.; Robey, R.C.; Skinner, N.G.; Tomlinson, C.D.; Goodbourn, S.; Skinner, M.A. Chicken interferome: avian interferon-stimulated genes identified by microarray and RNA-seq of primary chick embryo fibroblasts treated with a chicken type I interferon (IFN-α). *Vet. Res.* **2016**, *47*, 75. [CrossRef] [PubMed]
53. Kumar, N.; Mishra, B.; Mehmood, A.; Athar, M.; Mukhtar, M.S. Integrative Network Biology Framework Elucidates Molecular Mechanisms of SARS-CoV-2 Pathogenesis. *bioRxiv* **2020**. [CrossRef]
54. Pfaender, S.; Mar, K.B.; Michailidis, E.; Kratzel, A.; Hirt, D.; V'kovski, P.; Fan, W.; Ebert, N.; Stalder, H.; Kleine-Weber, H.; et al. LY6E impairs coronavirus fusion and confers immune control of viral disease. *bioRxiv* **2020**. [CrossRef]
55. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
56. Davis, S.; Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847. [CrossRef]
57. Bullard, J.H.; Purdom, E.; Hansen, K.D.; Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform.* **2010**, *11*, 94. [CrossRef]
58. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **2015**, *12*, 115. [CrossRef]
59. Zhu, A.; Ibrahim, J.G.; Love, M.I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **2019**, *35*, 2084–2092. [CrossRef]
60. Alvarez, J.M.; Riveras, E.; Vidal, E.A.; Gras, D.E.; Contreras-López, O.; Tamayo, K.P.; Aceituno, F.; Gómez, I.; Ruffel, S.; Lejay, L.; et al. Systems approach identifies TGA 1 and TGA 4 transcription factors as important regulatory components of the nitrate response of A rabidopsis thaliana roots. *Plant J.* **2014**, *80*, 1–13. [CrossRef]
61. Delgado-Chaves, F.M.; Gómez-Vela, F.; García-Torres, M.; Divina, F.; Vázquez Noguera, J.L. Computational Inference of Gene Co-Expression Networks for the identification of Lung Carcinoma Biomarkers: An Ensemble Approach. *Genes* **2019**, *10*, 962. [CrossRef]
62. Contreras-Lopez, O.; Moyano, T.C.; Soto, D.C.; Gutiérrez, R.A. Step-by-step construction of gene co-expression networks from high-throughput arabidopsis RNA sequencing data. In *Root Development*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 275–301.
63. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
64. Law, C.W.; Chen, Y.; Shi, W.; Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **2014**, *15*, R29. [CrossRef]
65. Genovese, C.R.; Roeder, K.; Wasserman, L. False discovery control with p-value weighting. *Biometrika* **2006**, *93*, 509–524. [CrossRef]
66. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **2006**, *1695*, 1–9.
67. Smoot, M.E.; Ono, K.; Ruscheinski, J.; Wang, P.L.; Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011**, *27*, 431–432. [CrossRef] [PubMed]
68. Gustavsen, J.A.; Pai, S.; Isserlin, R.; Demchak, B.; Pico, A.R. RCy3: Network biology using Cytoscape from within R. *F1000Research* **2019**, *8*, 1774. [CrossRef] [PubMed]
69. Li, W.; Wang, M.; Sun, J.; Wang, Y.; Jiang, R. Gene co-opening network deciphers gene functional relationships. *Mol. Biosyst.* **2017**, *13*, 2428–2439. [CrossRef] [PubMed]
70. Su, G.; Kuchinsky, A.; Morris, J.H.; States, D.J.; Meng, F. GLay: Community structure analysis of biological networks. *Bioinformatics* **2010**, *26*, 3135–3137. [CrossRef] [PubMed]
71. Morris, J.H.; Apeltsin, L.; Newman, A.M.; Baumbach, J.; Wittkop, T.; Su, G.; Bader, G.D.; Ferrin, T.E. clusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinform.* **2011**, *12*, 436. [CrossRef] [PubMed]
72. Doncheva, N.T.; Assenov, Y.; Domingues, F.S.; Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **2012**, *7*, 670. [CrossRef] [PubMed]
73. Flock, T.; Hauser, A.S.; Lund, N.; Gloriam, D.E.; Balaji, S.; Babu, M.M. Selectivity determinants of GPCR–G-protein binding. *Nature* **2017**, *545*, 317–322. [CrossRef] [PubMed]

74. Dovoedo, Y.; Chakraborti, S. Boxplot-based outlier detection for the location-scale family. *Commun. Stat. Simul. Comput.* **2015**, *44*, 1492–1513. [CrossRef]
75. Yang, J.; Rahardja, S.; Fränti, P. Outlier detection: how to threshold outlier scores? In Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing, Sanya, China, 19–21 December 2019; pp. 1–6.
76. Consortium, G.O. Gene ontology consortium: going forward. *Nucleic Acids Res.* **2015**, *43*, D1049–D1056. [CrossRef] [PubMed]
77. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* **2012**, *16*, 284–287. [CrossRef] [PubMed]
78. Huang, D.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44. [CrossRef]
79. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **2016**, *45*, D362–D368. [CrossRef]
80. Gibson, S.M.; Ficklin, S.P.; Isaacson, S.; Luo, F.; Feltus, F.A.; Smith, M.C. Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PLoS ONE* **2013**, *8*, e55871. [CrossRef]
81. Milenković, T.; Pržulj, N. Uncovering biological network function via graphlet degree signatures. *Cancer Inform.* **2008**, *6*, CIN-S680. [CrossRef]
82. Baron, S.; Fons, M.; Albrecht, T. Viral pathogenesis. In *Medical Microbiology*, 4th ed.; University of Texas Medical Branch at Galveston: Galveston, TX, USA, 1996.
83. Deng, X.; Chen, Y.; Mielech, A.M.; Hackbart, M.; Kesely, K.R.; Mettelman, R.C.; O'Brien, A.; Chapman, M.E.; Mesecar, A.D.; Baker, S.C. Structure-Guided Mutagenesis Alters Deubiquitinating Activity and Attenuates Pathogenesis of a Murine Coronavirus. *J. Virol.* **2020**. [CrossRef]
84. Khan, H.A.; Ahmad, M.Z.; Khan, J.A.; Arshad, M.I. Crosstalk of liver immune cells and cell death mechanisms in different murine models of liver injury and its clinical relevance. *Hepatobiliary Pancreat. Dis. Int.* **2017**, *16*, 245–256. [CrossRef]
85. Wu, D.; Wang, H.; Yan, W.; Chen, T.; Wang, M.; Han, M.; Wu, Z.; Wang, X.; Ai, G.; Xi, D.; et al. A disparate subset of double-negative T cells contributes to the outcome of murine fulminant viral hepatitis via effector molecule fibrinogen-like protein 2. *Immunol. Res.* **2016**, *64*, 518–530. [CrossRef]
86. Lewis, S.M.; Williams, A.; Eisenbarth, S.C. Structure and function of the immune system in the spleen. *Sci. Immunol.* **2019**, *4*. [CrossRef] [PubMed]
87. Oh, J.; Kim, J.Y.; Kim, H.S.; Oh, J.C.; Cheon, Y.H.; Park, J.; Yoon, K.H.; Lee, M.S.; Youn, B.S. Progranulin and a five transmembrane domain-containing receptor-like gene are the key components in receptor activator of nuclear factor κB (RANK)-dependent formation of multinucleated osteoclasts. *J. Biol. Chem.* **2015**, *290*, 2042–2052. [CrossRef] [PubMed]
88. Dougall, W.C.; Glaccum, M.; Charrier, K.; Rohrbach, K.; Brasel, K.; De Smedt, T.; Daro, E.; Smith, J.; Tometsko, M.E.; Maliszewski, C.R.; et al. RANK is essential for osteoclast and lymph node development. *Genes Dev.* **1999**, *13*, 2412–2424. [CrossRef]
89. Frattini, P.; Villa, C.; De Santis, F.; Meregalli, M.; Belicchi, M.; Erratico, S.; Bella, P.; Raimondi, M.T.; Lu, Q.; Torrente, Y. Autologous intramuscular transplantation of engineered satellite cells induces exosome-mediated systemic expression of Fukutin-related protein and rescues disease phenotype in a murine model of limb-girdle muscular dystrophy type 2I. *Hum. Mol. Genet.* **2017**, *26*, 3682–3698. [CrossRef]
90. Desmedt, S.; Desmedt, V.; Delanghe, J.; Speeckaert, R.; Speeckaert, M. The intriguing role of soluble urokinase receptor in inflammatory diseases. *Crit. Rev. Clin. Lab. Sci.* **2017**, *54*, 117–133. [CrossRef] [PubMed]
91. Zhao, X.; Zheng, S.; Chen, D.; Zheng, M.; Li, X.; Li, G.; Lin, H.; Chang, J.; Zeng, H.; Guo, J.T. LY6E Restricts the Entry of Human Coronaviruses, including the currently pandemic SARS-CoV-2. *bioRxiv* **2020**. [CrossRef]
92. yWorks. Available online: https://www.yworks.com/ (accessed on 16 July 2020.)
93. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 7, p. S7.
94. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef]

95. Gysi, D.M.; Voigt, A.; de Miranda Fragoso, T.; Almaas, E.; Nowick, K. WTO: An R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinform.* **2018**, *19*, 392. [CrossRef]
96. Meyer, P.E.; Lafitte, F.; Bontempi, G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinform.* **2008**, *9*, 461. [CrossRef]
97. Zhang, B.; Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*. [CrossRef] [PubMed]
98. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef] [PubMed]

Part IV

Conclusions and further work

Chapter 9

Outlook

We worked on a number of initiatives while this thesis was being developed, most of which centered on systems biology approaches in the context of biomedical research. We paid particular attention to the investigation of gene co-expression networks, which were rationally reconstructed so they may serve as models for potential disease-related mechanisms. Given that we often model interactions are indirect and that their participation in the disease may be ambiguous, gene-gene correlations discovered using GCN analysis may be carefully considered. However, in the context of data-driven disease module identification, GCN reconstruction is a potent tool to let the data "speak", agnostic of presumptions that hinder the discovery of novel disease traits.

The FAIR data stewardship principles, which highlight the need for data to be Findable, Accessible, Interoperable, and Reusable, are the basis for all of our research. We make a specific focus on how the proper instruments used in the proper manner under a rational viewpoint from a knowledgeable biological point of view, may genuinely supply the scientific community with valuable insights that can then be evaluated in the lab.

9.1 GCNs allow exploring disease mechanisms at the gene expression level

When I began my PhD research, I entered a new field with unformed concepts, so we worked to devise precise unified definitions in a framework in which we would be working for the next several years. As a result, we provided a general

background for the reconstruction of gene networks in Delgado-Chaves and Gómez-Vela [16]. We realize that minor modifications in nomenclature have occurred in recent years as a result of a lack of consistency and clarification of specific notions. Nevertheless, the main idea was to investigate various methods of reconstructing gene networks, with a particular emphasis on the mathematical modeling required in each case. Additionally, as part of the data-driven research stated in Section 1.1, we believe our review's relative success may be attributed in part to the emphasis placed on the knowledge database discovery approach.

Although new methods have emerged in recent years and some other approaches for describing biological relationships were not described because they were deemed outside the scope of the review, we described the types of biological data that could be considered suitable for the reconstruction of biological networks. This may be because, as discussed in Section 2.2, the field of systems biology is growing much broader than previously thought, and definitions frequently change. As an illustration of this, tools like *RNA-Magnet*, which infers cell-cell communication networks based on the expression of certain ligands and receptors, was only made possible by the development of single-cell RNA-Seq [205]. As a tool for analysis and prediction, other strategies like to those based on graph neural networks have also gained significance in areas including node classification, edge prediction, and graph classification [206]. As a result, the ability to represent biological interactions in a graph is only constrained by the availability of data, the technological progress and the algorithmic design.

We focus on the idea of gene regulatory networks in our review, although many of the methods we covered can also be used to infer gene co-expression networks. In our studies, we assessed whether gene-gene interactions might be used to predict potential physical interactions at the protein level or simply to identify functional modules in gene expression patterns. Overlying the layer of information pertaining to gene regulatory networks, i.e. TF-target gene interactions, might restrict the finding of novel linkages since the assessed associations are limited to TF-target interactions that have already been identified or predicted. Even in the case of model organisms like *Escherichia coli* their interactome is still incomplete. Already in Rajagopala et al. [207], Y2H experiments were conducted to reconstruct a PPI landscape of *E. coli* covering nearly 70% of the proteome. Given the case of the complex Human interactome, the corresponding landscape remains poorly understood. Additionally to this, biomolecules association upon disease change in unexpected and largely unknown ways, giving rise of disease-specific relationships

[208, 209]. Such disease mechanisms should then be predicted using data, and less prior knowledge. As a result, we opted to focus on gene co-expression networks, which are agnostic of the gene regulatory interactions documented in databases.

After reviewing the various techniques and mathematical frameworks, we decided to use correlation-based network inference, sometimes in conjunction with information-based metrics, as in the case of Gómez-Vela et al. [18]. The decision to use these methods was based on the understanding that these measures are powerful, competent in identifying biological interactions, and not too computationally demanding, enabling us to evaluate huge volumes of data at the appropriate runtimes.

Already in our review, we discuss the significance of systemic perturbations in the specific task of identifying disease mechanisms. This approach established the framework for our further research into the processes underlying disease, defined as a shift between a steady-state (health) and a disturbed state (disease). As a kind of feature selection for network inference, we discussed in our review the importance of limiting the number of nodes in the network to those genes of interest. This is exactly what we do when we reconstruct our networks using differentially-expressed genes in Delgado-Chaves et al. [210] and Delgado-Chaves et al. [19], in order to model the relationships that occur between these genes, which are supposed to be responsible for disease onset when they change. This is nothing more than an attempt to retrieve a disease mechanism that is not only composed of a collection of nodes, as in simple differential expression, but also attempts to determine the connections between these nodes in the healthy state vs the disease state. As an extension of this, we plan to work in other approaches based on differential networking approaches, detailed in Section 10.2.

Hence, with our review, we covered Objectives 1 and 2. Not only do we classify the primary techniques for reconstructing GRNs and GCNs in our review, but we also include the primary accessible public databases. We came to the conclusion that, for the specific objective of revealing disease mechanisms, we may choose bulk transcriptomics as a surrogate indicator of other omics. The widespread use of transcriptome profiling techniques, which results in more data and improved reproducibility, is another reason for employing expression data. From the many methods for reconstructing gene networks, we chose correlation-based and mutual-information based approaches, which allow processing large amounts of data, as is the case with expression data, at a relatively low computational cost. Given the few TF-target gene interactions that are currently understood, we have also

consider that the selection of GCNs over GRNs is more agnostic of prior knowledge, prompting the discovery of new disease-associated features.

9.2 Ensemble methods improve the robustness of GCN inference

We rapidly saw that there were benefits and drawbacks to the various metrics for computing co-expression in light of the previously-discussed ideas. In fact, available metrics represent the relationship between genes in different ways, thus capturing various sorts of interactions.

We compared microarray expression data from lung carcinoma patients with those from healthy controls in Delgado-Chaves et al. [210]. We employed an ensemble approach with Kendall, Spearman, and Blomqvist correlation coefficients, integrating differential expression analysis with GCN inference. The rationale behind this was to reduce the number of false positives by selecting those co-expression interactions that were supported by most metrics. A threshold was used to create the final network after the results of all metrics were averaged to provide a more reliable assessment of the modeled relationships. There is a lack of agreement about how to choose a threshold for correlation networks, as we previously commented in Section 4.2.1. In this instance, a threshold was chosen with the notion that correlations may be categorized as weak, medium, or strong. We attempted to solve this issue in Gómez-Vela et al. [18], where we compared the performance of different thresholds in retrieving the interactions of known networks.

We further explored ensemble strategies in Gómez-Vela et al. [18]. Contrary to Delgado-Chaves et al. [210], where we included only metrics based on correlation, we incorporated a mutual information-based metric in *EnGNet*. The goal here was to extract key relationships while maintaining the hypothesized networks' scale-free topology, a trait of biological networks covered in Section 3.2.2. *EnGNet* was the next logical step in our ensemble techniques, together with previous work from our lab on optimizing network architecture for enhanced interpretation while maintaining meaningful linkages [211]. We used normalized mutual information to find non-linear interactions in data, which conventional algorithms often ignore despite substantial evidence for this phenomenon in observable biological systems [212]. In contrast to the technique described in Delgado-Chaves et al. [210], co-

expression measures are subjected to a major voting strategy rather than being averaged per gene-pair. As a result, when at least two of the three measures discover a link, defined as exceeding a threshold, the association is retained in the inference. Each co-expression metric's threshold was determined by comparing it to the *GeneMANIA* [196] database of interactions.

The essential feature of *EnGNet* is its topology optimization technique, which may minimizes the complexity of GCNs, making them simpler to understand by wet lab scientists or clinicians. GCNs are pruned using a modification of the minimal spanning tree algorithm in this technique. Then, using a degree threshold, we identified hubs to which pruned relationships from the previous stage are re-added if they exceed a specific threshold. We also compared our tool to four others by utilizing well-characterized datasets and their corresponding "gold standards".

In the last part of this study, we used *EnGNet* in a dataset on post-traumatic stress disorder, yielding significant associations validated by matching the implicated genes to known biological processes. Although the generalization of the algorithm parameters may be further investigated, the relevance of *EnGNet* rests in its ability to reduce network size while maintaining biologically meaningful linkages. Some of these concerns have been addressed in *EnGNet 2.0* (Becchi, Delgado-Chaves and Gómez-Vela, pending publication), which incorporates another non-linear measure of co-expression. The thresholds for the individual measures are determined in the second edition of our algorithm depending on whether the metric significantly deviates from 0. The criterion for adding new linkages is based on the observed degree distribution, with hubs chosen if their degree is an upper outlier of the node degree distribution. We also allow users to specify the number of voters for the major voting technique in this second edition, resulting in more or less tight inferences.

Therefore, we have accomplished Objectives 3 and 4. We have investigated ensemble strategies that can overcome the limitations of different correlation-based and information-based techniques alone. The idea behind this was reconstructing more reliable networks that can more precisely identify the gene co-expression patterns, as measured by multiple metrics. By doing so, *EnGNet* not only evaluates gene co-expression using three different metrics, but also, thanks to its pruning step, it offers a result that is easy to interpret. When processing vast volumes of data, the latter is an important requirement because many other methods generate massive, densely connected networks. Furthermore, in order to make it usable by

people with little to no programming skills, we offer a basic application with a straightforward interface.

9.3 Disease modules and mechanisms can be uncovered directly from data

We mainly addressed the elucidation of disease modules and mechanisms in the applications we presented in Delgado-Chaves et al. [210] and Delgado-Chaves et al. [19]. In every case, we compared expression data from disease and healthy samples, in order to retrieve which are the key players that allow distinguishing between phenotypes.

As stated in Section 3.4, during disease onset, development, and maintenance, co-expression networks reorganize, generating modifications which are thought to be responsible for disease. Therefore, in an attempt to narrow the search space to gene interactions that are relevant to specific diseases, we used differential expression analysis to identify statistically significant changes in the expression of individual genes between the healthy and disease conditions. We further examined differentially-expressed genes by using them as input to rebuild GCNs. In this manner, we were able to assess the GCN for such genes in a steady-state scenario as well as their reconfigured GCN in a disease scenario.

Differential interactions presumably cause the change in phenotype. Traditional approaches like differential expression analysis examines statistically significant variations in gene expression levels without considering the interactions between genes, which could lead to a misleading outcome. With the idea that co-expressed genes are frequently involved in the same biological process, we may interpret differentially-expressed genes in the context of a functional module by combining differential expression with gene co-expression analysis. An extension of this approach would be the use of differential networking, detailed in Section 10.2, which we intend to cover in the future. Given that in our case, we evaluate differentially-expressed genes upon disease, the modules that we obtained may well be considered disease modules, as they are predicted to play a role in disease through their associated disease mechanism.

In Delgado-Chaves et al. [210], the experimental design involved reconstructing networks using the genes that showed differential expression between samples from healthy people and lung carcinoma patients. We obtained two GCNs that

model the interactions between differentially-expressed genes in the healthy and disease states, using the corresponding sample type. After that, we overlaid the two networks to identify which interactions only appeared in the healthy scenario and which ones were unique to the lung cancer scenario. This allowed us to distinguish the connections that were disrupted as a result of the disease as well as the biological processes that were enriched in each case (disease and control).

With such strategy, we could highlight potential biomarkers such *NCKAP1L* and *DMD* based on their mutation rate in samples from adenomas, adenocarcinomas, and squamous cell neoplasms with bronchus and lung as the primary site, by contrasting our networks to the Genomic Data Commons (GDC) portal of TCGA [193]. Despite the approach's limitations, we would want to emphasize in this work the potential of ensemble techniques combined with differential expression in identifying genes meaningful to a particular disease scenario. The establishment of genes that are linked to a disease, even indirectly, may result in their inclusion in gene panels used for lung cancer patient stratification and early detection.

We integrated differential expression analysis with our method *EnGNet* in Delgado-Chaves et al. [19] to examine viral-host interactions in the setting of a murine coronavirus infection, as a model for SARS-CoV-2. Note that this article dates back to the beginning of the pandemic, when human patient samples for COVID-19 were scarce. Certainly, there are solutions developed by now, but this article explains some unique problems that were being dealt with at the time, and might still be useful to virologists. The concept remained the same: how can we explore potential disease mechanisms upon coronavirus infection as well as what the deletion of a certain gene *Ly6E* may involve. We used the principles of perturbation, which were previously discussed in Delgado-Chaves and Gómez-Vela [16], by evaluating gene co-expression changes over the infection progression for two separate organs (spleen and liver) and two different scenarios, perturbed (*LyE6E* knockout) and unperturbed. We found that our approach was able to identify known viral immune system-related genes involved in the host response, as well as novel proteins that may play a role in viral-host interactions.

In this sense, a benefit of systems biology is that it can respond quickly to community requirements, in contrast to the inherently slower classic biomedical research. Computational biology methods were able to offer the first line of defense against the COVID-19 pandemic because data and technology were easily accessible, as well as because of our limited ability to reuse data in new contexts. Because of this, the pertinent study has to be viewed in light of the pandemic's emergency.

For instance, systems biology-based projects like CoVex [213], which begins with a disease module composed of SARS-CoV-2-related proteins, was developed to find COVID-19 therapeutic candidates. Of course, vaccinations and other advancements in technology came later, so we no longer require to find potential drugs to combat the COVID-19 pandemic.

The applications described above allowed us to meet Objective 5. In our experience, when comparing samples from healthy and diseased individuals, the combination of differential expression and differential co-expression analysis enables the study of the key network components that rewire upon disease. Given that the exact mechanisms of some of the diseases we investigated are unknown, we validated our findings by examining whether the biological processes represented by our networks made sense in the context of each condition. With the knowledge we gained from these case studies and thanks to data reusability, we will be able to employ systems biology to *in silico* predict disease-associated genes and pathways in a variety of scenarios, which can later be validated in wet labs. The main purpose of our research is fill the gap between academic discoveries and clinicians, making an impact in healthcare.

9.4 Other scientific contributions

Other contributions have been made as part of research communication throughout the past several years, either in the form of academic articles or in other ways. Such important contributions are listed here in a non-exhaustive manner.

9.4.1 Articles in collaboration with other research groups

Because of the experience in biotechnology and the computer science abilities I was able to acquire during my time as a PhD student, the collaboration with life science research groups was often immensely helpful. Such collaborations are shown below:

- Delgado-Chaves, F. M., Martínez-García, P. M., Herrero-Ruiz, A., Gómez-Vela, F., Divina, F., Jimeno-González, S. and Cortés-Ledesma, F., 2022. Data of transcriptional effects of the merbarone-mediated inhibition of TOP2. *Data in Brief*, 44, p.108499.

- Santamaría-Gómez, J., Rubio, M.Á., López-Igual, R., Romero-Losada, A.B., Delgado-Chaves, F.M., Bru-Martínez, R., Romero-Campero, F.J., Herrero, A., Ibba, M., Ochoa de Alda, J.A.G. and Luque, I., 2021. Role of a cryptic tRNA gene operon in survival under translational stress. *Nucleic acids research*, 49(15), pp.8757-8776.

As this manuscript repeatedly emphasizes, the need for cross-disciplinary cooperation in contemporary research is what gave rise to these partnerships. Prof. Dr. Felipe Cortés Ledesma from the Spanish National Cancer Research Centre (CNIO) directs the Topology and DNA Breaks Group, where he continues his research about how DNA topology shapes the genome's dynamics and organization, and how an imbalance in these processes can result in the appearance of pathological DNA breaks that threaten genome stability. In particular, we assisted their analysis of the transcriptional effects of TOP2 abnormal activity, which causes DNA double-strand breaks that can seriously compromise cell survival and genome integrity.

On the other hand, I continue collaborating with Prof. Dr. Ignacio Luque Romero from the Institute of Plant Biochemistry and Photosynthesis (IBVF), part of the Spanish National Research Council (CSIC) and the University of Seville, with whom I contribute bioinformatics analysis for basic research in cyanobacteria. More specifically, I assisted in genomic analyses of transfer RNA (tRNA), with relevant findings suggesting that the tRNA gene set in some bacteria may be divided into a housekeeping subset that continuously supports translation and an inducible subset that is normally silent but can operate under specific stress conditions.

Similar to bilingualism, one of the most useful skills I have developed over the years is the capacity to effectively communicate with both life and computer scientists and to bridge the gap between their disciplines. For the upcoming years, I intend to continue expanding my computer science knowledge while keeping one foot in the field of biomedicine, which I find most exciting.

9.4.2 Conference contributions

As part of my training as a scientific communicator, I took part in and gave talks at the conferences listed below. The website of each university that is hosting the conference can provide more details.

- *RExPO22 - 1st International Conference on Drug Repurposing*, Maastricht, The Netherlands, 2^{nd} - 3^{rd} September 2022.

- *Genomics and Transcriptomics, Integrated with Proteomics and Medical Informatics: learning the cornerstones of Systems Medicine (GTIPI)*, Mainz, Germany, May 29^{th} - July 3^{rd} 2022.

RExPO22 is the first in a series of worldwide conferences on drug repurposing. The conference was picked because it directly relates to understanding disease mechanisms. Professional societies, businesses, patent attorneys, legal and ethical experts, clinical research groups, regulators, payers, patient advocates, and many more collaborated on this conference. The conference's primary subjects were redefining disease, organ-agnostic medicine, network pharmacology, artificial intelligence, regulation, ethics & safety, precision medicine, and clinical trials. The following conference publications resulted from our participation at RExPO22:

- Delgado-Chaves, F.M., Oubounyt, M., Gómez-Vela, F.A., Divina, F., Zolotareva, O.I. and Baumbach, J., 2022. Differential network-based methods for the integration of omics data: overview and challenges. *RExPO22 - 1st International Conference on Drug Repurposing*, Maastricht, The Netherlands.

- Zolotareva, O., Isaeva, O.V., Hartung, M., Maier, A., Delgado-Chaves, F., Kaufmann, K.C., Savchik, A., Chervontseva, Z.S., Probul, N., Abisheva, A. and Zotova, E., 2022. DESMOND 2.0: Identification of differentially expressed biclusters for unsupervised patient stratification. *RExPO22 - 1st International Conference on Drug Repurposing*, Maastricht, The Netherlands.

The summer school and conference "Genomics and Transcriptomics, Integrated with Proteomics and Medical Informatics: Learning the Cornerstones of Systems Medicine" (GTIPI) was developed in response to recent advances in experimental techniques in biomedical research and treatment practice, which produced high throughput datasets and a variety of methods to process, model, and interpret such massive amounts of data. These methods sparked innovations in computational fields, such as biostatistics, bioinformatics, and computational biology. For the participants to actively acquire a broad range of fundamental skills in study design, data analysis, high throughput biology (genomics, transcriptomics, and proteomics - also at single cell resolution), and medical informatics, they provided lectures on the various cornerstone elements of systems medicine, supplemented by hands-on sessions with participatory learning (involving systems and data integration as well as applications of artificial intelligence).

9.4.3 Patent

We also patented EnGNet 1.0, a software for the reconstruction of scale-free gene co-expression networks using a three-way ensemble strategy, written in the programming language JAVA and vailable for Windows and Linux. EN SE18020 - Expedition: 18 Sept. 2020. This patent is registered in the Intellectual Property Registry by the Universidad Pablo de Olavide and its file can be found at:

https://www.upo.es/upotec/catalogo/salud/engnet-10-ensemble-and-greedy-gene-networks/

EnGNet 1.0 is the first application of the *EnGNet* method for the creation of biological models with a visual interface. We may emphasize the following as the primary benefits and contributions made by the application for the development of genetic networks:

- An ensemble technique is used in the method to get beyond the restrictions of using a single co-expression measurement to create the genetic network. The final network topology may also be optimized using this way.
- The methodology's outcomes demonstrate its potential in the area of biomarker identification and characterization.
- The approach used for model development (*EnGNet*) has proved superiority over typical methods of genetic network generation found in the literature, indicating an advancement in the field of Bioinformatics and the study of biological processes and diseases.

No internet connection or browsers are needed to use *EnGNet* 1.0 because it enables a stand-alone installation. The interface design was also created with the goal of maximizing simplicity and usability. As an example, it provides a "Log" panel within the same application where the user can view the various messages that report the status of the execution and a progress bar where the user can see what portion of the task has been completed.

Chapter 10

Ongoing projects and future directions

Unavoidable questions arose as the current thesis developed, opening doors for new study directions. We outline some of the issues we encountered in the sections that follow, along with suggested solutions and action plans for addressing these challenges.

10.1 Reconstruction of networks using large datasets and HPC

Many of the proposed methods, especially those employing ensemble strategies, have disproportionately long runtimes and might greatly benefit from high-performance computing (HPC) approaches.

The dimensionality of datasets, or the number of samples and biomolecules involved, is closely related to such runtimes. In our scenario, we primarily reconstruct gene co-expression networks using a subset of genes that limits functional interactions to those that may be connected to a particular phenotype or sample difference. However, including the co-expression of all genes, or even more so, of all gene isoforms or transcripts, would greatly expand the complexity of the data.

Due to his expertise in multi-GPU implementations for algorithm acceleration employing massive amounts of data, Dr. Aurelio López Fernandez, who received his PhD in Computer Science from Pablo de Olavide University, is a collaborator of ours. We are actively developing a CUDA implementation of *EnGNet* [18]. Compute Unified Device Architecture, or CUDA, refers to a parallel computing platform developed by *Nvidia* that enables programmers to write algorithms in a variant of

the C programming language called CUDA C for its parallel use on *Nvidia* graphic processing units.

Our initial findings indicate that CUDA-*EnGNet* performs better with large datasets, which frequently require unreasonable processing times or cannot be handled using conventional methods. This suggests that *EnGNet* may be applied to more difficult tasks like isoform-isoform networks or networks incorporating multiple omic layers.

10.2 *De novo* elucidation of disease mechanisms using differential networking approaches

GCNs may contain gene modules whose co-expressions vary significantly between conditions, which might represent biologically interesting findings that cannot be found by co-expression analysis alone. For example, in biomarker identification, genes that change their co-expression values with a significant number of neighbors, and between conditions, are more likely to be interesting [214].

It is feasible to identify the genes responsible for a broad variety of phenotypes in the case of GCNs, thanks to new methodologies like differential co-expression analysis (DCA) [215]. These networks comprise gene modules with co-expressions that are highly condition-dependent, revealing physiologically intriguing findings that cannot be obtained via co-expression analysis alone, as shown in Figure 10.1. For example, in biomarker identification, genes that change their co-expression values with a significant number of neighbors, and between conditions, are more likely to be relevant [214].

DCA is a growingly popular technique that goes beyond conventional co-expression networks [147]. With the understanding that genes are likely to be the regulators underlying phenotypic variations, DCA mehtods discover a subnetwork whose co-expression edges vary under multiple scenarios, such as disease states, tissue types, and developmental stages. For instance, disease-related genes usually have tissue-specific impairment, and both gene expression and regulation can be very tissue-specific. In this context, DCA, which can distinguish between common and tissue-specific co-expression characteristics, can be used to retrieve the interactions that are disease-specific taking into account tissue specificity.

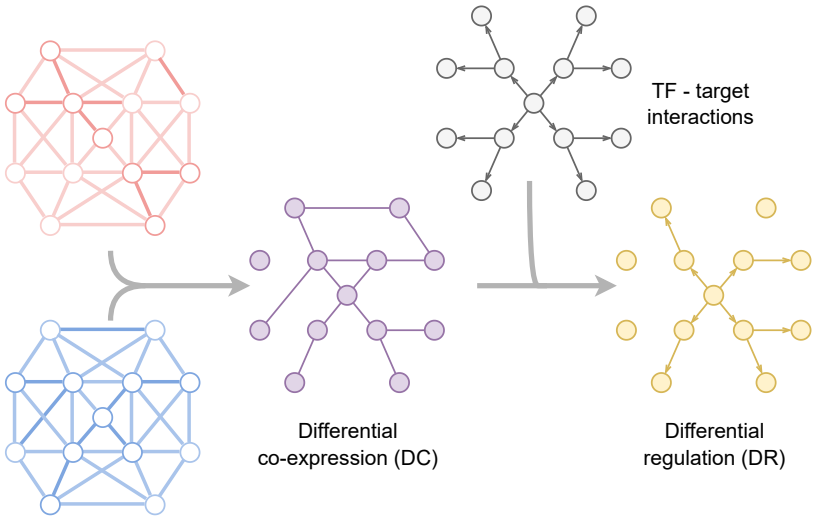


Fig. 10.1 Differential networking methodologies for uncovering putative disease mechanisms by contrasting networks formed from healthy and disease samples in blue and red, respectively. The appropriate differential gene regulatory network may be created by taking into account prior knowledge, such as known interactions between TFs and their targets.

10.3 Network-based integration of multi-omics data for comprehensive understanding of disease modules

Already in Delgado-Chaves and Gómez-Vela [16], we discussed the interest of integrating biological information of diverse nature upon the network inference process. Because of the complexity inherent to biology, the focus of biomedical research has shifted from single-omics analysis to the concurrent utilization of multi-omics datasets in order to obtain a holistic understanding of disease processes. Studies that address the integration of data from different omics are particularly relevant for studying complex diseases, which present alterations in several omics layers [216].

In the last decades, multiple methods for multi-omics integration arose together with the development of high-throughput technologies, According to Ebbels and Cavill [217], there are three main levels of data integration: (i) conceptual, in which omics are analyzed separately and resulting conclusions are matched; (ii) statistical, which tries finding associations between elements from different omics; and (iii) model-based, which apply computational or mathematical model mostly with predictive purposes. In this context, whereas the data integration level is considered higher on model-based approaches, conceptual and statistical integration methods provide easier to interpret results, especially for life scientists and clinicians [218].

Statistical methods have quickly become the most popular approach to multi-omics integration, Because they allow easy formulation of hypotheses [219, 59].

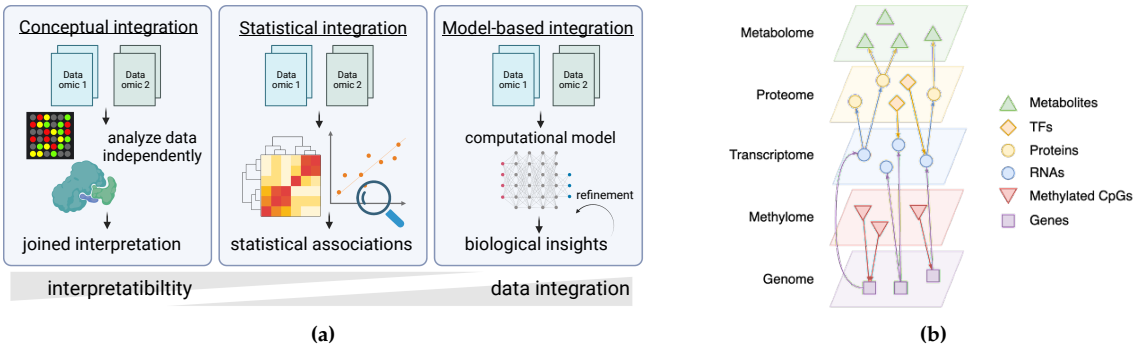


Fig. 10.2 (a) General classification of multi-omics integration approaches. The development of new technology and the availability of data has accelerated the transition of such techniques from reductionist conceptual approaches to sophisticated model-based methodologies (integrative). (b) Illustration of composite network with associations within and across multiple omic layers.

Within the framework of statistical models, network-based multi-omics integration approaches connect phenotypes to biological mechanisms from different omics layers, driven by data and/or prior knowledge (Figure 10.2b). Such approaches identify statistically significant associations between biomolecules in and across multiple omics, reconstructing an interconnected network, which is mined to get biological insights. BN inference has traditionally been performed on data sets corresponding to a single omics, which are usually homogeneous, rather than on complex data sets involving several omics sources, which are more heterogeneous [220]. However, two remaining important challenges in the analysis of biological information are (i) multi-omic data integration strategy and (ii) the detection of patterns across large-scale molecular networks.

Specifically in cancer research, composite BNs promise to identify new cancer-associated biomarkers, due to the emergence of emergent properties inherent to the connections between omics layers [221]. Cancer, now understood as a complex set of diseases, has attracted the interest of the scientific community, which has sought to address it from multiple omics approaches [222]. Due to its high relative incidence in the world population, one of the most studied cancer types is breast cancer [223]. These efforts have generated a large amount of publicly accessible multi-omics data, the volume of which varies according to the type of cancer. The data available for the different approaches that have been used to study this disease are a valuable source of information for their use in integration and subsequently a promising starting point [224]. Although still in their infancy, network-based

approaches for integrating a variety of omics data will provide new insights into the molecular mechanisms underlying the pathophysiology of diseases and can be used to support the diagnosis, prognosis and drug discovery, using many of the techniques developed for network biology [225].

10.4 Precision medicine and drug repurposing

Despite the investment over the past two decades in biomedical and pharmaceutical research and development, the annual number of new medicines authorized by the US Food and Drug Administration has remained mostly unchanged. Although there are numerous reasons that affect this low approval rate, one crucial and sometimes ignored one is the ongoing adherence to the conventional "one gene, one drug, one disease" concept of drug research. However, it can be argued that because drug targets do not work isolated from the complex interactome that compose the molecular machinery of the cell, each drug-target interaction must be investigated in an appropriate integrative context [84].

The process of drug repurposing involves the identification of new therapeutic uses for existing drugs. This can be accomplished through the study of a drug's mechanisms of action, side effects, and structure. Prof. Dr. Harald H.H.W. Schmidt from the University of Maastricht makes a compelling argument that both drug discovery and repurposing share the same research problem, which is caused by the fact that diseases are not mechanistically understood and are only treated symptomatically. For this reason, only this year has a new European initiative started, REPO4EU brings together such a team of collaborators in modern drug repurposing, which is becoming a cost-effective way to develop new treatments as it eliminates the need for expensive and time-consuming drug development [14].

By shedding light on the link between drugs and diseases, novel techniques, including network-based drug-disease proximity, might be effective tools for quick screening of possible new indications for authorized treatments or for previously unrecognized side events, an approach termed drug repurposing. Network-based drug repurposing usually takes disease modules or disease mechanisms as a starting point [122]. The premise behind network-based drug repurposing is that (i) drug targets for one disease may also be effective targets for another disease due to shared interactions and functional pathways revealed by the interactome, and (ii) drug targets for one disease may be localized in the corresponding disease module or sub-network within the overall interactome network.

Upon the identification of new drugs, the main idea is that if a drug interacts with a target, additional drugs that are similar to the first one are likely to interact with the same target. Chemical characteristics, pharmacological side effects, or treatment commonalities can all serve as indicators of similarity [84]. In a similar spirit, we can identify potential new targets for a certain drug based on target similarity, which can be estimated by evolutionary distance or target's functional similarity. Both approaches are represented in Figure 10.3a.

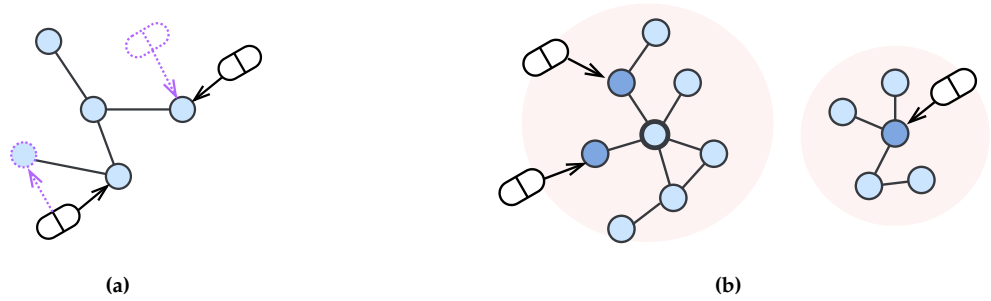


Fig. 10.3 (a) Prediction of a drug based on resemblance to a prescription drug and prediction of a target based on suspected target similarity. Predicted interactions are shown in purple. (b) A combination therapy approach in which specific targets (dark blue) in disease modules (red circles) are targeted by multiple drugs. A potential biomarker based on centrality metrics is highlighted in the left module.

With this idea, network-based drug-disease proximity approaches can be used for the rational design of combination therapies (Figure 10.3b). Due to the interdependencies of cellular and molecular effector components in biological systems, it is sometimes claimed that a "magic bullet" medicine used in monotherapy has effects that are far beyond its molecular targets [84, 122]. For this reason, combination therapy, which employ combinations of drugs to target different biological pathways, appear to be more successful than monotherapy. Because the dose of each active component when combined is lower than the dosage of each drug when used alone, combination therapy can help reduce the side effects of monotherapies. Thus, finding drug combinations that unite high effectiveness with low toxicity and could synergistically target disease mechanisms while reducing side effects is a hot topic in drug development research.

References

[1] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, 2018.

[2] Julien Venne, Ulrike Busshoff, Sebastian Poschadel, Robin Menschel, Nikolaos Evangelatos, Kranthi Vysyaraju, and Angela Brand. International consortium for personalized medicine: an international survey about the future of personalized medicine. *Personalized Medicine*, 17(2):89–100, 2020.

[3] Jennifer E Rood and Aviv Regev. The legacy of the human genome project. *Science*, 373(6562):1442–1443, 2021.

[4] Jeantine E Lunshof, Jason Bobe, John Aach, Misha Angrist, Joseph V Thakuria, Daniel B Vorhaus, Margret R Hoehe, and George M Church. Personal genomes in progress: from the human genome project to the personal genome project. *Dialogues in clinical neuroscience*, 2022.

[5] Ivo D Dinov. Volume and value of big healthcare data. *Journal of medical statistics and informatics*, 4, 2016.

[6] Davide Cirillo and Alfonso Valencia. Big data analytics for personalized medicine. *Current opinion in biotechnology*, 58:161–167, 2019.

[7] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[8] Timothy D Veenstra. Omics in systems biology: current progress and future outlook. *Proteomics*, 21(3-4):2000235, 2021.

[9] Fulvio Mazzocchi. Could big data be the end of theory in science? a few remarks on the epistemology of data-driven science. *EMBO reports*, 16(10):1250–1255, 2015.

[10] Disease and systems biology. URL https://www.omicscouts.com/en/disease-and-systems-biology.html.

[11] Harald H. H. W. Schmidt. *The End of Medicine As We Know It - and Why Your Health Has a Future*. Springer, May 2022. ISBN 978-3-030-95292-1. doi: 10.1007/978-3-030-95293-8.

[12] Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299–310, 2018.

References

[13] David Adams, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, Bernhard Boehm, Elias Campo, Andrea Caricasole, et al. Blueprint to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3):224–226, 2012.

[14] Harald HHW Schmidt. Repo4eu: The end of drug discovery as we know it. 2022.

[15] Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263, 2019.

[16] Fernando M Delgado-Chaves and Francisco Gómez-Vela. Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artificial intelligence in medicine*, 95:133–145, 2019.

[17] Fernando M Delgado-Chaves, Francisco Gómez-Vela, Miguel García-Torres, Federico Divina, and José Luis Vázquez Noguera. Computational inference of gene co-expression networks for the identification of lung carcinoma biomarkers: An ensemble approach. *Genes*, 10(12):962, 2019.

[18] Francisco Gómez-Vela, Fernando M Delgado-Chaves, Domingo S Rodríguez-Baena, Miguel García-Torres, and Federico Divina. Ensemble and greedy approach for the reconstruction of large gene co-expression networks. *Entropy*, 21(12):1139, 2019.

[19] Fernando M Delgado-Chaves, Francisco Gómez-Vela, Federico Divina, Miguel García-Torres, and Domingo S Rodriguez-Baena. Computational analysis of the global effects of ly6e in the immune response to coronavirus infection using gene networks. *Genes*, 11(7):831, 2020.

[20] Jana Zecha, Chien-Yun Lee, Florian P Bayer, Chen Meng, Vincent Grass, Johannes Zerweck, Karsten Schnatbaum, Thomas Michler, Andreas Pichlmair, Christina Ludwig, et al. Data, reagents, assays and merits of proteomics for sars-cov-2 research and testing. *Molecular & Cellular Proteomics*, 2020.

[21] Mohammed Uddin, Farah Mustafa, Tahir A Rizvi, Tom Loney, Hanan Al Suwaidi, Ahmed H Hassan Al-Marzouqi, Afaf Kamal Eldin, Nabeel Alsabeeha, Thomas E Adrian, Cesare Stefanini, et al. Sars-cov-2/covid-19: Viral genomics, epidemiology, vaccines, and therapeutic interventions. *Viruses*, 12(5):526, 2020.

[22] Iman Tavassoly, Joseph Goldfarb, and Ravi Iyengar. Systems biology primer: the basic methods and approaches. *Essays in biochemistry*, 62(4):487–500, 2018.

[23] Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. Big data analytics for preventive medicine. *Neural Computing and Applications*, 32(9):4417–4451, 2020.

[24] B Chen and AJ4785018 Butte. Leveraging big data to transform target selection and drug discovery. *Clinical Pharmacology & Therapeutics*, 99(3):285–297, 2016.

[25] Ben Lehner. Modelling genotype–phenotype relationships and human disease with genetic interaction networks. *Journal of Experimental Biology*, 210(9):1559–1566, 2007.

[26] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[27] Olena Morozova and Marco A Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, 2008.

[28] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1): D36–D42, 2012.

[29] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[30] Manuel Salto-Tellez and Ian A Cree. Cancer taxonomy: pathology beyond pathology. *European Journal of Cancer*, 115:57–60, 2019.

[31] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[32] Ulrike C Lange and Robert Schneider. What an epigenome remembers. *Bioessays*, 32 (8):659–668, 2010.

[33] Sergey Kurdyukov and Martyn Bullock. Dna methylation analysis: choosing the right method. *Biology*, 5(1):3, 2016.

[34] Michel Neidhart. *DNA methylation and complex human disease*. Academic Press, 2015.

[35] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[36] Parampreet Kaur, Ashima Singh, and Inderveer Chana. Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. *Archives of Computational Methods in Engineering*, 28(7):4595–4631, 2021.

[37] Rajeshwar Govindarajan, Jeyapradha Duraiyan, Karunakaran Kaliyappan, and Murugesan Palanisamy. Microarray and its applications. *Journal of pharmacy & bioallied sciences*, 4(Suppl 2):S310, 2012.

[38] Ruairi J Mackenzie. Rna-seq: Basics, applications and protocol. *Retrieved from Technology Networks: https://www. technologynetworks. com/genomics/articles/rna-seqbasics-applications-and-protocol-299461*, 2018.

[39] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1): 207–210, 2002.

[40] Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, et al. Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(suppl_1): D868–D872, 2009.

[41] Bruce Alberts. Molecular biology of the cell 5e. *Garland science*, pages 906–911, 2008.

[42] Penghao Wang and Susan R Wilson. Mass spectrometry-based protein identification by integrating de novo sequencing with database searching. In *BMC bioinformatics*, volume 14, pages 1–9. BioMed Central, 2013.

[43] Lloyd M Smith, Jeffrey N Agar, Julia Chamot-Rooke, Paul O Danis, Ying Ge, Joseph A Loo, Ljiljana Paša-Tolić, Yury O Tsybin, Neil L Kelleher, and Consortium for Top-Down Proteomics. The human proteoform project: defining the human proteome. *Science advances*, 7(46):eabk0734, 2021.

[44] Gustav N Sundell and Ylva Ivarsson. Interaction analysis through proteomic phage display. *BioMed research international*, 2014, 2014.

[45] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[46] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. Proteomics: technologies and their applications. *Journal of chromatographic science*, 55(2):182–196, 2017.

[47] Lennart Martens, Henning Hermjakob, Philip Jones, Marcin Adamski, Chris Taylor, David States, Kris Gevaert, Joël Vandekerckhove, and Rolf Apweiler. Pride: the proteomics identifications database. *Proteomics*, 5(13):3537–3545, 2005.

[48] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. Proteomexchange provides globally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3):223–226, 2014.

[49] Patroklos Samaras, Tobias Schmidt, Martin Frejno, Siegfried Gessulat, Maria Reinecke, Anna Jarzab, Jana Zecha, Julia Mergner, Piero Giansanti, Hans-Christian Ehrlich, et al. Proteomicsdb: a multi-omics and multi-organism resource for life science research. *Nucleic acids research*, 48(D1):D1153–D1163, 2020.

[50] David S Wishart. Current progress in computational metabolomics. *Briefings in bioinformatics*, 8(5):279–293, 2007.

[51] Kathleen A Lee-Sarwar, Jessica Lasky-Su, Rachel S Kelly, Augusto A Litonjua, and Scott T Weiss. Metabolome–microbiome crosstalk and human disease. *Metabolites*, 10 (5):181, 2020.

[52] Ute Roessner and Jairus Bowne. What is metabolomics all about? *Biotechniques*, 46(5): 363–365, 2009.

[53] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, Paula De Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, et al. Metabolights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(D1):D781–D786, 2013.

[54] David S Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, Daniel Johnson, Carin Li, Naama Karu, et al. Hmdb 4.0: the human metabolome database for 2018. *Nucleic acids research*, 46(D1): D608–D617, 2018.

[55] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

[56] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462, 2016.

[57] Upinder S Bhalla and Ravi Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.

[58] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[59] Srdjan Kesić. Systems biology, emergence and antireductionism. *Saudi journal of biological sciences*, 23(5):584–591, 2016.

[60] Trey Ideker. Systems biology 101—what you need to know. *Nature biotechnology*, 22 (4):473–475, 2004.

[61] Eddie Cano-Gamez and Gosia Trynka. From gwas to function: using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in genetics*, 11:424, 2020.

[62] So Yeon Kim, Hyun-Hwan Jeong, Jaesik Kim, Jeong-Hyeon Moon, and Kyung-Ah Sohn. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biology direct*, 14(1):1–13, 2019.

[63] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[64] WenJun Zhang. *Fundamentals of Network Biology*. World Scientific, 2018.

[65] Sravan Kumar Miryala, Anand Anbarasu, and Sudha Ramaiah. Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. *Gene*, 642:84–94, 2018.

[66] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Untangling statistical and biological models to understand network inference: the need for a genomics network ontology. *Frontiers in genetics*, 5:299, 2014.

[67] Frank Emmert-Streib and Matthias Dehmer. Networks for systems biology: conceptual connection of data and function. *IET systems biology*, 5(3):185–207, 2011.

[68] Medi Kori and Kazim Yalcin Arga. Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective. *PloS one*, 13(7), 2018.

[69] Marie Pier Scott-Boyer, Sébastien Lacroix, Marco Scotti, Melissa J Morine, Jim Kaput, and Corrado Priami. A network analysis of cofactor-protein interactions for analyzing associations between human nutrition and diseases. *Scientific reports*, 6(1):1–11, 2016.

[70] Robert J Schaefer, Jean-Michel Michno, and Chad L Myers. Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1860(1):53–63, 2017.

[71] Jingwen Yan, Shannon L Risacher, Li Shen, and Andrew J Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, 19(6):1370–1381, 2018.

[72] Elise AR Serin, Harm Nijveen, Henk WM Hilhorst, and Wilco Ligterink. Learning from co-expression networks: possibilities and challenges. *Frontiers in plant science*, 7: 444, 2016.

[73] Stephen K Burley, Helen M Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Ken Dalenberg, Jose M Duarte, Shuchismita Dutta, et al. Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research*, 47(D1):D464–D474, 2019.

[74] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

[75] Renu Goel, HC Harsha, Akhilesh Pandey, and TS Keshava Prasad. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Molecular BioSystems*, 8(2):453–463, 2012.

[76] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

[77] Jianfei Hu, Hee-Sool Rho, Robert H Newman, Jin Zhang, Heng Zhu, and Jiang Qian. Phosphonetworks: a database for human phosphorylation networks. *Bioinformatics*, 30(1):141–142, 2014.

[78] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[79] Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020.

[80] Mikaela Koutrouli, Evangelos Karatzas, David Paez-Espino, and Georgios A Pavlopoulos. A guide to conquer the biological network era using graph theory. *Frontiers in bioengineering and biotechnology*, page 34, 2020.

[81] NCBI Blast. Basic local alignment search tool. *Natl. Libr. Med. Natl. Cent. Biotechnol. Inf*, 43(D1):D6–D17, 2015.

[82] Peter C St. John, Jonathan Strutz, Linda J Broadbelt, Keith EJ Tyo, and Yannick J Bomble. Bayesian inference of metabolic kinetics from genome-scale multiomics data. *PLoS computational biology*, 15(11):e1007424, 2019.

[83] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906, 2008.

[84] Chuang Liu, Yifang Ma, Jing Zhao, Ruth Nussinov, Yi-Cheng Zhang, Feixiong Cheng, and Zi-Ke Zhang. Computational network biology: data, models, and applications. *Physics Reports*, 846:1–66, 2020.

[85] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

[86] Venkata P Satagopam, Margarita C Theodoropoulou, Christos K Stampolakis, Georgios A Pavlopoulos, Nikolaos C Papandreou, Pantelis G Bagos, Reinhard Schneider, and Stavros J Hamodrakas. Gpcrs, g-proteins, effectors and their interactions: human-gpdb, a database employing visualization tools and data integration techniques. *Database*, 2010, 2010.

[87] Maud Fagny, Joseph N Paulson, Marieke L Kuijjer, Abhijeet R Sonawane, Cho-Yi Chen, Camila M Lopes-Ramos, Kimberly Glass, John Quackenbush, and John Platig. Exploring regulation in tissues with eqtl networks. *Proceedings of the National Academy of Sciences*, 114(37):E7841–E7850, 2017.

[88] Hui Zhang, Yanchun Liang, Siyu Han, Cheng Peng, and Ying Li. Long noncoding rna and protein interactions: from experimental results to computational models based on network methods. *International journal of molecular sciences*, 20(6):1284, 2019.

[89] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[90] Princy Parsana, Claire Ruberman, Andrew E Jaffe, Michael C Schatz, Alexis Battle, and Jeffrey T Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome biology*, 20(1):1–6, 2019.

[91] Leon Danon, Ashley P Ford, Thomas House, Chris P Jewell, Matt J Keeling, Gareth O Roberts, Joshua V Ross, and Matthew C Vernon. Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on infectious diseases*, 2011, 2011.

[92] Tamara N Romanuk, Richard J Vogt, Angela Young, Constance Tuck, and Mather W Carscallen. Maintenance of positive diversity-stability relations along a gradient of environmental stress. *PloS one*, 5(4):e10378, 2010.

[93] Darren P Croft, Jens Krause, and Richard James. Social networks in the guppy (poecilia reticulata). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(suppl_6):S516–S519, 2004.

[94] Eva Delmas, Mathilde Besson, Marie-Hélène Brice, Laura A Burkle, Giulio V Dalla Riva, Marie-Josée Fortin, Dominique Gravel, Paulo R Guimarães Jr, David H Hembry, Erica A Newman, et al. Analysing ecological networks of species interactions. *Biological Reviews*, 94(1):16–36, 2019.

[95] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[96] Kwang-Il Goh and In-Geol Choi. Exploring the human diseasome: the human disease network. *Briefings in functional genomics*, 11(6):533–542, 2012.

[97] Abhijeet R Sonawane, Scott T Weiss, Kimberly Glass, and Amitabh Sharma. Network medicine in the age of biomedical big data. *Frontiers in Genetics*, 10:294, 2019.

[98] Evangelos Pafilis, Pier Luigi Buttigieg, Barbra Ferrell, Emiliano Pereira, Julia Schnetzer, Christos Arvanitidis, and Lars Juhl Jensen. Extract: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database*, 2016, 2016.

[99] Damian Szklarczyk, Alberto Santos, Christian Von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2016.

[100] Aris Gioutlakis, Maria I Klapa, and Nicholas K Moschonas. Pickle 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology. *PloS one*, 12(10):e0186039, 2017.

[101] Jérôme Kunegis. Konect: the koblenz network collection. In *Proceedings of the 22nd international conference on world wide web*, pages 1343–1350, 2013.

[102] Réka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21): 4947–4957, 2005.

[103] Raya Khanin and Ernst Wit. How scale-free are biological networks. *Journal of computational biology*, 13(3):810–818, 2006.

[104] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nature communications*, 7(1):1–13, 2016.

[105] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[106] Omar Odibat and Chandan K Reddy. Ranking differential hubs in gene co-expression networks. *Journal of bioinformatics and computational biology*, 10(01):1240002, 2012.

[107] Shuiming Cai, Zengrong Liu, and HC Lee. Mean field theory for biology inspired duplication-divergence network model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(8):083106, 2015.

[108] Minoo Ashtiani, Ali Salehzadeh-Yazdi, Zahra Razaghi-Moghadam, Holger Hennig, Olaf Wolkenhauer, Mehdi Mirzaie, and Mohieddin Jafari. A systematic survey of centrality measures for protein-protein interaction networks. *BMC systems biology*, 12 (1):1–17, 2018.

[109] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1): 1–27, 2003.

[110] Michael Schatz, Elliott Cooper-Balis, and Adam Bazinet. Parallel network motif finding. *Techinical report, University of Maryland Insitute for Advanced Computer Studies*, 2008.

[111] Yuriy Hulovatyy, Ryan W Solava, and Tijana Milenković. Revealing missing parts of the interactome via link prediction. *PloS one*, 9(3):e90073, 2014.

[112] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[113] Yujie Ye, Xin Kang, Jordan Bailey, Chunhe Li, and Tian Hong. An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. *PLoS computational biology*, 15(3):e1006855, 2019.

[114] Yadi Zhou, Junfei Zhao, Jiansong Fang, William Martin, Lang Li, Ruth Nussinov, Timothy A Chan, Charis Eng, and Feixiong Cheng. My personal mutanome: a computational genomic medicine platform for searching network perturbing alleles linking genotype to phenotype. *Genome biology*, 22(1):1–18, 2021.

[115] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47 (D1):D941–D947, 2019.

[116] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I Karras, Yang Wang, et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660, 2015.

[117] Mengge Huang, Zhenyu Zhong, Mengxin Lv, Jing Shu, Qiang Tian, and Junxia Chen. Comprehensive analysis of differentially expressed profiles of lncrnas and circrnas with associated co-expression and cerna networks in bladder carcinoma. *Oncotarget*, 7(30):47186, 2016.

[118] Edward R Dougherty. Validation of gene regulatory networks: scientific and inferential. *Briefings in bioinformatics*, 12(3):245–252, 2011.

[119] Ivan Brugere, Brian Gallagher, and Tanya Y Berger-Wolf. Network structure inference, a survey: Motivations, methods, and applications. *ACM Computing Surveys (CSUR)*, 51(2):1–39, 2018.

[120] Sara Barbosa, Bastian Niebel, Sebastian Wolf, Klaus Mauch, and Ralf Takors. A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints. *Biosystems*, 174:37–48, 2018.

[121] Sarvenaz Choobdar, Mehmet E Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, et al. Assessment of network module identification across complex diseases. *Nature methods*, 16(9):843–852, 2019.

[122] Cristian Nogales, Zeinab M Mamdouh, Markus List, Christina Kiel, Ana I Casas, and Harald HHW Schmidt. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends in pharmacological sciences*, 2021.

[123] CNAM Oldenhuis, SF Oosting, JA Gietema, and EGE De Vries. Prognostic versus predictive value of biomarkers in oncology. *European journal of cancer*, 44(7):946–953, 2008.

[124] Daphna Laifenfeld, David A Drubin, Natalie L Catlett, Jennifer S Park, Aaron A Van Hooser, Brian P Frushour, David de Graaf, David A Fryburg, and Renée Deehan. Early patient stratification and predictive biomarkers in drug discovery and development. In *Advances in Systems Biology*, pages 645–653. Springer, 2012.

[125] Theodosia Charitou, Kenneth Bryan, and David J Lynn. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*, 48(1): 1–12, 2016.

[126] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59, 2007.

[127] Dawn C Walker and Jennifer Southgate. The virtual cell—a candidate co-ordinator for 'middle-out'modelling of biological systems. *Briefings in bioinformatics*, 10(4):450–461, 2009.

[128] Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:CIN–S680, 2008.

[129] Rui Liu, Xiangdong Wang, Kazuyuki Aihara, and Luonan Chen. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Medicinal research reviews*, 34(3):455–478, 2014.

[130] G Pei, L Chen, and W Zhang. Wgcna application to proteomic and metabolomic data analysis. In *Methods in enzymology*, volume 585, pages 135–158. Elsevier, 2017.

[131] Siwei Chen, Jiebiao Wang, Ercument Cicek, Kathryn Roeder, Haiyuan Yu, and Bernie Devlin. De novo missense variants disrupting protein–protein interactions affect risk for autism through gene co-expression and protein networks in neuronal cell types. *Molecular autism*, 11(1):1–16, 2020.

[132] Jiaxing Lin, Meng Yu, Xiao Xu, Yutao Wang, Haotian Xing, Jun An, Jieping Yang, Chaozhi Tang, Dan Sun, and Yuyan Zhu. Identification of biomarkers related to cd8+ t cell infiltration with gene co-expression network in clear cell renal cell carcinoma. *Aging (Albany NY)*, 12(4):3694, 2020.

[133] Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical genomics*, pages 93–110. Springer, 2016.

[134] Tomas Hruz, Oliver Laule, Gabor Szabo, Frans Wessendorp, Stefan Bleuler, Lukas Oertle, Peter Widmayer, Wilhelm Gruissem, and Philip Zimmermann. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Advances in bioinformatics*, 2008, 2008.

[135] Mary Goldman, Brian Craft, Jingchun Zhu, and David Haussler. The ucsc xena system for cancer genomics data visualization and interpretation. *Cancer Research*, 77 (13_Supplement):2584–2584, 2017.

[136] Suhas V Vasaikar, Peter Straub, Jing Wang, and Bing Zhang. Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic acids research*, 46(D1): D956–D963, 2018.

[137] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[138] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, et al. Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, 31(1):68–71, 2003.

[139] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[140] Franziska Liesecke, Dimitri Daudu, Rodolphe Dugé de Bernonville, Sébastien Besseau, Marc Clastre, Vincent Courdavault, Johan-Owen De Craene, Joel Crèche, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, et al. Ranking genome-wide correlation measurements improves microarray and rna-seq based global and targeted co-expression networks. *Scientific reports*, 8(1):1–16, 2018.

[141] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews genetics*, 7(1): 55–65, 2006.

[142] Sipko van Dam, Thomas Craig, and João Pedro de Magalhães. Genefriends: a human rna-seq-based gene and transcript co-expression database. *Nucleic acids research*, 43 (D1):D1124–D1132, 2015.

[143] Ruqing Liang, Yaqin Zhi, Guizhi Zheng, Bin Zhang, Hua Zhu, and Meng Wang. Analysis of long non-coding rnas in glioblastoma for prognosis prediction using weighted gene co-expression network analysis, cox regression, and l1-lasso penalization. *OncoTargets and therapy*, 12:157, 2019.

[144] Koh Aoki, Yoshiyuki Ogata, and Daisuke Shibata. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology*, 48(3):381–390, 2007.

[145] F Alex Feltus, Stephen P Ficklin, Scott M Gibson, and Melissa C Smith. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an arabidopsis case study. *BMC systems biology*, 7(1):1–12, 2013.

[146] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.

[147] Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018.

[148] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1):1–14, 2016.

[149] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren, and Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96(1):86–103, 2009.

[150] Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata, and André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics*, 15(6):906–918, 2014.

[151] Franziska Liesecke, Johan-Owen De Craene, Sébastien Besseau, Vincent Courdavault, Marc Clastre, Valentin Vergès, Nicolas Papon, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, Olivier Pichon, et al. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Scientific reports*, 9(1): 1–16, 2019.

[152] Stefan R Maetschke, Piyush B Madhamshettiwar, Melissa J Davis, and Mark A Ragan. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in bioinformatics*, 15(2):195–211, 2014.

[153] Patrick E Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, 9(1):1–10, 2008.

[154] Yupeng Li, Stephanie A Pearl, and Scott A Jackson. Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in plant science*, 20(10):664–675, 2015.

[155] Marek Mutwil, Bjorn Usadel, Moritz Schutte, Ann Loraine, Oliver Ebenhoh, and Staffan Persson. Assembly of an interactive correlation network for the arabidopsis genome using a novel heuristic clustering algorithm. *Plant physiology*, 152(1):29–43, 2010.

[156] Jennifer H Wisecaver, Alexander T Borowsky, Vered Tzin, Georg Jander, Daniel J Kliebenstein, and Antonis Rokas. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell*, 29(5): 944–959, 2017.

[157] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9 (8):796–804, 2012.

[158] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):1–21, 2012.

[159] Francesca Petralia, Pei Wang, Jialiang Yang, and Zhidong Tu. Integrative random forest for gene regulatory network inference. *Bioinformatics*, 31(12):i197–i205, 2015.

[160] Diana Coman, Philipp Rütimann, and Wilhelm Gruissem. A flexible protocol for targeted gene co-expression network analysis. In *Plant Isoprenoids*, pages 285–299. Springer, 2014.

[161] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, page S7. Springer, 2006.

[162] Fan Zhu, Bharat Panwar, and Yuanfang Guan. Algorithms for modeling global and context-specific functional relationship networks. *Briefings in bioinformatics*, 17(4): 686–695, 2016.

[163] Homin K Lee, Amy K Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14(6): 1085–1094, 2004.

[164] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[165] George W Bassel, Hui Lan, Enrico Glaab, Daniel J Gibbs, Tanja Gerjets, Natalio Krasnogor, Anthony J Bonner, Michael J Holdsworth, and Nicholas J Provart. Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proceedings of the National Academy of Sciences*, 108(23): 9709–9714, 2011.

[166] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[167] Jörg Menche. The power of network visualizations. 2022.

[168] Patrik D'haeseleer. How does gene expression clustering work? *Nature biotechnology*, 23(12):1499–1501, 2005.

[169] Kazuki Saito, Masami Y Hirai, and Keiko Yonekura-Sakakibara. Decoding genes with coexpression networks and metabolomics–'majority report by precogs'. *Trends in plant science*, 13(1):36–43, 2008.

[170] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4 (1):44–57, 2009.

[171] Vijay K Ramanan, Li Shen, Jason H Moore, and Andrew J Saykin. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *TRENDS in Genetics*, 28(7):323–332, 2012.

[172] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, Reinhard Schneider, and Alfonso Valencia. Enrichnet: network-based gene set enrichment analysis. *Bioinformatics*, 28 (18):i451–i457, 2012.

[173] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009.

[174] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

[175] Palak Kaushal and Shailendra Singh. Network-based disease gene prioritization based on protein–protein interaction networks. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1):1–16, 2020.

[176] Shi Yu, Steven Van Vooren, Leon-Charles Tranchevent, Bart De Moor, and Yves Moreau. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, 24(16):i119–i125, 2008.

[177] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5): 537–544, 2006.

[178] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22(6):773–774, 2006.

[179] Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of gene ontology terms. *Bioinformatics*, 26 (18):i561–i567, 2010.

[180] Sasha F Levy and Mark L Siegal. Network hubs buffer environmental variation in saccharomyces cerevisiae. *PLoS biology*, 6(11):e264, 2008.

[181] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology*, 11 (4):e1004120, 2015.

[182] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

[183] Katie Ovens, B Frank Eames, and Ian McQuillan. Comparative analyses of gene co-expression networks: Implementations and applications in the study of evolution. *Frontiers in Genetics*, 12, 2021.

[184] Jinjin Tian, Jiebiao Wang, and Kathryn Roeder. Esco: single cell expression simulation incorporating gene co-expression. *Bioinformatics*, 37(16):2374–2381, 2021.

[185] S. Kelly and Michael Black. graphsim: An r package for simulating gene expression data from graph structures of biological pathways. *Journal of Open Source Software*, 5 (51):2161, Jul 2020. doi: 10.21105/joss.02161. URL http://dx.doi.org/10.21105/joss.02161.

[186] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[187] Dharmesh D Bhuva, Joseph Cursons, Gordon K Smyth, and Melissa J Davis. Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome biology*, 20(1): 1–21, 2019.

[188] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[189] Emma Pierson, GTEx Consortium, Daphne Koller, Alexis Battle, and Sara Mostafavi. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS computational biology*, 11(5):e1004220, 2015.

[190] Bahman Panahi, Mohammad Farhadian, and Mohammad Amin Hejazi. Systems biology approach identifies functional modules and regulatory hubs related to secondary metabolites accumulation after transition from autotrophic to heterotrophic growth condition in microalgae. *Plos one*, 15(2):e0225677, 2020.

[191] Wenyuan Li, Lijun Wang, Yue Wu, Zuyi Yuan, and Juan Zhou. Weighted gene co-expression network analysis to identify key modules and hub genes associated with atrial fibrillation. *International Journal of Molecular Medicine*, 45(2):401–416, 2020.

[192] Jianing Tang, Deguang Kong, Qiuxia Cui, Kun Wang, Dan Zhang, Yan Gong, and Gaosong Wu. Prognostic genes of breast cancer identified by gene co-expression network analysis. *Frontiers in oncology*, 8:374, 2018.

[193] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[194] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. The human protein atlas—a tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(4):387–393, 2008.

[195] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1): D529–D541, 2019.

[196] Max Franz, Harold Rodriguez, Christian Lopes, Khalid Zuberi, Jason Montojo, Gary D Bader, and Quaid Morris. Genemania update 2018. *Nucleic acids research*, 46(W1): W60–W64, 2018.

[197] Karla Tonelli Bicalho Crosara, Eduardo Buozi Moffa, Yizhi Xiao, and Walter Luiz Siqueira. Merging in-silico and in vitro salivary protein complex partners using the string database: a tutorial. *Journal of proteomics*, 171:87–94, 2018.

[198] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943, 2016.

[199] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

[200] Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1): D789–D798, 2015.

[201] Simon Jupp, Tony Burdett, Catherine Leroy, and Helen E Parkinson. A new ontology lookup service at embl-ebi. *SWAT4LS*, 2:118–119, 2015.

[202] Anita Sathyanarayanan, Rohit Gupta, Erik W Thompson, Dale R Nyholt, Denis C Bauer, and Shivashankar H Nagaraj. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Briefings in bioinformatics*, 21(6):1920–1936, 2020.

[203] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012.

[204] David A Orlando, Siobhan M Brady, Jeremy D Koch, José R Dinneny, and Philip N Benfey. Manipulating large-scale arabidopsis microarray expression data: identifying dominant expression patterns and biological process enrichment. *Plant Systems Biology*, pages 57–77, 2009.

[205] Chiara Baccin, Jude Al-Sabah, Lars Velten, Patrick M Helbling, Florian Grünschläger, Pablo Hernández-Malmierca, César Nombela-Arrieta, Lars M Steinmetz, Andreas Trumpp, and Simon Haas. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature cell biology*, 22(1):38–48, 2020.

[206] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12, 2021.

[207] Seesandra V Rajagopala, Patricia Sikorski, Ashwani Kumar, Roberto Mosca, James Vlasblom, Roland Arnold, Jonathan Franca-Koh, Suman B Pakala, Sadhna Phanse, Arnaud Ceol, et al. The binary protein-protein interaction landscape of escherichia coli. *Nature biotechnology*, 32(3):285–290, 2014.

[208] Florian R Greten and Sergei I Grivennikov. Inflammation and cancer: triggers, mechanisms, and consequences. *Immunity*, 51(1):27–41, 2019.

[209] Ezio Laconi, Fabio Marongiu, and James DeGregori. Cancer as a disease of old age: changing mutational and microenvironmental landscapes. *British journal of cancer*, 122(7):943–952, 2020.

[210] Fernando M Delgado-Chaves, Francisco Gómez-Vela, Miguel García-Torres, Federico Divina, and José Luis Vázquez Noguera. Computational inference of gene co-expression networks for the identification of lung carcinoma biomarkers: An ensemble approach. *Genes*, 10(12):962, 2019.

[211] Francisco Gómez-Vela, Domingo S Rodriguez-Baena, and José Luis Vázquez-Noguera. Structure optimization for large gene networks based on greedy strategy. *Computational and mathematical methods in medicine*, 2018, 2018.

[212] Christoph D Rau, Nicholas Wisniewski, Luz D Orozco, Brian Bennett, James Weiss, and Aldons J Lusis. Maximal information component analysis: a novel non-linear network analysis method. *Frontiers in genetics*, 4:28, 2013.

[213] Sepideh Sadegh, Julian Matschinske, David B Blumenthal, Gihanna Galindez, Tim Kacprowski, Markus List, Reza Nasirigerdeh, Mhaned Oubounyt, Andreas Pichlmair, Tim Daniel Rose, et al. Exploring the sars-cov-2 virus-host-drug interactome for drug repurposing. *Nature communications*, 11(1):1–9, 2020.

[214] Hussain Ahmed Chowdhury, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. (differential) co-expression analysis of gene expression: a survey of best practices. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(4):1154–1173, 2019.

[215] Robersy Sanchez and Sally A Mackenzie. Integrative network analysis of differentially methylated and expressed genes for biomarker identification in leukemia. *Scientific reports*, 10(1):1–16, 2020.

[216] Manik Vohra, Anu Radha Sharma, Padmalatha S Rai, et al. Snps in sites for dna methylation, transcription factor binding, and mirna targets leading to allele-specific gene expression and contributing to complex disease risk: a systematic review. *Public Health Genomics*, pages 1–16, 2020.

[217] Timothy MD Ebbels and Rachel Cavill. Bioinformatic methods in nmr-based metabolic profiling. *Progress in nuclear magnetic resonance spectroscopy*, 4(55):361–374, 2009.

[218] Hiromi WL Koh, Damian Fermin, Christine Vogel, Kwok Pui Choi, Rob M Ewing, and Hyungwon Choi. iomicspass: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ systems biology and applications*, 5(1):1–10, 2019.

[219] Rachel Cavill, Danyel Jennen, Jos Kleinjans, and Jacob Jan Briedé. Transcriptomic and metabolomic data integration. *Briefings in bioinformatics*, 17(5):891–901, 2016.

[220] Katsuyuki Yugi, Hiroyuki Kubota, Atsushi Hatano, and Shinya Kuroda. Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. *Trends in biotechnology*, 34(4):276–290, 2016.

[221] Tonmoy Das, Geoffroy Andrieux, Musaddeque Ahmed, and Sajib Chakraborty. Integration of online omics-data resources for cancer research. *Frontiers in Genetics*, 11, 2020.

[222] Sarah S Knox. From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell International*, 10(1):1–13, 2010.

[223] Bingbing Xie, Zifeng Yuan, Yadong Yang, Zhidan Sun, Shuigeng Zhou, and Xiangdong Fang. Mobcdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine. *Breast cancer research and treatment*, 169(3): 625–632, 2018.

[224] Zahra Momeni, Esmail Hassanzadeh, Mohammad Saniee Abadeh, and Riccardo Bellazzi. A survey on single and multi omics data mining methods in cancer data classification. *Journal of Biomedical Informatics*, 107:103466, 2020.

[225] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

Appendix A

Conclusiones

De acuerdo a la normativa de la Universidad Pablo de Olavide para optar a la Mención Internacional, se ha realizado una estancia predoctoral en un país de habla no hispana, por lo que se presenta la tesis en inglés y se presentan al menos resumen y conclusiones en español.

Durante el desarrollo de esta tesis trabajamos en varias iniciativas, la mayoría de las cuales se centraron en enfoques de biología de sistemas en el contexto de la investigación biomédica. Prestamos especial atención a la investigación de las redes génicas de coexpresión (GCNs), que se reconstruyeron de forma racional para que pudieran servir como modelos de posibles mecanismos relacionados con la enfermedad. Dado que a menudo las interacciones de los modelos son indirectas y que su participación en la enfermedad puede ser ambigua, las correlaciones gen-gen descubiertas mediante el análisis de GCN deben ser consideradas cuidadosamente. Sin embargo, en el contexto de la identificación de módulos de enfermedad basada en datos, la reconstrucción de la GCN es una potente herramienta para dejar que los datos "hablen", sin tener en cuenta las presunciones que dificultan el descubrimiento de nuevas características de la enfermedad.

Los principios de administración de datos FAIR (por sus siglas en inglés), que destacan la necesidad de que los datos sean localizables, accesibles, interoperables y reutilizables, son la base de toda nuestra investigación. Nos centramos específicamente en cómo los instrumentos apropiados, utilizados de la manera adecuada y bajo un punto de vista racional desde un enfoque biológico bien informado, pueden proporcionar realmente a la comunidad científica valiosas ideas que pueden ser evaluadas en el laboratorio.

A.1 Las GCNs permiten explorar los mecanismos de las enfermedades a nivel de la expresión génica

Cuando comencé mi investigación de doctorado, entré en un campo nuevo con conceptos no formados, por lo que trabajamos para solidificar definiciones unificadas y precisas en un marco en el que trabajaríamos durante los siguientes años. Como resultado, proporcionamos una base general para la reconstrucción de redes de genes en Delgado-Chaves y Gómez-Vela [16]. Somos conscientes de que en los últimos años se han producido modificaciones en la nomenclatura como consecuencia de la falta de consistencia y la aclaración de nociones específicas. No obstante, la idea principal era investigar varios métodos de reconstrucción de redes de genes, con un énfasis particular en el modelado matemático requerido en cada caso. Además, como parte de la investigación impulsada por los datos indicada en la Sección 1.1, creemos que el relativo éxito de nuestra revisión puede atribuirse en parte al énfasis puesto en el enfoque de descubrimiento basado en bases de datos de conocimiento.

Aunque en los últimos años han surgido nuevos métodos y algunos otros enfoques para describir las relaciones biológicas no se describieron porque se consideraron fuera del ámbito de la revisión, describimos los tipos de datos biológicos que podrían considerarse adecuados para la reconstrucción de redes biológicas. Esto puede deberse a que, como se discute en la Sección 2.2, el campo de la biología de sistemas está creciendo mucho más de lo que se pensaba, y las definiciones cambian con frecuencia. Como ejemplo de esto, herramientas como *RNA-Magnet*, que infiere redes de comunicación célula-célula basándose en la expresión de ciertos ligandos y receptores, sólo fue posible gracias al desarrollo de la secuenciación de RNA*single-cell* [205]. Como herramienta de análisis y predicción, otras estrategias como las basadas en redes neuronales de grafos también han ganado importancia en áreas que incluyen la clasificación de nodos, la predicción de bordes y la clasificación de grafos [206]. En consecuencia, la capacidad de representar las interacciones biológicas en un grafo sólo está limitada por la disponibilidad de datos, el progreso tecnológico y el diseño algorítmico.

En nuestra revisión nos centramos en la idea de las redes de regulación génica (GRNs), aunque muchos de los métodos que tratamos también pueden utilizarse para inferir redes génicas de coexpresión. En nuestros estudios, evaluamos si las interacciones entre genes podrían utilizarse para predecir posibles interacciones físicas a nivel de proteínas o simplemente para identificar módulos funcionales en

los patrones de expresión de los genes. La superposición de la capa de información relativa a las redes de regulación génica, es decir, las interacciones entre factores de transcripción (TF) y genes diana, podría restringir el hallazgo de nuevos vínculos, ya que las asociaciones evaluadas se limitan a las interacciones TF-diana que ya han sido identificadas o predichas. Incluso en el caso de organismos modelo como *Escherichia coli* su interactoma está todavía incompleto. Ya en Rajagopala et al. [207], se realizaron experimentos Y2H para reconstruir un conjunto de interacciones proteína-proteína (PPI) de *E. coli* que abarca casi el 70% del proteoma. En el caso del complejo interactoma humano, el interactoma correspondiente sigue siendo poco conocido. Además, la asociación de biomoléculas en la enfermedad cambia de forma dramática y en gran medida desconocida, dando lugar a relaciones específicas de la enfermedad [208, 209]. Estos mecanismos de la enfermedad deben predecirse a partir de los datos, y en menor medida a partir del conocimiento previo. Como resultado, optamos por centrarnos en las redes génicas de coexpresión, que son agnósticas de las interacciones reguladoras de genes documentadas en las bases de datos.

Después de revisar las diversas técnicas y marcos matemáticos, decidimos utilizar la inferencia de redes basada en la correlación, a veces junto con métricas basadas en teoría de la información, como en el caso de Gómez-Vela et al. [18]. La decisión de utilizar estos métodos se basó en el entendimiento de que estas medidas son potentes, competentes en la identificación de interacciones biológicas, y no demasiado exigentes computacionalmente, lo que nos permite evaluar enormes volúmenes de datos en los tiempos de ejecución adecuados.

Ya en nuestra revisión, discutimos la importancia de las perturbaciones sistémicas en la tarea específica de identificar los mecanismos de las enfermedades. Este enfoque estableció el marco para nuestra investigación posterior sobre los procesos subyacentes a la enfermedad, definidos como un cambio entre un estado estable (salud) y un estado perturbado (enfermedad). Como una especie de selección de atributos para la inferencia de redes, en nuestra revisión discutimos la importancia de limitar el número de nodos en la red a aquellos genes de interés. Esto es exactamente lo que hacemos cuando reconstruimos nuestras redes utilizando genes diferencialmente expresados en Delgado-Chaves et al. [210] y Delgado-Chaves et al. [19], con el fin de modelar las relaciones que se producen entre estos genes, que se supone que son responsables de la aparición de la enfermedad cuando cambian. Esto no es más que un intento de recuperar un mecanismo de enfermedad que no sólo se compone de una colección de nodos, como en la simple expresión diferencial,

sino que también intenta determinar las conexiones entre estos nodos en el estado sano frente al estado de enfermedad. Como extensión de esto, planeamos trabajar en otras aproximaciones basadas en enfoques de redes diferenciales, detalladas en la Sección 10.2.

Por lo tanto, con nuestra revisión, cubrimos los objetivos 1 y 2. En nuestra revisión no sólo clasificamos las principales técnicas para reconstruir GRNs y GCNs, sino que también incluimos las principales bases de datos públicas accesibles. Llegamos a la conclusión de que, para el objetivo específico de revelar los mecanismos de la enfermedad, podemos elegir la transcriptómica masiva como un indicador sustituto de otras ómicas. El uso generalizado de las técnicas de elaboración de perfiles del transcriptoma, que da lugar a un mayor número de datos y a una mejor reproducibilidad, es otra razón para emplear los datos de expresión. De entre los muchos métodos para reconstruir redes de genes, elegimos los enfoques basados en la correlación y en la información mutua, que permiten procesar grandes cantidades de datos, como es el caso de los datos de expresión, a un coste computacional relativamente bajo. Dadas las escasas interacciones TF-gen diana que se conocen actualmente, también hemos considerado que la selección de GCNs sobre GRNs es más agnóstica del conocimiento previo, impulsando el descubrimiento de nuevas características asociadas a la enfermedad.

A.2 Los métodos *ensemble* mejoran la robustez de la inferencia GCN

Rápidamente vimos que había ventajas e inconvenientes en las distintas métricas para computar la coexpresión a la luz de las ideas discutidas anteriormente. De hecho, las métricas disponibles representan la relación entre los genes de diferentes maneras, capturando así varios tipos de interacciones.

Comparamos los datos de expresión de microarrays de pacientes con carcinoma de pulmón con los de controles sanos en Delgado-Chaves et al. [210]. Empleamos un enfoque de conjunto con los coeficientes de correlación de Kendall, Spearman y Blomqvist, integrando el análisis de expresión diferencial con la inferencia de GCN. La razón de ser de este método era reducir el número de falsos positivos seleccionando aquellas interacciones de coexpresión que estaban respaldadas por la mayoría de las métricas. Se utilizó un umbral para crear la red final después de promediar los resultados de todas las métricas para proporcionar una evaluación

más fiable de las relaciones modeladas. Hay una falta de acuerdo sobre cómo elegir un umbral para las redes de correlación, como comentamos anteriormente en la Sección 4.2.1. En este caso, se eligió un umbral con la noción de que las correlaciones pueden clasificarse como débiles, medias o fuertes. Intentamos resolver esta cuestión en Gómez-Vela et al. [18], donde comparamos el rendimiento de diferentes umbrales en la recuperación de las interacciones de las redes conocidas.

Además, exploramos las estrategias *ensemble* en Gómez-Vela et al. [18]. Al contrario que en Delgado-Chaves et al. [210], donde sólo incluimos métricas basadas en la correlación, incorporamos una métrica basada en la información mutua en *EnGNet*. El objetivo era extraer las relaciones clave manteniendo la topología libre de escala, una característica de las redes biológicas que se trata en la Sección 3.2.2. *EnGNet* fue el siguiente paso lógico en nuestras técnicas *ensemble*, junto con el trabajo anterior de nuestro laboratorio sobre la optimización de la arquitectura de la red para mejorar la interpretación, manteniendo vínculos significativos [211]. Utilizamos la información mutua normalizada para encontrar interacciones no lineales en los datos, que los algoritmos convencionales suelen ignorar a pesar de la evidencia sustancial de este fenómeno en los sistemas biológicos observables [212]. A diferencia de la técnica descrita en Delgado-Chaves et al. [210], las medidas de coexpresión se someten a una estrategia de votación mayoritaria en lugar de ser promediadas por par de genes. Como resultado, cuando al menos dos de las tres medidas descubren un vínculo, definido como la superación de un umbral, la asociación se mantiene en la inferencia. El umbral de cada métrica de coexpresión se determinó comparándolo con la base de datos de interacciones *GeneMANIA* [196].

La característica esencial de *EnGNet* es su técnica de optimización de la topología, que puede minimizar la complejidad de las GCNs, haciéndolas más sencillas de entender por los científicos del laboratorio húmedo o incluso médicos. Los GCNs se podan utilizando una modificación del algoritmo de árbol de extensión mínima en esta técnica. A continuación, utilizando un umbral de grado, identificamos núcleos a los que se vuelven a añadir las relaciones podadas de la etapa anterior si superan un umbral específico. También comparamos nuestra herramienta con otras cuatro utilizando conjuntos de datos bien caracterizados y sus correspondientes "estándares de oro".

En la última parte de este estudio, utilizamos *EnGNet* en un conjunto de datos sobre el trastorno de estrés postraumático, produciendo asociaciones significativas validadas por la coincidencia de los genes implicados con procesos biológicos

conocidos. Aunque la generalización de los parámetros del algoritmo puede ser investigada más a fondo, la relevancia de *EnGNet* descansa en su capacidad para reducir el tamaño de la red mientras mantiene vínculos biológicamente significativos. Algunas de estas preocupaciones se han abordado en *EnGNet 2.0* (Becchi, Delgado-Chaves y Gómez-Vela, pendiente de publicación), que incorpora otra medida no lineal de coexpresión. Los umbrales para las medidas individuales se determinan en la segunda edición de nuestro algoritmo dependiendo de si la métrica se desvía significativamente de 0. El criterio para añadir nuevos enlaces se basa en la distribución de grados observada, eligiendo los nodos si su grado es un valor superior de la distribución de grados de los nodos. También permitimos a los usuarios especificar el número de votantes para la técnica de votación mayoritaria en esta segunda edición, lo que da lugar a inferencias más o menos ajustadas.

Por lo tanto, hemos logrado los objetivos 3 y 4. Hemos investigado estrategias de ensamblaje que puedan superar las limitaciones de las diferentes técnicas basadas en la correlación y en la información por sí solas. La idea detrás de esto era reconstruir redes más fiables que puedan identificar con mayor precisión los patrones de coexpresión de genes, medidos por múltiples métricas. De este modo, *EnGNet* no sólo evalúa la coexpresión génica mediante tres métricas diferentes, sino que, gracias a su paso de poda, ofrece un resultado fácil de interpretar. Cuando se procesan grandes volúmenes de datos, esto último es un requisito importante porque muchos otros métodos generan redes masivas y densamente conectadas. Además, para que pueda ser utilizado por personas con escasos o nulos conocimientos de programación, ofrecemos una aplicación básica con una interfaz sencilla.

A.3 Descubriendo módulos y mecanismos de la enfermedad directamente a partir de datos

Abordamos principalmente la elucidación de módulos y mecanismos de enfermedad en las aplicaciones que presentamos en Delgado-Chaves et al. [210] y Delgado-Chaves et al. [19]. En todos los casos, comparamos los datos de expresión de las muestras de la enfermedad y de las muestras sanas, con el fin de recuperar cuáles son los actores clave que permiten distinguir entre los fenotipos.

Como se indica en la Sección 3.4, durante el inicio, el desarrollo y el mantenimiento de la enfermedad, las redes de coexpresión se reorganizan, generando

modificaciones que se consideran responsables de la enfermedad. Por lo tanto, en un intento de reducir el espacio de búsqueda a las interacciones de los genes que son relevantes para las enfermedades específicas, utilizamos el análisis de expresión diferencial para identificar los cambios estadísticamente significativos en la expresión de los genes individuales entre las condiciones sanas y la enfermedad. Además, examinamos los genes con expresión diferencial utilizándolos como entrada para reconstruir las GCN. De este modo, pudimos evaluar la GCN de dichos genes en un escenario de estado estacionario, así como su GCN reconfigurado en un escenario de enfermedad.

Es de suponer que las interacciones diferenciales causan el cambio de fenotipo. Los enfoques tradicionales, como el análisis de expresión diferencial, examinan las variaciones estadísticamente significativas en los niveles de expresión de los genes sin tener en cuenta las interacciones entre ellos, lo que podría llevar a un resultado engañoso. Con la idea de que los genes coexpresados están frecuentemente implicados en el mismo proceso biológico, podemos interpretar los genes diferencialmente expresados en el contexto de un módulo funcional combinando la expresión diferencial con el análisis de coexpresión génica. Una extensión de este enfoque sería el uso de redes diferenciales, detallado en la Sección 10.2, que pretendemos cubrir en el futuro. Dado que en nuestro caso, evaluamos los genes diferencialmente expresados en la enfermedad, los módulos que obtuvimos bien pueden considerarse módulos de enfermedad, ya que se predice que desempeñan un papel en la enfermedad a través de su mecanismo de enfermedad asociado.

En Delgado-Chaves et al. [210], el diseño experimental consistió en reconstruir redes utilizando los genes que mostraban una expresión diferencial entre las muestras de personas sanas y las de pacientes con carcinoma de pulmón. Obtuvimos dos GCNs que modelan las interacciones entre los genes diferencialmente expresados en los estados de salud y enfermedad, utilizando el tipo de muestra correspondiente. Después, superpusimos las dos redes para identificar qué interacciones sólo aparecían en el escenario sano y cuáles eran exclusivas del escenario de cáncer de pulmón. Esto nos permitió distinguir las conexiones que se alteraban como resultado de la enfermedad, así como los procesos biológicos que se enriquecieron en cada caso (enfermedad y control).

Con dicha estrategia, pudimos destacar algunos potenciales biomarcadores como *NCKAP1L* y *DMD* en base a su tasa de mutación en muestras de adenomas, adenocarcinomas y neoplasias de células escamosas con bronquios y pulmón como sitio primario, contrastando nuestras redes con el portal Genomic Data Commons

(GDC) del TCGA [193]. A pesar de las limitaciones del enfoque, queremos destacar en este trabajo el potencial de las técnicas de *ensemble* combinadas con la expresión diferencial en la identificación de genes significativos para un escenario de enfermedad concreto. El establecimiento de genes vinculados a una enfermedad, aunque sea de forma indirecta, puede dar lugar a su inclusión en paneles de genes utilizados para la estratificación de pacientes con cáncer de pulmón y su detección temprana.

Integramos el análisis de expresión diferencial con nuestro método *EnGNet* en Delgado-Chaves et al. [19] para examinar las interacciones entre el virus y el huésped en el marco de una infección por coronavirus murino, como modelo del SARS-CoV-2. Hay que tener en cuenta que este artículo se remonta al principio de la pandemia, cuando las muestras de pacientes humanos para COVID-19 eran escasas. Ciertamente, ahora hay soluciones desarrolladas, pero este artículo explica algunos problemas singulares que se estaban tratando en ese momento, y que podrían seguir siendo útiles para los virólogos. El concepto seguía siendo el mismo: cómo podemos explorar los posibles mecanismos de la enfermedad tras la infección por coronavirus, así como lo que puede suponer la deleción de un determinado gen *Ly6E*. Utilizamos los principios de la perturbación, que fueron discutidos previamente en Delgado-Chaves and Gómez-Vela [16], mediante la evaluación de los cambios de coexpresión de genes a lo largo de la progresión de la infección para dos órganos separados (bazo e hígado) y dos escenarios diferentes, perturbado (*LyE6E* knockout) y sin perturbar. Encontramos que nuestro enfoque fue capaz de identificar genes conocidos relacionados con el sistema inmune viral que participan en la respuesta del huésped, así como nuevas proteínas que pueden desempeñar un papel en las interacciones entre el virus y el huésped.

En este sentido, una de las ventajas de la biología de sistemas es que puede responder rápidamente a las necesidades de la comunidad, en contraste con la inherentemente más lenta investigación biomédica clásica. Los métodos de la biología computacional pudieron ofrecer la primera línea de defensa contra la pandemia de COVID-19 porque los datos y la tecnología eran fácilmente accesibles, así como por nuestra limitada capacidad para reutilizar los datos en nuevos contextos. Por ello, el estudio pertinente debe considerarse a la luz de la emergencia de la pandemia. Por ejemplo, proyectos basados en la biología de sistemas como *CoVex* [213], que parte de un módulo de enfermedad compuesto por proteínas relacionadas con el SARS-CoV-2, se desarrolló para encontrar candidatos terapéuticos para el COVID-19. Por supuesto, las vacunas y otros avances tecnológicos llegaron más tarde, por

lo que ya no necesitamos encontrar posibles fármacos para combatir la pandemia de COVID-19.

Las aplicaciones descritas anteriormente nos permitieron cumplir el objetivo 5. Según nuestra experiencia, cuando se comparan muestras de individuos sanos y enfermos, la combinación de análisis de expresión diferencial y coexpresión diferencial permite estudiar los componentes clave de la red que se reconfiguran con la enfermedad. Dado que se desconocen los mecanismos exactos de algunas de las enfermedades que investigamos, validamos nuestros hallazgos examinando si los procesos biológicos representados por nuestras redes tenían sentido en el contexto de cada enfermedad. Con los conocimientos que obtuvimos de estos estudios de casos y gracias a la reutilización de datos, podremos emplear la biología de sistemas para *in silico* predecir los genes y las vías asociadas a la enfermedad en una variedad de escenarios, que posteriormente podrán ser validados en laboratorios húmedos. El objetivo principal de nuestra investigación es llenar el vacío existente entre los descubrimientos académicos y los clínicos, logrando un impacto en la atención sanitaria.

A.4 Otras contribuciones científicas

A lo largo de los últimos años se han realizado otras contribuciones en el marco de la comunicación de la investigación, ya sea en forma de artículos académicos o de otro tipo. Estas importantes contribuciones se enumeran aquí de forma no exhaustiva.

A.5 Artículos en colaboración con otros grupos de investigación

Debido a la experiencia en biotecnología y a las habilidades informáticas que pude adquirir durante mi etapa de estudiante de doctorado, la colaboración con grupos de investigación en ciencias de la vida fue a menudo inmensamente útil. Dichas colaboraciones se muestran a continuación:

- Delgado-Chaves, F. M., Martínez-García, P. M., Herrero-Ruiz, A., Gómez-Vela, F., Divina, F., Jimeno-González, S. y Cortés-Ledesma, F., 2022. Datos de los efectos transcripcionales de la inhibición de TOP2 mediada por merbarona. *Datos en Breve*, 44, p.108499.

- Santamaría-Gómez, J., Rubio, M.Á., López-Igual, R., Romero-Losada, A.B., Delgado-Chaves, F.M., Bru-Martínez, R., Romero-Campero, F.J., Herrero, A., Ibba, M., Ochoa de Alda, J.A.G. y Luque, I., 2021. Papel de un operón genético de ARNt críptico en la supervivencia bajo estrés traslacional. *Investigación sobre ácidos nucleicos*, 49(15), pp.8757-8776.

Como se subraya repetidamente en este manuscrito, la necesidad de cooperación interdisciplinaria en la investigación contemporánea es lo que dio lugar a estas asociaciones. El Prof. Dr. Felipe Cortés Ledesma, del Centro Nacional de Investigaciones Oncológicas (CNIO), dirige el Grupo de Topología y Rupturas de ADN, donde continúa su investigación sobre cómo la topología del ADN da forma a la dinámica y organización del genoma, y cómo un desequilibrio en estos procesos puede dar lugar a la aparición de rupturas patológicas del ADN que amenazan la estabilidad del genoma. En concreto, colaboramos en su análisis de los efectos transcripcionales de la actividad anómala de TOP2, que provoca roturas de doble cadena de ADN que pueden comprometer seriamente la supervivencia celular y la integridad del genoma.

Por otro lado, sigo colaborando con el Prof. Dr. Ignacio Luque Romero del Instituto de Bioquímica Vegetal y Fotosíntesis (IBVF), perteneciente al Consejo Superior de Investigaciones Científicas (CSIC) y a la Universidad de Sevilla, con quien contribuyo al análisis bioinformático para la investigación básica en cianobacterias. Más concretamente, he colaborado en los análisis genómicos del ARN de transferencia (ARNt), con hallazgos relevantes que sugieren que el conjunto de genes de ARNt en algunas bacterias puede estar dividido en un subconjunto de mantenimiento que apoya continuamente la traducción y un subconjunto inducible que normalmente está silenciado pero que puede operar bajo condiciones específicas de estrés.

Al igual que el bilingüismo, una de las habilidades más útiles que he desarrollado a lo largo de los años es la capacidad de comunicarme eficazmente tanto con los científicos de la vida como con los de la computación y de salvar la distancia entre sus disciplinas. Para los próximos años, tengo la intención de seguir ampliando mis conocimientos de informática sin dejar de tener un pie en el campo de la biomedicina, que me parece de lo más apasionante.

A.6 Contribuciones en conferencias

Como parte de mi formación como comunicador científico, he participado e impartido charlas en las conferencias que se indican a continuación. El sitio web de cada universidad que acoge la conferencia puede proporcionar más detalles.

- RExPO22 - 1st International Conference on Drug Repurposing, Maastricht, Países Bajos, 2-3 septiembre de 2022.
- Genómica y transcriptómica, integradas con la proteómica y la informática médica: aprendizaje de las piedras angulares de la medicina de sistemas (GTIPI), Maguncia, Alemania, 29 de mayo - 3 de julio de 2022.

RExPO22 es la primera de una serie de conferencias mundiales sobre reutilización de medicamentos. La conferencia fue elegida porque está directamente relacionada con la comprensión de los mecanismos de las enfermedades. En esta conferencia colaboraron sociedades profesionales, empresas, abogados de patentes, expertos legales y éticos, grupos de investigación clínica, reguladores, pagadores, defensores de los pacientes y muchos más. Los temas principales de la conferencia fueron la redefinición de la enfermedad, la medicina de diagnóstico de órganos, la farmacología de redes, la inteligencia artificial, la regulación, la ética y la seguridad, la medicina de precisión y los ensayos clínicos. Las siguientes publicaciones de la conferencia fueron resultado de nuestra participación en RExPO22:

- Delgado-Chaves, F.M., Oubounyt, M., Gómez-Vela, F.A., Divina, F., Zolotareva, O.I. y Baumbach, J., 2022. Métodos basados en redes diferenciales para la integración de datos ómicos: visión general y desafíos. textitRExPO22 - 1st International Conference on Drug Repurposing, Maastricht, The Netherlands.
- Zolotareva, O., Isaeva, O.V., Hartung, M., Maier, A., Delgado-Chaves, F., Kaufmann, K.C., Savchik, A., Chervontseva, Z.S., Probul, N., Abisheva, A. y Zotova, E., 2022. DESMOND 2.0: Identificación de biclusters de expresión diferencial para la estratificación no supervisada de pacientes. textitRExPO22 - 1st International Conference on Drug Repurposing, Maastricht, The Netherlands.

La escuela de verano y la conferencia "Genómica y transcriptómica, integradas con proteómica e informática médica: Aprendiendo los fundamentos de la medicina de sistemas" (GTIPI) se desarrolló en respuesta a los recientes avances en las técnicas experimentales en la investigación biomédica y la práctica del tratamiento,

que producen conjuntos de datos de alto rendimiento y una variedad de métodos para procesar, modelar e interpretar tales cantidades masivas de datos. Estos métodos provocaron innovaciones en campos como la bioestadística, la bioinformática y la biología computacional. Para que los participantes adquieran activamente una amplia gama de conocimientos fundamentales en el diseño de estudios, el análisis de datos, la biología de alto rendimiento (genómica, transcriptómica y proteómica -también con resolución unicelular-) y la informática médica, se impartieron conferencias sobre los distintos elementos fundamentales de la medicina de sistemas, complementadas con sesiones prácticas de aprendizaje participativo (que incluían la integración de sistemas y datos, así como aplicaciones de la inteligencia artificial).

A.7 Patente

También hemos patentado *EnGNet* 1.0, un programa informático para la reconstrucción de redes de coexpresión de genes libres de escala utilizando una estrategia de conjunto de tres vías, escrito en el lenguaje de programación JAVA y disponible para Windows y Linux. ES SE18020 - Expedición: 18 de septiembre de 2020. Esta patente está inscrita en el Registro de la Propiedad Intelectual por la Universidad Pablo de Olavide y su expediente se encuentra en:

https://www.upo.es/upotec/catalogo/salud/engnet-10-ensemble-and-greedy-gene-networks/

EnGNet 1.0 es la primera aplicación del método *EnGNet* para la creación de modelos biológicos con interfaz visual. Podemos destacar como principales beneficios y aportaciones de la aplicación para el desarrollo de redes genéticas las siguientes:

- En el método se utiliza una técnica de ensemble para superar las restricciones de utilizar una única medida de coexpresión para crear la red genética. La topología final de la red también puede optimizarse de esta manera.
- Los resultados de la metodología demuestran su potencial en el área de identificación y caracterización de biomarcadores.
- El enfoque utilizado para el desarrollo del modelo (*EnGNet*) ha demostrado su superioridad sobre los métodos típicos de generación de redes genéticas encontrados en la literatura, lo que indica un avance en el campo de la Bioinformática y el estudio de los procesos biológicos y las enfermedades.

No se necesita conexión a Internet ni navegadores para utilizar *EnGNet* 1.0 porque permite una instalación autónoma. El diseño de la interfaz también se creó con el objetivo de maximizar la simplicidad y la facilidad de uso. Como ejemplo, ofrece un panel de "Log" dentro de la misma aplicación donde el usuario puede ver los distintos mensajes que informan del estado de la ejecución y una barra de progreso donde el usuario puede ver qué parte de la tarea se ha completado.