Development and implementation of a methodology for the inference of a *Streptomyces coelicolor* gene regulatory network from genomic and transcriptomic data

Dolly Andrea Zorro Aranda

Tesis doctoral presentada para optar al título de Doctora en Ingeniería Química

Tutor

Julio Augusto Freyre González, Doctor (PhD) en Ciencias Bioquímica

Universidad de Antioquia

Facultad de Ingeniería

Doctorado en Ingeniería Química

Medellín, Antioquia, Colombia

2022

Cita	Zorro Aranda, 2022 [1]
Referencia Estilo IEEE (2020)	[1] Zorro Aranda, D. A.,"Development and implementation of a methodology for the inference of a *Streptomyces coelicolor* gene regulatory network from genomic and transcriptomic data", [Tesis doctoral]. Universidad de Antioquia, Medellín, Colombia, 2022.

Doctorado en Ingeniería Química, Cohorte VII.

Grupo de Investigación Bioprocesos.

Biblioteca Carlos Gaviria Díaz

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes

Decano/Director: Jesús Francisco Vargas Bonilla

Jefe departamento: Lina María González Rodríguez

ABSTRACT

Streptomyces coelicolor A3(2) is a model microorganism for the study of Streptomycetes, antibiotic production, and secondary metabolism in general. However, little effort to globally study its transcription has been made even though *S. coelicolor* has an outstanding variety of regulators among bacteria. In this work, we aim to reconstruct a Gene Regulatory Network (GRN) for *S. coelicolor*. For this, we manually curated experimentally validated gene regulatory interactions from which we reconstruct a curated network. Next, based on this curation, we inferred a complete regulation network applying different mathematical methods from two different approaches. One approach was motif detection in DNA sequences and the other one was an inference from transcriptomic data. Further, we analyze the structural properties and functional architecture of both curated and inferred networks. And we compared them to assess the reliability of the predictions. From this analysis, we proposed the functional annotation and biological function for some genes of *S. coelicolor*. Moreover, we proposed the Natural Decomposition Approach as a methodology for the assessment of GRN inference. Finally, we present applications for the curated and inferred networks. The curated networks were deposited in the Abasy Atlas database while the inferences and additional information are available in the supplementary file.

"The important thing is not to stop questioning.
Curiosity has its own reason for existence.
One cannot help but be in awe when he contemplates
the mysteries of eternity, of life, of the marvelous structure of reality.
It is enough if one tries merely to comprehend
a little of this mystery each day." — Albert Einstein

ACKNOWLEDGMENTS

I would like to express my gratitude to all the people who help me to carry out this project. I am very much thankful to *Professor Julio Freyre* for his guidance and support, and especially because without them I would not have been able to carry out successfully my PhD. I am also thankful to all the people in the FreyreLab, for their companionship and support during my time in Mexico and after that. I especially thank *Juan* for his priceless support during the completion of my project. I also would like to extend my thanks to the University of Antioquia and the people there who guide me during my whole PhD. I am especially thankful to *Professor Felipe Bustamante* for its vital rol in the completion of my PhD. I also would like to thank *Carlos Andres*, *Ana Maria* and *Juan Fernando* for their moral support which helped me to endure the difficult times and enjoy the good ones. Finally, I would like to thank my family for their support and patience in this very long process.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACT Actinorhodin

ANOVA Analysis of Variance

AUPR	Area Under the Precision-Recall
AUROC	Area Under the Receiver Operating Characteristic
BGC	Biosynthetic Gene Cluster
CDA	Calcium-Dependent Antibiotic
ChIP	Chromatin Immunoprecipitation
CLR	Context Likelihood of Relatedness
CPK A	Coelimycin A
CSR	Cluster-Situated Regulator
DACA	DNA Affinity Capture Assay
DNA	Deoxyribonucleic Acid
DREAM	Dialogue on Reverse Engineering Assessment and Methods
ECF	Extra-Cytoplasmic Function
EM	Expectation-Maximization
FIMO	Find Individual Motif Occurrences
FP	False Positive
FPR	False Positive rate
FN	False Negative
GENIE3	Gene Network Inference with Ensemble of trees
GEO	Gene Expression Omnibus
GOA	Gene Ontology Annotation
GR	Global Regulator
GRN	Gene Regulatory Network
GS	Gold Standard
KS	Kolmogorov-Smirnov
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
MCC	Matthews Correlation Coefficient

MDscan Motif Discovery Scan

MEME Multiple EM For Motif Elicitation

MLE Maximum Likelihood Estimator

MRMR Maximum Relevance/Minimum Redundancy

mRNA messenger RNA

NCBI National Center for Biotechnology Information

NDA Natural Decomposition Approach

PCA Principal Component Analysis

PFM Position Frequency Matrix

PPM Position Probability Matrix

PR Precision-Recall

PWM Position Weight Matrix

qPCR Quantitative real-time PCR

RED Undecylprodigiosi

RMA Robust Multi-chip Averaging

RNA Ribonucleic Acid

RNA-Seq RNA sequencing

ROC Receiver Operating Characteristic

RT Reverse Transcription

SARP Streptomyces Antibiotic Regulatory Protein

TCS Two-Component System

TF Trancription Factor

TFBS Transcription Factor-Binding Site

TG Target Gene

TIGRESS Trustful Inference of Gene REgulation using Stability Selection

TP True Positive

TPR True Positive rate

TN True Negative

tRNA transfer RNA

1

INTRODUCTION

1.1 INTRODUCTION

Streptomyces is a genus of Gram-positive bacteria that is abundant in the soil, giving it its characteristic odor of wet earth after the rain, courtesy of one of its secondary metabolites, the geosmin [1]. Streptomycetes have a very complex life cycle unique among gram-positive bacteria, comprising a morphological differentiation of diverse cell types. The process begins with the germination of a spore in a suitable environment producing one or more hyphae. These hyphae grow by tip extension forming the vegetative (or substrate) mycelium, which resemblances more to filamentous fungi than other bacteria. In response to the environmental condition, specialized aerial hyphae emerge from the mycelium. Many of them go to produce prespores, which differentiate into mature spores and disperse in the environment (see Figure 1.1) [2].

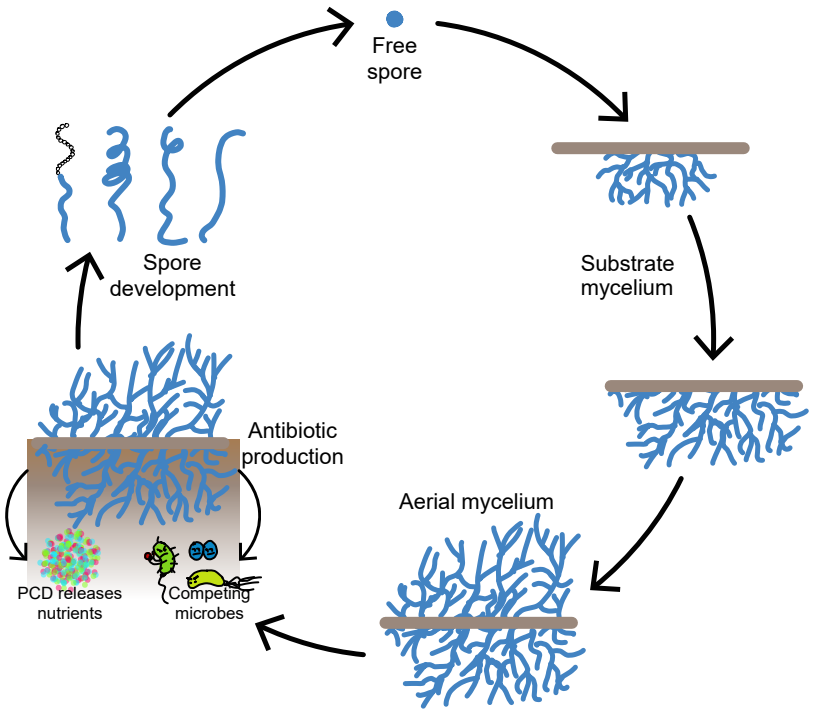


Figure 1.1: Life cycle of *Streptomyces coelicolor*.

Related to its complex life cycle and its natural habitat, where the microorganism faces a great chemical, physical and biological stress; streptomycetes produce a high variety of secondary metabolites that

help them to survive in this hostile environment, which can be classified as [3]:

SIDEROPHORES: These are small molecules excreted by most of the bacteria with a high affinity for ferric iron, which is essential for cell respiration and metabolism; and due to its aqueous insolubility, its normality unavailable for the cell. Siderophores form a complex with ferric iron, which is actively transported into the cell, dissociated in the cell interior so the ferric iron can be used as a cofactor in several cellular processes.[4]

ANTIBIOTICS: Most of the antibiotics are produced when, due to substate exhaustion, vegetative mycelium degrades to provide the nutrients necessary for the development of aerial mycelium. They serve as a defense against other microorganisms that might be attracted by sugars, amino acids, and other small molecules, produced during this process [4]. Antibiotics can be also key components in symbiotic interactions of the bacteria with other organisms, such as fungus, plants, and others; where they prevent infections caused by other microorganisms [3].

SPORE PIGMENTS: Most streptomycetes have pigmented spores. The purpose of this might be to provide a higher resistance of the cell walls to enzymatic digestion, either by the own cell or by other organisms. Also, in some cases, the pigments increase slightly the UV protection of the spore [3].

These secondary metabolites have great applicability in the pharmaceutical industry. Streptomycetes produce about half the antibiotics used clinically among other biochemical compounds, such as antifungals, antivirals, anticancer agents, and immunosuppressives [5]. From genome sequencing usually, around 20-30 Biosynthetic Gene Clusters (BGCs) for diverse secondary metabolites are found, most of them different among species [1], [6], which suggest there is a high number of them to be characterized and used in the health industry. Nevertheless, their industrial production is still quite challenging; since most of them are naturally produced under specififc environmental conditions, most of them unknown, different from the ones in the laboratory [7]. Just to give an example, in submerged liquid cultures, like the ones used in industrial fermentations, aerial mycelium is not formed. This is the stage where antibiotics are produced [1].

Novel experimental technologies, such as new cultivation strategies, improved screening techniques, and new genetic engineering tools are applied in the industry to overcome this difficulty in the production of new biotechnological products [8]. However, their biosynthesis in the cell is caused by certain environmental signals, which trigger the expression of diverse genes responsible for morphological differentiation and secondary

metabolite production. To conserve energy and resources, the expression of these genes is controlled by complex processes of regulation at different levels [9]. The specific regulatory processes of secondary metabolism in streptomycetes are still not fully understood. Therefore, to properly apply these novel experimental techniques, especially in the case of genetic manipulation, a deeper understanding of the whole cellular regulation process might be highly advantageous [10]. Here is where systems biology becomes handy since it allows us to create a model for the study of cellular regulation as a whole system, instead of disconnected individual components.

For the study of streptomycetes regulation, we choose to focus on *Streptomyces coelicolor* A3(2), which along the text will be referred as *S. coelicolor*; nevertheless, this strain is properly a *Streptomyces violaceoruber* [11], as also is the strain *Streptomyces lividans* 66 [12]. *S. coelicolor* has been the model microorganism for the study of secondary metabolism and morphological differentiation in streptomycetes [5], [8]. It was early known by its production of the blue-pigmented antibiotic Actinorhodin (ACT), the red-pigmented antibiotic Undecylprodigiosi (RED), and the Calcium-Dependent Antibiotic (CDA); nevertheless, its genome sequencing revealed one of the largest genomes in bacteria, with more than 20 BGCs [6], such as the Coelimycin A (CPK A), the precursor of yellow coelimycins P1 (yCPK) and P2 [13], that was later characterized.

1.2 REGULATION OF SECONDARY METABOLISM IN *streptomyces coelicolor* A3(2)

There are two main processes of regulation in the cell. One controls the amount of the protein, and the other one its activity. First genes are transcribed into messenger RNA (mRNA), which is then translated into a protein. The amount of protein is controlled at either the transcription stage, by the amount of mRNA produced, or at the translation stage, by the amount of mRNA translated. Afterward, the activity of the protein is regulated post-translationally, by covalent modification, degradation, feedback inhibition, and interactions with other proteins (see Figure 1.2) [9].

In this work, we focus on transcriptional regulation since is the first and principal process of regulation in the cell. It is important to have clarity over this step, before introducing other elements in the model of the cell regulation. Transcription of DNA to RNA starts when the RNA polymerase recognizes and binds to the initiation site on the DNA, or promoter. In Bacteria, promoters are recognized by the sigma factors, which are a subunit of RNA polymerase holoenzyme [9]. Different molecular mechanisms appear to guarantee the proper distribution of RNA polymerase among the different promoters, such as the promoter

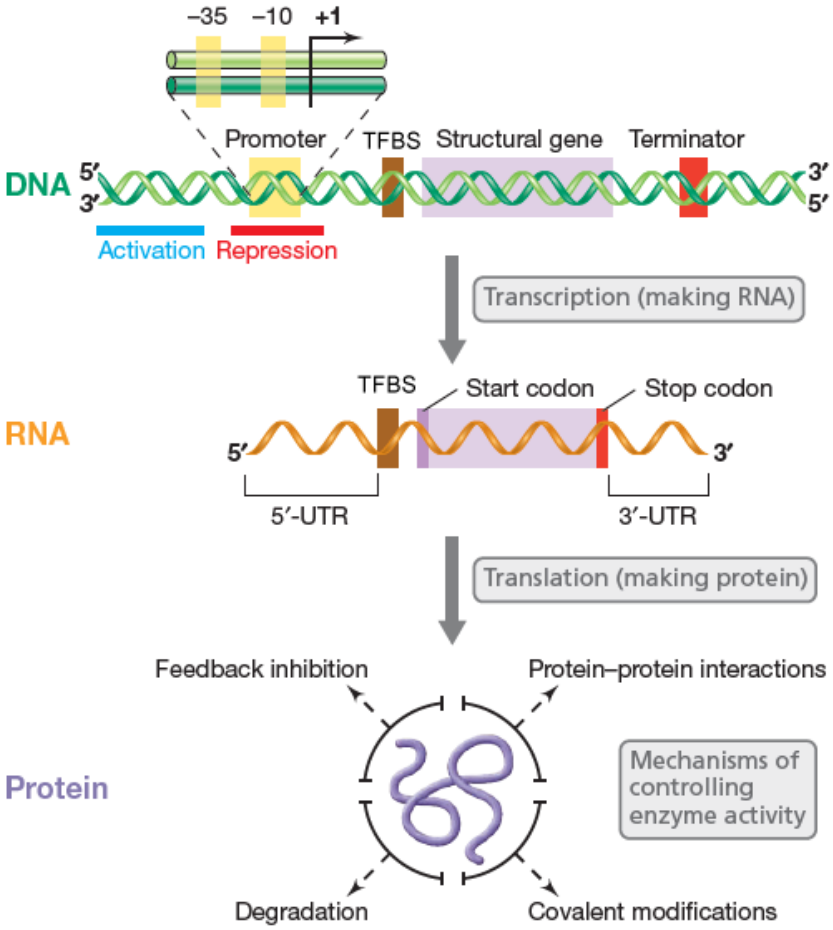


Figure 1.2: Regulation of protein activity. Modified from Madigan *et al.* [9]

DNA sequences, sigma factors, and transcription factors [14]. The basal control comes from the promoter sequence; in bacteria, there are nearly 2000 different promoter sequences, which causes an uneven distribution of the RNA polymerase, where some of them function more efficiently than the others. These differences in the promoter sequences provide useful control over a high number of genes, however, this control is static and does not respond to environmental changes [14]. On the other hand, sigma factors, as was mentioned before, are essential for the recognition of the promoters. Bacteria usually have one main sigma factor that recognizes most promoters, which is called the "housekeeping" sigma factor, and it is responsible for the transcription of the genes needed during cellular growth. However, others accumulate in response to a specific environmental stress, helping the cell to transcribe the genes needed to counter it [15]. Finally, Trancription Factors (TFs) modulates the binding of the RNA polymerase, binding to the Transcription Factor-Binding Site (TFBS) in the DNA. TFs can repress the transcription, interfering with the RNA polymerase, or activate it, helping in the

LOCAL TFS
regulate genes of a specific biological process
GLOBAL TFS
regulate genes from diverse biological processes

recruitment of the RNA polymerase; TFs can have one or both effects. Another molecular mechanism is the strength of the effect on the protein concentration and binding affinity. Strong TFBSs function with lower concentrations of TFs, whereas weak required higher ones. Moreover, local TFs usually have high-affinity TFBSs, while global TFs are less specific, binding to diverse TFBSs. TFs have two important domains related to its regulatory function. One function by ligand-binding as a signal sensor, where the ligand is usually a metabolite or a physicochemical signal from the environment. The other is the responsive component (DNA-binding domain), interacting with the TFBS in the DNA. The most common one, in bacteria, is the helix-turn-helix domain. In the case of bacteria, usually, one protein possesses both components, nevertheless, in the case of Two-Component Systems (TCSs), one protein functions as a sensor, phosphorylating the other one, which functions as the responsive regulator [16].

Years of study of secondary metabolism in *S. coelicolor* have revealed very complex processes of regulation both at a global and a cluster-specific level. This comes from the ability of *S. coelicolor* to grow in soil, where a proper response to diverse external stimuli is essential for its survival. Just for starting, the sequencing of *S. coelicolor*, among its 7846 annotated ORFs (SCO0001– SCO7846) contains revealed a high number (965) of proteins with predicted regulatory function, from which 65 are sigma factor, an exceptional number for bacteria, and from them, 51 are Extra-Cytoplasmic Function (ECF) sigma factors [17]. These specific sigma factors are involved in the response to various environmental stresses; the variety of them may account for the independent regulation of diverse stress response regulons [6]. It also counts with a high amount of TCSs; 85 sensor kinases and 79 response regulators were identified, along with regulators from known families such as LysR, GntR, IclR, and MerR, among others [6]. Besides many putative DNA-binding proteins that seem to not belong to any characterized family of regulators in bacteria [6].

Secondary metabolism usually takes place when the microorganism senses a nutrient deprivation, or other environmental changes, which trigger a complex regulatory response. This causes a higher transcription rate of stress response genes, while expression of genes not highly required in periods of slow growth is reduced [18]. First, as it was mentioned before, a morphological differentiation takes place (see Figure 1.1). This process is mainly controlled by two groups of genes; *bld* genes which are essential for the formation of aerial hyphae, and *whi* genes which are essential for the sporulation of the aerial hyphae [17]. Since secondary metabolites are produced during this stage, these genes are part of their regulation pathways [19]. On the other hand, a secondary metabolite is usually produced by a clustered group of genes that are categorized as a BGC. These BGCs are regulated by two types of mechanisms: global

regulators that are usually TCSs, and pathway-associated regulatory proteins [20].

For many years, it was considered that global regulators activate BGC through its pathway associate regulators; however, it has been demonstrated that they can bind directly to the promoter of the biosynthetic genes [13]. Therefore, there are many regulatory signals confluent to activate a BGC instead of a well-defined regulatory cascade [1]. At a global level, a shortage of nitrogen or phosphate is one of the main causes of the *S. coelicolor* morphological differentiation and secondary metabolites production. This since its main nutirent source is vegetation, which is rich in carbon and poor in nitrogen and phosphate [21]. The response to changes in nitrogen availability is mainly mediated by the orphan response regulator GlnR [20]. Under nitrogen starvation, GlnR activates the transcription of genes involved in nitrogen assimilation, amino acid biosynthesis, and secondary metabolism, among other processes [22]. In the case of phosphate, the response to its limited availability is controlled by the TCS PhoR-PhoP, where PhoP is the response regulator. PhoP also binds to the promoter of the *glnR* gene and many genes related to nitrogen metabolism; nevertheless, it has not been proved control of GlnR over the *phoP* gene [22]. PhoP also has another type of cross-regulation with the global regulator AfsR; which seems to be highly related to the production of the two main antibiotics in *S. coelicolor*: Actinorhodin (ACT) and Undecylprodigiosi (RED). This cross-regulation shows a deep interconnection between primary and secondary metabolism, and the complexity of BGCs activation [20]. Besides these global regulators, many others are involved in secondary metabolism, where most of them are TCSs [23].

Even though most TCSs are located outside the BGCs, there are some which are part of the clusters as the TCS AbsA1-AbsA2 [20]. The *absA* operon, which encodes the sensor kinase AbsA1 and the response regulator AbsA2, is located within the Calcium-Dependent Antibiotic (CDA) BGC, regulating its production. Nevertheless, it also has a strong regulatory effect over other antibiotics such as ACT and RED [20]. These regulatory genes, which are part of the BGCs are defined as Cluster-Situated Regulators (CSRs), usually control its own BGC. Nevertheless, as in the case of the TCS AbsA1-AbsA2, they have been prove also to control different clusters [1], [18]. Part of these CSRs are the regulatory activators of each cluster, which bind directly to the promoters upstream the BGC; and in the case of streptomycetes, they are defined as Streptomyces Antibiotic Regulatory Proteins (SARPs). The main SARPs of *S. coelicolor* are: CpkO (also known as KasO) for CPK A, CdaR for CDA, RedZ and RedD for RED and ActII-orf4 for ACT [13]. These SARPs are activated by global regulators and in response to nutrient depletion, through a signal transduction cascade [19]. The regulatory interactions described above and the lack of understanding

of the complete regulatory processes in *S. coelicolor* are what hinder the overproduction of a desired metabolite through the activation of the BGC by genetic engineering.

1.3 ORIGINAL CONTRIBUTION

The purpose of this work was to define the best methodology for the reconstruction of a GRN, which serves as a model of the regulation of *Streptomyces coelicolor*, and the following objectives were proposed:

GENERAL OBJECTIVE: To develop a methodology for the inference of gene regulatory networks from genomic and transcriptomic data to build a proper gene regulatory network of Streptomyces coelicolor, which eventually will help to improve the modeling of the microorganism metabolism through its integration into a metabolic model.

SPECIFIC OBJECTIVES:

- To collect and curate regulatory interactions experimentally discovered from literature for *S. coelicolor*
- To develop a new methodology for the inference of gene regulatory networks from genomic and transcriptomic data.
- To apply the methodology developed to infer a gene regulatory network for *S. coelicolor* and assess it with respect to the curated network, studying its functional architecture and system-level elements.
- To apply an integration method to the inferred regulatory network and a metabolic network to improve the modeling of secondary metabolism in S. coelicolor.

1.4 ORGANIZATION OF THE THESIS

This document is organized into seven chapters, including the Introduction (Chapter 1) and the Conclusions (Chapter 7). Each chapter has an introduction, which connects each chapter with the previous ones; its methodology; and its results. The chapters are:

CHAPTER 2. *Gene Regulatory Network Reconstruction*: In this chapter we present the collection and curation of regulatory interactions of *S. coelicolor*. Then we reconstruct diverse curated networks according to their source. Finally, we analyze the structural properties and the functional architecture of these networks.

CHAPTER 3. *Gene Regulatory Network Inference from Transcriptomics:* In this chapter, we infer a GRN from transcriptomic data from different sources. We apply 7 different mathematical methods, from which two are proposed in this work. Finally, we assess the inference applying the curated network as GS.

CHAPTER 4. *Gene Regulatory Network Inference from Genomics:* In this chapter, we infer a GRN from the genome of *S. coelicolor*. We apply 3 different methods for motif discovery in DNA sequence. Then, we assess the inference and performed a statistical validation to complement the curated networks.

CHAPTER 5. *Assessment of the Inferred Gene Regulatory Networks:* In this chapter, we compared the structural properties and functional architecture of the curated and inferred networks, to perform a thorough assessment of the inference.

CHAPTER 6. *Biotechnological application of Inferred Gene Regulatory Networks:* In this chapter, we present some of the possible applications of the curated and inferred networks.

In the supplementary file can be found most of the tables of this work, along with the final inferred networks.

2

GENE REGULATORY NETWORK RECONSTRUCTION

2.1 INTRODUCTION

Biology, traditionally, aimed to understand living organisms from the detailed knowledge for each one of the components constituting it, even to the molecular level. Nevertheless, with the increasing level of biological data and the advances in the computation field, it becomes evident that the behavior of the systems present in every living organism cannot be characterized by simple fundamental laws, instead, they work through complex and non-linear interactions between the different molecules, such as DNA, RNA, proteins and small molecules [24], [25]. Here is where systems biology appears as a new field in biology to help us understand the biological process at a system-level [26]. This approach allows us to understand a biological process in-depth, revealing its structure, dynamics, and control, besides allowing us to apply methods for the design and modification of the systems [26]. For the study of some biological processes such as metabolism and regulation at a system level, biological networks it is a proper initial approach for its representation. This since the genes, proteins, and other molecules; and their interactions can be represented as graphs, moreover, graph theory can be applied to them to reveal new aspects of the biological processes that are not evident from the study of the individual components [24].

2.2 GENE REGULATORY NETWORKS

Gene Regulatory Networks (GRNs) represent cellular regulation as a graph, where the nodes or vertexes are the genes, and the links are the regulatory interactions among them. The number of nodes is denoted by N and is the same as the size of the network. In this case, as this is an initial model of *S. coelicolor* regulation, we focus solely on transcriptional interactions among Trancription Factors (TFs) and Target Genes (TGs) (see Figure 2.1). TFs are combined with the gene encoding them, thus the network is solely among genes. These networks have some specific structural features, which are presented below [27], [28].

2.2.1 *Gene Regulatory Network are scale-free*

Traditionally, networks have been represented and analyzed as random graphs, in which links are placed randomly among the nodes. As nodes

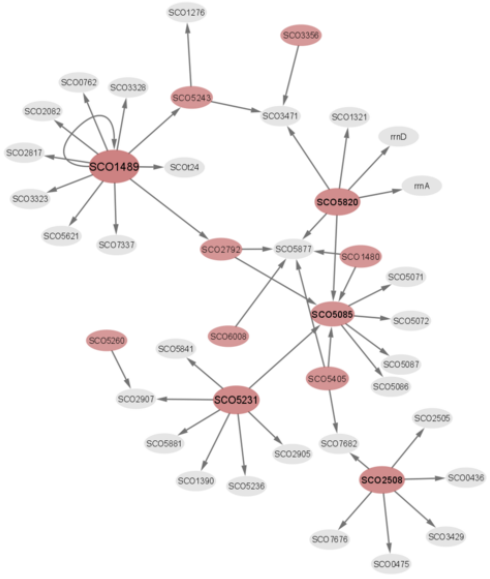
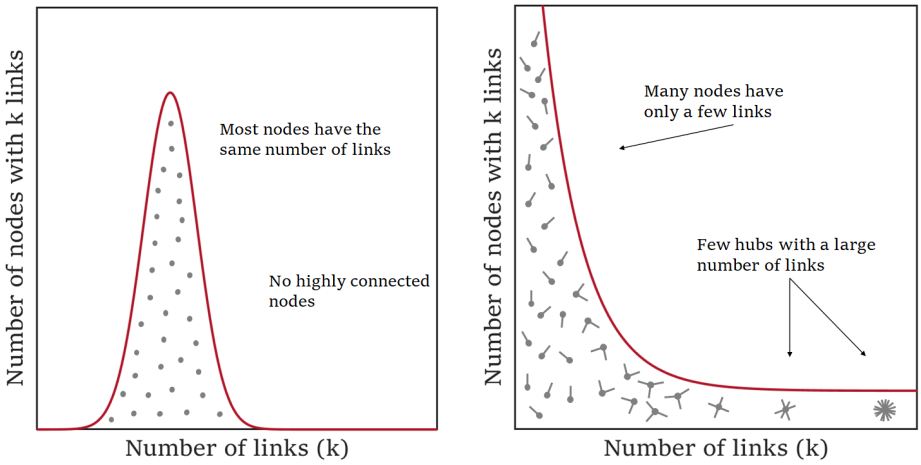


Figure 2.1: Gene Regulatory Network. Red nodes are TFs and grey nodes are TGs.

are placed randomly, most of the nodes have roughly the same degree, close to the average degree $\langle k \rangle$. Thus, the degree distribution follows a Poisson distribution with a peak at $P(\langle k \rangle)$ (see Figure 2.2a). Nevertheless, Barabási *et al.* [29], after analyzing diverse networks reconstruct from real data, proposed that the degree distribution of most of these networks, including biological ones, follow a power law $p(k) \sim k^{-\alpha}$ (see Figure 2.2b), instead of a Poisson distribution. The authors denote this type of network as scale-free since they maintain this property at different stages of their development. The difference in the degree distributions comes from that, in random networks, we find few nodes with low connectivity and practically none with high connectivity; however, in scale-free networks, most of the nodes have low connectivity, and some nodes with very high connectivity are present. This is due to two aspects: first, while random networks are considered to have a constant number of nodes, scale-free networks growth constantly due to the addition of new nodes; second, while in random networks the probability of connection among the nodes is uniform, scale-free networks have shown that new nodes prefer to connect to highly connected ones. These highly connected nodes are also known as hubs. The value of α reveals the role of the hubs in the system. The characteristics of scale-free network are present in networks with an $2 < \alpha < 3$. For $\alpha = 2$ a hub-and-spoke, where all the nodes connect to a single central hub. For an $\alpha > 3$ these characteristics vanish, and the network starts behaving as a random network. And for an $\alpha < 2$ the network has scale-free properties as long as there are multi-links present, which means two or more links between the same two nodes or self-loops. [30]

DEGREE *The degree of a node, or connectivity, is the number of connections it has to other nodes. It is denoted as k_i for the i^{th} node. In the case of directed networks, the incoming connections are denoted as k^{in}, the outgoing as k^{out}, and the total degree $k_i = k_i^{in} + k_i^{out}$.*

DEGREE DISTRIBUTION *Denoted as $p(k)$, is the probability of having exactly a degree k for a randomly selected node.*



(a) Poisson (b) Power law

Figure 2.2: Random vs. Scale-free Networks. Modified from Barabási *et al.* [30].

2.2.2 *Gene Regulatory Network are ultra-small world*

PATH LENGTH *is the distance in terms of links between two nodes.*

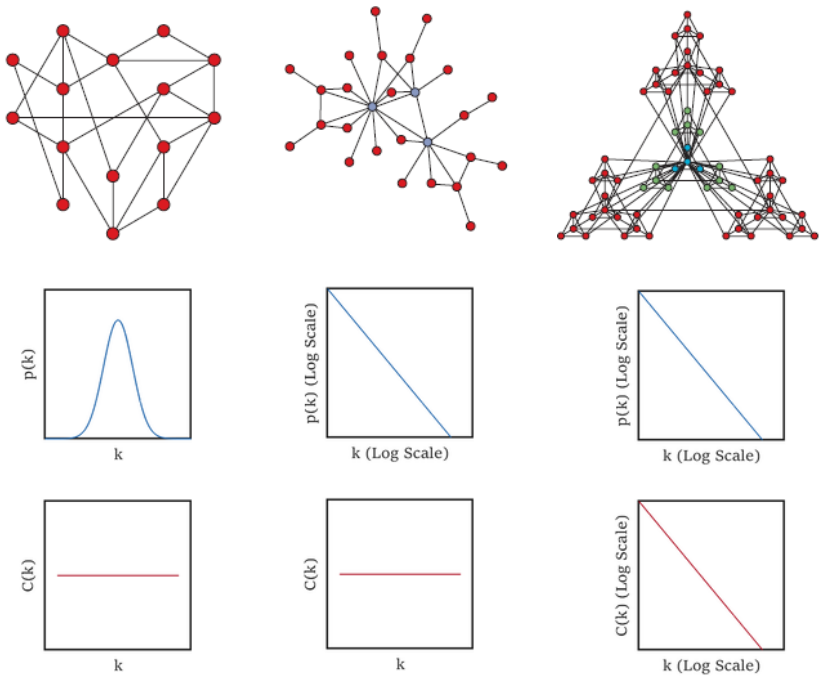
The small-world phenomenon is present in all networks, and it states that any two nodes are connected at a small distance compared to the size of the network. More specifically that $\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$, which means that the average path length depends on $\ln N$ which is much smaller than the size of the network N. In the case of scale-free networks, the presence of hubs allows the nodes to be at a very short distance from all the other ones to the order of $\ln \ln N$ for an $2 < \alpha < 3$, causing the effect of ultra-small world. [30]

2.2.3 *Gene Regulatory Network are hierarchical modular*

CLUSTERING COEFFICIENT *denoted as C_i, measures the interconnection among the neighbors of a node. C_i can vary from 0 to 1, where a $C_i = 0$ implies that none of the neighbors connects to each other, and a $C_i = 1$ that all the neighbors are connected.*

In networks eventually are present communities. These are groups of nodes that have a higher probability of connecting to each other than to nodes outside the community. The concept of communities in networks is very important in the study of biological networks since it is known that, in the cell, molecules form functional modules which focus on specific cellular functions. Nevertheless, the presence of communities in biological networks might be contradictory to the fact that they are scale-free, which implies that most of the genes are connected to a few hubs. In this case, the hierarchical modular network model conciliates both characteristics. This model consists of small communities that form larger communities, which in turn are combined again in much larger communities (see Figure 2.3c). This model produces a scale-free network with an exponent $\alpha = 2.1$. In this model the clustering coefficient of the node depends on its degree; the higher the degree the smaller the

clustering coefficient $C(k) \sim k^{-1}$. This means that small degree nodes reside in highly connected communities, while hubs are linked to different communities having a small cluster coefficient. [30]



(a) Random (b) Scale-free (c) Hierarchical Modular

Figure 2.3: Hierarchical Modular Network. Modified from Barabási *et al.* [30].

2.2.4 *Natural Decomposition Approach*

The Natural Decomposition Approach (NDA) is a framework for the characterization of system-level components of the GRNs based on their intrinsic global properties. The nodes, in this case, genes are classified into four categories: global regulators, modular genes, intermodular genes, and basal machinery genes. They interact with each other in the following way: global regulators coordinate both the basal machinery of the cell, genes whose products are essential for the cell maintenance (DNA and RNA polymerases, transfer RNAs (tRNAs) and its charging enzymes, ribosomal proteins and RNAs, etc.); and local systems (modules) which carry specific biological processes and are defined by modular genes. Meanwhile intermodular genes integrates diverse modules in response to environmental changes [31], [32].

As it was mentioned before, in the case of hierarchical modular networks, the distribution of its cluster coefficient follows a power law, $C(k) \sim k^{-1}$ (see Figure 2.4), which can be divided into two zones: one

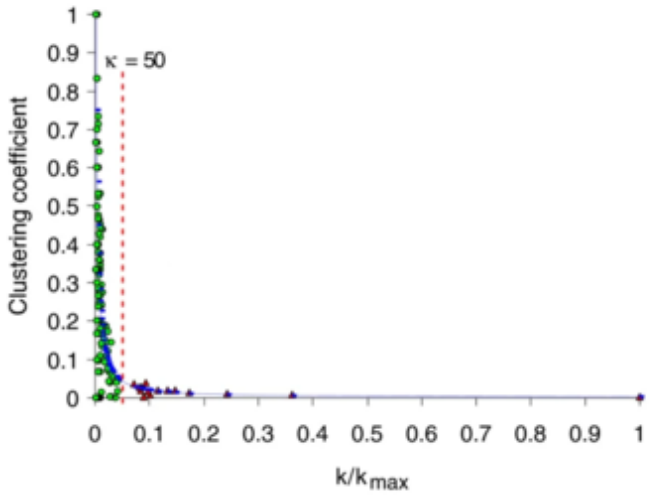


Figure 2.4: Clustering coefficient distribution $C(k)$, and calculated κ value. The blue line represents the $C(k)$ distribution. The dashed red line indicates the κ value obtained for this $C(k)$ distribution. Red triangles represent hierarchical nodes, while green circles indicate modular nodes. Modified from Freyre-González *et al.* [31].

where we have nodes with a small degree in highly connected communities, or modules; and hubs, or in this case global regulators, with a high degree and low cluster coefficient, connecting to different modules. To characterize the nodes as hubs we inferred the equilibrium point where the cluster coefficient distribution diverges; this inflection point is computed as $\frac{dC(k)}{dK} = -1$. From solving this equation, we find the connectivity (κ) of this inflection point and we characterized nodes with higher connectivity as global regulators. [31]

These global regulators act as coordinators of the cellular processes; when they are removed, the network separates into nodes connected in modules and isolated nodes. These isolated nodes are characterized as the basal machinery of the cell, as the others are characterized as modular. Usually, there is a much larger module than the others which is cataloged as a mega-module. When only genes encoding TFs are considered in the network reconstruction, there is no evidence of this mega-module, suggesting the presence of sub-modules and an element of connection in the mega-module. These nodes that connect the sub-modules and are non-TFs are therefore characterized as intermodular genes. [31]

2.3 METHODOLOGY

2.3.1 *Collection and Curation of Transcriptional Regulatory Interactions*

2.3.1.1 *Data Collection*

We review thoroughly literature related to *Streptomyces coelicolor* A3(2) to identify its transcriptional regulatory interactions. First, we performed a quest in Google Scholar and PubMed with the keywords "*Streptomyces coelicolor*" AND "transcriptional" AND "regulation" and different variations of them. In each of the papers, key information was identified such as the microorganism strain and mutations, studied genes, experiments performed, and their experimental conditions. In the case where different experiments were referend in the paper, or it was a review, the references were followed to the original research paper.

2.3.1.2 *Data Curation*

The collected regulatory interactions were standardized and organized in a table with the following information:

- TF name: Gene name.
- TF locus tag: Gene locus tag as stated in the paper. For papers published before *S. coelicolor* genome sequencing a locus tag was assigned according to the paper information and databases such as StrepDB[1], UniProt[2] [33], and BioCyc[3] [34], among others.
- TF description: Gene biological function according NCBI database[4] [35].
- TG name: Gene name.
- TG locus tag: Same as TF locus tag.
- Experiment: Experiments performed in the research that support the regulatory interaction; names were standardized and summarized in seek of clarity.
- Evidence: Strongest experimental evidence that supports the regulatory interaction. See Section 2.3.1.3.

1 http://strepdb.Streptomyces.org.uk/
2 https://www.uniprot.org/
3 https://www.biocyc.org/
4 https://www.ncbi.nlm.nih.gov/

- Regulatory Function: Regulatory function of the TF over the TG according to the experiment performed: activation, repression, or unknown.
- Evidence Classification: Certainty of the direct interaction; classified as "strong" or "weak" . See Section 2.3.1.3.
- PubMed ID: PubMed ID of the research paper.
- DOI: DOI of the research paper.
- Year of Publication: Year of publication of the research paper.
- Notes: Comments or clarification about the interaction.

2.3.1.3 *Evidence Classification*

Following the evidence classification scheme proposed by RegulonDB[5] [36], [37], we classified the interactions as "strong" or "weak" according to the methodology of the experiments performed to identify the transcriptional regulatory interaction. Experiments performed to identify promoters or binding sites were annotated but not considered in the interaction classification.

"STRONG": Evidence of a highly probable direct physical interaction between the TF and the TG; usually classical experiments.

"WEAK" : Evidence of interaction between the TF and the TG; however, there is no certainty whether it is direct or not; usually a high-throughput protocol.

Some of the evidences of the regulatory interactions curated are:

- Binding of Purified Proteins: Experiments that identify nucleic-acid binding proteins performed with purified proteins [38]. We considered this to be strong evidence.
- *in vitro* Transcription Assay: The study of transcriptional regulation process of a specific TG and TF performed *in vitro* [39]. We considered this to be a "strong" evidence.
- Binding Affinity by Bead-based Assays: Experiments that use Beads in the process to isolate *in vivo* DNA sequences that have binding affinity with the TF, such as ChIP [40] and DACA [41]. As it is not possible to prove direct binding, we consider this evidence to be weak.

5 http://regulondb.ccg.unam.mx/evidenceclassification

- Binding of Cellular Extracts: Experiments that identify nucleic-acid binding proteins performed with cell extracts [38]. As there is no certainty of the proteins present, we considered this to be weak evidence.
- Gene Expression Analysis: Experiments that quantify gene expression through the measurement of mRNA [40]. We consider this to be weak evidence.
- Proteomic Analysis: Experiments that quantify gene expression through the measurement protein levels. We consider this to be weak evidence.

2.3.2 *Gene Regulatory Network Reconstruction*

The curated interactions were combined along previously curated from two different databases; the first one was provided to us by the DBSCR team[6]; and the second one was retrieved from RegTransBase [42], which is available at the Abasy Atlas database [43]. Information for similar interactions was combined and all interactions were classified according to their strongest evidence (see page 15). The regulatory function was represented as "+" for activation and "-" for repression. In the case in which there is evidence for both effects and are at the same evidence level ("strong" or "weak"), the regulatory function is represented as "+/-". If there is evidence for both effects, but at different evidence levels, we keep the regulatory function with the strongest evidence. Finally, when the regulatory function is unknown is represented as "? ".

2.3.3 *Gene Regulatory Network Structural Properties*

2.3.3.1 *Estimation of the Degree Exponent* (α)

Plotting the node degree distribution will give us an initial idea of the structural properties of the networks. There we can see if the network is scale-free (see Section 2.2.1), or hierarchical modular (see Section 2.2.3). First, we compute the probability as $p(k) = \frac{N_k}{N}$, where N_k is the number of nodes with degree k. Because of the difference in the degrees, which can be of many orders of magnitudes (see Figure 2.5a), it is more suitable to plot the distribution in a log scale instead of a linear scale. As a power-law distribution in a log-log plot will be a straight line, we will be able to compute the degree exponent from a linear regression. Nevertheless, simply plotting the probability in a log-log plot, which is linear binning, is not the most appropriate to compute the degree exponent computation. This since, we have a large number of nodes with

6 http://dbscr.hgc.jp/

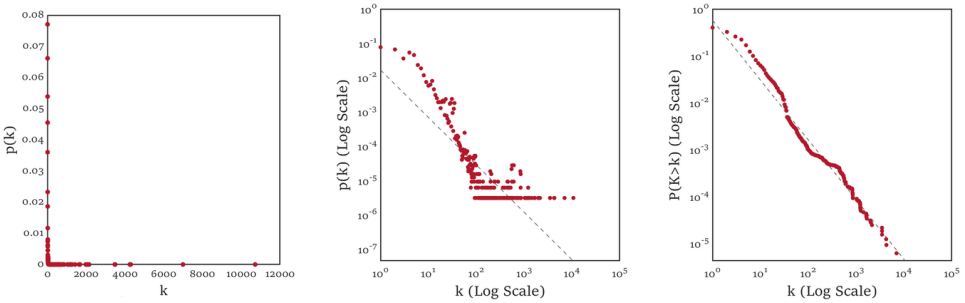
a small degree allowing us to do a proper estimation of the probability of these degrees. However, in the case of large degrees, we have very few nodes to perfomr a proper estimation of the probability, which will bias the fit of the linear regression (see Figure 2.5b). An alternative is to plot the complementary cumulative distribution in a manner of improving the statistical relevance of the nodes with high degrees (see Figure 2.5c). The cumulative distribution of a power law is

$$P(k) = Ck^{-\alpha+1} \tag{2.1}$$

Then, applying natural logarithm to plot it in logarithmic scale is

$$\ln P(k) = (1-\alpha)\ln k + constant \tag{2.2}$$

Thus, from a linear regression of the cumulative distribution, we can compute α [30]. This is the same procedure to estimate the exponent from the $C(k)$, just plotting the clustering coefficient instead of the probability.



(a) Linear Scale (b) Linear Binning (c) Cumulative

Figure 2.5: Estimation of the Degree Exponent from the degree distribution plot. Modified from Barabási *et al.* [30].

Nevertheless, real systems hardly fit a pure power law, they usually present two recurring features (see Figure 2.6):

LOW-DEGREE SATURATION is a flattened probability for small degrees, which means that there are fewer small degree nodes than expected for the case of a pure power law.

HIGH-DEGREE CUTOFF is a rapid drop in the region of high degrees, which means fewer high degree nodes and a smaller maximum degree than expected in the case of a pure power law.

Therefore, we should verify first that the distribution is a power law, applying the Kolmogorov-Smirnov (KS) test to compare the degree distribution to the power law and other heavy-tailed distributions such as lognormal or exponential. The KS test (D) measures the distance

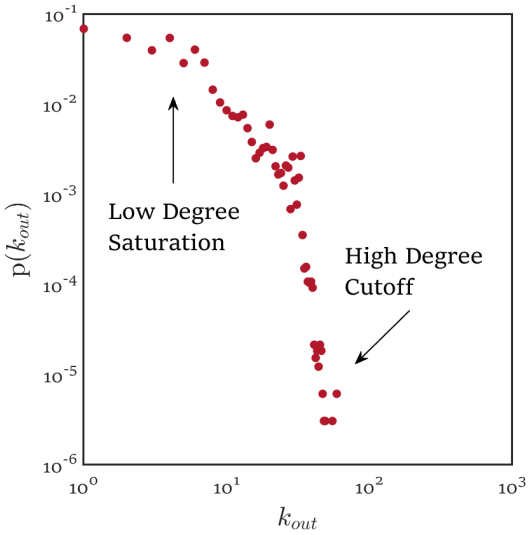


Figure 2.6: Real Degree Distribution. Modified from Barabási *et al.* [30].

between two distribution functions as the maximum value of the absolute difference between two cumulative distribution functions [44]. Therefore, to compare a data set cumulative probability distribution $S_N(x)$ to a known cumulative probability distribution $P(x)$, the KS statistics is

$$D = \max_{-\infty<x<\infty} |S_N(x) - P(x)| \tag{2.3}$$

After assuring that the distribution follows a power law, we can compute the Maximum Likelihood Estimator (MLE) of the parameter $\hat{alpha}$ ("ˆ" represents the estimator) as

$$\hat{\alpha} \simeq 1 + N \left[\sum_{i=1}^{N} \ln \frac{k_i}{k_{min} - \frac{1}{2}} \right]^{-1} \tag{2.4}$$

where k_{min} is the lowest degree where the distribution fits a power law; this since, as it was mentioned before, real systems present a low-degree saturation and k_{min} is the bound of this region. As this value is unknown, the estimation of the parameter $\hat{\alpha}$ is solved numerically through an iterative process introducing as initial values the α and k_{min} estimated from the log-log plot. The whole process and algorithm in different programming languages can be found in Barabási *et al.* [30][7] and Clauset *et al.* [45].

2.3.3.2 *Natural Decomposition Approach Computation*

As it was mentioned before, the NDA classified the genes in four categories (see Section 2.2.4). The process of the NDA is as follows: First, the GRN

7 http://networksciencebook.com/

is represented as a direct graph where the edges go from TFs to TGs. From this graph, the clustering coefficient C_i for each node is computed and we plot its distribution depending on the out-degree k_{out}. If we have various degrees with the same clustering coefficient, we compute the degree mean; thus, we plot the distribution of

$$C\langle k^{out}\rangle = \gamma\langle k^{out}\rangle^{-\alpha} \tag{2.5}$$

From this distribution, we can now compute the inflection point κ, solving the derivative as

$$\kappa = \sqrt[\alpha+1]{\alpha\gamma} \cdot \langle k^{out}\rangle_{max} \tag{2.6}$$

Then we start we node classification in the following way: First, nodes

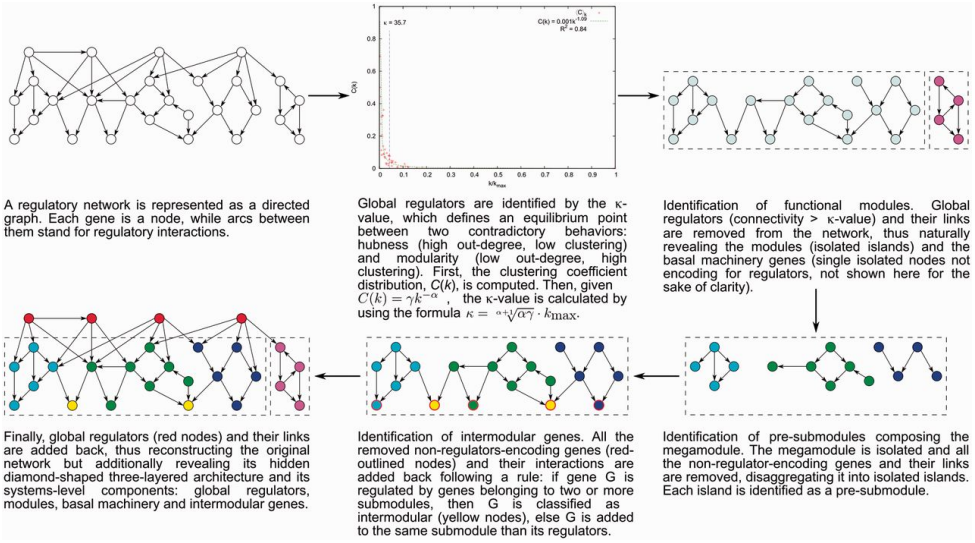
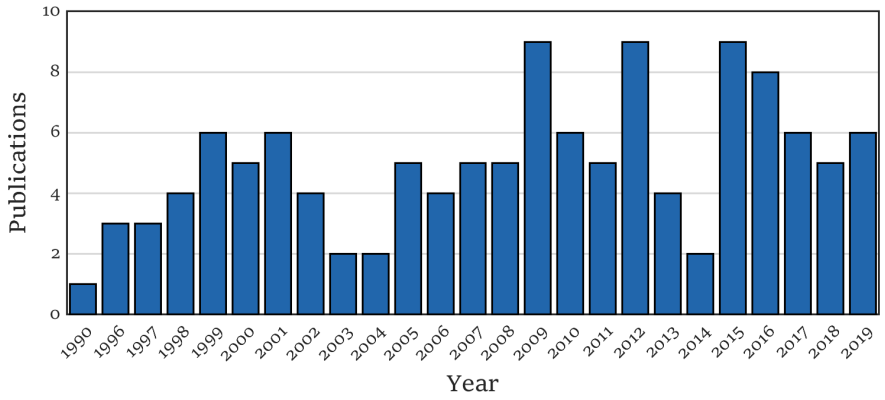


Figure 2.7: Natural Decomposition Approach. Modified from Freyre-González *et al.* [32].

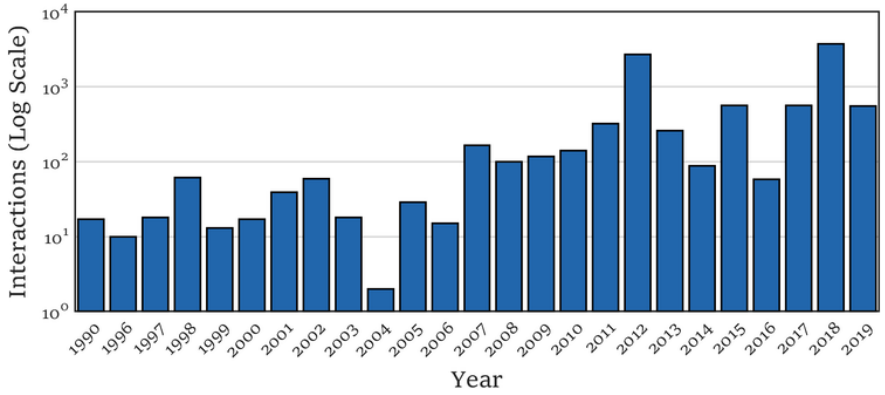
with $k^{out} > \kappa$ are denoted as global regulators. They, along with their interactions, are removed from the graph, leaving us disconnected groups of genes (modules) and isolated genes not encoding for TFs (basal machinery). Afterward, the mega-module is identified. From it, all non-TFs-encoding genes are removed, revealing isolated groups of TFs, which are the pre-submodules. Subsequently, non-TFs-encoding genes are reintegrated to the submodules according to their TFs: if its TFs belong to different submodules, then the gene is categorized as intermodular or is added to the submodule of its TFs otherwise (see Figure 2.7).

2.4 RESULTS

2.4.1 *Collection and Curation*



(a) Number of publications per year



(b) Number of interactions reported per year

Figure 2.8: Curation from literature of transcriptional regulatory interactions for *Streptomyces coelicolor* A3(2).

The first step was to reconstruct a GRN for *Streptomyces coelicolor* A(3)2 from experimental data. For this, we first search for papers related to transcriptional regulation in *S. coelicolor* in PubMed and Google Scholar. We collected and curated a total of 124 papers, covering 29 years (from 1990 to July 2019) (see Figure 2.8a). From these papers, relevant information related to the transcriptional regulation interaction was retrieved, organized, and standardized (see Section 2.3.1.2). The curation can be found in the supplementary file (Table 1). We collected a total of 9714 regulatory interactions among 5331 genes, some of which were repeated 2 or more times. The TFs which are more studied in these papers are the ones encoded by *phoP* (SCO4230), *glnR* (SCO4159), and the sigma factor encoded by *sigR* (SCO5216). The complete list of the number of publications in which a TF is present and the most studied interactions can be found in the supplementary file (Table 3).

We believe this is an important guide for scientists who are designing new experiments in *S. coelicolor*. We noticed a significant increase in the number of interactions reported and papers publish (seeFigure 2.8) following the complete sequencing of the *S. coelicolor* genome, in the year 2002. This since the standardization of the genes facilitates the study and reporting of the regulatory interactions.

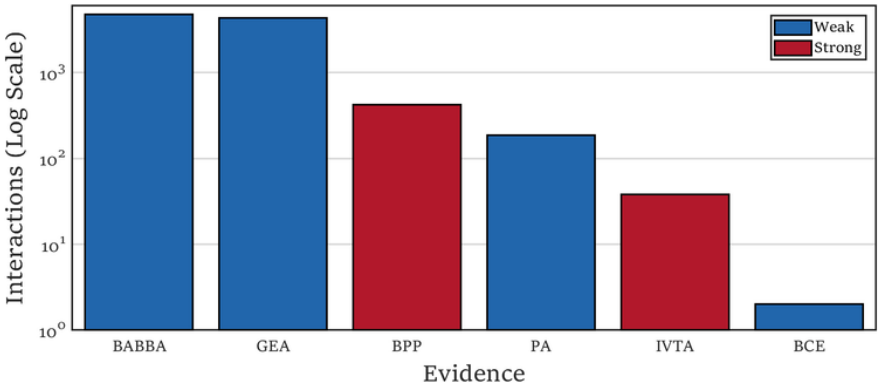


Figure 2.9: Number of interactions for each type of evidence (see Section 2.3.1.3).

As part of the standardization of the regulatory interactions, we segregate them according to the experimental methodology applied to infer the interaction. The classification was performed based on the RegulonDB scheme, where we designate the interaction as "strong" and "weak" according to the experiment performed (see Section 2.3.1.3). "Strong" evidence is assigned to experiments that prove a physical regulatory interaction among the TF and the TG, and "weak" when there is no evidence of direct interaction (see Figure 2.9). The complete list of experiments present in the curation and their classification can be found in the supplementary file (Table 2). From this curation, we reconstruct two networks (see Section 2.3.2):

- *Curated_FL* with a total of 9454 unique interactions, from which ∼5% (438/9454) are "strong".
- *Curated_FL(cS)* with the 438 "strong" interactions from *Curated_FL*. *cS* means that are interactions which are categorized as "strong" in the curation.

Afterward, we gathered the interactions curated along with curations previously reported. First with the ones reported in RegTransBase, which is now available at the Abasy atlas database. Then to the ones reported in DBSCR, which was shared to us in an XML file by the authors of the database. We follow the same process for these curations that to our own. First. we classified the interactions as "strong" and "weak" and then 3 networks were reconstructed:

- *Curated_DBSCR* with the 341 interactions from DBSCR where the ∼34% (115/341) were classified as "strong" interactions.
- *Curated_DBSCR(S)* with the 115 "strong" interactions form DBSCR.
- *Curated_RTB* with the 330 interactions of RegTransBase, all of which were categorized as "weak" since there were no information of the experiments performed.

Then, merging all these networks, we obtained 2 final networks:

- *Curated_FL-DBSCR-RTB* which is the merging of *Curated_FL*, *Curated_DBSCR*, and *Curated_RTB* with a total of 9707 unique interactions for 5386 genes. This is the most extensive experimentally based GRN up to date.
- *Curated_FL(cS)-DBSCR(S)*with the "strong" interactions from the meta-curated network *Curated_FL-DBSCR-RTB*. This network consists of 480 interactions for 387 genes.

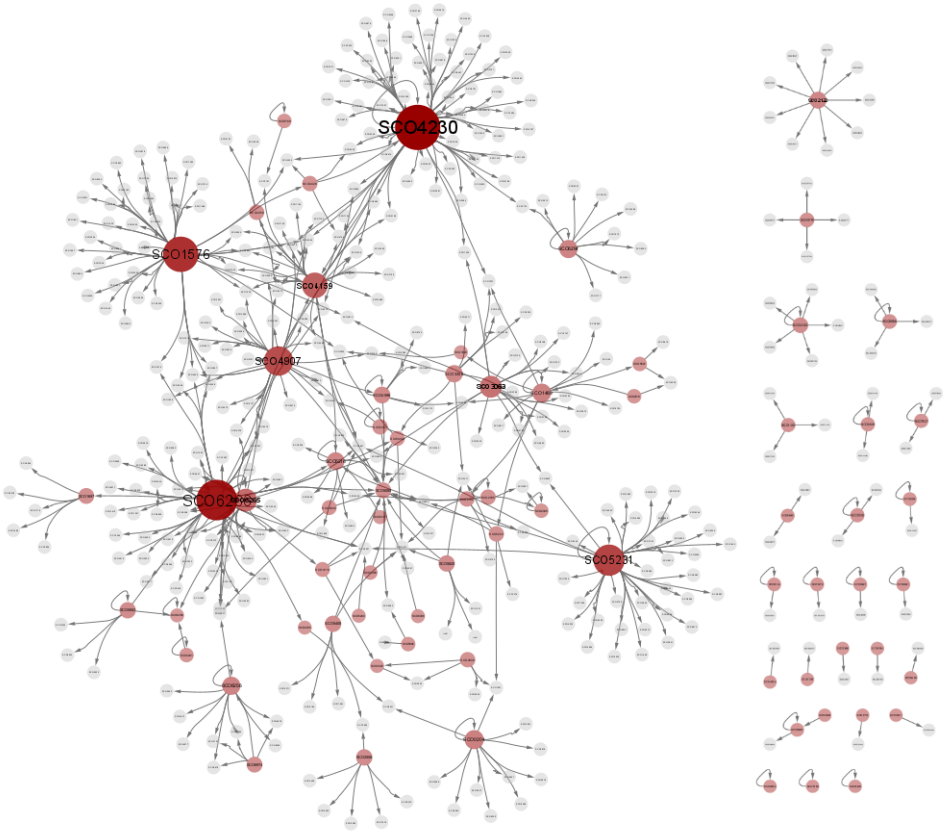
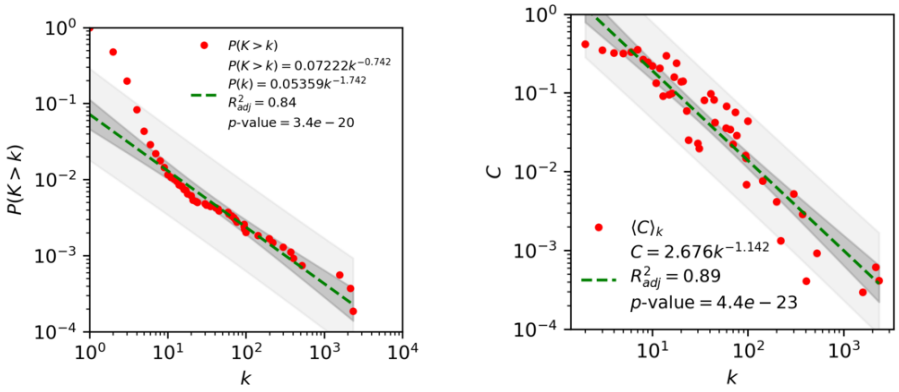


Figure 2.10: Diagram of the Curated Network *Curated_FL(S)-DBSCR(S)*.

A total of 7 GRNs for *Streptomyces coelicolor* A(3)2 were reconstructed, and their complete description can be found in Table A.1.

2.4.2 *Structural Properties of the meta-curated network*

Next, we checked the structural properties of the curated networks. The complete structural properties can be found in the supplementary file (Table 9). This curated network fulfills the main structural characteristics of a biological network. First, all have a low network density ($< 2\%$), where the largest networks (*Curated_FL* and *Curated_FL-DBSCR-RTB*) have the smallest densities ($\sim 0.1\%$), and the smallest networks have the largest densities. The average path length for all the networks is smaller than the logarithm of their network size, is in the same order as $\ln \ln(N)$ which indicated that they are ultra-small world networks (see Section 2.2.2). Then, we plot the cumulative probability of the degree distribution ($P(k)$). For the meta-curated network *Curated_FL-DBSCR-RTB*, the distribution is close to a straight line in a log-log plot, which indicates that the distribution seems to follow a power-law with and $\alpha = 1.74$ (see Figure 2.11a). Thus, we considered this network to be scale-free (see Section 2.2.1). As it was mentioned before, a network with an $\alpha < 2$ can be considered scale-free if there are multi-links present, such as self-loops, like in this network. From the distribution of the other curated networks, we concluded that all can be considered scale-free (see Figures A.1 to A.6).



(a) Network degree distribution

(b) Network clustering coefficient distribution

Figure 2.11: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the meta-curated network *Curated_FL-DBSCR-RTB*.

Knowing that the networks seem scale-free, we next checked if they were hierarchical modular (see Crefsec:hmod). For this, we plotted the cumulative distribution of the clustering coefficient ($C(K)$). For the case of the meta curated network *Curated_FL-DBSCR-RTB*, in a log-log plot, the distribution is close to a straight line with an $\alpha = 1.1$, which indicated that the network is hierarchical modular (see Figure 2.11b). The same

was for all the other curated networks. These three characteristics, small world, scale-free, and hierarchical modular have been previously observed in diverse bacterial networks [27]. The structural properties of GRNs for other bacteria can be found at the Abasy Atlas Database.

To assure that these networks were scale-free, we compute the KS distance between their degree distribution and several similar probability functions: power law, truncated power law, log-normal, stretched exponential, and exponential. All the networks have the smallest distance to a power-law distribution, except to the *Curated_DBSCR(S)* which have similar distances to a power law and a log-normal. Assuring that all the distributions are power-law then we recompute the α through a maximum-likelihood estimation (see Section 2.3.3.1). All the coefficients were between 2 and 3, which corroborated that all the curated networks are scale-free. The KS distances and coefficients can be found in the supplementary file (Table 8).

2.4.3 *Natural Decomposition Approach of the meta-curated network*

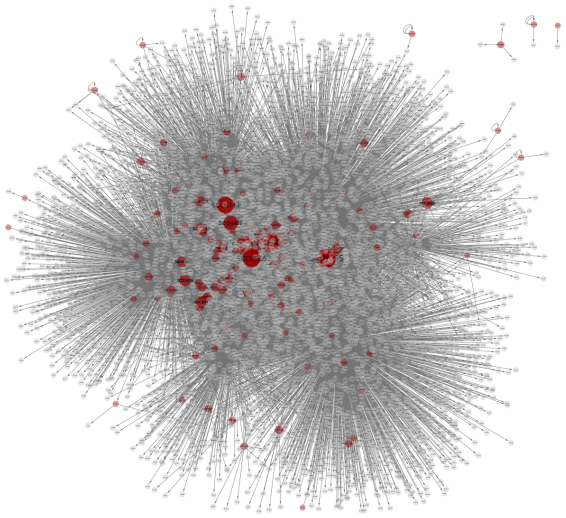


Figure 2.12: Diagram of the curated network *Curated_FL-DBSCR-RTB*.

Knowing that the meta-curated network *Curated_FL-DBSCR-RTB* is scale-free, which means that hubs (Global Regulators (GRs)) are present in the GRN; we can categorize the genes applying the NDA methodology (see Section 2.2.4) in four structural classes: GRs, modular genes, intermodular genes, and basal machinery. After the κ is computed (see Section 2.2.4), we found the GRs and removed them from the network leaving separated subgraphs (modules) (see Figure 2.13), where a mega module is present, and disconnected genes (basal machinery). From this mega module, intermodular genes are found. We decided to study the curated network *Curated_FL-DBSCR-RTB* since is the most complete

one. The NDA analysis revealed 20 GRs, 0.37% of the 5386 genes present in the network, 502 modular genes (9.32%), 18 intermodular genes (0.33%), and 4846 basal machinery genes (89.97%). The cathegorization for each gene can be found in the supplementary file (Table 5).

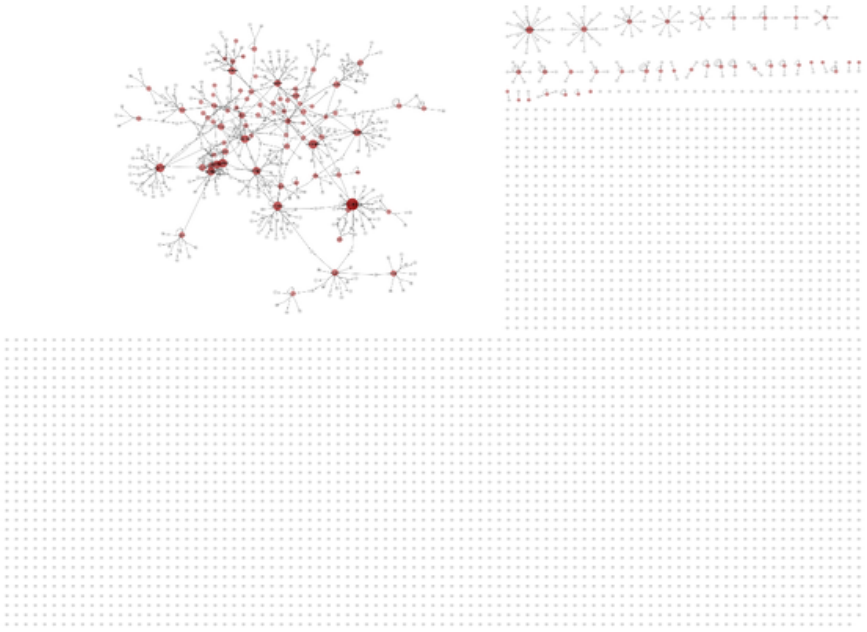


Figure 2.13: Diagram of the curated network *Curated_ FL-DBSCR-RTB* after removing the predicted Global Regulator by the NDA.

2.4.3.1 *Global Regulators*

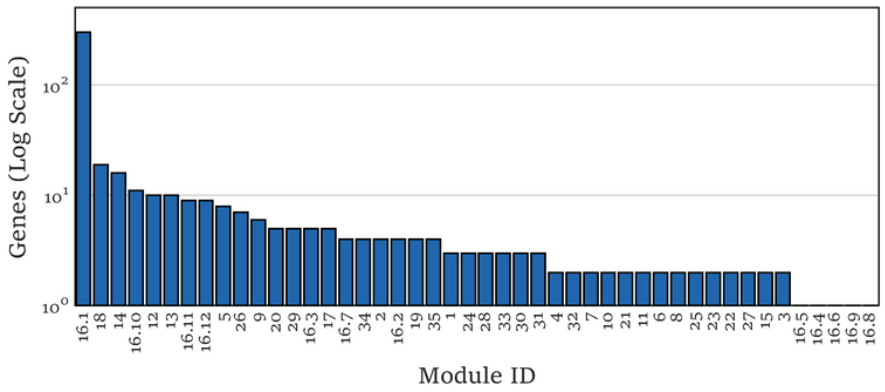
As it was mentioned before, 20 GRs were identified by the NDA. Nine of which were reported in the review about GRs in *Streptomyces* published by Martín *et al.* [22]. The review provides a list of genes considered as global and wide-domain regulators, due to the hundreds of genes they regulate and the multiple effects they produce. In an additional literature search, we found an additional 20 genes reported individually either as global or pleiotropic regulators. From these 13 were categorized as GRs by the NDA. The complete list of publications can be found in the supplementary file (Table 4). The nine GRs previously reported in the review are the TFs encoded by *argR* (SCO1576), *absA2* (SCO3226), *phoP* (SCO4230), *afsS* (SCO4425), *abrC3* (SCO4596), *dasR* (SCO5231), *absC* (SCO5405), *ndgR* (SCO5552), and *scbR* (SCO6265). *phoP* is the gene with the highest out-connectivity in the meta-curated network *Curated_ FL-DBSCR-RTB*. PhoP is a response regulator from the TCS PhoR–PhoP. It has been experimentally identified to act as GR in vivo controlling phosphate scavenging systems and cell wall/extracellular polymer biosynthesis [46]. The other 10 genes that were classified as GRs in the review, were not identified as such by the NDA. The reason for these

false negatives is the criteria used by the author to their classification since it is done by their capability to regulate genes from multiple pathways (wide-domain regulators) or the regulation of hundreds of genes. Therefore, in an incomplete GRN, TFs controlling genes from multiple pathways but with a few TGs in the network will not be identified as GRs by the NDA. This since a high out-connectivity and low clustering coefficient of the gene are, by definition, features required to be classified as GR. Following, we describe the GRs or pleiotropic regulators that were reported individually. The sigma factor encoded by SigR (SCO5216) was recently reported as GR controlling DNA repair, protein quality control, thiol homeostasis, sulfur metabolism, ribosome modulation, and DNA repair [47]. ScbR2 (SCO6286) has been identified to regulate morphological differentiation and stress response through a plethora of genes across the *S. coelicolor* genome, suggesting a global-level regulation [47]. The ECF sigma factor encoded by SCO4117 has been previously reported as a pleiotropic regulator that controls secondary metabolism and morphogenesis [48]. The gene SCO5283 has just been described to encode the cognate response regulator of the TCS SCO5282/SCO5283, having a pleiotropic effect in glycolysis, gluconeogenesis, stress-signaling pathways, proteins secretion, and cell envelope metabolism [49]. Rok7B7 (SCO6008) has been found to control carbon catabolite repression, antibiotic biosynthesis, xylose utilization, and morphological development [50]. The TF encoded by SCO7173 has been reported as pleiotropic regulators of phosphate starvation response and actinorhodin biosynthesis [51]. The TF encoded by SCO5785 is a response regulator related to antibiotic synthesis, sporulation, and several ribosomal gene [52]. Aor1 (SCO2281) has been recently described as a global regulator, orphan response regualtor containing REC and HTH domains, which act as a positive regulator of antibiotic production of ACT, RED, and CDA; and of the genes involved in morphological differentiation [53]. WblA (SCO3579) has been reported as a pleiotropic regulator of various antibiotic pathways, the formation of aerial hyphae, and response to oxidative stress[54]. HrdB (SCO5820) is known to be the housekeeping sigma factor of *S. coelicolor* and to be essential for its survival, thus affecting a great number of biological processes and genes [55]. The only gene predicted as GR that has not been reported as such is SCO3356, which codes for the ECF sigma factor SigE and has only been reported as a coordinator of the cell wall integrity system [56], [57]. Nevertheless, the maintenance of the wall integrity carries diverse biological functions, which might imply a pleiotropic effect on the cell.

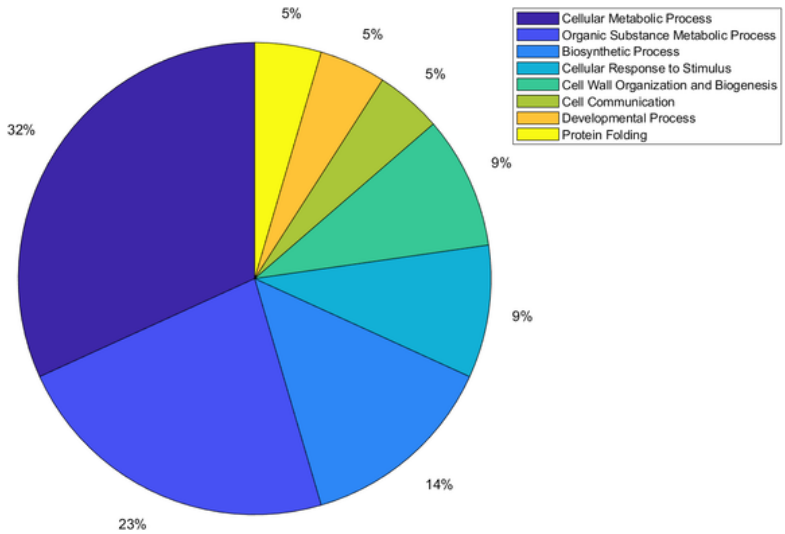
2.4.3.2 *Modular Genes*

There are in total 46 modules and submodules (see Figure 2.14a). There is a mega module (Module 16), which is dived into 12 submodules. The largest of the submodules has 313 genes, which is more than half of

the modular genes. During the process of the NDA, these modules were annotated by computing their functional enrichment [43]. The full list of modules and their annotation can be found in the supplementary file (Table 6a). From the 46 modules and submodules in the GRN, 26% (12/46) are annotated. Most of the modules annotated are related to cellular metabolism, organic substances metabolism, and biosynthetic processes (see Figure 2.14b), which are fundamental processes for every cell. There are also modules related to the response to stimulus, which is one of the main characteristics of *S. coelicolor*, its ability to respond and adapt to environmental change. From the annotation of these modules, through the systems-level guilt-by-association strategy, we were able to suggest the annotation for 79 genes that were not annotated before in Gene Ontology Annotation (GOA) [58] (Supplementary file Table 6b). This is one of the main applications of the NDA methodology.



(a) Number of genes per module



(b) Distribution of the modules per biological function

Figure 2.14: Modules of the meta-curated network *Curated_FL-DBSCR-RTB*.

2.4.3.3 *Intermodular Genes*

Intermodular genes integrate the regulatory response of different modules, which means they coordinate different biological processes in the cell. This analysis also revealed 18 intermodular genes. Five genes are predicted as intermodular genes. Two genes *glnA* (SCO2198) *glnII* (SCO2210), encode glutamine synthase and the *amtB-glnK-glnD* (SCO5583-85) operon encodes an ammonium transporter, a PII protein, and an adenylyltransferase. These five genes are known to be mediators between the nitrogen and phosphate metabolism through the binding to their promoter of the GlnR (SCO4159), the major regulator of nitrogen metabolism, and PhoP (SCO4230) the principal regulator of phosphate metabolism [59]–[61]. Moreover, glnII is also involved in the onset of mycelial differentiation, which might suggest a role in the regulation of secondary metabolism [62]. Additionally, an intrinsic role of glutamine synthetase in secondary metabolism has been also suggested in *S. lividans* [63]. This suggests that both genes might have in role in the coordination of nitrogen and phosphate metabolism, along with the secondary metabolism. Another gene, *ssgB* (SCO1541) a homolog of the sporulation gene *ssgA*, and its product has been suggested as a key regulator of the process of growth cessation before sporulation-specific cell division, affecting cell sporulation along with actinorhodin production [64]. Regarding another gene cluster *agl3EFG* (SCO7165-67) classified as intermodular, Hillerich *et al.* [65] suggested it have the function of carbohydrate transport. This cluster appears to be regulated by GlnR (SCO4159) and Agl3R (SCO7168), a GntR family transcriptional regulator. While GlnR is known to regulate the genes involved in nitrogen metabolism [66], many GntR family regulators have been proved to have a role of repression in carbon metabolism [65]. This might suggest that this cluster has a role in coordinating nitrogen and carbon metabolism, as it has been shown for other processes such as nitrogen and phosphate metabolism [59]. Other intermodular genes are the *actII-orf2* and *actII-orf3* (SCO5083-84), which are part of the BGC of ACT antibiotic. Both genes are regulated by *actII-orf4* (SCO5085) (the SARP of ACT), AfsS (SCO4425) and SCO7173. AfsS integrated the response to phosphate limitation, through PhoP, and the response to unknown stimuli, through AfsR (SCO4426) [67]. SCO7173 is also involved in phosphate metabolism and also affects the biosynthesis of ACT, which might be achieved through these intermodular genes [52].

3

GENE REGULATORY NETWORK INFERENCE FROM TRANSCRIPTOMICS

3.1 INTRODUCTION

We curated a high number of regulatory interactions inferred experimentally from the literature. As a result of this curation, we were able to reconstruct a GRN for *Streptomyces coelicolor* A(3)2. Nevertheless, the meta-curated GRN covers only the $\sim 65\%$ (5386/7825) of its genome and has $\sim 41\%$ (9707/23908) of the expected total regulatory interactions (see Section 3.2.4). Considering that many of these interactions are indirect effects, we are still missing a great portion of the regulatory interactions among all the genes in the microorganism. However, with the high-throughput technologies developed in recent years, the reconstruction of a complete GRNs through computational inferred regulatory interaction appears as an appealing alternative. From these high-throughput technologies, we can obtain genome-wide expression profiles, measured as the amount of RNA transcripts related to each gene, at different conditions. Then, through mathematical, statistical, and computational tools, we can infer regulatory influences among RNA transcripts [68]. Moreover, the reconstruction of a regulation model would be advantageous for properly analyzing the vast amount of information obtained by high-throughput data [69]. Here we infer a GRN from transcriptomic data, through seven different methods, and assess their performance considering the curated network *Curated_FL(cS)-DBSCR(S)* as the Gold Standard (GS).

3.2 METHODOLOGY

3.2.1 *Data Extraction*

We collect transcriptomic data for *S. coelicolor* from two different sources. First from the NCBI Gene Expression Omnibus (GEO) [70], and second from the COLOMBOS database [71]. COLOMBOS is a compendium of microarray and RNA sequencing (RNA-Seq) data from NCBI GEO and ArrayExpress. From there we obtain a dataset of 371 samples of microarray data for 8239 genes. In the NCBI GEO database, to June of 2019, there were 121 samples from RNA-Seq data and 888 samples from microarrays. Nevertheless, not all RNA-Seq datasets were processed the same, and the microarray data came from different platforms. The list of datasets of RNA-Seq and their normalization can be found in the supplementary file (Table 7a). For RNA-Seq we take the higher amount

of data that have the same normalization. Thus, we selected the series GSE132487 and GSE132488, which came from the same study, for a total of 54 samples. In the case of microarrays, as there is not an approved methodology for platform integration for GRN inference, we chose to work with the gene expression data from only one platform. The two largest ones were the GPL4908 for spotted cDNA with 238 samples and the GPL9417 for Affymetrix with 137 samples. We decided to use the data from the Affymetrix platform, due to higher confidence in its measurements[72]–[74] and the tools available for its handling.

For the COLOMBOS and the RNA-Seq data, we worked with the data provided without any further processing. For the Affymetrix data, we downloaded the raw data form NCBI GEO with the *GEOquery* package [75]. Then we performed a Robust Multi-chip Averaging (RMA) normalization with the *affy* package [76]. The RMA is a method developed specifically for Affymetrix normalization [77]. Then we identified a batch effect in the data after the normalization through a guided Principal Component Analysis (PCA) [78]. The batch effect is data variations among groups of samples analyzed together (batches) and is related to experimental features that are not biological, such as performing different experiments in different machines [79]. This analysis was performed with the *gPCA* package. Afterward, we correct this batch effect with the ComBat method [79] from the *sva* package [80]. All the packages are available in Bioconductor for R[1]. As most of the articles curated are from plasmid-free strains of *S. coelicolor*, we selected only the genes in the chromosome. Thus, after filtering all the datasets, we finally obtained for Affimetrix the gene expression data of 7738 genes, for COLOMBOS the data of 6952 genes, and for RNA-Seq the data for 7824 genes, of the 7846 genes of the chromosome according to NCBI Genome [81].

3.2.2 *Network Inference*

Marbach *et al.* [82] performed a comprehensive assessment of over 30 GRN inference methods as part of the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, and it was called the DREAM5 challenge. This challenge with *in silico* data and biological data from different microorganisms. They did not find a method that performed the best consistently through the different data sets. Nevertheless, there were some methods that stuck out, such as TIGRESS, CLR, GENIE3, two-way ANOVA and Inferelator. In previous works at FreyreLab with different microorganisms, we also found these methods to be among the best, plus MRNET, which was not evaluated in the DREAM challenge. Therefore, we selected these methods to perform a GRN inference of *Streptomyces coelicolor* A3(2). Moreover, we proposed

1 https://www.bioconductor.org/

a variation for the two-way ANOVA method and a new method for GRN inference, which will be described below.

These methods can be divided into two categories, co-expression methods, and influence methods. The first category is a methodology that finds genes with similar expression profiles, which will imply that they are expressed simultaneously. If one of these genes is a TF, we assumed that there is a regulatory interaction between the TF and the other gene. This is a rough assumption since the TF and the gene might also be part of the same regulon and are regulated by a second TF. Also, in the case that both are TFs we will have to count the interactions on both sides since there is no certainty in the direction of the regulatory interaction. The second one is a methodology that measured the influence of one gene expression profile over the other one. In this case, we have certainty in the regulatory interaction and a clear direction of this interaction.

3.2.2.1 *Co-expression Methods*

- Context Likelihood of Relatedness (CLR) applied the relevance network approach. This method scores the similarity between the expression profiles of two genes through mutual information. This metric identifies statistical dependence between two variables, without assuming linearity, continuity, and other properties, as in the case of the correlation metric. The relevance network is then refined through a background correction step, where indirect influences and false correlations are eliminated. This is done through the computation of the statistical likelihood of each score obtained by mutual information compared to the distribution of all the scores, taking the ones that are significantly higher than the background distribution. [83]

- MRNET is based also on the mutual information metric. The method applied Maximum Relevance/Minimum Redundancy (MRMR) which selects the variables with the highest mutual information score with respect to a given gene (maximum relevance). These variables are selected from the ones that are maximally mutual independents (minim redundancy). The purpose of this is to minimize the inference of indirect interactions since they will have redundant information with respect to the direct ones. [84]

Both methods are implemented in R and are available in the *minet* package [85] of Bioconductor.

3.2.2.2 *Influence Methods*

- Gene Network Inference with Ensemble of trees (GENIE3) divided the problem in N number of subproblems, where N is the number

of genes in the data. Each subproblem aims to determine the regulators of the given gene, and it is independent of the other ones. As the main purpose is to identify genes that directly influence the expression of the gene, the subproblem can be addressed as a feature selection problem. Therefore GENIE3 applied the embedded feature ranking mechanism of tree-based ensemble methods. [86]

- Inferelator applied standard regression and model shrinkage (LASSO) to infer regulatory influences among genes, considering the expression levels of the regulators and the interactions among them. Inferelator considers a regression problem where they want to predict the expression level of a given gene from the expression level of its regulators. There Least Absolute Shrinkage and Selection Operator (LASSO) shrink or set to 0 some of the coefficients of the regression. This is to reduce the variance of the predicted values to improve the accuracy of predictions. Also, it will allow us to easily identify the variables with the strongest effect, in this case, the regulators with the higher influence. [87]

- Trustful Inference of Gene REgulation using Stability Selection (TIGRESS) applied Least Angle Regression (LARS) along with stability selection. LARS is highly related to LASSO regression, from which they select a set of regulators. The stability selection process is to run several times LARS resampling in each run the data and the variables. TIGRESS finally select the regulators that were selected with the higher frequency across all the runs. [88]

We used the Matlab implementation for GENIE3 and TIGRESS and the R implementation for Inferelator.

3.2.2.3 *Proposed Methods*

FRIEDMAN: We proposed a modification of the method proposed by Küffner *et al.* [89]. Here a new score for GRN inference derived from the Analysis of Variance (ANOVA) is suggested. This score id a non-linear correlation coefficient which described the likelihood of interaction between a TF and a TG. A two-way ANOVA is used to model a dependent variable (TG expression) as a response of two independent variables C and G, as well as the error. In this case, C is the effect across different experimental conditions of the differential expression and G is the similarity in the expression profiles of the TG and the TF evaluated. The null hypothesis for a two-way ANOVA is that there is no significant difference in means of C, G, and their interaction. Thus, the sum of squared deviations (SS) is divided into four components:

$$SS_T = SS_C + SS_G + SS_{CG} + SS_{error} \tag{3.7}$$

Where a high SS_C represents a strong difference in the expressions among conditions. A high SS_G represents a strong difference in the expression of both genes. And a high SS_{CG} indicates that the two effects are liked, which means strong differences among conditions due to strong differences between the genes. Therefore, the strength of association (η_+^2) is proportional to the fraction of SS_C of the total sum S_T, *i. e.* the fraction of the total variance that corresponds to the difference in the expression among conditions.

$$\eta_+^2 = \frac{SS_C}{SS_T} \tag{3.8}$$

In contrast to other correlation coefficients, η_+^2 do not identify negative correlations. Thus, reversing the signs of the TF expression profile, we can compute η_-^2. And compute the final η^2 as

$$\eta^2 = \max(\eta_+^2, \eta_-^2) \tag{3.9}$$

However, ANOVA has specific requirements to perform a proper application of the metric. One of them is that the distributions are assumed to be normal [90]. This might not be accurate in the case of gene expression profiles. Therefore, we proposed to compute the non-linear correlation coefficient from a Friedman Test instead of a two-way ANOVA. This is a non-parametric alternative since it does not assume normality [91]. The computation of η^2 is the same as in Equations (3.7) to (3.9). The algorithm was implemented in Matlab, both the ANOVA and the Friedman method.

STATMODEL: Based on the Statmodel method [92], we proposed a novel method for GRN inference. This method is an alternative tool for statistical modeling and analysis of experiments than ANOVA. Since this methodology allows us to determine the influence of independent factors (TFs expression) over a response variable (TG expression, we found it as a proper tool for GRN inference. For this purpose, we proposed a modification of the methodology initially presented and a suggested score for the interaction reliability. Statmodel presents some advantages concerning ANOVA, such as no assumption of the data distributions and the minimization of the variance of the residual error probability model, reducing the chances of over-fitting. Moreover, this methodology reduces the spurious effects minimizing the number of predictor variables, which is mainly what we look for in GRN inference.

The method is based on the following model where a response variable (Y) is represented as a linear combination of the predictor variables (X_i)

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \tag{3.10}$$

where β_i is the coefficient for each one of the predictors, β_0 is the independent coefficient and ε is the random error of the model. To find the best model representing the response variable, we minimize the error variance maintaining the model unbiased and parsimonious. This is achieved through the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \{Var(\varepsilon), n\} \\ &\text{s. to } E(\varepsilon) = 0 \end{aligned} \tag{3.11}$$

where ε is obtain form Equation (3.10) as

$$\varepsilon = Y - \beta_0 - \sum_{i=1}^{n} \beta_i (X_i) \tag{3.12}$$

To accomplish the restriction $E(\varepsilon) = 0$ and reduce the effect caused by different orders of magnitudes among the variable, we can rewrite the Equation (3.10) as

$$\frac{Y - \langle Y \rangle}{\max(Y) - \min(Y)} = \sum_{i=1}^{n} \beta_i^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon^* \tag{3.13}$$

which is obtained by replacing

$$\beta_0 = \langle Y \rangle - \sum_{i=1}^{n} \beta_i \langle X_i \rangle \tag{3.14a}$$

$$\beta_i = \beta_i^* \frac{\max(Y) - \min(Y)}{\max(X_i) - \min(X_i)} \tag{3.14b}$$

$$\varepsilon = \varepsilon^* (\max(Y) - \min(Y)) \tag{3.14c}$$

Considering that the expected value of the average value of a random variable X is

$$E(\langle X \rangle) = E(X) \tag{3.15}$$

From Equations (3.13), (3.14c) and (3.15)

$$\begin{aligned} E(\varepsilon) &= (\max(Y) - \min(Y))E(\varepsilon^*) \\ &= E(Y - \langle Y \rangle) - \sum_{i=1}^{n} \beta_i^* \frac{\max(Y) - \min(Y)}{\max(X_i) - \min(X_i)} \\ &\quad E(X_i - \langle X_i \rangle) \\ &= 0 \end{aligned} \tag{3.16}$$

Therefore, the restriction in Equation (3.11) is fulfilled and the predictions of the model can be considered unbiased.

Moreover, to obtain a parsimonious model, we should reduce the number of parameters of the model. This can be done through a hypothesis test to find the coefficients β_i^* that are significantly different from zero, as following

$$\begin{aligned} H_0 &: \beta_i^* = 0 \\ H_a &: \beta_i^* \neq 0 \end{aligned} \tag{3.17}$$

To perform this test without making assumptions of the distributions of the variables, we can rewrite Equation (3.13) as

$$\begin{aligned} \frac{Y - \langle Y \rangle}{\max(Y) - \min(Y)} &= \beta_i^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} \\ &\quad + \sum_{i \neq j} \beta_j^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon^* \\ &= \beta_i^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon_i^* \end{aligned} \tag{3.18}$$

where

$$\varepsilon_i^* = \sum_{i \neq j} \beta_j^* \frac{X_i - \langle X_i \rangle}{\max(X_i) - \min(X_i)} + \varepsilon^* \tag{3.19}$$

is a random variable consolidating all the effects different from X_i, maintaining the restriction $E(\varepsilon_i^*) = 0$.

Therefore, we can evaluate each predictor variable independently. Considering to subgroups from the data, one positive $X_i \geq \langle X_i \rangle$ and one negative $X_i \leq \langle X_i \rangle$, we can perform the hypothesis test. If the expected value of the standardized response variable $\left(\frac{Y - \langle Y \rangle}{\max(Y) - \min(Y)}\right)$ for the positive group is different form the one for the negative group (Equation (3.20), the coefficient β_j^* is considered to be significantly different from zero.

$$\begin{aligned} H_0 &: E(Y | X_i \geq \langle X_i \rangle) = E(Y | X_i \leq \langle X_i \rangle) \\ H_a &: E(Y | X_i \geq \langle X_i \rangle) \neq E(Y | X_i \leq \langle X_i \rangle) \end{aligned} \tag{3.20}$$

To perform the hypothesis test, first, we find the probability distribution that better adjusts to Y. Then, we evaluate if the data for the smallest subgroup is adjusted to this probability distribution. To evaluate this, we applied a χ^2 test in Matlab. The coefficient β_j^* is considered to be significantly different from zero if the alternative hypothesis cannot be rejected. This means that the null hypothesis Is rejected by the χ^2 test. We selected the smallest group since it has less statistical power for the hypothesis rejection [93]. Thus, a hypothesis rejected with the smallest group will be rejected with the other one. Then, the non-zero β_i^* can be computed as

$$\beta_i^* = \frac{Cov(Y, X_i)}{Var(X_i)} \cdot \frac{\max(Y) - \min(Y)}{\max(X_i) - \min(X_i)} \tag{3.21}$$

This β_j^* gives us the minimum $Var(\varepsilon_j^*)$.

For the GRN inference, we proposed as a score of reliability of the interaction the $-\log(\text{p-value})$ of the χ^2 test. This, since we look for the TFs that most affect the expression of the evaluated TG, and according to the method should be the ones that produce the most different distributions of Y between both subgroups. Thus, they are the ones that have the smaller p-values of the χ^2 test.

We run all inference methods for each one of the gene expressions datasets and take as a list of TFs the ones obtained from the meta-curated network *Curated_ FL-DBSCR-RTB*. We use this list since those are the genes that are experimentally proven to be TFs. In the case of co-expression methods, the interactions were filtered so at least one TF is present. In the case that both genes are TFs the interaction is duplicated in the opposite sense since there is no certainty of the direction of the regulation. The output of the inference methods is a list of interactions ranked according to their score, from the highest to the lowest. The score is different for each method since it depends on the methodology applied.

3.2.3 *Community Networks*

As it was mentioned before, the DREAM project [82] performed an evaluation of a high variety of methods, from different methodologies. They were not able to find a tendency in their performance over the different data sets. This aspect was also seen in previous projects at the Freyre Lab. This since each method applied different mathematical strategies, each one with different assumptions and biases [69]. Therefore, each methodology highlights different aspects of the network. The integration of the inferred networks presented a robust performance across different data sets since it takes the advantages of each methodology while reducing the biases [69], [82]. Thus, we decided the integrate the inferred networks into one community network applying the Borda count method.

We considered Borda since the score is different for each methodology and it would be difficult to integrate them. Meanwhile, in the Borda method [82] we considered the position of the interactions in each one of the ranked lists obtained. The final position of the interaction is the average position across the lists. If an interaction is missing in one list, its position in the list is the total number of interactions +1. Thus, missing interactions have a higher penalization in larger lists. Interactions with the same score, have the same position in the list.

3.2.4 *Network Refinement*

From the inference methods, we obtained thousands of interactions, with, sometimes, all regulators interacting with almost all the genes. It is established that most regulators interact with genes related to a specific biological function. Therefore, a methodology must reduce this large amount of interaction to a more reasonable number. Campos *et al.* [94] performed an extensive study about GRNs properties with all the networks available, at that moment 71, at the Abasy Atlas database [43]. They identified a constraint in the network density, which they later applied in the prediction of the number of regulatory interactions in a complete GRN. vWhen the density (d) of all the networks was represented as a function of the number of nodes (n), a decreasing tendency to a specific value was perceived. They found that this tendency followed a power law $d \sim n^{-\alpha}$ with $\alpha \approx 1$. As the density is related to the number of interactions (I), this model was reformulated as $I \sim n^{-\gamma}$ with $\gamma = 2-\alpha$[2]. Assuming n as the total number of genes in the genome, we would be able to predict from the model the total number of interactions of the GRN. The exact value of α is computed considering the totality of the networks and is updated as new networks are added to the database. Thus, the total number of interactions of a microorganism might slightly change over time. [94]

NETWORK DENSITY *describes the fraction of actual interactions with respect to the number of all possible interactions among the nodes*

3.2.5 *Assessment*

To perform an initial assessment of the inferred networks, we computed the Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall (AUPR) for each one of the networks using *Curated_FL(cS)-DBSCR(S)* as the GS. First, for each inferred network a confusion matrix is computed (see Figure 3.1, which is the base for these and other metrics. There each one of the inferred interactions is classified in four categories [95]–[97]:

- True Positives (TPs) describes the inference of an interaction presented in the GS.

2 https://abasy.ccg.unam.mx/

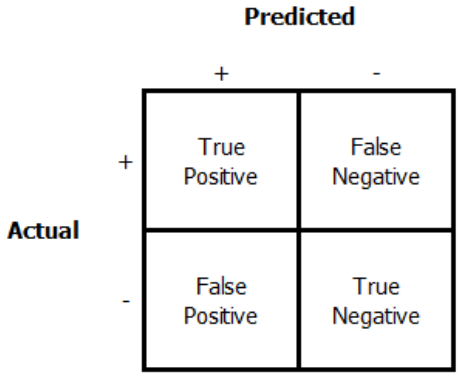


Figure 3.1: Confusion Matrix.

- False Positives (FPs) describes an inference interaction that is not part of the GS.
- True Negatives (TNs) describes an interaction that is neither inferred nor part of the GS.
- False Negatives (FNs) described an interaction that is not inferred, however, is present in the GS.

A Receiver Operating Characteristic (ROC) curve portray the relative trade-off between the benefits, in this case the TPs interactions (y), and the cost, the FPs interactions (x) [95].

$$TPR = \frac{TP}{TP + FN} \tag{3.22}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.23}$$

We can build a curve computing the True Positive rate (TPR) and False Positive rate (FPR) of the predictions taking different thresholds for the score, having a point for each one [95]. We considered only as inferred (positive) interactions those with a score above the threshold. In this case, we take each score value as a different threshold. For some methods, this implies adding one interaction at a time, while for others, where many interactions have the same score, is more than one interaction.

The point $(0, 0)$ is the point where no interaction is classified as positive (part of the GS), so there is neither FP, neither TP. The opposite will be the point $(1, 1)$ where all the interactions are considered positive. The point $(0, 1)$ is a perfect inference since all the interactions are TP. Thus, a point is better than another if it has a higher TPR, a lower FPR, or both. Any point in the diagonal line $y = x$, the baseline, represents

predictions performed randomly, and point below this line are predictions worse than random predictions. [95]

A Precision-Recall (PR) curve is built by plotting the precision against the recall. The recall, the same as the proportion of the GS total interactions retrieved, is computed as the TPR (see Equation (3.22)). While precision, the proportion of the GS in the total predicted interactions, is computed as follow [97]

$$Precision = \frac{TP}{TP + FP} \tag{3.24}$$

The point $(1, 1)$ represents a perfect prediction since the prediction will be the same as the GS. The ideal will be a PR curve that goes towards that point [96]. While the baseline of the ROC curve is fixed to the line $y = x$. In the case of the PR curve change for each prediction. It is determined by the ratio of interaction in the GS (positives (P)) and the interactions outside of it (negatives (N)) [97], as

$$y = \frac{P}{P + N} \tag{3.25}$$

As curves are difficult to compare, it is advisable to represent the curve with a scalar value [96]. Therefore, we compute the AUROC and AUPR for all the predictions to make the assessment easily comparable. We considered as the total space of the problem only the interactions among the genes present in the GS. This since to compute the TN and the FN, we need to consider all the possible interactions among all the genes, that are not in the GS. There are 480 interactions for 387 genes (see Table A.1), then if we consider the 7846 genes of *S. coelicolor*, the space of the problem will be much larger. This will produce a larger proportion of FPs since the method might be inferring actual interactions that have been not experimentally proved yet [98].

Even though we computed both metrics, we focused primarily on the PR. This since it has been shown that this metric is more informative than the ROC in imbalance problems [97]. As the GS has only 480 interactions and all the possible interactions will be a permutation with repetitions of the 387 genes, which is equal to 149769 possible interactions. Therefore, the positive fraction of the GS is much smaller than the negative fraction, causing this to be an imbalance problem.

PR is more informative in imbalance sets since it focuses on the correctly predicted GS interactions, while ROC is a trade-off of the TPRs and the FPRs. However, a high acTPR might depend on a small set of FN, and a low FPR might depend on a large set of TN, Thus, the aim of the assessment which is to evaluate the method that better inferred the GS is a straightforward result of the PR metric, while in the ROC metric will be more challenging to discern. [97]

3.3 RESULTS

To complement the curated network, we collected high-throughput data of *Streptomyces coelicolor* to infer the missing regulatory interactions. We performed a GRN inference from transcriptomic data, from different sources. First, we used the microarray data consisting of 371 samples available from the COLOMBOS Database [71]. Second, we obtain microarray and RNA-Seq data from NCBI GEO [70]. As there is not a consensus method for microarray data integration from different platforms, we decide to take the platform with the largest dataset (137 samples), which was an Affymetrix platform. This data was taken raw, and RMA normalization was performed [77] and the data was batch-effect corrected [79]. For the case of RNA-Seq, the data is available with different types of normalization for each series. Therefore, we decided to take the largest dataset as well (54 samples) (see Section 3.2.1). We used seven methods for GRN inference based on the gene expression data: CLR, Friedman, GENIE3, Inferelator, MRNET, Statmodel, and TIGRESS (see Section 3.2.2.2). Since most of the experiments in the curation were performed on the *S. coelicolor* A3(2) strain M145, and other plasmid-free strains, we restricted the inference to interactions among genes of the chromosome.

The inference form expression data was performed over the 5 datasets available, to select the best datast for the inference:

- COLOMBOS (*colombos*)
- RNA-Seq (*rnaseq*)
- Affymetrix raw (*raw*)
- Affymetrix with RMA normalization (*rma*)
- Affimetrix with batch-effect correction (*rmabatch*)

To provide insights on the quality of the predictions, the inferred GRNs were assessed using the network *Curated_FL(cS)-DBSCR(S)* as GS, and the AUPR were computed for each one of the networks (see Figure 3.2). We assessed the inferred GRN based mostly on the AUPR since, as it was mentioned before, it is more informative for imbalanced datasets (see Section 3.2.5). Before the assessment to prune the inferred GRN to its predicted number of interactions, which at the moment, with the networks available in Abasy, is 23908 interactions (see Section 3.2.4). This number might be slightly different if new networks are added to the database. From this evaluation, we notice that, despite a large amount of data, the prediction with the dataset from COLOMBOS performs poorly than the other data sets. From the Affymetrix data, the *rmabatch* set performed the best with a very similar result to the *rnaseq* dataset.

However, the latter comes from one unique study, while the Affymetrix data comes from different studies with more diverse data. Then, we selected the Affymetrix *rmabatch* as our dataset for the final inference from transcriptomic data. This, since we believe a more diverse data will allow us to identify a higher quantity of regulatory interactions. Moreover, due to the overly poor performance of ANOVA across the data sets, we decided not to consider it for the GRN inference.

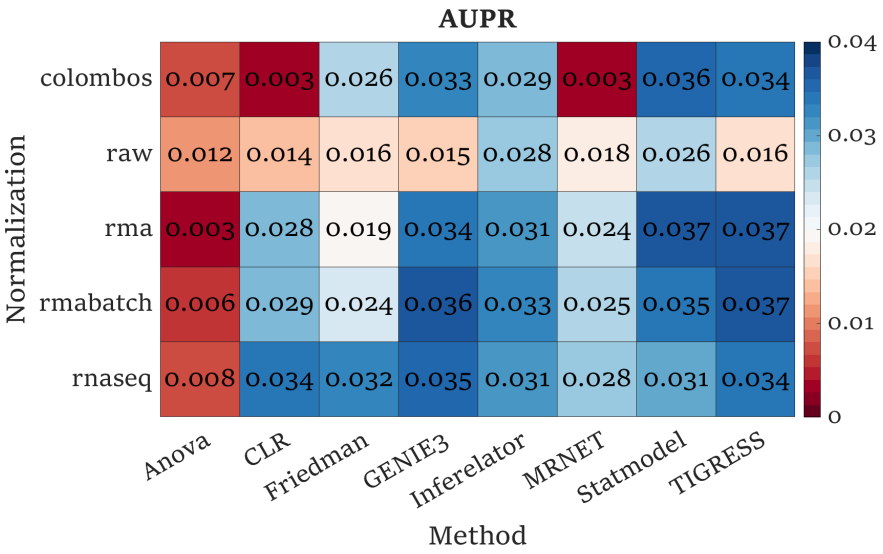


Figure 3.2: AUPR for inference by transcriptomics with different gene expression datasets.

According to the AUPR, TIGRESS performed better, followed closely by GENIE3 and Inferelator. According to the AUROC Stamodel performed better, follow closely by GENIE3 and TIGRESS. All these methods measure the influence of the expression of a TF over the expression of a TG (see Figure 3.3). This indicated that this is a better methodology for GRN inference than the co-expression methods, where correlations among the expression of the gene are measured. The correlations might be misleading since a TF and a TG regulated by a same TF might have a high correlation, but it might not be a direct interaction among them. Nevertheless, all methods have a very similar performance, which becomes difficult to assure which methods performed the best. This might be to the very small size of the GS, which only has 387 interactions, corresponding to ∼1.6 % of the 23908 regulatory interactions expected in the complete regulatory network of *S. coelicolor*. However, using the meta-curated GRN *Curated_FL-DBSCR-RTB* as GS could produce spurious results in the evaluation since indirect regulatory interactions are not adequate to assess causal interactions. Therefore, since it is not possible to properly discern which inferred GRN, is the most accurate; we decided to build a community network (see Section 3.2.3), to which we will be referring to as *Inferred_Exp*.

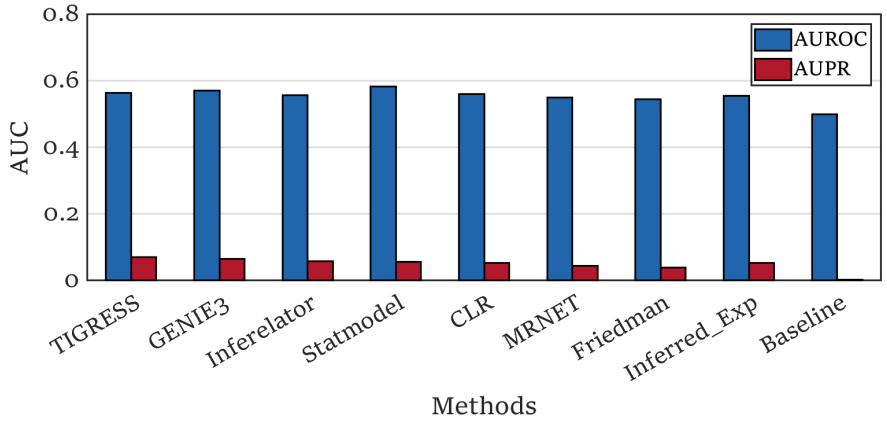
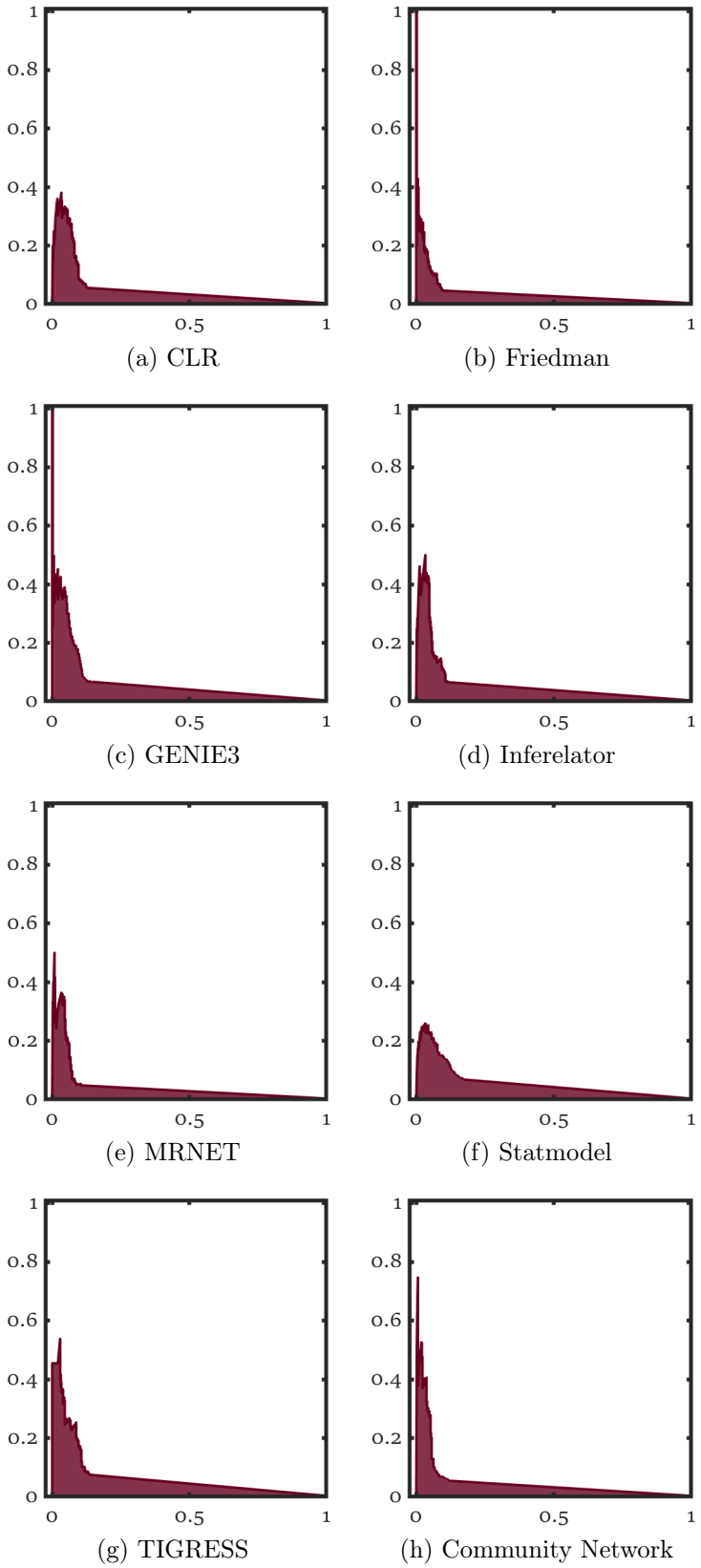


Figure 3.3: AUPR and AUROC for each method of inference by transcriptomics.

From the PR curves (see Figure 3.4) we see that for the methods Friedman, GENIE3 and somehow for TIGRESS, the highest precision of curves is at the beginning, which are the interactions in the top of the list. Meanwhile, for the others, the curve starts at low precision and increases after. Some have a more regular tendency such as CLR and Statmodel, while others have a more irregular form like MRNET. This is directly related to the scoring process of the methods. A good scoring will place the TP interactions at the top of the list. Nevertheless, these curves show us that the TP are scattered across the whole list of inferred interactions. On one side, when pruning the GRNs, we might be losing many TP interactions. On the other side, as the proportion of FP interactions is increasing the precision is decreasing. Thus, an inferred network with TP interactions at the top of the list will have high values of precision (see Equation (3.24)). Meanwhile, the same TP interactions in other positions in the list will produce low values of precision, which will affect the final AUPR. Therefore, a proper scoring process is highly relevant in the GRN inference methodology. In the case of the community network, the curve has a descending tendency, which indicates that this integration is a good filter and reordering of the TP interactions (see Figure 3.4h).



(a) CLR (b) Friedman

(c) GENIE3 (d) Inferelator

(e) MRNET (f) Statmodel

(g) TIGRESS (h) Community Network

Figure 3.4: Precision-Recall (PR) curves for the inference by transcriptomics.

4

GENE REGULATORY NETWORK INFERENCE FROM GENOMICS

4.1 INTRODUCTION

Previously, we reconstructed a complete GRN from regulatory interaction inferred computationally from gene expression data. However, given what seems a low goodness-of-fit between the curated and the inferred GRN, we chose to study a different approach in the search for a more reliable inferred GRN. To clarify, we are not certain that the low goodness-of-fit of the inferred network is due to the poor performance of the inference methods, it could be also because of the incompleteness of the GS applied in the evaluation. Thus, the inference of a GRN with a different methodology and data could also shed some light on the deficiencies in the methods for inference of GRNs. In this case, we select to reconstruct a GRN for *Streptomyces coelicolor* A(3)2 from genomic data, which means performed a regulon extension through sequence motif discovery [99]. A sequence motif is a short sequence pattern of nucleotides in the DNA that is recurring through the whole genome. They are assumed to have a biological function, usually related to protein binding sites, although others are related to processes at RNA level [100]. Here, we are focused solely on sequence-specific binding sites for TF. The consensus sequence of the Transcription Factor-Binding Sites (TFBSs) is described through a Position Frequency Matrix (PFM), which is a representation of how often each base appears in each position of each sequence [100]. Then, if we normalized the count by the number of sequences, we would have a Position Probability Matrix (PPM), which shows us the probability of each base in each position. Therefore, from the curated network, we separate the TGs for each of the TFs. For each one of these regulons, we predict the sequence motif for the TFBSs and compute their PPM with diverse *de novo* motif discovery tools (see Figure 4.1) [101]. Then, using Find Individual Motif Occurrences (FIMO) as a motif scanning tool, we compute a Position Weight Matrix (PWM) and with it we search new TGs with a similar TFBS through the whole genome [102].

4.2 METHODOLOGY

4.2.1 *Sequences Retrieval*

As an initial step, we retrieved the sequences of the upstream and the initial part of the coding regions of all genes in the genome of *Streptomyces*

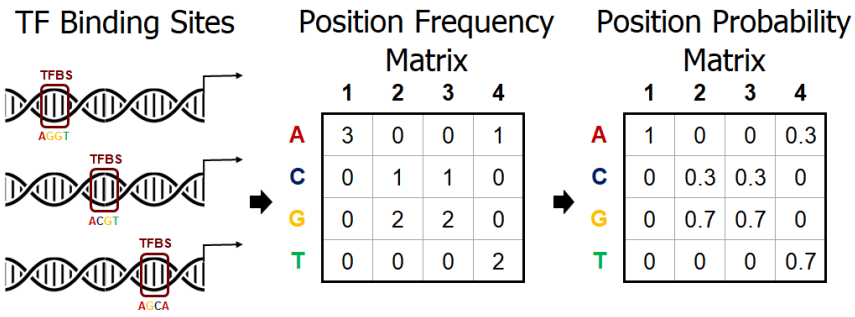


Figure 4.1: Motif representation as a PPM.

coelicolor. This is since activators usually bind upstream of the gene promoter, while repressors bind overlapping the promoter to prevent the binding of RNA polymerase. Thus the TFBS of repressors might slightly overlap with the coding region [103]. These sequences were obtained from the RSAT "retrieve sequence" online tool[1] [104], as a specific region of each gene from the genome of *S. coelicolor* reported in the GenBank[2] [81]. There we selected to retrieve the non-overlapping region -300 bp to +50 bp, which are the limits usually applied for bacteria [105], [106]. These values are the base pairs from the position of the star of the ORF, which means that the first nucleotide of the start codon is the coordinate 0 (see Figure 4.2). Non-overlapping means that the sequence obtained does not take information from other genes, thus, if the intergenic region is less than 300 bp, we take only the intergenic region.

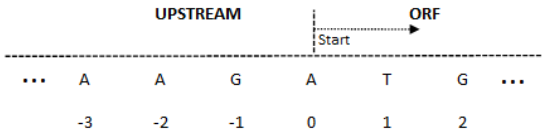


Figure 4.2: Upstream Sequence. Modified from RSAT website [104].

4.2.2 *Motif Discovery*

For the motif discovery, we took as a starting point the network *Curated_FL(cS)-DBSCR(S)* since we are certain of the physical binding of the TFs to their TGs. For each one of the TFs present in the network, we selected the upstream regions of its TGs. Then, we attempt to find *de novo* the motifs present in these upstream regions. We assumed the motif to be the binding site sequence of the TF selected. TFBSs are usually between ∼12 to 30 bp; thus, finding a motif of this size among the selected genes,

1 http://embnet.ccg.unam.mx/rsat/
2 https://www.ncbi.nlm.nih.gov/genbank/

where for each one we have 350 bp is not a straightforward task. Since there is not an established methodology for motif discovery, we decided to apply three different tools with different methodologies. We selected the tools based on the following criteria: First, we look for tools that were public, well-documented, and that they need only sequence and no other information such as gene expression data or phylogenetic relationships. It was also important that the tools could be used from the command line since it would facilitate the automatization of the process. Moreover, we selected the tools with the best performance in the best-curated microorganisms in previous works performed at the FreyreLab. The tools finally selected are the following:

- Multiple EM For Motif Elicitation (MEME) looks for motifs by searching for sites in the input sequences that are remarkably similar to one other site or more, applying the Expectation-Maximization (EM) algorithm. MEME searches for the most "significant" motifs, where significance is determined by the length of the pattern, the number of times that appear, and the degree of similarity among all the appearances. To give a measurable value of significance, MEME employs a statistical objective function based on the information content of the motif. MEME gives each motif an E-value, which is an approximation of how many motifs will be found by chance if the basis in the input sequences were shuffled. Thus, a small E-value indicates that the motif has a very low probability of being a random sequence. [107]

- BioProspector applies Gibbs sampling technique to look for regulatory sequence motifs in the upstream region of genes. BioProspector employs zero to third-order Markov background models, with user-supplied parameters or parameters inferred from a sequence file. The significance of each motif discovered is determined using a Monte Carlo method to estimate a motif score distribution. BioProspector also modifies the motif model used in previous Gibbs samplers to allow gapped motifs and motifs with palindromic patterns. [108]

- Motif Discovery Scan (MDscan) combines two commonly used motif search techniques, word enumeration and position-specific weight matrix updating. Since these sequences have a higher signal-to-noise ratio, the assumption is to look for similar terms in sequences more likely to include the motif first. Words in each similarity group can start a PPM, which can then be updated and refined using the entire input sequences. Even though MDscan aims to find motifs in ChIP experiments, the authors state that can be applied to a group of sequences that can be hypothesized to have abundant motif sites. [109]

4.2.3 *Motif Scanning*

As a motif scanning tool, we decided to use FIMO, since it presented the best results in other bacteria in previous works performed at FreyreLab. FIMO take as input motifs represented as PPMs. Then transformed the PPM into a PWM computing its elements as log-likelihoods using a background model specific to the microorganism. The background is the expected proportion of each nucleotide in the genome. For the case of *S. coelicolor* as it has a high proportion of GC (72%), the background is A: 0.14 T:014 G:0.36 C:0.36 instead of the standard A:0.25 T:0.25 G:0.25 C:0.25. The score of a match sequence is computed as the sum of the values in the matrix for each nucleotide in each position. These scores are transformed in *P-values*, assuming a zero-order null model in which sequences are generated at random according to a background specific to the microorganism. This p-value is the probability that a random sequence will have an equal or better score than the match sequence. Then a q-value is computed, which is defined as the false discovery rate, in the case the match sequence is cataloged as significant. [102]

Therefore, we introduce the PPMs in FIMO along with the upstream regions of all the genes. This is to find possible match sequences. BioProspector and MDscan have as an outcome a list of sites that correspond to the motif of the TF. In this case, we applied the sites2meme tool from the MEME Suite [110] to transform the list of sites into a PPM. We took as the final match sequence those with a p-value smaller than the threshold $1e^{-4}$.

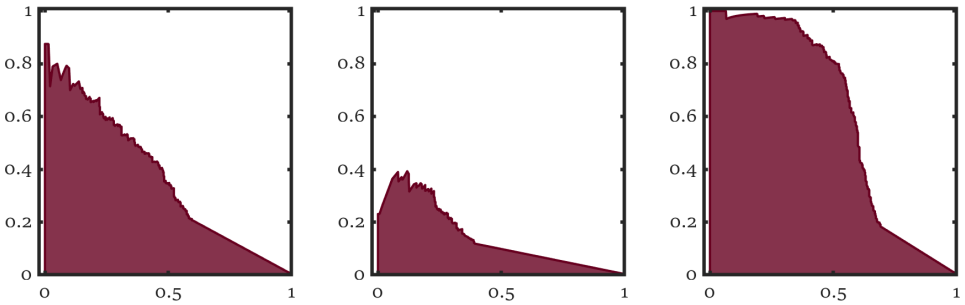
4.2.4 *Network Reconstruction and Inference Assessment*

Next, we take the downstream operon of the upstream region with the match sequences. In the case the match sequence is in an interoperon region, we take only the downstream genes. Thus, assuming that the motif predicted corresponds to the binding site of the TF studied, we link these genes to the TF. We consider as the confidence score of the interaction the p-value computed by FIMO. We merged all the interactions in a list ranked by the confidence score. As in the previous chapter, we reconstruct a community network from the three methods for motif discovery, applying the Borda approach (see Section 3.2.3). Then we prune the list to the expected regulatory interaction for *S. coelicolor* (23908) (see Section 3.2.4). Finally, with this list of interactions, we reconstruct the inferred network by binding site prediction for *S. coelicolor*. Then we proceed to compute the AUPR and the AUROC (see Section 3.2.5) of the three methods along with the community network to assess the performance of the inference by binding sites prediction.

4.2.5 *Statistical validation of ChIP data*

The sequence motif analysis, like the one performed in this chapter, can be used to pinpoint the precise position in the potential target regions obtained by ChIP. Moreover, trough the corroboration of the presence of a TFBS, we can validate the interactions obtained by ChIP as "strong" interactions instead of "weak" as defined in the curation (see Section 2.3.1.3). [36]

4.3 RESULTS



(a) BioProspector (b) MDScan (c) MEME

Figure 4.3: Precision-Recall (PR) curves for the inference by genomics.

For the GRN inference from genomics, we performed a regulon reconstruction, through the *de novo* prediction of TFBSs and linked them to downstream genes. The regulon reconstruction was based on the network *Curated_FL(cS)-DBSCR(S)* using three methods for motif discovery: MEME, Bioprospector, and MDscan (see Section 4.2.2). Moreover, we applied FIMO as the scanning method over the genome. As in the case of the inference by transcriptomics, we restricted the inference to interactions among genes of the chromosome. Next, we assessed the inferred networks considering again *Curated_FL(cS)-DBSCR(S)* as the GS (see Section 3.2.5). From the AUPR, it is evident that in general GRN inference from genomics performed better than the inference from transcriptomics (see Figure 4.3). MEME performed better than the other methods and Bioprospector also had a good performance. Meanwhile, MDscan did not perform well considering the results of the other two methods. Nevertheless, as the GS was used as prior for the regulon extension, it might provide an advantage for the network inferred by genomics. Moreover, because of its methodology, inference by genomics predicts direct regulatory interactions. Thus, there is a higher probability that these interactions are present in the GS, as it is formed only by direct interactions. Meanwhile, the inference by transcriptomics predicts, apart from direct regulatory interactions, also indirect ones, which are

not present in the GS. Therefore, it is expected that the inferred GRN by genomics have a higher AUPR than by transcriptomics. However, due to the size of the GS, we are not completely confident of these results. Thus, we decided to also reconstruct a community network, which will be designated as *Inferred_BSs*.

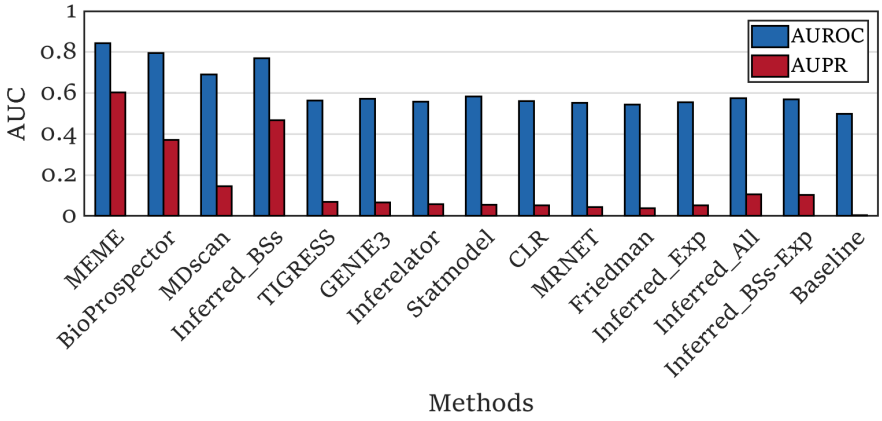


Figure 4.4: AUROC and AUPR for all inferred and community networks.

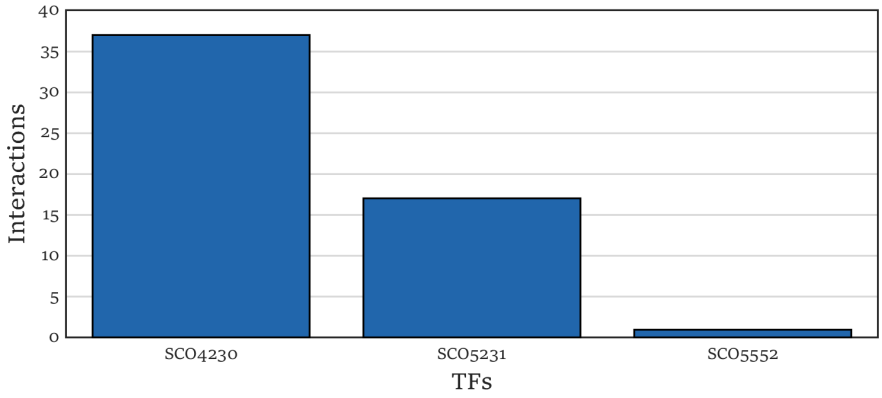


Figure 4.5: Statistically validated interactions from ChIP experiments by TF.

As an approach of integration of both inference methodologies, we decided to reconstruct another two community networks. The first one, *Inferred_BSs-Exp*, is a community network of both networks *Inferred_Exp* and *Inferred_BSs*. The second one, *Inferred_All*, is a community network of all individual inferred networks. We considered only three inferred networks from transcriptomics to balance the number of networks from both methodologies. We selected the three methods with the better performance, considering the AUPR and the AUROC. These methods were: GENIE3, Statmodel and TIGRESS. Then we also assess these community networks with the same GS (see Figure 4.4). *Inferred_BSs* outperformed the rest of the community networks at both AUPR and AUROC. Meanwhile, both integration communities present

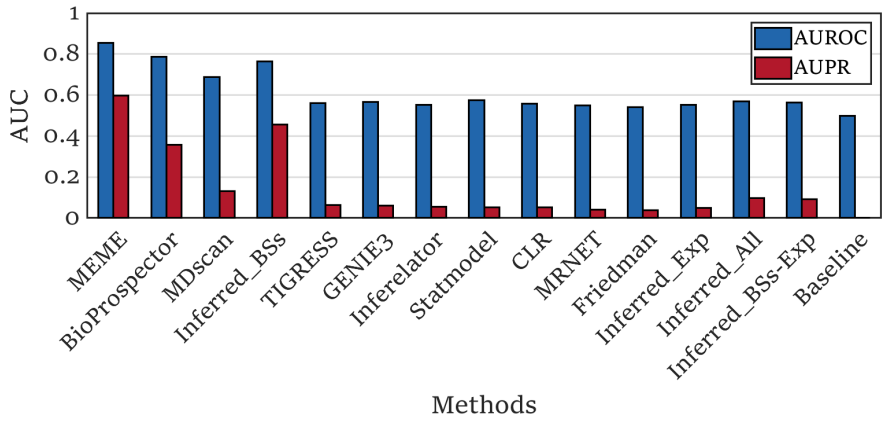


Figure 4.6: AUROC and AUPR for all inferred and community networks.

very similar results in both metrics. However, MEME outperformed all inferred and community networks at both metrics.

Given the outstanding performance of MEME, we applied it to perform a statistical validation of "weak" interactions supported by ChIP-data (see Section 4.2.5). A total of 55 "weak" interactions were reclassified as "strong" (see Figure 4.5 and supplementary file Table 7c). PhOP (SCO4230) was the TF with the most interactions validated. One of these interactions (SCO4230-SCO4878) was already reported as "strong" in the DBSCR database (*Curated_DBSCR(S)*). These statistically validated interactions were added to the networks *Curated_FL(cS)* and *Curated_FL(cS)-DBSCR(S)* to reconstruct the *Curated_FL(S)* and *Curated_FL(S)-DBSCR(S)*. We reassessed the inferred network predictions with *Curated_FL(S)-DBSCR(S)* as GS and the results remained virtually the same (see Figure 4.6).

5

ASSESSMENT OF THE INFERRED GENE REGULATORY NETWORKS

5.1 INTRODUCTION

In previous chapters we performance an GRN inference from transcriptomic and from genomics. The AUROC and the AUPR metrics revealed that the inference from genomics performed better. However, there are some aspects related to this assessment that we should consider. First, these metrics depend primarily on the raking of the inferred interactions. Where a high rank for a false positive will heavily penalize the final metric score. As we have such a limited GS with $\sim 2\%$ (480/23908) of the final interactions, we have a very large proportion of missing interactions from the GS taken as false positive. Moreover, the size of the GS itself causes the assessment to be not very trustworthy. Therefore, we decided to assess the structural properties of the inferred GRNs, comparing them to the ones from the curated networks. This aimed to complement the initial assessment, evaluating which inferred GRN possess the most similar structural properties of a biological network.

5.2 METHODOLOGY

5.2.1 *Cluster Analysis of Structural Properties*

Additionally, to the structural properties presented in Chapter 2, we compute other properties, for both curated and inferred networks. The properties considered were:

- Percentage of Regulators: Percentage of genes regulating other genes (TFs).
- Directed regulatory interactions: All the interactions in the network since GRNs are directed graphs.
- Self-regulations: Genes regulating themselves.
- Maximum out-connectivity: The largest number of genes regulated by one regulator.
- Network density: See Section 3.2.4.
- Average short path length: See Section 2.2.2.
- Network diameter: The longest of the shortest path lengths [28].

- Average clustering coefficient: See Section 2.2.3.
- Weakly connected components: A directed subgraphs where all the genes are connected by undirected edges [111].
- Genes in the giant component: The largest connected subgraph, where all the nodes have a path that connects them to each one of the other nodes. In a graph, some subgraphs might be disconnected from the rest of the nodes.
- Feedforward circuits, complex feedforward circuits, and 3-Feedback loops: Network motifs (small subgraphs repeated at a higher frequency than random ones) usually found in biological networks. These motifs are often related to biological functions. [28], [31]
- P(k) R^2adj: The R^2 of the linear regression. See Section 2.3.3.1.
- C(k) R^2adj: The R^2 of the linear regression. See Section 2.3.3.1.

Then we build a clustered map of the vector of properties for each network. This is represented as in a heatmap with dendrograms (see Figure 5.1). This since, due to the number of properties and networks, this type of representation allows us to have a clearer a more straightforward picture of the whole data. The dendrogram is built through hierarchical clustering. There the networks are assembled in a tree, where networks with a similar vector of properties are linked by short branches, and as the similarities decrease, the length of the branches increase [112]. First, we compute the pairwise Pearson correlation among the vectors of structural properties, and this value is represented in the heatmap. The networks were then clustered based on their Euclidean distance applying as linkage method Ward's method.

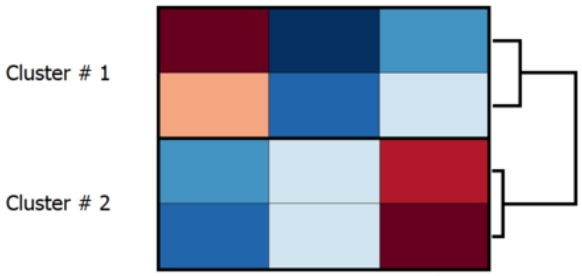


Figure 5.1: Example of a clustered map.

5.2.2 *Network Dissimilarity*

In addition to the previous analysis, we also compute the dissimilarity measure (D) proposed by Schieber *et al.* [113]. This measure identifies and

quantifies the differences in the structural topology among networks. This is done through the comparison of the probability distribution functions of the connectivity distance of the nodes. There are three distances considered in a three-term function: a) the distance distribution of the network, which captures the global topology; b) the distance distribution of each node, to compare the connectivity of each node; and c) the centrality of the nodes, to analyze how this connectivity occurs. This last term is not indispensable, but it is necessary in the case of comparing networks of different sizes (in terms of nodes). This was computed with the parameters proposed by the authors (0.45, 0.45, 0.1) and then clustered in the same way that the Pearson correlation of the vector of properties.

5.2.3 *Natural Decomposition Approach*

To complement the topological comparison of the inferred GRNs, we classified the genes of all networks in the four categories of the NDA (see Section 2.2.4): global regulators, modular genes, intermodular genes, and basal machinery. Next, we compared the sets of each category among networks applying the Simpson's similarity index [114], [115] (see Equation (5.26)). This, as its name state, measures the similarity between two sets, telling us if their composition is similar concerning the smallest set. This means that if one of the sets is a subset of the other one, we will obtain a score of 1, and from two different sets, we will obtain a score of 0. It is important that it is measured concerning the smallest subset since curated networks are way smaller than inferred ones. Therefore, the subsets are all different sizes. Other indexes, such as Jaccard measure if all the elements of one set are presented in the other one. Thus, in this case, these indexes might not be informative, since it is not possible to have a complete set of an inferred network in a curated one. These pairwise indexes were also clustered as in both previous analyses.

$$
\begin{aligned}
SI &= \frac{a}{\min(S_1, S_2)} \\
&\text{where} \\
a &= \text{Elements present in both sets} \\
S_1 &= \text{Size of set 1} \\
S_2 &= \text{Size of set 2}
\end{aligned}
\tag{5.26}
$$

Additionally, we assessed the global regulators inferred by the NDA in the curated and inferred networks based in the literature. As GS for the assessment, we considered the regulators reported as global or pleiotropic. We based the GS in the review published by Martín *et al.* [22] and complemented it with regulators reported in individual publications. To

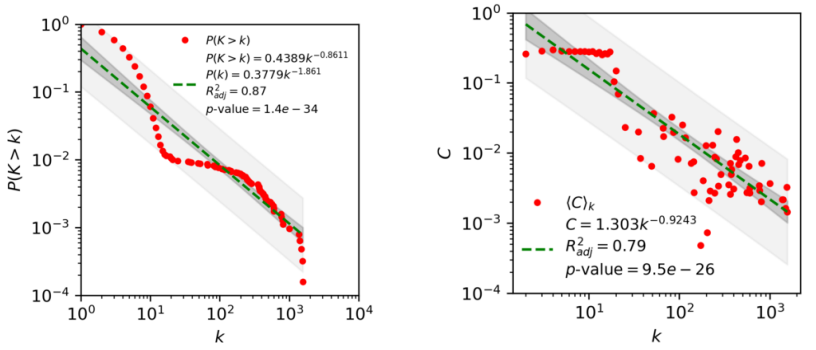
perform the evaluation, we compute the precision (see Equation (3.24)), Matthews Correlation Coefficient (MCC) (see Equation (5.27)) and F_1-score (see Equation (5.28)) of the global regulators inferred for each network [116]. These metrics are based on the classification presented in Section 3.2.5.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{5.27}$$

$$F_1\text{-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{5.28}$$

5.3 RESULTS

5.3.1 *Assessment by their Structural Properties*

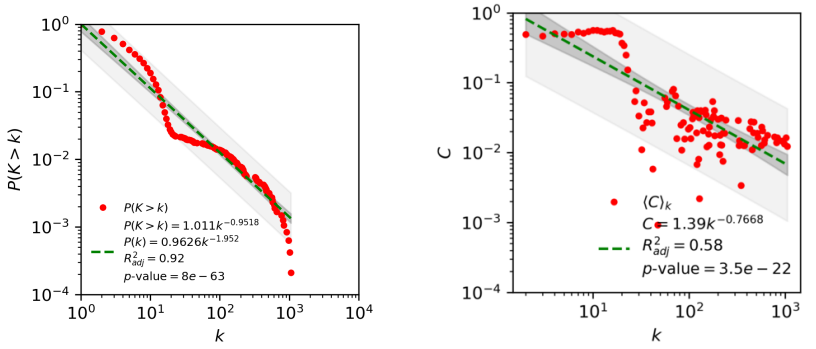


(a) Network degree distribution

(b) Network clustering coefficient distribution

Figure 5.2: Cumulative distribution of the network node degree ($P(K)$) and clustering coefficient ($C(K)$) of the inferred network *Inferred_ BSs*.

Even though the AUPR and AUROC metrics allow the assessment of the predictions, both metrics heavily rely on the ranking of the predicted interactions. Moreover, the GS is not complete and missing interactions would be still classified as false positives, decreasing the score more as higher their ranking is. Therefore, we also decided to assess the inferences in terms of their structural properties and compared them against the curated networks to compensate for such drawbacks. Note that this approach has its caveats. The global structural properties of the network might be different once the GS is complete, this can be approached by comparing the predictions to all the curated networks,



(a) Network degree distribution

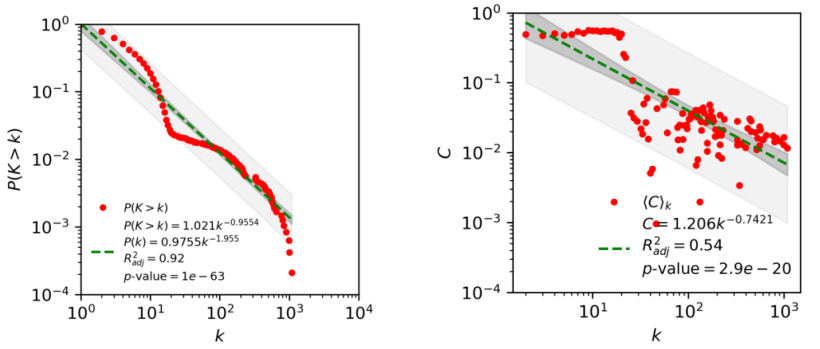
(b) Network clustering coefficient distribution

Figure 5.3: Cumulative distribution of the network node degree ($P(K)$) and clustering coefficient ($C(K)$) of the inferred network *Inferred_Exp*.

each of them with different completeness. Also, two networks could have the same topology with different node entities. For this reason, we use the topological assessment in complement with the AUPR and AUROC metrics to identify the best prediction.

One of the main characteristics of biological networks is that they are scale-free and hierarchically modular. Same characteristics that our curated networks have been proved to possess (see Section 2.4.2). Therefore, as an initial approach, we asked whether the inferred networks are also scale-free. First, to compute the α of the degree distribution for the inferred networks (see Section 2.3.3.1), we performed a robust linear regression over a log-log plot of the complementary cumulative degree distribution and corrected the exponent accordingly (see Figures 5.2 to 5.5). All degree distribution seems to follow a power law according to the adjusted coefficient. Nevertheless, the data points in that *Inferred_All* appear to be divided into three regions with different tendencies, instead of the two that are present in the other networks. Usually, this type of network is divided into two regions, the region of the nodes (genes) with a low degree, and the one with nodes with a high degree (see Section 2.3.3.1). The appearance of a third region might be a consequence of merging networks inferred with methods with different approaches. This could affect the structural properties of the merged networks, while communities from the same approach appear to have more similar structural properties. In the case of *Inferred_BSs-Exp*, the construction of communities ahead by each approach creates more compatible networks in terms of structure that can be conveniently mixed.

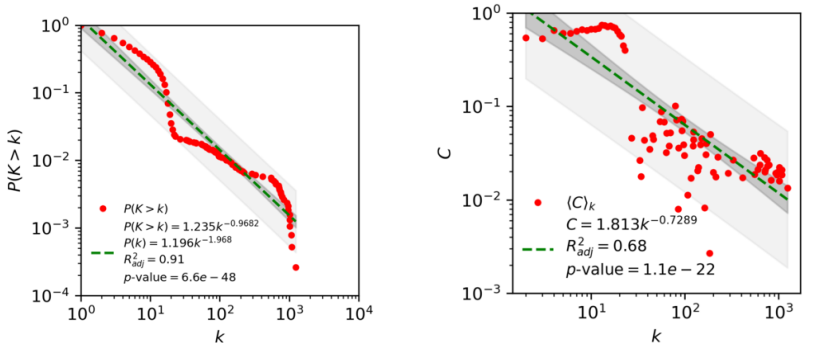
Next, we wanted to confirm that their degree distribution followed a power law, so we performed a KS test between each potential power law



(a) Network degree distribution

(b) Network clustering coefficient distribution

Figure 5.4: Cumulative distribution of the network node degree ($P(K)$) and clustering coefficient ($C(K)$) of the inferred network *Inferred_ BSs-Exp*.



(a) Network degree distribution

(b) Network clustering coefficient distribution

Figure 5.5: Cumulative distribution of the network node degree ($P(K)$) and clustering coefficient ($C(K)$) of the inferred network *Inferred_ All*.

of the inferred networks and alternative fat-tailed probability distributions (see Section 2.3.3.1). The values can be found in the supplementary file (Table 8a). We found that the degree distribution of the inferred networks adjusted better to a power-law distribution than to an alternative distribution. Then, we computed a maximum likelihood estimation for the exponent of the power law and found that most of them are between two and three, except for *Inferred_ All*. This shows that it is an anomalous scale-free network (Supplementary file Table 8b), perhaps due to the mixing of networks with diverse structural properties. Nevertheless, we could consider all inferred networks to be scale-free. However, we wanted to check the other properties of scale-free networks. All the properties can be found in the supplementary file (Table 9). The four community networks have small average shortest path lengths

and a high clustering coefficient. *Inferred_BSs* has the smallest average short path length, while *Inferred_All* has the highest average clustering coefficient. Scale-free networks also present an ultra-small world effect, which implies that the average path length is proportional to $\ln(\ln(N))$ (see Section 2.2.2). This is the case for all the inferred networks. Another characteristic of GRNs is their hierarchical modularity. In a scale-free network, this implies that the clustering coefficient depending on the degree follows a power law whit a coefficient close to 1 (see Section 2.2.3). *Inferred_BSs* has the exponent closest to -1 (0.92), with the best R^2. Even though *Inferred_BSs* seems to be the network that behaves closest to a GRN, all networks have similar values, which makes it difficult to discern the most reliable inferred network.

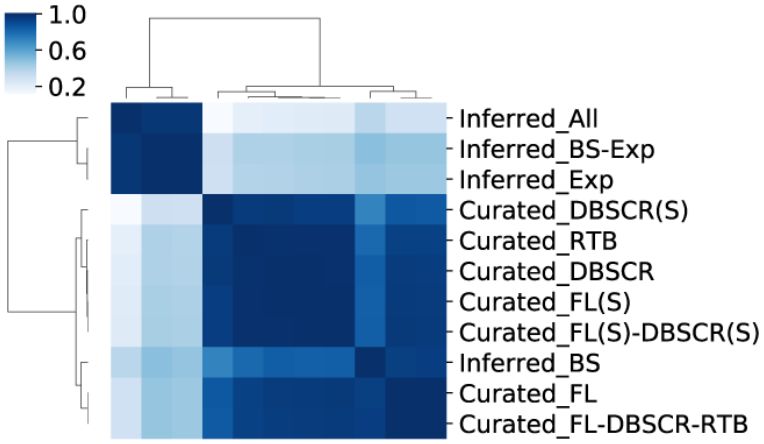


Figure 5.6: Cluster map of pairwise Pearson correlation coefficient of the profile of structural properties.

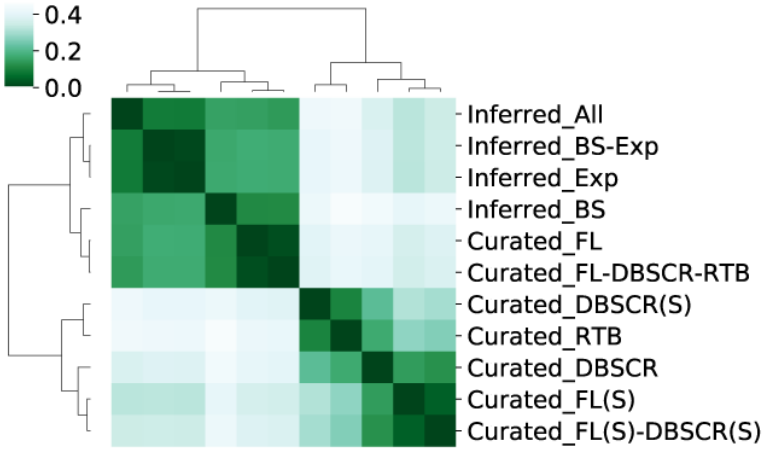


Figure 5.7: Cluster map of pairwise dissimilarity measure (D) of the networks.

To perform a more thorough comparison of their structural graph properties, we include several others. We clustered the vectors of structural properties for the curated and community-inferred networks (see Figure 5.6). The clustering partitions the networks into two major groups. The first one contains the curated networks and *Inferred_BSs*, while the second group contains the other inferred networks. The first group is in turn also divided into two groups: one with the two largest curated networks and *Inferred_BSs*, and the other one with the remaining curated networks. The reason for this may be due to the size of the networks (see Table A.1).

To reduce the network size influence we used the network dissimilarity measure (see Section 5.2.2). We considered the third term which makes the distance measure robust to graph size in terms of the number of nodes (genes) (see Figure 5.7). Even with this metric, the two largest curated networks were clustered with the inferred networks. This might be a consequence of the high fraction of maximum out-connectivity and structural genes in the largest curated networks, like those found in the inferred networks. This shows that the inferred networks are similar to the most complete curated networks in terms of structure, suggesting their reliability.

5.3.2 *Assessment by their Natural Decomposition Approach components*

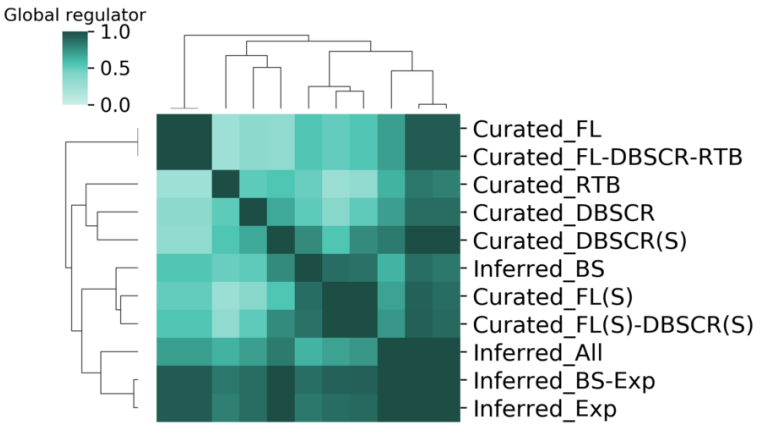


Figure 5.8: Cluster map of the pairwise Simpson's similarity index of the GR.

We compared all curated and community inferred networks based on Simpson´s similarity index of the four components proposed by the NDA: global regulators, modular genes, intermodular genes, and the basal machinery (see Section 5.2.3). The number of genes predicted in each category can be found in the supplementary file (Table 10). When comparing the global regulators (see Figure 5.8), there is not a distinct division among the networks. *Inferred_BSs-Exp* and *Inferred_Exp*

have a similar correlation with all the curated networks, slightly higher with *Curated_ DBSCR(S)*, *Curated_ FL*, and *Curated_ FL-DBSCR-RTB*. These two inferred networks have the highest amount of GR, 116, and 114 respectively; thus, the other GRs predicted could be easily a subset of them. In the case of *Inferred_ BSs*, it has the highest correlation with the "strong" networks. This is expected since these networks have only direct regulatory interactions, as the interactions predicted in *Inferred_ BSs*, while there is no evidence of direct regulation for transcriptomic-based inferred interactions. This can affect the measurement of the effect of GR over the rest of the genes since the GRs which regulates a few targets from different processes might not be predicted as such in the "strong" networks. However, when their indirect influence is represented in the network, their ranking as GRs is noticeable.

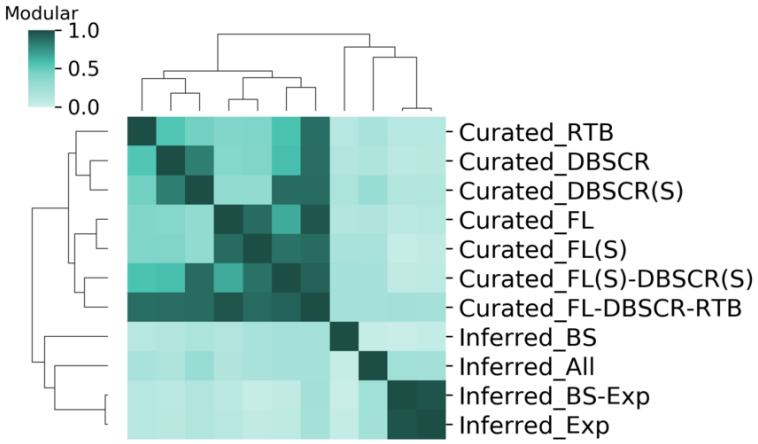


Figure 5.9: Cluster map of the pairwise Simpson's similarity index of the modular genes.

When analyzing the modular genes, there are two major groups (see Figure 5.9): the major group, with the curated networks, is divided into two subgroups, one with *Curated_ RTB*, *Curated_ DBSCR*, and its "strong" version *Curated_ DBSCR(S)*, and the second subgroup contains the integration proposed in this work. Interestingly, the meta-curated network *Curated_ FL-DBSCR-RTB* correlates very well with all the networks it contains, from which we could deduce that modular genes are conserved despite the addition of new regulatory interactions. In the second group, composed of the inferred networks, we can see there is not a high correlation among them. *Inferred_ BSs*is the closest to the curated networks, while *Inferred_ Exp* and *Inferred_ BSs-Exp* have a high correlation. This tells us that the interactions in *Inferred_ Exp* have a larger influence on the module configuration of *Inferred_ BSs-Exp* than *Inferred_ BSs*. The difference between the curated and inferred networks might come from the fact that inferred networks have a greater number

of GRs and a much lower number of modular genes when compared with the curated networks.

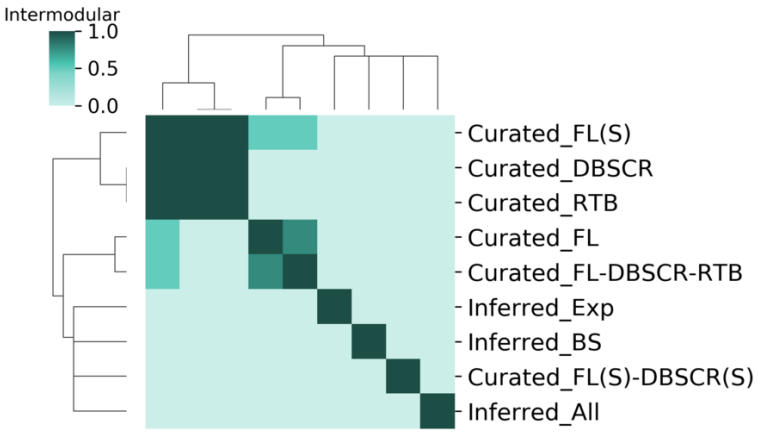


Figure 5.10: Cluster map of the pairwise Simpson's similarity index of the intermodular genes.

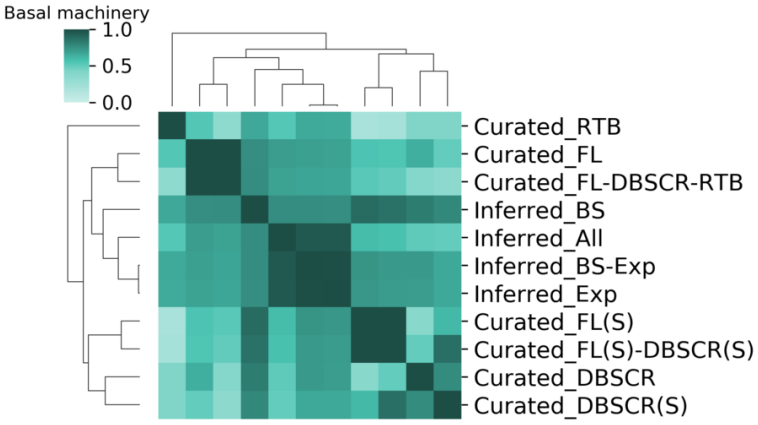
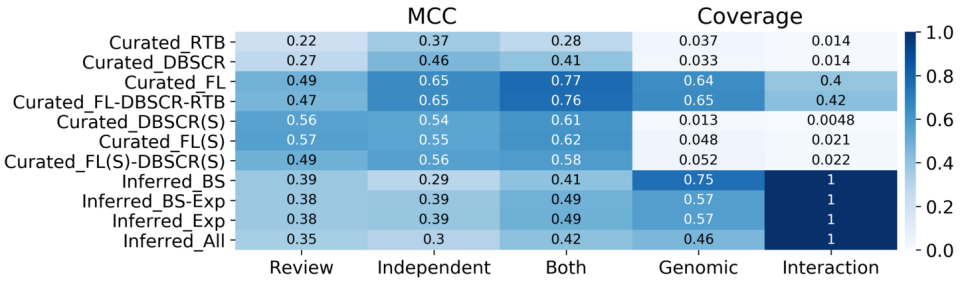


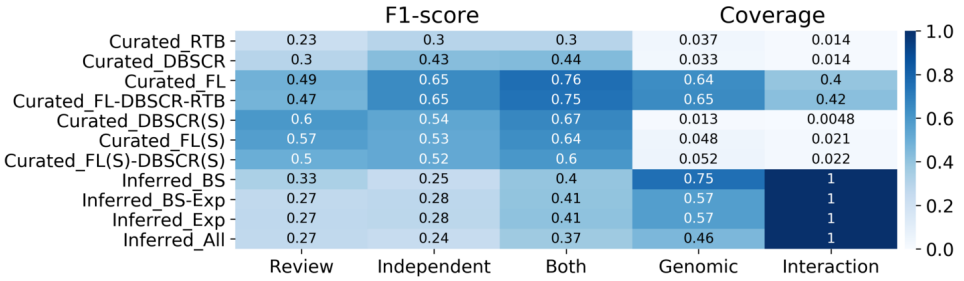
Figure 5.11: Cluster map of the pairwise Simpson's similarity index ofthe basal machinery.

Intermodular genes are the less conserved NDA class (see Figure 5.10). There is overlap only among the smallest curated networks, all share the intermodular de gene SCO5877, which appears as a TF in the other curated networks. Moreover, an overlap among *Curated_ FL* and *Curated_ FL-DBSCR-RTB*, which share most of the interactions. Thus, is expected that they also share most of the intermodular genes. Note that the networks *Curated_ DBSCR(S)* and *Inferred_ BSs-Exp* are not included in the clustering since they did not present any intermodular genes. Finally, when analyzing the basal machinery (seeFigure 5.11), the larger curated networks are grouped on one side, next to the inferred networks, and finally the smallest curated networks with *Curated_ RTB* as an outgroup. Even though *Inferred_ BSs* is grouped with the other

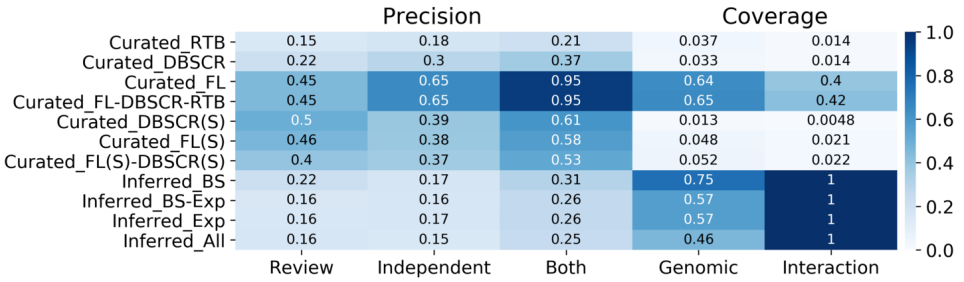
inferred networks, it has a higher correlation with *Curated_FL(S)* and *Curated_FL(S)-DBSCR(S)*, which again evidence the similarity among these three networks.



(a) MCC



(b) F_1-Score



(c) Precision

Figure 5.12: Assessment of the Global Regulators predicted by the NDA.

Because of the NDA algorithm, the identification of global regulators is a key step in the classification of every node in the GRN. Previously, it has been reported a high overlap between the predictions of global regulators by the NDA and those reported in the literature for *E. coli* [31], *B. subtilis* [117], and *C. glutamicum* [32]. We used the set of GRs reported by Martín *et al.* [22], besides those reported in independent articles, and the union of both sets. This can be found in the supplementary file (Table 4). Then we assessed the predictions of the GR using the MCC (see Figure 5.12a). We used the MCC score as it is more informative and reliable than F_1-score for binary classification evaluation [116] but using the F_1-score and the precision we obtained consistent results (see Figure 5.12b). *Curated_FL* and *Curated_FL-DBSCR-RTB* have the

best performance in GR prediction. However, the "strong" networks have a slightly smaller score even having much less genomic coverage. This shows that the GRs are very robust to perturbations in the network as previously shown [31]. On the other hand, despite the high coverage of the inferred networks, their performance of the predictions was poor. This could be, as it was mentioned before, due to the great amount of GRs predicted by these networks, which would cause a high proportion of false positives, affecting the score. *Inferred_BSs* produced the most conservative prediction (lowest false-positives rate) among the inferred network (see Figure 5.12c).

6

BIOTECHNOLOGICAL APPLICATION OF INFERRED GENE REGULATORY NETWORKS

6.1 INTRODUCTION

The GRNs presented in the previous chapters have multiple applications in different fields. Apart from the results already presented where we suggested annotations, NDA component characterization, and interactions for different genes, we decided to apply the meta-curated and the inferred network to demonstrate their applicability. First, we performed a comparison of the meta-curated network *Curated_FL(S)-DBSCR(S)* against a curated network of *Corynebacterium glutamicum*. There we complete the curated networks with the curated interactions of the other ones. Moreover, we compare the NDA characterization of orthologs presented in both networks. Second, from the inferred network *Inferred_BSs*, we found genes that might regulate one or more of the SARPs, since they are responsible for the production of secondary metabolites of interest in the biochemical industry. These proposed regulators might help to elucidate novel regulation and metabolic processes of the senary metabolism. Besides, they can be targets for genetic modification, to increment the yields of these metabolites.

6.2 RESULTS

6.2.1 *Comparative analysis with* Corynebacterium glutamicum

The diamond-shaped structure identified by the NDA is conserved between *E. coli* and *B. subtilis* [117]. As an application of the meta-curated network, we studied the conservation of its system-level components, comparing it against the *C. glutamicum* network, which is phylogenetically related to *S. coelicolor*, and a model organism for the study of GRNs [32]. We applied the regulogs analysis [118] with one-to-one ortholog relationships to alleviate network incompleteness and make them comparable. As prior networks, we used *Curated_FL(S)-DBSCR(S)* (534 interactions) for *S. coelicolor* and 196627_v2020_s21_eStrong from Abasy Atlas (2941 interactions) for *C. glutamicum* [119]. considering only the interactions between two genes both mapping to a locus tag. We used MEME to construct a PWM for every TF with at least three TGs using their upstream sequences. These sequences were defined as the non-overlapping regions of up to -300 to +50 bp with respect to the translation start codon and were obtained with retrieve-seq from

RSAT (see Section 4.2.1). Then, we used FIMO with the PWM of the TFs from *S. coelicolor* to find individual occurrences with a p-value $< 1e^{-4}$ in the upstream sequences of *C. glutamicum*. The same was done in the opposite direction. With this, we seek to alleviate network incompleteness by extrapolating known interactions from an organism to the other [118]. Predicted interactions were sorted by p-value and, in the case of redundant interactions, only the best scoring result was conserved. Afterward, we used Orthofinder to find one-to-one ortholog relationships between both microorganisms. We used OrthoFinder due to its high accuracy [120]. We obtained a total of 188 GRN-wide orthologous relationships from a total of 995 1:1 orthologs. The orthologs were used to further filter FIMO predictions to conserve interactions in which both TF and TG have a one-to-one orthologous relationship in the other organism. We considered the original "strong" network interactions at the beginning of the interactions list. After the regulogs analysis, we ended up with 2966 interactions in *C. glutamicum* and 692 interactions in *S. coelicolor*.

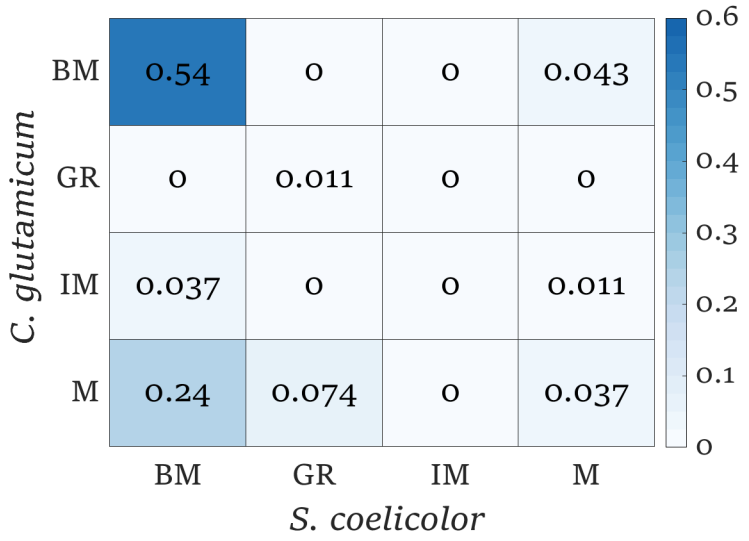


Figure 6.1: Conservation of the systems-level components between *S. coelicolor* and *C. glutamicum*.

The NDA was applied to both expanded GRNs to identify ortholog systems and only the genes with one-to-one orthologs in the other organism's network (GRN-wide orthologs) were considered in the analysis. Then, computed the fraction of the GRN-wide orthologous in each combinatory relationship between the components (see Figure 6.1). We found that most of the GRN-wide orthologous (54%) are classified as basal machinery in both microorganisms. This is expected since 73% and 74% of the genes correspond to the basal machinery in the complemented networks of *C. glutamicum* and *S. coelicolor* respectively. Moreover, the distribution of the genes in the chromosome of *S. coelicolor* shows a central core, where are genes likely related to primary functions such as

DNA replication, transcription, translation, and amino-acid biosynthesis; and likely non-essential genes such as secondary metabolism are in the chromosome arms [6]. More than 59% (111/188) of the GRN-wide orthologs conserved the same class in both microorganisms showing high conservation of the NDA classification.

Figure 6.2: Simpson similarity index of NDA between *S. coelicolor* and *C. glutamicum*.

We studied the pairwise Simpson similarity index between the four classes between the two microorganisms to remove the problem of the imbalanced classes in both microorganisms (see Figure 6.2). GRs is the class with the highest conservation rate, the orthologs of seven of the eight GRs in *C. glutamicum* are also GRs in *S. coelicolor*. The conservation between the same class in the two microorganisms is also high for the basal machinery, while poor for the modular genes. For the case of intermodular genes, even though the networks were complemented with information from the other network, they are not conserved at all. Previous work reported intermodular genes as the least conserved of the system-level components [32]. Intermodular genes are the most likely responsible for giving the GRN flexibility and increasing evolvability by scouting different combinations of regulatory interactions between physiological functions, so the organism could adapt better to environmental changes [117]. These results agree with a previous analysis of the robustness of the NDA to a random node and edge remotion showing GRs and intermodular genes as the most and least conserved classes, respectively (see Figure 8 in Freyre-González *et al.* [32]).

On the other hand, 24% of the GRN-wide orthologs that are modular genes in *C. glutamicum* were classified in *S. coelicolor* as basal machinery. This could be due to three possible reasons [117]: i) the basal machinery genes in *S. coelicolor* are misclassified and further research is needed to find the missing regulatory interactions that will integrate some of the

basal machinery genes into a module. ii) The GRs controlling *C. glutamicum* genes are not yet identified as such. iii) Genes in *S. coelicolor* need a more direct regulation because of their physiological function (high plasticity of transcriptional regulation). The previous comparison between *S. coelicolor*, *Mycobacterium tuberculosis*, and *Corynebacterium diphtheriae* showed a synteny among the whole chromosome of these last two micro microorganisms and the core of the chromosome of *S. coelicolor* [6]. *C. glutamicum* is phylogenetically closely related to *M. tuberculosis* and *C. diphtheriae*, with roughly similar genome size. Therefore, a similar result would be expected. Furthermore, as more classical experiment data become available, new regulations for the currently basal machinery would turn those genes into the modular class. However, a deeper analysis of diverse factors such as genome size, the niche of the microorganisms, and a wider range of microorganisms are required to further study the robustness of the NDA analysis.

6.2.2 *Prediction of new Trancription Factors for the most studied Streptomyces Antibiotic Regulatory Proteins*

Even though other similar studies suggest the integration of inference approaches as the most suitable methodology for GRN reconstruction [121], [122], because of the analysis performed in this paper, we consider *Inferred_BSs* as the most reliable inferred network. From the evaluation against the GS, where it presented the highest AUPR and AUROC among the community networks, through its structural properties along with the NDA analysis where it showed the most similar configuration to a GRN reconstruct from biological experiments. Moreover, it has the largest genomic coverage among all the networks, which would be advantageous for a deeper study of transcriptional regulation in *S. coelicolor*. Therefore, we decided to use *Inferred_BSs* to further study the regulation of the SARPs of the most studied antibiotics in *S. coelicolor*: ActII–orf4, *redD*/RedZ, CpkO (also known as KasO), and CdaR, which regulate the production of ACT, RED, CPK A, and CDA, respectively [23]. A total of 13 new interactions for the SARPs were predicted, providing us a great opportunity to find new targets to manipulate the *S. coelicolor* antibiotic production. Next, we describe some of the TFs predicted for the SARPs:

- For *actII-orf4* (SCO5085) only one regulator was inferred, MacR (SCO2120) which is the response regulator of the TCS MacRS. This TCS has been proved to activate ACT production. Nevertheless, a ChIP-qPCR analysis was not able to prove an *in vivo* interaction between MacR and *actII-orf4*, although a direct binding was not tested as *Inferred_BSs* predicted [123].

- For *redD* (SCO5877) two new regulators were predicted, LipR (SCO0712) and *actII-ORF4* (SCO5085). LipR is related to AfsR (SCO4426) [124], homolog to the SARPs, and activator of the ACT and RED production [125]. Moreover, its mutant affects ACT production [124], which makes it plausible to affect RED production as well. It has been suggested that *actII-ORF4* might regulate the production of other antibiotics [7], which could be by binding directly to their CSR.

- For *redZ* (SCO5881) five new regulators were predicted, among them is GluR (SCO5778) which has been shown to affect RED production. Nevertheless, it has been shown that GluR does not bind directly to *redZ*, thus it could be more an indirect regulation [126]. Another one, StgR (SCO2964) has been shown, by an RT-qPCR experiment, to be a repressor of *redD* [127], thus this repression could be through the direct binding to *redZ*. HpdA (SCO2928) and HpdR (SCO2935) are related to Tyrosine catabolism, which produces important precursors for antibiotic biosynthesis [128]. Moreover, HpdA has been shown to activate *actII-ORF4*, therefore might have a more direct role in RED production.

- In the case of *cdaR* (SCO3217), we have four predicted regulators, among them, OsdR (SCO0204) and RamR (SCO6685). Both are related to the response to stress and the development of *S. coelicolor* [129], [130]. SsgR (SCO3925) regulates the sporulation and morphological differentiation [131]. These all processes are highly related to antibiotic production.

- Finally, for *cpkO/kasO* (SCO6280) six new regulators were inferred, among them OsdR (SCO0204), LipR (SCO0712), and StgR (SCO2964) were described before. Another one is NnaR (SCO2958), which regulates spore formation and antibiotic production [132].

The complete list of predicted interactions for the SARPs can be found in the supplementary file (Table 11).

7

CONCLUSIONS AND OUTLOOK

This work has three main outcomes. First, a meta-curated GRN for *Streptomyces coelicolor* A3(2) (*Curated_FL-DBSCR-RTB*) was reconstructed from a collection and curation of regulatory interactions experiments in literature and databases. This network is the most complete up to date in terms of genome coverage and number of interactions. The size of this network allows us to have a better analysis of its structural properties and therefore of their NDA components. Thanks to it, we were able to identify 20 global regulators, of which 95% (19/20) have already been reported as global or pleiotropic regulators; 18 intermodular genes, some of them are already known for their involvement in different metabolic pathways; and 46 modules and submodules, allowing to propose the function for 79 genes without previous functional annotation. Moreover, this network helps us to complement the GRN of *Corynebacterium glutamicum* and to compare the function of their orthologs in both microorganisms.

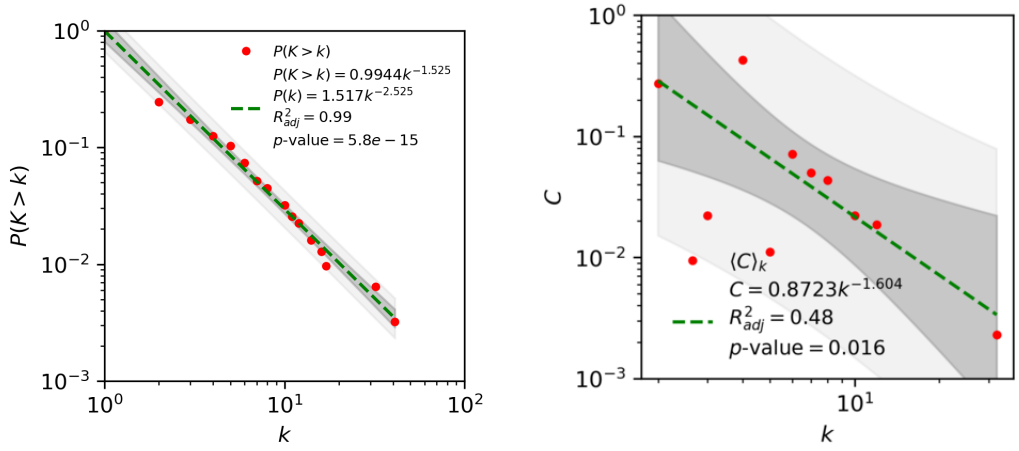
Second, we inferred a GRN applying different strategies for it. From this work, we perceive that the inference from genomics has an outstanding performance over the inference from transcriptomic data. In the inference from genomics, we based our inference on the curated network. However, we only consider the "strong" interactions as the GS of the evaluation, which were quite fewer than the complete meta-curated network. We are aware that this might present a bias in the assessment. Therefore, we believe the metrics AUPR and AUROC are helpful tools for the judgment of individual inference methods in both methodologies (genomics and transcriptomics). Nevertheless, it does not seem appropriate for the judgment between the methodologies, and between the methods of methodology integration.

And third, we found in the NDA a valuable tool for GRN inference assessment. In the first part of the work, we present that GRNs have defined structural properties. These are related to gene organization in the regulation process. Therefore, is highly important for an inferred GRN to present these same structural properties, to assure the reliability of the inferred interactions. However, a comparison of the properties did not provide proper discrimination of the inferred networks. Here is where NDA appears as an additional tool for the structural comparison, focusing on the system-level components instead of their properties. In the comparison of the NDA components among, we concluded that the inferred network from genomics (*Inferred_BSs*) has a structure closer to a GRN. Then, we applied this network to suggest new regulators for the SARPs. Most of them are known to indirectly affect antibiotic

metabolism or to be related to secondary metabolism and morphological differentiation. Therefore, this inferred network could be applied in the design of experiments in *S. coelicolor* secondary metabolism and regulation in general. Moreover, it could be used also for the modeling and computational analysis of *S. coelicolor* metabolism and regulation. Thus, this network could have many applications in diverse research fields.

This work was an initial approach to the modeling of *S. coelicolor* and in general cell regulation. There is still plenty of room, both in the complete reconstruction of a GRN for *S. coelicolor* and in the GRN inference. It is imperative the realization of classical experiments to identify "strong"interactions. As we stated before, there is only known around 2% of the direct interaction expected for *S. coelicolor*. High-throughput is also a valuable tool since provides us with a wider view of the regulatory processes. Nevertheless, some of these interactions are indirect effects, and from only these experiments it is impossible to know the actual regulation path. A larger "strong" network will result in a more proper characterization of the NDA components and a more reliable prediction of the biological function of each gene. Moreover, we would be able to infer a more accurate GRN. Also, we will be very helpful in the development of methodologies that allow the cross-validation of "weak" interaction into "strong" as the statistical validation of Chromatin Immunoprecipitation (ChIP) experiments applied here. In this way, we would be able to reconstruct the "strong" network at a higher rate. In the field of network inference, there are also plenty of things to accomplish for a proper inference. We stated the importance of a good score procedure, however, we perceived that some of the methods have a deficient process in this matter. Improvements in the scoring process are necessary. Also, it is important the development of integration methods, both in data and in methodologies. From the use of COLOMBOS data, we perceived that if the transcriptomic data from different sources are not properly integrated, its application could cause more drawbacks than advantages. In methodologies is also important a proper integration of predictions. Intuitively the integration of genomics and transcriptomics seems an adequate approach since we have data for *in vitro* and *in vivo* interactions. However, in practice, we saw that the integration by Borda was not a proper methodology. The integration methods need to maintain the network structure. Finally, the development of mathematical methods with a more biological background will be a very important resource for an accurate GRN inference.

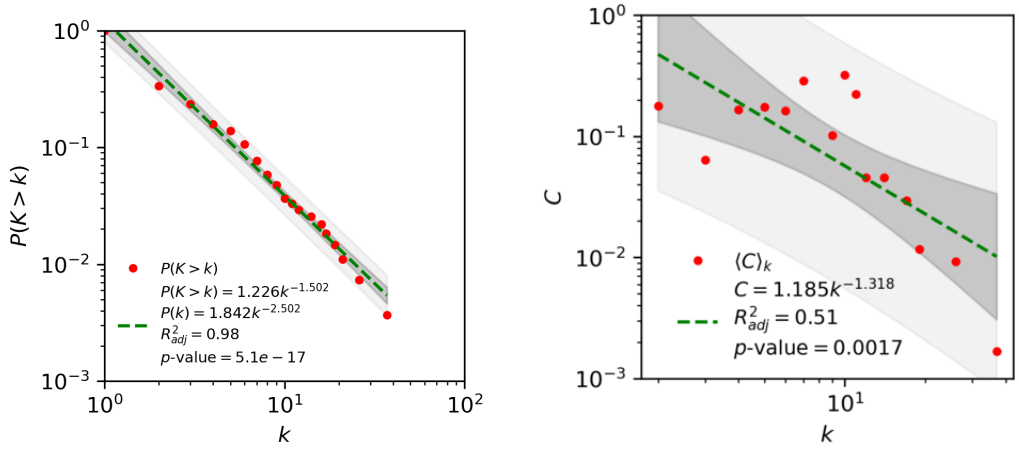
APPENDIX A



(a) Network degree distribution

(b) Network clustering coefficient distribution

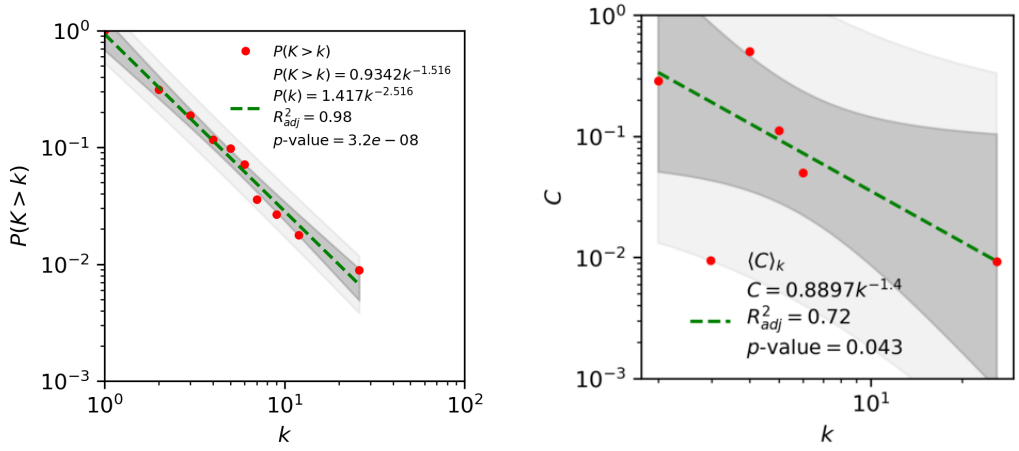
Figure A.1: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the curated network *Curated_ RTB*.



(a) Network degree distribution

(b) Network clustering coefficient distribution

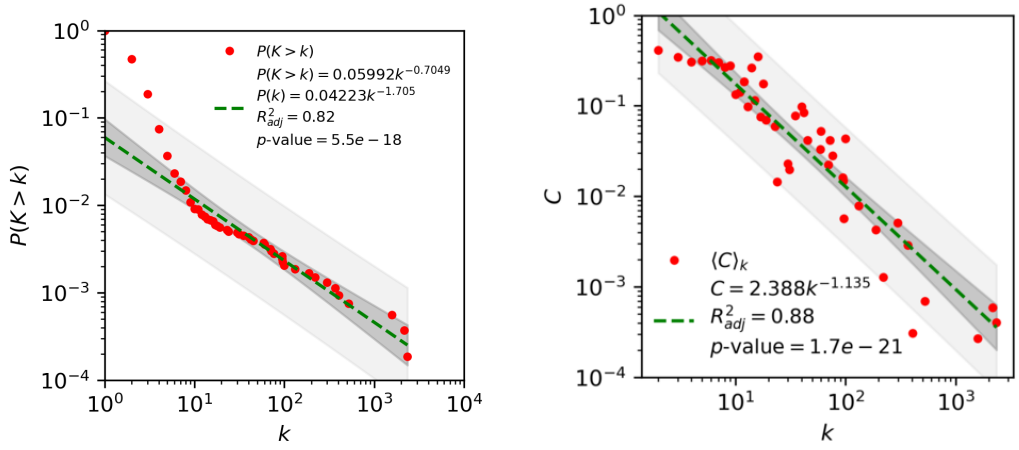
Figure A.2: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the curated network *Curated_ DBSCR*.



(a) Network degree distribution

(b) Network clustering coefficient distribution

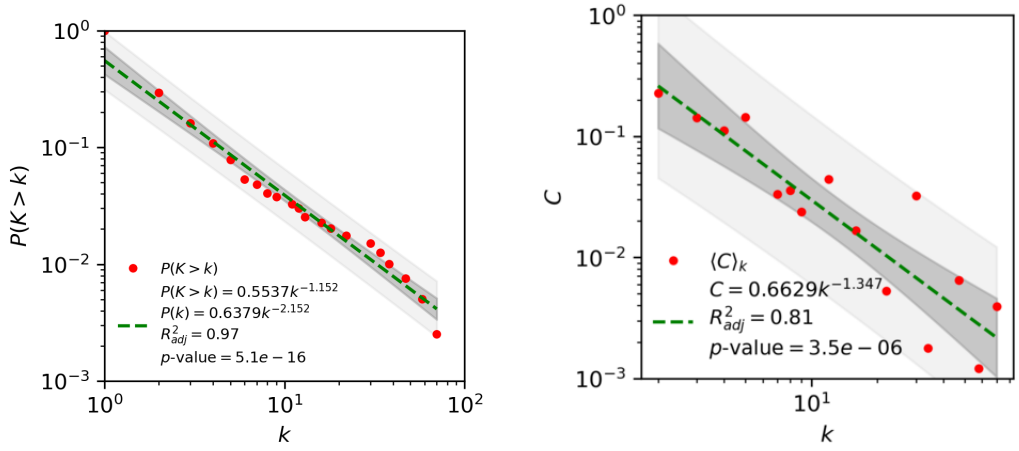
Figure A.3: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the curated network *Curated_ DBSCR(S)*.



(a) Network degree distribution

(b) Network clustering coefficient distribution

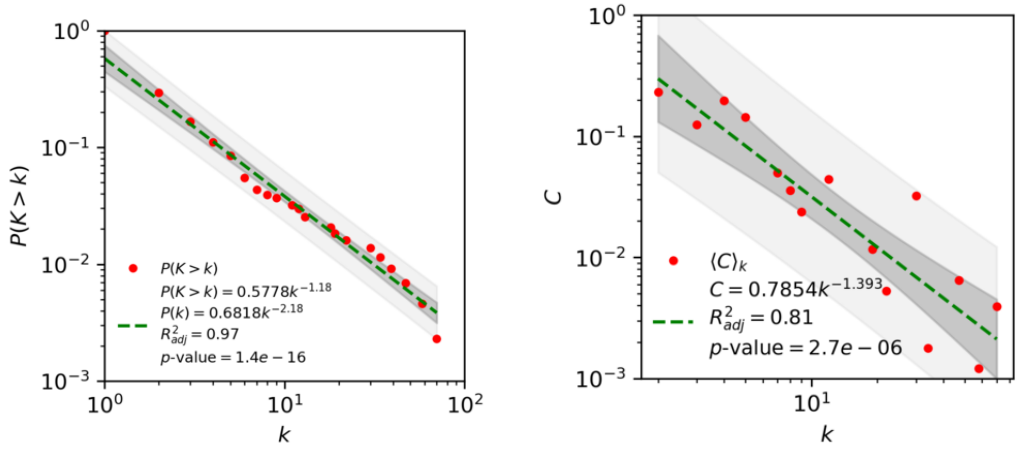
Figure A.4: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the curated network *Curated_ FL*.



(a) Network degree distribution

(b) Network clustering coefficient distribution

Figure A.5: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the curated network *Curated_FL(S)*.



(a) Network degree distribution

(b) Network clustering coefficient distribution

Figure A.6: Cumulative distribution of the network node degree ($P(k)$) and clustering coefficient ($C(k)$) of the curated network *Curated_FL(S)-DBSCR(S)*.

Table A.1: Description of the curated and inferred networks in this work

NETWORK	ABASY ID	GENES	INTERACTIONS	DESCRIPTION
Curated_ RTB	100226_v2015_sRTB13	311	330	Network from RegTransBase database
Curated_ DBSCR	100226_v2015_sDBSCR15	273	341	Network from Database of transcriptional regulation in *Streptomyces coelicolor* and its closest relatives.
Curated_ DBSCR(S)	100226_v2015_sDBSCR15_eStrong	112	115	Filtration of interactions with strong evidence from the DBSCR network.
Curated_ FL	100226_v2019_sFL	5331	9454	Network from the collection and curation performed for this work.
Curated_ FL(cS)	Not Reported	347	438	Filtration of interactions with strong evidence from the FL network (cS=curated strong)
Curated_ FL(S)	100226_v2019_sFL_eStrong	396	493	Filtration of interactions with strong evidence from the FL network along with statistically validated interactions.
Curated_ FL-DBSCR-RTB	100226_v2019_sFL-DBSCR15-RTB13	5386	9707	Meta-curation of RTB, DBSCR and FL networks.
Curated_ FL(cS)-DBSCR(S)	Not Reported	387	480	Filtration of interactions with strong evidence from the meta-curated network.
Curated_ FL(S)-DBSCR(S)	100226_v2019_sFL-DBSCR15_eStrong	435	534	Filtration of interactions with strong evidence from meta-curated networks along with statistically validated interactions.
Inferred_ BSs	Available as a supplementary file	6263	23908	Inferred GRN from binding sites prediction.
Inferred_ Exp	Available as a supplementary file	4739	23908	Inferred GRN from transcriptomic data.
Inferred_ BSs-Exp	Available as a supplementary file	4763	23908	Community network from *Inferred_ BSs* and *Inferred_ Exp* .
Inferred_ All	Available as a supplementary file	3804	23908	Community network from all the inference methods.

BIBLIOGRAPHY

[1] G. Liu, K. F. Chater, G. Chandra, G. Niu, and H. Tan, "Molecular regulation of antibiotic biosynthesis in *Streptomyces*," *Microbiology and Molecular Biology Reviews*, vol. 77, no. 1, pp. 112–143, Mar. 1, 2013, ISSN: 1092-2172, 1098-5557. DOI: 10.1128/MMBR.00054-12. [Online]. Available: https://mmbr.asm.org/content/77/1/112.

[2] J. R. McCormick and K. Flärdh, "Signals and regulators that govern *Streptomyces* development," *FEMS Microbiology Reviews*, vol. 36, no. 1, pp. 206–231, Jan. 1, 2012, ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2011.00317.x. [Online]. Available: https://academic.oup.com/femsre/article/36/1/206/535804.

[3] K. F. Chater, S. Biró, K. J. Lee, T. Palmer, and H. Schrempf, "The complex extracellular biology of *Streptomyces*," *FEMS Microbiology Reviews*, vol. 34, no. 2, pp. 171–198, Mar. 1, 2010, ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2009.00206.x. [Online]. Available: https://academic.oup.com/femsre/article/34/2/171/473215.

[4] G. L. Challis and D. A. Hopwood, "Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100 Suppl 2, pp. 14 555–14 561, Nov. 25, 2003, ISSN: 0027-8424. DOI: 10.1073/pnas.1934677100.

[5] P. A. Hoskisson and G. P. v. Wezel, "*Streptomyces coelicolor*," *Trends in Microbiology*, vol. 27, no. 5, pp. 468–469, May 1, 2019, ISSN: 0966-842X, 1878-4380. DOI: 10.1016/j.tim.2018.12.008. [Online]. Available: https://www.cell.com/trends/microbiology/abstract/S0966-842X(18)30284-1.

[6] S. D. Bentley, K. F. Chater, A.-M. Cerdeño-Tárraga, *et al.*, "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* a3(2)," *Nature*, vol. 417, no. 6885, pp. 141–147, May 2002, ISSN: 1476-4687. DOI: 10.1038/417141a. [Online]. Available: https://www.nature.com/articles/417141a.

[7] T. C. McLean, B. Wilkinson, M. I. Hutchings, and R. Devine, "Dissolution of the disparate: Co-ordinate regulation in antibiotic biosynthesis," *Antibiotics*, vol. 8, no. 2, Jun. 18, 2019, ISSN: 2079-6382. DOI: 10.3390/antibiotics8020083. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6627628/.

[8] Y. Mast and E. Stegmann, "Actinomycetes: The antibiotics producers," *Antibiotics*, vol. 8, no. 3, p. 105, Sep. 2019. DOI: 10.3390/antibiotics8030105. [Online]. Available: https://www.mdpi.com/2079-6382/8/3/105.

[9] M. T. Madigan, J. M. Martinko, K. S. Bender, D. H. Buckley, D. A. Stahl, and T. Brock, *Brock Biology of Microorganisms*, 14th. Boston: Pearson, 2014, 1032 pp., ISBN: 978-0-321-89739-8.

[10] Z. Xu, Y. Wang, K. F. Chater, *et al.*, "Large-scale transposition mutagenesis of *Streptomyces coelicolor* identifies hundreds of genes influencing antibiotic biosynthesis," *Applied and Environmental Microbiology*, vol. 83, no. 6, Mar. 15, 2017, ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.02889-16. [Online]. Available: https://aem.asm.org/content/83/6/e02889-16.

[11] D. A. Hopwood, "Forty years of genetics with *Streptomyces*: From in vivo through in vitro to in silico," *Microbiology (Reading, England)*, vol. 145 (Pt 9), pp. 2183–2202, Sep. 1999, ISSN: 1350-0872. DOI: 10.1099/00221287-145-9-2183.

[12] K. Duangmal, A. C. Ward, and M. Goodfellow, "Selective isolation of members of the *Streptomyces violaceoruber* clade from soil," *FEMS Microbiology Letters*, vol. 245, no. 2, pp. 321–327, Apr. 1, 2005, ISSN: 0378-1097. DOI: 10.1016/j.femsle.2005.03.028. [Online]. Available: https://academic.oup.com/femsle/article/245/2/321/562370.

[13] B. Bednarz, M. Kotowska, and K. J. Pawlik, "Multi-level regulation of coelimycin synthesis in *Streptomyces coelicolor* a3(2)," *Applied Microbiology and Biotechnology*, vol. 103, no. 16, pp. 6423–6434, Aug. 1, 2019, ISSN: 1432-0614. DOI: 10.1007/s00253-019-09975-w. [Online]. Available: https://doi.org/10.1007/s00253-019-09975-w.

[14] D. F. Browning and S. J. W. Busby, "The regulation of bacterial transcription initiation," *Nature Reviews Microbiology*, vol. 2, no. 1, pp. 57–65, Jan. 2004, ISSN: 1740-1534. DOI: 10.1038/nrmicro787. [Online]. Available: https://www.nature.com/articles/nrmicro787.

[15] M. S. Paget, "Bacterial sigma factors and anti-sigma factors: Structure, function and distribution," *Biomolecules*, vol. 5, no. 3, p. 1245, Sep. 2015. DOI: 10.3390/biom5031245. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4598750/.

[16] E. Balleza, L. N. López-Bojorquez, A. Martínez-Antonio, *et al.*, "Regulation by transcription factors in bacteria: Beyond description," *Fems Microbiology Reviews*, vol. 33, no. 1, pp. 133–151, Jan. 2009, ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2008.00145.x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2704942/.

[17] J. Kormanec, B. Sevcikova, R. Novakova, D. Homerova, B. Rezuchova, and E. Mingyar, "The complex roles and regulation of stress response σ factors in *Streptomyces coelicolor*," in *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*, John Wiley & Sons, Ltd, 2016, pp. 328–343, ISBN: 978-1-119-00481-3. DOI: 10.1002/9781119004813.ch29. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119004813.ch29.

[18] H. U. van der Heul, B. L. Bilyk, K. J. McDowall, R. F. Seipke, and G. P. v. Wezel, "Regulation of antibiotic production in actinobacteria: New perspectives from the post-genomic era," *Natural Product Reports*, vol. 35, no. 6, pp. 575–604, Jun. 20, 2018, ISSN: 1460-4752. DOI: 10.1039/C8NP00012C. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2018/np/c8np00012c.

[19] A. Romero-Rodríguez, I. Robledo-Casados, and S. Sánchez, "An overview on transcriptional regulators in *Streptomyces*," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1849, no. 8, pp. 1017–1039, Aug. 1, 2015, ISSN: 1874-9399. DOI: 10.1016/j.bbagrm.2015.06.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874939915001303.

[20] J.-F. Martín and P. Liras, "Engineering of regulatory cascades and networks controlling antibiotic biosynthesis in *Streptomyces*," *Current Opinion in Microbiology*, vol. 13, no. 3, pp. 263–273, Jun. 2010, ISSN: 1879-0364. DOI: 10.1016/j.mib.2010.02.008.

[21] D. A. Hodgson, "Primary metabolism and its control in streptomycetes: A most unusual group of bacteria," *Advances in Microbial Physiology*, vol. 42, pp. 47–238, Jan. 1, 2000, ISSN: 0065-2911. DOI: 10.1016/S0065-2911(00)42003-5. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0065291100420035.

[22] J. F. Martín, F. Santos-Beneit, A. Sola-Landa, and P. Liras, "Cross-talk of global regulators in *Streptomyces*," in *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*, F. J. d. Bruijn, Ed., John Wiley & Sons, Inc., 2016, pp. 257–267, ISBN: 978-1-119-00481-3. DOI: 10.1002/9781119004813.ch22. [Online]. Available: http://onlinelibrary.wiley.com.ezproxy.unal.edu.co/doi/10.1002/9781119004813.ch22/summary.

[23] S. Chen, G. Zheng, H. Zhu, *et al.*, "Roles of two-component system AfsQ1/q2 in regulating biosynthesis of the yellow-pigmented coelimycin p2 in *Streptomyces coelicolor*," *FEMS Microbiology Letters*, vol. 363, no. 15, D. Clarke, Ed., fnw160, Aug. 2016, ISSN: 1574-6968. DOI: 10.1093/femsle/fnw160. [Online]. Available: https://academic.oup.com/femsle/article-lookup/doi/10.1093/femsle/fnw160.

[24] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, Nov. 1, 2005, ISSN: 0021-9533, 1477-9137. DOI: 10.1242/jcs.02714. [Online]. Available: https://jcs.biologists.org/content/118/21/4947.

[25] E. Almaas, "Biological impacts and context of network theory," *Journal of Experimental Biology*, vol. 210, no. 9, pp. 1548–1558, May 1, 2007, ISSN: 0022-0949, 1477-9145. DOI: 10.1242/jeb.003731. [Online]. Available: https://jeb.biologists.org/content/210/9/1548.

[26] H. Kitano, *Foundations of Systems Biology*. MIT Press, 2001, 297 pp., ISBN: 978-0-262-11266-6.

[27] A.-L. Barabási and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, Feb. 2004, ISSN: 1471-0064. DOI: 10.1038/nrg1272. [Online]. Available: https://www.nature.com/articles/nrg1272.

[28] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, "A guide to conquer the biological network era using graph theory," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 34, 2020, ISSN: 2296-4185. DOI: 10.3389/fbioe.2020.00034. [Online]. Available: https://www.frontiersin.org/article/10.3389/fbioe.2020.00034.

[29] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 15, 1999, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.286.5439.509. [Online]. Available: https://science.sciencemag.org/content/286/5439/509.

[30] A.-L. Barabási and M. Pósfai, *Network Science*, Edición: 1. Cambridge, United Kingdom: Cambridge University Press, Aug. 5, 2016, 475 pp., ISBN: 978-1-107-07626-6. DOI: 10.1063/PT.3.3526. [Online]. Available: http://networksciencebook.com/.

[31] J. A. Freyre-González, J. A. Alonso-Pavón, L. G. Treviño-Quintanilla, and J. Collado-Vides, "Functional architecture of escherichia coli: New insights provided by a natural decomposition approach," *Genome Biology*, vol. 9, no. 10, R154, Oct. 27, 2008, ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-10-r154. [Online]. Available: https://doi.org/10.1186/gb-2008-9-10-r154.

[32] J. A. Freyre-González and A. Tauch, "Functional architecture and global properties of the *Corynebacterium glutamicum* regulatory network: Novel insights from a dataset with a high genomic coverage," *Journal of Biotechnology*, vol. 257, pp. 199–210, Sep. 10, 2017, ISSN: 1873-4863. DOI: 10.1016/j.jbiotec.2016.10.025.

[33] T. U. Consortium, "UniProt: A worldwide hub of protein knowledge," *Nucleic Acids Research*, vol. 47, pp. D506–D515, D1 Jan. 8, 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gky1049. [Online]. Available: https://academic.oup.com/nar/article/47/D1/D506/5160987.

[34] P. D. Karp, R. Billington, R. Caspi, *et al.*, "The BioCyc collection of microbial genomes and metabolic pathways," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1085–1093, Jul. 19, 2019, ISSN: 1467-5463. DOI: 10.1093/bib/bbx085. [Online]. Available: https://academic.oup.com/bib/article/20/4/1085/4084231.

[35] N. R. Coordinators, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 44, pp. D7–D19, Database issue Jan. 4, 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkv1290. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702911/.

[36] V. Weiss, A. Medina-Rivera, A. M. Huerta, *et al.*, "Evidence classification of high-throughput protocols and confidence integration in RegulonDB," *Database: The Journal of Biological Databases and Curation*, vol. 2013, Jan. 17, 2013, ISSN: 1758-0463. DOI: 10.1093/database/bas059. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3548332/.

[37] A. Santos-Zavaleta, H. Salgado, S. Gama-Castro, *et al.*, "RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* k-12," *Nucleic Acids Research*, vol. 47, pp. D212–D220, D1 Jan. 8, 2019, ISSN: 1362-4962. DOI: 10.1093/nar/gky1077.

[38] J. A. Stead and K. J. McDowall, "Two-dimensional gel electrophoresis for identifying proteins that bind DNA or RNA," *Nature Protocols*, vol. 2, no. 8, pp. 1839–1848, Aug. 2007, ISSN: 1750-2799. DOI: 10.1038/nprot.2007.248. [Online]. Available: https://www.nature.com/articles/nprot.2007.248.

[39] X. Yang and C. Ma, "In vitro transcription assays and their application in drug discovery," *Journal of Visualized Experiments : JoVE*, no. 115, Sep. 20, 2016, ISSN: 1940-087X. DOI: 10.3791/54256. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5092058/.

[40] D. P. Clark and N. J. Pazdernik, "Chapter 19 - analysis of gene expression," in *Molecular Biology (Second Edition)*, D. P. Clark and N. J. Pazdernik, Eds., Boston: Academic Press, Jan. 1, 2013, pp. 581–614, ISBN: 978-0-12-378594-7. DOI: 10.1016/B978-0-12-378594-7.00019-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123785947000196.

[41] S.-S. Park, B. J. Ko, and B.-G. Kim, "Mass spectrometric screening of transcriptional regulators using DNA affinity capture assay," *Analytical Biochemistry*, vol. 344, no. 1, pp. 152–154, Sep. 1, 2005, ISSN: 0003-2697. DOI: 10.1016/j.ab.2005.05.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003269705003945.

[42] M. J. Cipriano, P. N. Novichkov, A. E. Kazakov, *et al.*, "RegTransBase – a database of regulatory sequences and interactions based on literature: A resource for investigating transcriptional regulation in prokaryotes," *BMC Genomics*, vol. 14, no. 1, p. 213, Apr. 2, 2013, ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-213. [Online]. Available: https://doi.org/10.1186/1471-2164-14-213.

[43] J. M. Escorcia-Rodríguez, A. Tauch, and J. A. Freyre-González, "Abasy atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1228–1237, Jan. 1, 2020, ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.05.015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2001037020302786.

[44] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*, 2nd edition. Cambridge ; New York: Cambridge University Press, Oct. 30, 1992, 994 pp., ISBN: 978-0-521-43108-8.

[45] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, Nov. 4, 2009, ISSN: 0036-1445. DOI: 10.1137/070710111. [Online]. Available: https://epubs.siam.org/doi/10.1137/070710111.

[46] N. E. E. Allenby, E. Laing, G. Bucca, A. M. Kierzek, and C. P. Smith, "Diverse control of metabolism and other cellular processes in *Streptomyces coelicolor* by the PhoP transcription factor: Genome-wide identification of in vivo targets," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9543–9556, Oct. 2012, ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gks766. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks766.

[47] X. Li, J. Wang, S. Li, J. Ji, W. Wang, and K. Yang, "ScbR- and ScbR2-mediated signal transduction networks coordinate complex physiological responses in *Streptomyces coelicolor*," *Scientific Reports*, vol. 5, no. 1, p. 14 831, Dec. 2015, ISSN: 2045-2322. DOI: 10.1038/srep14831. [Online]. Available: http://www.nature.com/articles/srep14831.

[48] M. T. López-García, P. Yagüe, N. González-Quiñónez, B. Rioseras, and A. Manteca, "The SCO4117 ECF sigma factor pleiotropically controls secondary metabolism and morphogenesis in *Streptomyces coelicolor*," *Frontiers in Microbiology*, vol. 9, p. 312, Feb. 21, 2018, ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.00312. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fmicb.2018.00312/full.

[49] E. E. Arroyo-Pérez, G. González-Cerón, G. Soberón-Chávez, D. Georgellis, and L. Servín-González, "A novel two-component system, encoded by the SCO5282/SCO5283 genes, affects *Streptomyces coelicolor* morphology in liquid culture," *Frontiers in Microbiology*, vol. 10, p. 1568, Jul. 9, 2019, ISSN: 1664-302X. DOI: 10.3389/fmicb.2019.01568. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2019.01568/full.

[50] X. Lu, X. Liu, Z. Chen, *et al.*, "The ROK-family regulator rok7b7 directly controls carbon catabolite repression, antibiotic biosynthesis, and morphological development in *Streptomyces avermitilis*," *Environmental Microbiology*, May 25, 2020, ISSN: 1462-2920. DOI: 10.1111/1462-2920.15094.

[51] S. M. Guerra, A. Rodríguez-García, J. Santos-Aberturas, *et al.*, "LAL regulators SCO0877 and SCO7173 as pleiotropic modulators of phosphate starvation response and actinorhodin biosynthesis in *Streptomyces coelicolor*," *PLoS ONE*, vol. 7, no. 2, M. Otto, Ed., e31475, Feb. 20, 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0031475. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0031475.

[52] D. Rozas, S. Gullón, and R. P. Mellado, "A novel two-component system involved in the transition to secondary metabolism in *Streptomyces coelicolor*," *PLoS ONE*, vol. 7, no. 2, M. Polymenis, Ed., e31760, Feb. 9, 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0031760. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0031760.

[53] S. Antoraz, S. Rico, H. Rodríguez, *et al.*, "The orphan response regulator aor1 is a new relevant piece in the complex puzzle of *Streptomyces coelicolor* antibiotic regulatory network," *Frontiers in Microbiology*, vol. 8, p. 2444, Dec. 12, 2017, ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02444. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fmicb.2017.02444/full.

[54] H.-N. Lee, J.-S. Kim, P. Kim, H.-S. Lee, and E.-S. Kim, "Repression of antibiotic downregulator WblA by AdpA in streptomyces coelicolor," *Applied and Environmental Microbiology*, vol. 79, no. 13, pp. 4159–4163, Jul. 1, 2013, ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.00546-13. [Online]. Available: https://journals.asm.org/doi/10.1128/AEM.00546-13.

[55] M. J. Buttner, K. F. Chater, and M. J. Bibb, "Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of streptomyces coelicolor a3(2)," *Journal of Bacteriology*, vol. 172, no. 6, pp. 3367–3378, Jun. 1990, ISSN: 0021-9193, 1098-5530. DOI: 10.1128/jb.172.6.3367-3378.1990. [Online]. Available: https://journals.asm.org/doi/10.1128/jb.172.6.3367-3378.1990.

[56] H.-J. Hong, M. S. B. Paget, and M. J. Buttner, "A signal transduction system in streptomyces coelicolor that activates the expression of a putative cell wall glycan operon in response to vancomycin and other cell wall-specific antibiotics: *S. coelicolor* σ^{E}," *Molecular Microbiology*, vol. 44, no. 5, pp. 1199–1211, May 23, 2002, ISSN: 0950382X, 13652958. DOI: 10.1046/j.1365-2958.2002.02960.x. [Online]. Available: http://doi.wiley.com/10.1046/j.1365-2958.2002.02960.x.

[57] M. I. Hutchings, H.-J. Hong, E. Leibovitz, I. C. Sutcliffe, and M. J. Buttner, "The σ^{E} cell envelope stress response of *Streptomyces coelicolor* is influenced by a novel lipoprotein, CseA," *Journal of Bacteriology*, vol. 188, no. 20, pp. 7222–7229, Oct. 2006, ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.00818-06. [Online]. Available: https://journals.asm.org/doi/10.1128/JB.00818-06.

[58] E. Camon, M. Magrane, D. Barrell, *et al.*, "The gene ontology annotation (GOA) database: Sharing knowledge in uniprot with gene ontology," *Nucleic Acids Research*, vol. 32, pp. D262–266, Database issue Jan. 1, 2004, ISSN: 1362-4962. DOI: 10.1093/nar/gkh021.

[59] J. F. Martín, A. Sola-Landa, F. Santos-Beneit, L. T. Fernández-Martínez, C. Prieto, and A. Rodríguez-García, "Cross-talk of global nutritional regulators in the control of primary and secondary metabolism in *Streptomyces*: Cross-talk of global regulators," *Microbial Biotechnology*, vol. 4, no. 2, pp. 165–174, Mar. 2011, ISSN: 17517915. DOI: 10.1111/j.1751-7915.2010.00235.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1751-7915.2010.00235.x.

[60] A. Rodríguez-García, A. Sola-Landa, K. Apel, F. Santos-Beneit, and J. F. Martín, "Phosphate control over nitrogen metabolism in *Streptomyces coelicolor*: Direct and indirect negative control of *glnR*, *glnA*, *glnII* and amtB expression by the response regulator PhoP," *Nucleic Acids Research*, vol. 37, no. 10, pp. 3230–3242, Jun. 2009, ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkp162. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp162.

[61] Y. Wang, X.-F. Cen, G.-P. Zhao, and J. Wang, "Characterization of a new GlnR binding box in the promoter of *amtB* in *Streptomyces coelicolor* inferred a PhoP/GlnR competitive binding mechanism for transcriptional regulation of *amtB*," *Journal of Bacteriology*, vol. 194, no. 19, pp. 5237–5244, Oct. 2012, ISSN: 0021-9193. DOI: 10.1128/JB.00989-12. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3457235/.

[62] J. S. Rokem, A. E. Lantz, and J. Nielsen, "Systems biology of antibiotic production by microorganisms," *Natural Product Reports*, vol. 24, no. 6, pp. 1262–1287, Nov. 21, 2007, ISSN: 1460-4752. DOI: 10.1039/B617765B. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2007/np/b617765b.

[63] C. Lai, J. Xu, Y. Tozawa, Y. Okamoto-Hosoya, X. Yao, and K. Ochi, "Genetic and physiological characterization of *rpoB* mutations that activate antibiotic production in *Streptomyces lividans*," *Microbiology,*, vol. 148, no. 11, pp. 3365–3373, 2002, ISSN: 1350-0872, DOI: 10.1099/00221287-148-11-3365. [Online]. Available: https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-148-11-3365.

[64] B. J. Keijser, E. E. Noens, B. Kraal, H. K. Koerten, and G. P. Wezel, "The *Streptomyces coelicolor ssgB* gene is required for early stages of sporulation," *FEMS Microbiology Letters*, vol. 225, no. 1, pp. 59–67, Aug. 2003, ISSN: 03781097, 15746968. DOI: 10.1016/S0378-1097(03)00481-6. [Online]. Available: https://academic.oup.com/femsle/article-lookup/doi/10.1016/S0378-1097(03)00481-6.

[65] B. Hillerich and J. Westpheling, "A new GntR family transcriptional regulator in *Streptomyces coelicolor* is required for morphogenesis and antibiotic production and controls transcription of an ABC transporter in response to carbon source," *Journal of Bacteriology*, vol. 188, no. 21, pp. 7477–7487, Nov. 1, 2006, ISSN: 0021-9193. DOI: 10.1128/JB.00898-06. [Online]. Available: http://jb.asm.org/cgi/doi/10.1128/JB.00898-06.

[66] Y. Tiffert, P. Supra, R. Wurm, W. Wohlleben, R. Wagner, and J. Reuther, "The *Streptomyces coelicolor* GlnR regulon: Identification of new GlnR targets and evidence for a central role of GlnR in nitrogen metabolism in actinomycetes: The *Streptomyces coelicolor* GlnR regulon," *Molecular Microbiology*, vol. 67, no. 4, pp. 861–880, Jan. 7, 2008, ISSN: 0950382X. DOI: 10.1111/j.1365-2958.2007.06092.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1365-2958.2007.06092.x.

[67] F. Santos-Beneit, A. Rodríguez-García, A. Sola-Landa, and J. F. Martín, "Cross-talk between two global regulators in *Streptomyces*: PhoP and AfsR interact in the control of *afsS*, *pstS* and *phoRP* transcription: Cross-talk of PhoP and AfsR in *Streptomyces*," *Molecular Microbiology*, vol. 72, no. 1, pp. 53–68, Mar. 19, 2009, ISSN: 0950382X, 13652958. DOI: 10.1111/j.1365-2958.2009.06624.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1365-2958.2009.06624.x.

[68] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, "Gene regulatory network inference: Data integration in dynamic models-a review," *Bio Systems*, vol. 96, no. 1, pp. 86–103, Apr. 2009, ISSN: 1872-8324. DOI: 10.1016/j.biosystems.2008.12.004.

[69] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 717–729, Oct. 2010, ISSN: 1740-1526. DOI: 10.1038/nrmicro2419. [Online]. Available: http://www.nature.com/nrmicro/journal/v8/n10/full/nrmicro2419.html.

[70] E. Clough and T. Barrett, "The gene expression omnibus database," in *Statistical Genomics: Methods and Protocols*, ser. Methods in Molecular Biology, E. Mathé and S. Davis, Eds., New York, NY: Springer, 2016, pp. 93–110, ISBN: 978-1-4939-3578-9. DOI: 10.1007/978-1-4939-3578-9_5. [Online]. Available: https://doi.org/10.1007/978-1-4939-3578-9_5.

[71] M. Moretto, P. Sonego, N. Dierckxsens, *et al.*, "COLOMBOS v3.0: Leveraging gene expression compendia for cross-species analyses," *Nucleic Acids Research*, vol. 44, pp. D620–D623, Database issue Jan. 4, 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkv1251. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702885/.

[72] Y. Woo, J. Affourtit, S. Daigle, *et al.*, "A comparison of cDNA, oligonucleotide, and affymetrix GeneChip gene expression microarray platforms," *Journal of Biomolecular Techniques : JBT*, vol. 15, no. 4, pp. 276–284, Dec. 2004, ISSN: 1524-0215. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2291701/.

[73] C. L. Yauk, M. L. Berndt, A. Williams, and G. R. Douglas, "Comprehensive comparison of six microarray technologies," *Nucleic Acids Research*, vol. 32, no. 15, e124, 2004, ISSN: 0305-1048. DOI: 10.1093/nar/gnh123. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC516080/.

[74] R. D. Canales, Y. Luo, J. C. Willey, *et al.*, "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, Sep. 2006, ISSN: 1546-1696. DOI: 10.1038/nbt1236. [Online]. Available: https://www.nature.com/articles/nbt1236.

[75] S. Davis and P. S. Meltzer, "GEOquery: A bridge between the gene expression omnibus (GEO) and BioConductor," *Bioinformatics (Oxford, England)*, vol. 23, no. 14, pp. 1846–1847, Jul. 15, 2007, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm254.

[76] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "Affy—analysis of affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, Feb. 12, 2004, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg405. [Online]. Available: https://doi.org/10.1093/bioinformatics/btg405.

[77] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix GeneChip probe level data," *Nucleic Acids Research*, vol. 31, no. 4, e15, Feb. 15, 2003, ISSN: 0305-1048. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC150247/.

[78] S. E. Reese, K. J. Archer, T. M. Therneau, *et al.*, "A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis," *Bioinformatics (Oxford, England)*, vol. 29, no. 22, pp. 2877–2883, Nov. 15, 2013, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt480.

[79] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 1, 2007, ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. [Online]. Available: https://doi.org/10.1093/biostatistics/kxj037.

[80] J. T. Leek, W. E. Johnson, H. S. Parker, *et al.*, *Sva: Surrogate variable analysis*, 2019. DOI: 10.18129/B9.bioc.sva.

[81] D. A. Benson, M. Cavanaugh, K. Clark, *et al.*, "GenBank," *Nucleic Acids Research*, vol. 41, pp. D36–42, Database issue Jan. 2013, ISSN: 1362-4962. DOI: 10.1093/nar/gks1195.

[82] D. Marbach, J. C. Costello, R. Küffner, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, Aug. 2012, ISSN: 1548-7091. DOI: 10.1038/nmeth.2016. [Online]. Available: http://www.nature.com/nmeth/journal/v9/n8/full/nmeth.2016.html.

[83] J. J. Faith, B. Hayete, J. T. Thaden, *et al.*, "Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLOS Biology*, vol. 5, no. 1, e8, Jan. 9, 2007, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0050008. [Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050008.

[84] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, no. 1, p. 79 879, 2007, ISSN: 1687-4145. DOI: 10.1155/2007/79879. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3171353/.

[85] P. E. Meyer, F. Lafitte, and G. Bontempi, "Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information," *BMC Bioinformatics*, vol. 9, p. 461, Oct. 29, 2008, ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-461. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2630331/.

[86] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PloS One*, vol. 5, no. 9, Sep. 28, 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0012776.

[87] R. Bonneau, D. J. Reiss, P. Shannon, *et al.*, "The inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biology*, vol. 7, no. 5, R36, 2006, ISSN: 1474-760X. DOI: 10.1186/gb-2006-7-5-r36.

[88] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "TIGRESS: Trustful inference of gene REgulation using stability selection," *BMC Systems Biology*, vol. 6, no. 1, p. 145, Nov. 22, 2012, ISSN: 1752-0509. DOI: 10.1186/1752-0509-6-145. [Online]. Available: https://doi.org/10.1186/1752-0509-6-145.

[89] R. Küffner, T. Petri, P. Tavakkolkhah, L. Windhager, and R. Zimmer, "Inferring gene regulatory networks by ANOVA," *Bioinformatics*, vol. 28, no. 10, pp. 1376–1382, May 15, 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts143. [Online]. Available: https://academic.oup.com/bioinformatics/article/28/10/1376/212009.

[90] R. E. Walpole, R. H. Myers, S. L. Myers, and K. E. Ye, *Probability & Statistics for Engineers & Scientists, MyLab Statistics Update*, 9 edition. Boston: Pearson, Mar. 17, 2016, 816 pp., ISBN: 978-0-13-411585-6.

[91] J. I. E. Hoffman, "Chapter 26 - analysis of variance II. more complex forms," in *Biostatistics for Medical and Biomedical Practitioners*, J. I. E. Hoffman, Ed., Academic Press, Jan. 1, 2015, pp. 421–447, ISBN: 978-0-12-802387-7. DOI: 10.1016/B978-0-12-802387-7.00026-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128023877000263.

[92] H. Hernandez, "Statistical modeling and analysis of experiments without ANOVA," *ForsChem Research Reports*, vol. 5, 2018. DOI: 10.13140/RG.2.2.21499.00803. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.21499.00803.

[93] H. Hernandez, "Parameter identification using standard transformations: An alternative hypothesis testing method," *ForsChem Research Reports*, vol. 4, 2018. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.14895.02728.

[94] A. I. Campos and J. A. Freyre-González, "Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions," *Scientific Reports*, vol. 9, no. 1, p. 3618, Mar. 6, 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-39866-z.

[95] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, vol. 27, no. 8, pp. 861–874, Jun. 1, 2006, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2005.10.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[96] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2013, pp. 451–466, ISBN: 978-3-642-40994-3. DOI: 10.1007/978-3-642-40994-3_29.

[97] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, Mar. 4, 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118432. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432.

[98] C. Siegenthaler and R. Gunawan, "Assessment of network inference methods: How to cope with an underdetermined problem," *PLOS ONE*, vol. 9, no. 3, e90481, Mar. 6, 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0090481. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090481.

[99] D. A. Rodionov, "Comparative genomic reconstruction of transcriptional regulatory networks in bacteria," *Chemical reviews*, vol. 107, no. 8, pp. 3467–3497, Aug. 2007, ISSN: 0009-2665. DOI: 10.1021/cr068309+. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2643304/.

[100] P. D'haeseleer, "What are DNA sequence motifs?" *Nature Biotechnology*, vol. 24, no. 4, pp. 423–425, Apr. 2006, ISSN: 1546-1696. DOI: 10.1038/nbt0406-423. [Online]. Available: https://www.nature.com/articles/nbt0406-423.

[101] P. D'haeseleer, "How does DNA sequence motif discovery work?" *Nature Biotechnology*, vol. 24, no. 8, pp. 959–961, Aug. 2006, ISSN: 1546-1696. DOI: 10.1038/nbt0806-959. [Online]. Available: https://www.nature.com/articles/nbt0806-959.

[102] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: Scanning for occurrences of a given motif," *Bioinformatics (Oxford, England)*, vol. 27, no. 7, pp. 1017–1018, Apr. 1, 2011, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr064.

[103] H. Li, V. Rhodius, C. Gross, and E. D. Siggia, "Identification of the binding sites of regulatory proteins in bacterial genomes," *Proceedings of the National Academy of Sciences*, vol. 99, no. 18, pp. 11 772–11 777, Sep. 3, 2002, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.112341999. [Online]. Available: https://www.pnas.org/content/99/18/11772.

[104] N. Nguyen, B. Contreras-Moreira, J. A. Castro-Mondragon, *et al.*, "RSAT 2018: Regulatory sequence analysis tools 20th anniversary," *Nucleic Acids Research*, vol. 46, W209–W214, W1 Jul. 2, 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gky317. [Online]. Available: https://doi.org/10.1093/nar/gky317.

[105] B. Liu, C. Zhou, G. Li, *et al.*, "Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses," *Scientific Reports*, vol. 6, no. 1, p. 23 030, Mar. 15, 2016, ISSN: 2045-2322. DOI: 10.1038/srep23030. [Online]. Available: https://www.nature.com/articles/srep23030.

[106] X. Chen, A. Ma, A. McDermaid, *et al.*, "RECTA: Regulon identification based on comparative genomics and transcriptomics analysis," *Genes*, vol. 9, no. 6, May 30, 2018, ISSN: 2073-4425. DOI: 10.3390/genes9060278. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6027394/.

[107] T. L. Bailey, "Discovering novel sequence motifs with MEME," *Current Protocols in Bioinformatics*, vol. 00, no. 1, pp. 2.4.1–2.4.35, 2003, ISSN: 1934-340X. DOI: https://doi.org/10.1002/0471250953.bi0204s00. [Online]. Available: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi0204s00.

[108] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 127–138, 2001, ISSN: 2335-6928.

[109] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nature Biotechnology*, vol. 20, no. 8, pp. 835–839, Aug. 2002, ISSN: 1087-0156. DOI: 10.1038/nbt717.

[110] T. L. Bailey, M. Boden, F. A. Buske, *et al.*, "MEME suite: Tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, W202–W208, suppl_2 Jul. 1, 2009, ISSN: 0305-1048. DOI: 10.1093/nar/gkp335. [Online]. Available: https://doi.org/10.1093/nar/gkp335.

[111] A. Dharwadker and S. Pirzada, *Graph Theory*. Institute of Mathematics, Aug. 8, 2011, 482 pp., ISBN: 978-1-4662-5499-2.

[112] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863–14 868, Dec. 8, 1998, ISSN: 0027-8424, 1091-6490. [Online]. Available: https://www.pnas.org/content/95/25/14863.

[113] T. A. Schieber, L. Carpi, A. Díaz-Guilera, P. M. Pardalos, C. Masoller, and M. G. Ravetti, "Quantification of network structural dissimilarities," *Nature Communications*, vol. 8, no. 1, p. 13 928, Jan. 9, 2017, ISSN: 2041-1723. DOI: 10.1038/ncomms13928. [Online]. Available: https://www.nature.com/articles/ncomms13928.

[114] G. G. Simpson, "Mammals and the nature of continents," *American Journal of Science*, vol. 241, no. 1, pp. 1–31, 1943, ISSN: 0002-9599. DOI: 10.2475/ajs.241.1.1. [Online]. Available: https://www.mendeley.com/catalogue/d5c1779e-5559-318f-a2a3-e32a783dfb2e/.

[115] M. J. Warrens, "Similarity measures for 2×2 tables," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 4, pp. 3005–3018, Jan. 1, 2019, ISSN: 1064-1246. DOI: 10.3233/JIFS-172291. [Online]. Available: https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs172291.

[116] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2, 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7. [Online]. Available: https://doi.org/10.1186/s12864-019-6413-7.

[117] J. A. Freyre-González, L. G. Treviño-Quintanilla, I. A. Valtierra-Gutiérrez, R. M. Gutiérrez-Ríos, and J. A. Alonso-Pavón, "Prokaryotic regulatory systems biology: Common principles governing the functional architectures of bacillus subtilis and escherichia coli unveiled by the natural decomposition approach," *Journal of Biotechnology*, vol. 161, no. 3, pp. 278–286, Oct. 2012, ISSN: 01681656. DOI: 10.1016/j.jbiotec.2012.03.028. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168165612002982.

[118] W. B. Alkema, "Regulog analysis: Detection of conserved regulatory networks across bacteria: Application to staphylococcus aureus," *Genome Research*, vol. 14, no. 7, pp. 1362–1373, Jun. 14, 2004, ISSN: 1088-9051. DOI: 10.1101/gr.2242604. [Online]. Available: http://www.genome.org/cgi/doi/10.1101/gr.2242604.

[119] J. M. Escorcia-Rodríguez, A. Tauch, and J. A. Freyre-González, "*Corynebacterium glutamicum* regulation beyond transcription: Organizing principles and reconstruction of an extended regulatory network incorporating regulations mediated by small RNA and protein-protein interactions," Systems Biology, preprint, Jan. 8, 2021. DOI: 10.1101/2021.01.07.423633. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2021.01.07.423633.

[120] D. M. Emms and S. Kelly, "OrthoFinder: Phylogenetic orthology inference for comparative genomics," *Genome Biology*, vol. 20, no. 1, p. 238, Dec. 2019, ISSN: 1474-760X. DOI: 10.1186/s13059-019-1832-y. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y.

[121] D. Marbach, S. Roy, F. Ay, *et al.*, "Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks," *Genome Research*, vol. 22, no. 7, pp. 1334–1349, Jul. 2012, ISSN: 1088-9051. DOI: 10.1101/gr.127191.111. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3396374/.

[122] A. Lihu and Holban, "A review of ensemble methods for de novo motif discovery in ChIP-seq data," *Briefings in Bioinformatics*, vol. 16, no. 6, pp. 964–973, Nov. 1, 2015, ISSN: 1467-5463. DOI: 10.1093/bib/bbv022. [Online]. Available: https://doi.org/10.1093/bib/bbv022.

[123] M. Liu, P. Zhang, Y. Zhu, *et al.*, "Novel two-component system MacRS is a pleiotropic regulator that controls multiple morphogenic membrane protein genes in *Streptomyces coelicolor*," *Applied and Environmental Microbiology*, vol. 85, no. 4, M. A. Elliot, Ed., e02178–18, /aem/85/4/AEM.02178–18.atom, Dec. 7, 2018, ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.02178-18. [Online]. Available: http://aem.asm.org/lookup/doi/10.1128/AEM.02178-18.

[124] F. Valdez, G. González-Cerón, H. M. Kieser, and L. Servın-González, "The streptomyces coelicolor a3(2) lipAR operon encodes an extracellular lipase and a new type of transcriptional regulator the GenBank accession numbers for the sequences described in this paper are AF009336 and u03114.," *Microbiology*, vol. 145, no. 9, pp. 2365–2374, Sep. 1, 1999, ISSN: 1350-0872, 1465-2080. DOI: 10.1099/00221287-145-9-2365. [Online]. Available: https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-145-9-2365.

[125] B. Aigle, A. Wietzorrek, E. Takano, and M. J. Bibb, "A single amino acid substitution in region 1.2 of the principal sigma factor of *Streptomyces coelicolor* a3(2) results in pleiotropic loss of antibiotic production," *Molecular Microbiology*, vol. 37, no. 5, pp. 995–1004, Sep. 2000, ISSN: 0950-382X, 1365-2958. DOI: 10.1046/j.1365-2958.2000.02022.x. [Online]. Available: http://doi.wiley.com/10.1046/j.1365-2958.2000.02022.x.

[126] L. Li, W. Jiang, and Y. Lu, "A novel two-component system, GluR-GluK, involved in glutamate sensing and uptake in *Streptomyces coelicolor*," *Journal of Bacteriology*, vol. 199, no. 18, I. B. Zhulin, Ed., e00097–17, /jb/199/18/e00097–17.atom, Sep. 15, 2017, ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.00097-17. [Online]. Available: http://jb.asm.org/lookup/doi/10.1128/JB.00097-17.

[127] X.-M. Mao, Z.-H. Sun, B.-R. Liang, *et al.*, "Positive feedback regulation of *stgR* expression for secondary metabolism in *Streptomyces coelicolor*," *Journal of Bacteriology*, vol. 195, no. 9, pp. 2072–2078, May 1, 2013, ISSN: 0021-9193. DOI: 10.1128/JB.00040-13. [Online]. Available: http://jb.asm.org/cgi/doi/10.1128/JB.00040-13.

[128] H. Yang, L. Wang, Z. Xie, Y. Tian, G. Liu, and H. Tan, "The tyrosine degradation gene hppD is transcriptionally activated by HpdA and repressed by HpdR in streptomyces coelicolor, while hpdA is negatively autoregulated and repressed by HpdR," *Molecular Microbiology*, vol. 65, no. 4, pp. 1064–1077, Aug. 2007, ISSN: 0950-382X, 1365-2958. DOI: 10.1111/j.1365-2958.2007.05848.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2007.05848.x.

[129] M. Urem, T. van Rossum, G. Bucca, *et al.*, "OsdR of *Streptomyces coelicolor* and the dormancy regulator DevR of *Mycobacterium tuberculosis* control overlapping regulons," *mSystems*, vol. 1, no. 3, M. Traxler, Ed., e00014–16, /msys/1/3/e00014–16.atom, Jun. 28, 2016, ISSN: 2379-5077. DOI: 10.1128/mSystems.00014-16. [Online]. Available: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00014-16.

[130] K. T. Nguyen, J. M. Willey, L. D. Nguyen, L. T. Nguyen, P. H. Viollier, and C. J. Thompson, "A central regulator of morphological differentiation in the multicellular bacterium *Streptomyces coelicolor*," *Molecular Microbiology*, vol. 46, no. 5, pp. 1223–1238, 2002, ISSN: 1365-2958. DOI: 10.1046/j.1365-2958.2002.03255.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.2002.03255.x.

[131] B. A. Traag, G. H. Kelemen, and G. P. Van Wezel, "Transcription of the sporulation gene *ssgA* is activated by the IclR-type regulator SsgR in a whi-independent manner in *Streptomyces coelicolor* a3(2): Developmental regulation of *ssgRA* in *S. coelicolor*," *Molecular Microbiology*, vol. 53, no. 3, pp. 985–1000, Jun. 10, 2004, ISSN: 0950382X, 13652958. DOI: 10.1111/j.1365-2958.2004.04186.x. [Online]. Available: http://doi.wiley.com/10.1111/j.1365-2958.2004.04186.x.

[132] R. Amin, J. Reuther, A. Bera, W. Wohlleben, and Y. Mast, "A novel GlnR target gene, *nnaR*, is involved in nitrate/nitrite assimilation in *Streptomyces coelicolor*," *Microbiology*, vol. 158, pp. 1172–1182, Pt_5 May 1, 2012, ISSN: 1350-0872, 1465-2080. DOI: 10.1099/mic.0.054817-0. [Online]. Available: https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.054817-0.