Doctoraatsproefschrift nr. 943 aan de faculteit Bio-ingenieurswetenschappen van de K.U.Leuven

ENSEMBLE METHODS FOR BACTERIAL NETWORK INFERENCE

Riet DE SMET

Promotor:
Prof. dr. ir. K. Marchal (promotor)
Prof. dr. ir. B. De Moor (co-promotor)

Leden van de examencommissie:
Prof. dr. ir. J. Martens (voorzitter)
Dr. ir. J. Ramon
Prof. dr. ir. J. Vanderleyden
Dr. K. McDowall (University of Leeds, U.K.)
Dr. T. Michoel (University of Freiburg, Germany)

Proefschrift voorgedragen tot het behalen van de graad van Doctor in de Bio-ingenieurswetenschappen

December 2010

ISBN 978-90-8826-175-6
D/ 2010/11.109/58

Voorwoord

Gezegend met het genetisch materiaal van een sociaal agoge en een economist, besloot ik al op jonge leeftijd om tegen de wetten van de klassieke genetica in te gaan en voor een carrière als wetenschapper te kiezen. Deze ietwat eigenzinnige keuze, gezien mijn genetische mix, heeft blijkbaar toch opgebracht en nu ik op het punt sta om een belangrijke mijlpaal in deze nog prille wetenschappelijke carrière te behalen, immers een doctoraat, is het hoog tijd om een hele resem mensen te bedanken voor hun steun en bijdrage in deze verwezenlijking.

De persoon die zonder twijfel het meest heeft bijgedragen tot dit boekje is Kathleen Marchal. Kathleen, enorm bedankt om mij de kans te geven om niet alleen aan dit doctoraat te kunnen beginnen, maar ook om het tot een succesvol einde te brengen. Al was ik niet altijd overtuigd van mijn eigen werk je zag er altijd potentieel in en wist me steeds te overtuigen door te gaan op mijn (of ook soms jou) elan. Niet alleen was je een goede mentor maar ook naast het werk zorgde je voor de nodige ontspanning en je was altijd aangenaam gezelschap om mee op congres te gaan.

Verder wens ik ook mijn co-promotor Prof. Bart De Moor en mijn assessoren Prof. Jos Vanderleyden en Prof. Iven Van Mechelen te bedanken voor hun opvolging van mijn doctoraatswerk.

I would also like to thank the members of my Examination Committee: Prof. Jos Vanderleyden, Dr. Jan Ramon, Dr. Tom Michoel and Dr. Kenneth McDowall. Thanks for your valuable comments and suggestions, which significantly improved my PhD thesis. Special thanks to Dr. Kenneth McDowall for crossing the Channel in order to attend my defense!

Een bioinformaticus wordt vaak geacht een beetje een manusje van alles te zijn: zowel de biologie, de wiskunde en statistiek als computertechnische kunde dient zijn/haar deel te zijn. Het is echter onmogelijk om specialist te zijn op al deze uiteenlopende vlakken. Gelukkig heb ik tijdens dit doctoraat het genoegen gehad om met verscheidene mensen samen te werken die elk heel kundig zijn in hun eigen domein. Initieel heeft Thomas Dhollander mij ingewijd in het domein van de (query-gebaseerde) biclustering. Geen gemakkelijke taak gezien mijn bijna onbestaande kennis van genexpressiedata, clustering en bayesiaanse statistiek. Op het vlak van netwerkinferentie algoritmen heb ik dan weer enorm veel opgestoken van Tom Michoel en Anagha Joshi. Bedankt Tom om me toen onder jullie vleugels te nemen. Die periode dat ik met jullie heb samengewerkt heeft in grote mate mijn interesse en invulling van de rest van mijn doctoraat bepaald. Verder ook veel dank aan Kim Hermans, Sigrid De Keersmaecker en Jos Vanderleyden om mij in te wijden in de biologie achter de *Salmonella* biofilmen. Kim, nog eens bedankt voor het kritisch nalezen van mijn *Salmonella* schrijfselen.

De vele collega's die zowel op ESAT als op CMPG de revue passeerden zorgden telkens opnieuw voor een aangename werksfeer. Karen, bedankt om me de eerste maanden in ESAT de weg te tonen en voor de leuke en vruchtbare samenwerkingen. Valerie, het leven is niet mals voor je geweest, maar ik hoop van ganser harte dat ook jij binnenkort je doctoraat mag afleggen en dat het leven je daarna mag toelachen want niemand verdient dit meer dan jij. Bedankt om naast een goede collega ook een goede vriendin te zijn. Thanks Carolina for being such a considerate colleague and friend. It's a pity we never really collaborated together, but who knows in future ... Verder, in onwillekeurige volgorde bedankt: Lore, Yan, Fu, Ivan, Hui, Pieter, Lyn, Inge, Sunny, Peyman, Kristof, Aminael, Abeer, Pieter Tim, Shi, Roland, Anneleen, Olivier, Daniela, Ernesto, Thomas, Wout and Leo.

Geen inspanning zonder ontspanning. Zo was er het jaarlijkse skireisje. Bij Kathleen aan een doctoraat beginnen gebeurt immers onder lichte dwang: je zal minstens een keer in je doctoraatsloopbaan meegegaan zijn op de jaarlijkse skitrip. In het eerste jaar heb ik de druk nog wat kunnen afhouden, maar vanaf het tweede jaar stond ik ook op latten tussen medelotgenoten. En 't was plezant! Gedurende een ganse week stortten we ons in het spoor van vakdeskundige Prof. Y. Van de Peer op het empirisch onderzoek van waaghalzerij op de skipiste. En om de dag af te sluiten werd er des avonds dan nog eens duchtig op weerwolven gejaagd, ten minste door diegenen die hun ogen nog konden openhouden. Merci, Cindy, Yves en Kathleen om op jaarlijkse basis dit reisje te organiseren. Naast het jaarlijkse skireisje kon ik voor wekelijkse ontspanning terecht bij de loopmaatjes van de Hagelandse Running Club. Onvoorstelbaar hoe verkwikkend een uurtje 'waggelen' kan zijn. En verder zorgden ook de reisjes en de veel te schaarse afspraken met de vriend(inn)en voor de nodige ontspanning: merci Emmy, Marjolein, Annelies, Ineke, Steven, Griet en alle andere 'trek'-kameraadjes voor de lekkere etentjes, babbelgelegenheden, fiets- en wandeltochten, leuke reisjes ...

Het is dankzij mijn ouders dat ik in de eerste plaats de basis heb kunnen leggen om aan dit doctoraat te beginnen. Zij hebben me altijd gesteund in mijn keuzes en ze hebben me alle kansen gegeven om me zelf mijn weg te laten zoeken in dit leven. Sanne, ik wens jou en Hendrik enorm veel succes toe daar in het verre, verre Nieuw-Zeeland. En vergeet niet af en toe eens terug naar huis te komen zodanig dat we op een zondagmiddag nog eens allemaal samen aan tafel kunnen zitten.

Lieve Pascal, als een ietwat computeronkundige bioinformaticus is het een zegen om met een computerwetenschapper onder hetzelfde dak te wonen. Maar gelukkig gaat onze relatie veel verder dan enkele linuxcommando's en kan ik niet alleen bij je terecht met mijn gekke wetenschappelijke ideeën maar ook voor een luisterend oor, een lach, een traan of een bemoedigende knuffel. Bedankt voor alle steun en liefde de voorbije jaren.

Riet

Abstract

Within microorganisms the Transcriptional Regulatory Network (TRN) plays an important role in maintaining cellular homeostasis under changing environmental conditions. Therefore understanding the structure and dynamics of this network is fundamental for understanding and ultimately predicting organism behavior. With the emergence of the microarray technology genome wide data has become available that provides snapshots of the activity of the TRN. An important computational challenge is to infer or reverse-engineer the structure and dynamics of this TRN from available data.

The computational problem of inferring TRNs from gene expression data is however underdetermined: multiple equivalent solutions exist that each explain the data equally well. Ensemble methods provide an elegant way for dealing with this problem of underdetermination by considering multiple equivalent solutions and by reinforcing those solutions that are repeatedly retrieved. In this thesis we present different ensemble strategies to improve upon and extend the scope of existing methods to infer the TRN from gene expression data.

In a first part we focus on module detection or the detection of sets of coexpressed genes from gene expression data. In particular we develop an ensemble strategy for existing query-based biclustering methods in order to extend their application to input sets that are heterogeneous in their expression profiles. As such the method can be used to interrogate gene expression compendia for experimentally derived gene lists, as is illustrated on an *Escherichia coli* and *Salmonella* Typhimurium case study.

In a second part we focus on inference of the TRN itself. Here, we first present Stochastic LeMoNe. This method uses a stochastic optimization approach to output multiple equivalent outcomes of the network inference problem. By using ensemble averaging we demonstrate that both module detection and inference of the transcriptional program can be improved. Further we illustrate that by making certain assumptions on the inference problem, Stochastic LeMoNe is biased towards making correct predictions for only subparts of the TRN. Building upon this observation, we categorized existing network inference methods according to their conceptual differences and illustrated how these differences result in distinct methods highlighting different parts of the TRN.

Korte inhoud

In micro-organismen speelt het Transcriptioneel Regulatorisch Netwerk (TRN) een belangrijke rol in de aanpassing aan veranderende ongevingsomstandigheden. Bijgevolg is het ontrafelen van de structuur en de dynamiek van dit netwerk essentieel om het gedrag van een organisme te begrijpen en ultiem te voorspellen. Microroosterdata verlenen inzicht in de activiteit van het TRN door genexpressie te profileren onder verscheidene omgevingsomstandigheden. Een belangrijke computationele uitdaging is om de structuur en dynamiek van het TRN te infereren van deze data.

Inferentie van het TRN van genexpressiedata is echter ondergedetermineerd: meerdere oplossingen bestaan die elk de data even goed verklaren. *Ensemble* methoden voorzien een elegante manier om met dit probleem van onderdeterminatie om te gaan door meerdere equivalente oplossingen te beschouwen en zo oplossingen te bekrachtigen die herhaaldelijk worden geïnfereerd. In deze thesis beschrijven we verschillende *ensemble* methoden om zowel bestaande netwerkinferentie methoden te verbeteren als hun toepassingsdomein uit te breiden.

In het eerste deel focussen we op de detectie van sets van genen die coexpressie vertonen. In het bijzonder ontwikkelen we een *ensemble* strategie voor bestaande query-gebaseerde biclusteringsmethoden om hun toepassing uit te breiden naar input sets die heterogeen zijn in hun expressieprofiel. Zodoende kunnen deze methoden worden toegepast om genexpressie compendia te interrogeren voor experimenteel bekomen genlijsten. De praktische toepasbaarheid van deze methode werd aangetoond op zowel een *Escherichia coli* als *Salmonella* Typhimurium *case study*.

In het tweede deel van deze thesis staat inferentie van het TRN zelf centraal. Hier introduceren we Stochastische LeMoNe. Deze methode incorporeert een stochastische optimisatiestrategie om meerdere equivalente oplossingen van het netwerkinferentie probleem te bekomen. We demonstreren dat door deze stochastische optimisatie te koppelen aan een *ensemble* strategie er zowel op het vlak van moduledetectie als inferentie van het transcriptioneel regulatorisch programma een betere performantie bekomen wordt. Verder illustreren we ook dat aangezien Stochastische LeMoNe welbepaalde veronderstellingen maakt op het netwerkinferentie probleem, deze methode een zekere neiging vertoont naar het voorspelen van welbepaalde subonderdelen van het TRN. Steunend op deze observatie, categoriseren we netwerkinferentie methoden volgens hun conceptuele verschillen en illustreren we hoe deze verschillen resulteren in voorspellingen die complementair zijn.

Abbreviations and terminology

Abbreviations

AIC	Akaike Information Criterion
AP	Affinity Propagation
ChIP-chip	Chromatin Immunoprecipitation (ChIP) on a microarray (chip)
CLR	Context Likelihood of Relatedness
COALESCE	Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction
DISTILLER	Data Integration System To Identify Links in Expression Regulation
DNA	deoxyribonucleic acid
DREAM	Dialogue on Reverse Engineering Assessments and Methods
ECM	Extracellular Matrix
eQTL	expression Quantitative Trait Loci
GEO	Gene Expression Omnibus
GO	Gene Ontology
GPS	Gene Promoter Scan
GR	Gene Recommender
ISA	Iterative Signature Algorithm

LeMoNe	Learning Module Networks
LPS	Lipopolysaccharide
MCL	Markov Clustering
MEM	Multi-Experiment Matrix
mRNA	messenger RNA
NCA	Network Component Analysis
NI	Network Inference
NMI	Normalized Mutual Information
PPI	Protein-Protein Interaction
QDB	Query-Driven Biclustering
RNA	ribonucleic acid
SA	Signature algorithm
SEREND	Semi-supervised Regulatory Network Discoverer
SIRENE	Supervised Inference of Regulatory Networks
SPELL	Serial Pattern of Expression Levels Locator
sRNA	Small RNA
TF	Transcription Factor
TOM	Topological Overlap Matrix
TRN	Transcriptional Regulatory Network

Terminology

Classification In classification, properties or features of known targets and non-targets of a regulator are derived from high-throughput data and used to construct a classifier function, *i.e.* a mathematical function that describes the relationship between the class labels (being a target versus being a non-target) and the corresponding properties of the high-throughput data. These classifier functions can then be used to predict for novel genes whether or not they are a target of the studied TF based on their data properties.

Cross-validation Statistical technique that assesses the performance of a predictive modelling method by estimating the extent to which a model fitted on a certain dataset by the method, can also predict the observations made on an independent dataset (or the generalizability of a model).

***De novo* motif detection** Computational strategy to identify transcription factor binding sites without any prior information on how the binding site should look like. It relies on certain subsequences being statistically overrepresented in a set of coregulated genes.

Guilt-by-association principle Using the assumption that genes with similar functions exhibit similar expression patterns, the function of an unknown gene can be inferred from the function of annotated genes that are coexpressed with the unknown gene.

Module inference Identifying groups of coexpressed genes from gene expression data based on **clustering** or **biclustering** algorithms. Clustering methods group genes with similar expression patterns across all conditions, while biclustering methods combine the selection of coexpressed genes sets with a condition selection step in order to infer the set of conditions relevant to the bicluster genes.

Motif Transcription factor binding site or specific sequence tag located in a gene's promoter region that is recognized by a TF.

Operon A genomic segment of consecutive genes that are all under control of the same promoter. Operons occur typically within prokaryotes where they usually group functional related genes, as such allowing for a coordinated transcription of these genes.

Precision-recall curve Customarily used method that compares precision versus recall to evaluate the performance of an algorithm. The **precision** is the proportion of correctly inferred interactions according to an external standard on the total number of predictions made. The **recall** is the degree to which the total number of existing interactions in the real network has been covered by the predictions.

Search space The **search space** of a problem consists of all possible solutions that need to be evaluated to find the most optimal one according to preset criteria. In most inference problems the number of possible solutions is prohibitively large and can not be enumerated exhaustively. In those cases an **optimization strategy** is used to screen the search space in a clever way that allows finding the optimal (or almost optimal solution) without having to evaluate all possible solutions.

Top-down network inference Refers here to the reverse engineering or the *de novo* reconstruction of the structure of biological networks on a genomewide scale by exploiting high-throughput data. **Bottom-up** regulatory network inference, in contrast, is meant to construct a quantitative model from the data (both high- as low-throughput) by using a known mathematically formalized connectivity network as input. Estimating the kinetic parameters of this model from the data allows modeling the dynamic behavior of the network.

Underdetermined computational problem A high number of possible solutions (large search space) in combination with limited availability of experimental data results in finding many solutions that all explain the data equally, so no unique best solution can be found.

Table of contents

Chapter 1

General introduction

1.1 Context of the thesis

1.1.1 The system's biology era

Organisms adapt quickly and in a precise manner to changing environmental conditions. A long-standing question in molecular biology has been to unveil the mechanistic underpinnings on the level of single genes or even single nucleotides that explain and ultimately predict the organism's behavior (*phenotype*). In a reductionist approach organism behavior was studied on a gene-by-gene basis, *e.g.* one would render a gene inactive (knock-out) and then study the effects of this knock-out on the organism's phenotype. However, pioneering work by Jacob and Monod [1] revealed that genes do not work in isolation but are instead part of regulatory circuits in which regulatory proteins (*transcription factors*), encoded by regulatory genes, control the expression of structural genes by physically binding to the promoter regions of these genes. This network of transcription factors (TFs) and their corresponding *target genes* is further referred to as the *transcriptional regulatory network* (TRN). The observation that genes are parts of complex networks consisting of genes and proteins, has launched the idea that organism behavior can not be explained by separate gene behavior but should rather be studied by considering the network of cellular components, their mutual interactions and their interaction with the environment. In short, organism behavior should be studied at the system's level. This conviction that system's level understanding is crucial to understanding organism's behavior is central to the scientific field of *systems biology*.

Early attempts at a system's level understanding of biology, however, suffered from inadequate data on which to base the theories. It was only due to some major technological breakthroughs in the mid nineties that it was possible to study regulatory networks on a cellular scale. The first of these breakthroughs allowed generating the organism's gene list by sequencing whole genomes for a relatively low price in a matter of months [2]. This progress was followed by the development of another novel high-throughput technology: the *microarrays* [3]. Given a list of genes, microarrays allow to measure simultaneously the expression of an organism's complete gene set (the *transcriptome*) under a plethora of experimental conditions. As such a snapshot of the activity of the TRN can be obtained. After genome sequencing, DNA microarray analysis has become the most widely used source of genome-scale data [3] and microarrays have increasingly been carried out in biological and medical research to address a wide range of problems [4-6].

The emergence of *high-throughput* technologies, such as genome sequencing technologies and microarrays, has lead to an explosion of complex and noisy data. To understand the underlying biology of these data, systems biology is relying on an intimate integration of both mathematical and biological methods. The novel field of *bioinformatics* or *computational biology* is concerned with the development of data mining tools that are specifically designed to translate complex biological data into novel biological insights and that can be used interchangeably with experimental procedures to validate the predictions. This field covers a broad range of biological topics, such as gene function prediction, cis-regulatory motif detection, network prediction, gene evolution etc.

The subject of this thesis falls within this interdisciplinary field of bioinformatics. We particularly focus on computational methods which aim to reverse-engineer or *infer* the TRN from gene expression data (microarray data). This inference problem is a tremendous computational task which consists both of collecting, preprocessing and storing available gene expression data as developing computational tools to translate these data into appropriate wiring diagrams, representing the TRN. Whereas we discuss the problem of collecting gene expression data briefly in this introductory chapter, within this thesis we mainly focus on the computational tools for network inference themselves. In particular we present different computational approaches and discuss their biological applications.

1.1.2 Transcriptional regulatory network

A genome consists out of thousands of genes, each of them serving as templates for protein production which perform a variety of physiological and structural functions within the cell. As cells need to maintain homeostasis within constantly changing environments, tight regulation of protein production from the genome under changing extracellular and intracellular conditions is key to survival. This regulation is manifested at different levels: the transcriptional, translational and post-translational level. At the transcriptional level *transcription factors* (TFs) represent the cellular environmental state by changing rapidly from inactive to active molecular states in response to changing extracellular or intracellular conditions. These activated TFs then bind to specific stretches of DNA, corresponding to the promoter region of a certain gene, and promote transcription of this gene into *messenger RNA* (mRNA). This mRNA is further translated into proteins that can act upon the environment. In *Escherichia coli* there are about 300 TFs, which each act upon different environmental states, and which regulate the production of about 4500 proteins. Remark that TFs themselves are also proteins, encoded by genes and whose transcription is often regulated by a different set of TFs.

The Transcriptional Regulatory Network (TRN) describes all regulatory transcription interaction within the cell. This TRN can be represented as a graph in which the nodes represent the genes and directed edges (*i.e.* edges with a defined direction) point from one node to another, indicating that the first gene codes for a transcription factor that regulates expression of the second gene (Figure 1- 1a). As transcription factors can act both as repressors (*i.e.* suppresses gene expression) and activators (*i.e.* promotes gene expression), the edges are signed. This network is not densely connected, but is instead *sparse*: each TF modulates the expression of a limited set of *target genes* and the expression of each gene is under control of one or few TFs. TF-gene interactions are *condition-dependent*: some interactions might be present in some experimental conditions but absent in others [7;8]. Therefore different graphs might be drawn depending on the environmental context. In practice, however, a graphical representation of the TRN often represents all possible gene-TF interactions and therefore hides any contextual information.

As a cell consists of thousands of genes each controlled by one or multiple out of dozens of TFs the theoretical possible number of wirings between genes and TFs is dazzling high. TRNs are however surprisingly well-structured. Shen-orr *et al.* [9], for instance, observed that TRNs are built from recurring interaction patterns, called *network motifs* (Figure 1-1b). These network motifs represent patterns amongst genes and regulatory proteins in the network that are present more frequently in biological networks than in random networks. Hence these motifs are assumed to have biological functions: they are postulated to be basic information processing elements aimed at for instance speeding up a certain transcriptional response [10].

An additional important structural feature of TRNs is its modularity [11-13]: most biological functions are carried out by specific groups of genes and proteins that can be separated into *functional modules*. Modules consist of a set of nodes within the TRN that are strongly functionally related and whose function is clearly separated from those of genes of other modules. Such functional modularity is mainly achieved by joint regulation of the genes within a module by a common set of TFs (also called the *transcriptional program*). Consequently modularity exists at the transcriptional level [14]: genes within a module are coexpressed and hence modules can also be considered as sets of nodes which show strong coexpression interaction with each other, but only scarce

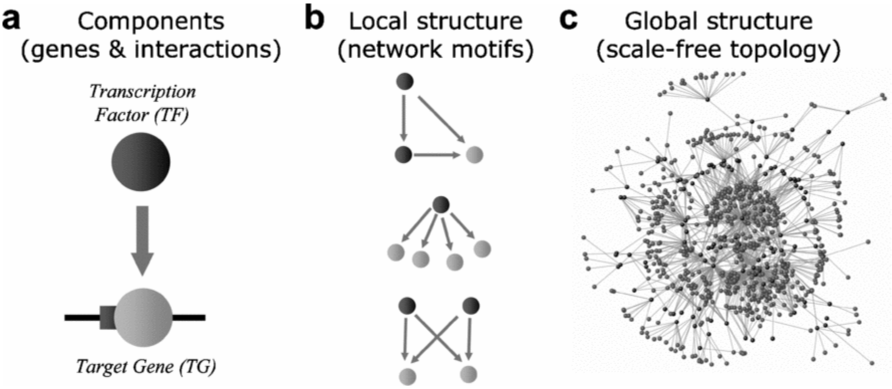


Figure 1- 1 Representation of the TRN at different scales. a) and b) represent basic units of the TRN, with a) referring to a single TF-target gene interaction and b) to network motifs. c) represents the complete TRN, or all interactions between target genes and their cognate TFs. Figure taken from [15].

coexpression interactions with nodes outside the module. Modularity allows orchestrating a coordinated response of a set of genes to changing environmental conditions. As genes might participate into multiple cellular functions, modules are not static cellular entities: depending on the environmental conditions genes might participate into different modules. This property of modularity of the TRN can be exploited by computational biologists to facilitate the task of modeling transcriptional regulation from high-throughput data.

Within this thesis we focus on the TRNs of the model bacteria *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. Although the TRN forms only a fraction of the total regulatory system (*i.e.* it ignores protein-protein interaction and protein-metabolite interactions), it represents a major level of regulation in prokaryotes: it allows bacteria to alter their gene expression and to adapt to novel environmental conditions. As the TRNs of bacteria are considered to be less complex than their eukaryotic counterparts, the networks of model-bacteria are well-characterized and therefore constitute excellent test cases for mathematical tools aimed at inferring the transcriptional regulatory network. In particular the *Escherichia coli* regulatory network is estimated to be one of the most complete TRNs of all organisms, and therefore this network is often used as a benchmark for computational tools. In addition, different bacteria cause diseases and therefore understanding the molecular mechanisms underlying infection and survival of these pathogens might contribute towards a better disease management.

1.1.3 From microarrays to gene expression compendia

The collection of all mRNA present in a cell at a certain stage is referred to as the *transcriptome*. Revealing this transcriptome allows gaining insight into the functions of the individual genes and their interrelationships and on a more global scale it constitutes a principle source of information on the activity of the TRN. Whereas in the pre-systems biology era it was a laborious process to measure the expression or mRNA production of a few genes simultaneously, microarrays have facilitated this by parallelizing their measurement. Indeed, they measure the whole transcriptome quantitatively on one chip.

Different microarray platforms for measuring gene expression exist, such as Affymetrix, Agilent, Codelink or in-house microarrays (see [16] for a review). Each different platform requires its own optimized sample preparation, labeling, hybridization and scanning protocol, and concomitantly also a specific normalization procedure. Normalization of the raw, extracted intensities aims to remove consistent and systematic sources of variation to ensure comparability of the measurements, both within and across arrays.

Microarray experiments are made publicly available in specialized databases such as Gene Expression Omnibus [17], Stanford microarray database [18] or ArrayExpress [19]. To ensure exchangeability of these data, data submitted to these databases should be compliant to the "Minimum Information About a Microarray Experiment (MIAME)" standard [20]. The MIAME standard enforces a careful description of the conditions under which the microarray experiment was performed, such as the genetic background of the used strains, the used media, growth conditions, triggering factors, etc. It does, however, not specify

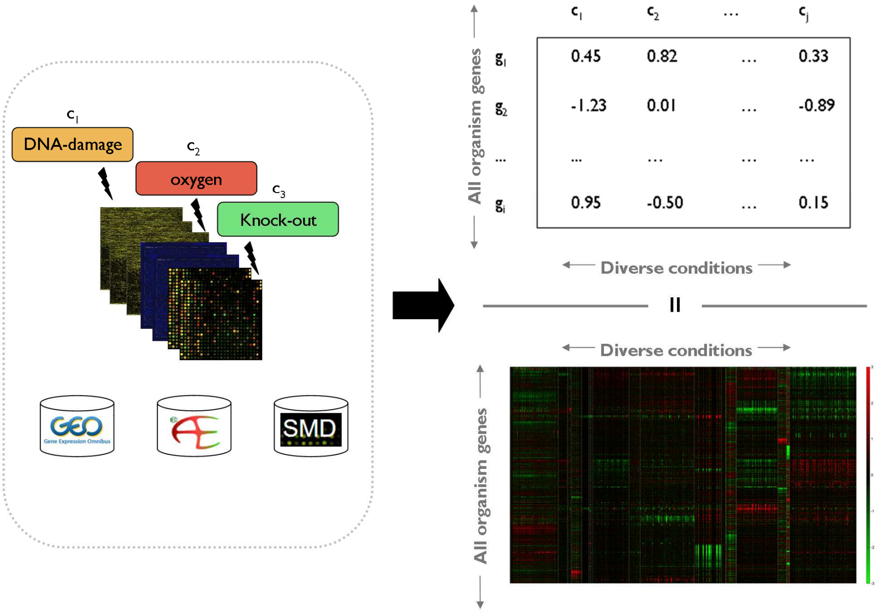


Figure 1- 2 Gene expression compendia combine all the publicly available expression data for a certain organism. Expression data is generally stored in public repositories such as Gene Expression Omnibus [17], Stanford microarray database [18] or ArrayExpress [19]. A gene expression compendium can conceptually be seen as a matrix with each element corresponding to the expression value of a gene measured on a certain array (condition) (upper right). These compendia can be visualized as heatmaps (lower right) with shades of red (overexpression) and green (underexpression) representing the gene expression values.

the format in which this meta-information should be presented. As a result, extracting data and information from these public microarray databases remains tedious and for a large part relies on manual curation: information is not only stored in different formats and data models, but is also redundant, incomplete and/or inconsistent. To fully exploit the large resource of information offered by these public databases, ideally all these data should be available as large species-specific gene expression *compendia*: matrices that for each of the organism's genes (rows) contains the microarray expression values for all conditions (columns) in which microarrays were performed (Figure 1- 2). The construction of such compendia from gene expression data stored in public repositories can be performed in a semi-automated process. Single-platform compendia combine all data on a particular organism that were obtained from one specific platform. Most single-platform compendia focus on Affymetrix data as this is considered one of the more robust and reproducible platforms [21;22]. The Many Microbe Microarrays Database (M3D) [23], for instance, offers Affy-based compendia for three microbial organisms. Cross-platform compendia, on the other hand, include data from different platforms and require more specialized normalization procedures to combine data from both one and two channel microarrays [8;24;25].

1.1.4 Mining the gene expression information

Presently large collections of public gene expression data are available in gene expression compendia for model prokaryotes such as *Escherichia coli* (about 1500 arrays) and *Salmonella* Typhimurium (about 800 arrays)[24]. The success of the microarray technology therefore does not only depend on clever design and polished protocol, but also on the successful analysis of very large data sets to translate complex and noisy data into biological insights. Pioneering work with this respect was accomplished by Eisen *et al.* [26], who proposed hierarchical *clustering* as a means to identify patterns within the data. The idea of clustering is simple: genes (or patients) with similar expression behavior across a range of conditions (or genes) are grouped together. As similarity in expression indicates functional relatedness or joint regulation by a similar set of transcription factors (TF), clustering is a convenient way to transfer a dataset containing thousands of genes into a few dozen of biologically meaningful entities (the clusters). As clustering is exploratory in nature and therefore requires little or no previous knowledge on the data, it is often used. Indeed, many different clustering algorithms, such as k-means and self-organizing maps, have been developed and applied to gene expression data to solve a range of biological problems. Clustering is for instance often used to infer the functional roles of genes [26], to classify tumor samples [4] or as a first step for the *de novo* detection of cis-regulatory elements [27]. With the ever-growing number of publicly available gene expression data, the data sets get more complex and more heterogeneous in their conditions. Consequently, clustering of these data becomes problematic as the presence of conditions in the data set under which the genes are not coherently transcriptionally regulated will reduce the signal-to-noise level of the data and complicate identifying sets of coexpressed genes. Therefore *biclustering* methods [28;29] have been developed to combine a search for coexpressed genes with a condition selection step to identify the conditions under which the genes are coexpressed, *i.e.* the conditions in which the joint transcriptional program of the bicluster genes is active.

Clustering and biclustering methods both take advantage of the modular structure of the TRN: they infer modules of coexpressed genes which often correspond to separate functional units (Figure 1- 3). They reveal the correlations or dependencies between genes without revealing the cause of the relationship. Therefore methods have been developed to infer the transcriptional regulatory networks (TRNs) from gene expression data [8;30-35]. These methods go one step beyond (bi)clustering and infer causality relationships in the network by also identifying the transcriptional programs that describe how transcription factors (TFs) cause the observed changes in expression of their cognate target genes. In particular the TRN can be represented as a graph in which nodes represent either the transcription factors (TFs) or the target genes or bi(clusters) (Figure 1- 3) (see section 1.1.2). Edges are directed as they reflect a causal relationship: they indicate that an observed correlation in expression pattern between nodes is caused by a node corresponding to a TF regulating a node that corresponds to a target gene. A transcriptional program corresponds to a set of TFs sharing the same set of target genes, ideally under a similar subset of conditions (Figure 1- 3).

Applying these inference procedures on public data sets of well-studied model organisms has largely improved our global understanding of TRNs. In bacteria, simple regulons that constitute only a few operons, show expression modularity. The operon organization seems crucial to

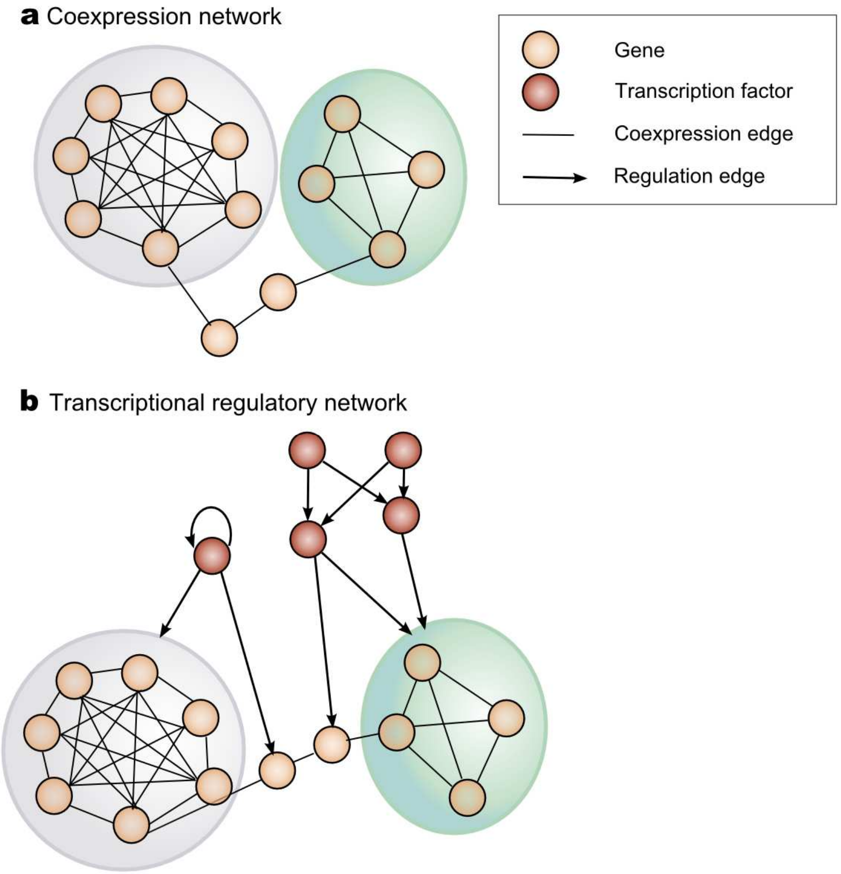


Figure 1- 3 Methods for module inference such as clustering and biclustering methods assume that the TRN is represented as a coexpression network (a). Hence the aim of these methods is to derive cliques of coexpressed genes, or sets of genes that are all mutually coexpressed. These modules are indicated by colored ovals in the figure. Methods that infer the TRN (b), in contrast, also aim to infer the causal regulators that explain gene coexpression.

preserve this modular level of coexpression under some conditions, while under other conditions, the presence of intra-operonic promoters breaks up this modularity [25;36;37]. In addition, complex regulation involving multiple regulators, generally results in single genes showing highly specific expression behavior that is not shared with that of other genes [8]. When focusing on the role of the transcriptional program, Zare *et al.* [38] observed that not only global transcription factors (TFs), but also local regulators in *E. coli* respond to a range of different conditions. In addition, many TFs are being active in similar conditions and thus trigger similar sets of genes, suggesting either redundancy in their functionality or an intricate cooperation between different TFs to mediate a common response [38].

Several notable examples have set the stage for adopting inference methods in daily laboratory practice. Kohanski *et al.* [39] unveiled the unprecedented link between protein mistranslation and the reaction to reactive oxygen species in response to antibiotics treatment by combining network inference with experimental evidence in *E. coli*. Yoon *et al.* [40] used a similar approach, to unravel the complex network regulating host-pathogen interactions in *Salmonella* Typhimurium, and Bonneau *et al.* [41] also used a combination of network inference and experimental data to chart the transcriptional network of the archeon *Halobacterium salinarum* for the first time. Computationally inferred interactions thus offer a useful resource to put experimental findings in a more global context by finding novel interactions that remained unveiled, by unfolding links between the pathway under investigation and other cellular processes or by identifying the conditions under which a favourite regulator is being active.

1.1.5 Ensemble methods for network inference

Under the assumption that each gene is regulated by only one regulator, inferring the interaction network in *E. coli* would imply testing the individual links between approximately 4500 genes and each of the 300 known and predicted regulators [42], resulting in 4500*300 tests. When also taking into account the existence of combinatorial regulation (*i.e.* cases in which binding of multiple TFs is necessary to control gene transcription) and feedback loops, the theoretical number of combinations can no longer be exhaustively enumerated. This means that the number of possible solutions is prohibitively large and clever algorithms strategies are needed to screen them in a time-efficient way. Also, module inference or finding the best combination of genes and conditions that define a coexpressed gene set according to preset criteria is combinatorially prohibitive. This large number of possible solutions (or the large search space), together with the restricted number of independent data points and the relatively low information content of the available data [43;44] turns TRN and module inference into an *underdetermined problem* with different solutions being possible that all explain the data equally well.

Because of the large search space, finding the most optimal solution to a module or network inference problem is non-trivial and optimization algorithms often result in suboptimal solutions that all approximate the true global optimal solution but differ slightly from each other [45]. Therefore within both the community of machine learning as clustering it has been suggested that it is more suitable to consider an ensemble of solutions than simply searching for a single optimal solution. The idea behind such *ensemble-based strategies* (also called *consensus approaches*) is that each prediction only corresponds to an approximation of the real underlying solution and that therefore predictions that are repeatedly inferred by different methods from the same data can be better statistically motivated. Ensemble methods have been applied in a diversity of biological contexts: such as motif detection [46-49], protein fold prediction [50;51], classification of tumor samples [52], clustering of gene expression data [53-57], RNA secondary structure prediction [58], clustering of PPI-data [59;60], gene function prediction [61;62], network inference [32;63] etc. In many of these cases the ensemble methods have been shown to perform at least as accurate or to outperform single solutions of the optimization problem (*e.g.* [46;51;56;59;64]).

Ensemble-based methods usually run over two different steps: (1) ensemble creation and (2) aggregation of the outputs in the ensemble (Figure 1- 4). For the first step it is important that the ensemble of solutions generated from the data set are accurate and in addition as diverse as possible. The reasoning behind this diversity-assumption is that each prediction should make errors on different instances, which can then be filtered out in the aggregation step. Different strategies have been developed to obtain such a diverse ensemble of solutions. A first approach uses the same algorithm on the same data set to generate an ensemble. Hereby, often subsampling or bootstrapping of the dataset is used in order to diversify the predictions made from this dataset (*e.g.* [52;54;56]). Alternatively, algorithms can be used that depending on the initialization and parameter settings converge to different local optima in order to obtain diversity in the outcomes (*e.g.* [32;48;53]). Yet another approach is to combine the outcomes of different algorithms applied to the same dataset, in stead of using the outcomes obtained by the same algorithm (*e.g.* [46;47;50;51;61;64]). Finally, it is also possible to create an ensemble by considering the outcomes for algorithm run on different data sets (also called data integration or data fusion) (*e.g.* [62]).

Once the ensemble of solutions is generated usually a *consensus* solution is extracted from this ensemble (step 2). Depending on the application here also different approaches exist. For instance, in case of clustering, often a new similarity matrix (the *consensus matrix*) is constructed representing the similarity of the clustered entities across the ensemble of clusterings generated in step 1. This new similarity matrix can be clustered to obtain *consensus clusters* [53-55;59]. Alternatively, when the output consists of lists which rank the predictions according to a score (as is often the case in a machine learning context) *majority voting* can be used, to produce a *consensus list* in which predictions that are repeatedly ranked highly across the ensemble get a high rank in the consensus list (*e.g.* [32;52;56]). However, also other approaches have been presented that for instance cast this aggregation step into a classification problem [51;64]. In this thesis we will further explore the usage of ensemble methods in the context of network inference and module inference.

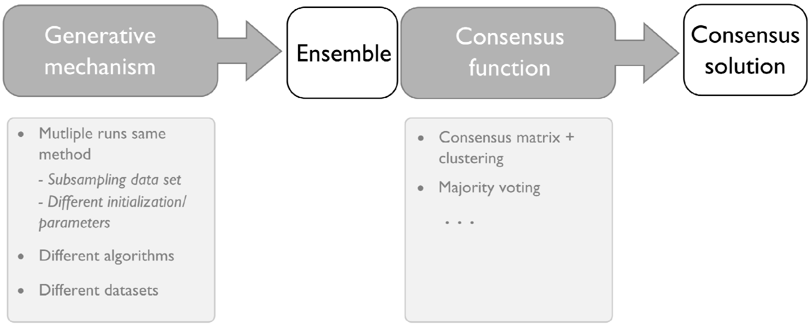


Figure 1- 4 Schematic overview of the ensemble approach. This approach consists out of two key steps: (1) generation of an ensemble of predictions through a 'generative mechanism' and (2) aggregation of the predictions by a 'consensus function'. There are possible ways to obtain as well an ensemble of solutions as a consensus solution from the ensemble. The outcome of the ensemble approach is the consensus solution.

1.2 Aim and deliverables of the thesis

Central to this thesis is the existence of genomewide expression compendia that implicitly assess transcriptional regulation on a genomewide scale in a plethora of conditions. Whereas bioinformaticians have continued to propose new algorithms to improve module detection (or (bi)clustering) and network inference from these compendia, here we aim to improve upon existing algorithms by drawing from concepts of ensemble learning. In particular, we discuss two distinct cases where ensemble strategies were introduced to solve distinct problems: one example in module detection and one in network inference.

First, we focus on query-based biclustering tools to explore gene expression compendia for genes coexpressed to genes of interest to a certain researcher. Whereas such tools have proven to be useful when exploring such compendia for single genes or sets of genes (the *query*) that are mutually tightly coexpressed, they fail when applied to query-sets that are heterogeneous in their expression profiles. This severely limits the applicability of these methods, as it could for instance be interesting for a user to view its own experimental data – which is often heterogeneous in its expression profiles - within these expression compendia. To circumvent this problem and to render query-based biclustering methods applicable to such more complicated query-sets, we introduce in Chapter 3 a generic ensemble framework for query-based biclustering. In particular we present a split-and-merge strategy in which each gene from the query-set is treated separately as input of a query-based biclustering algorithm. The outputs are then statistically merged in an ensemble biclustering framework to remove redundancy amongst the outputs and to allow for easy interpretation of the genes within the resulting biclusters.

Secondly, in Chapter 5, we introduce a network inference method, LeMoNe, which incorporates a stochastic framework in combination with ensemble averaging to improve upon regulatory network inference. By combining multiple equivalent outcomes of the network inference problem into an ensemble averaged network, reliability scores can be assigned to the inferred interactions. We illustrate that these scores do indeed prioritize known biological TF-gene interactions by these methods.

Finally, different groups have continued to produce new network inference methods at a staggering rate, each time claiming that theirs is better than previously published counterparts. In Chapter 5 and 6, in stead of giving a global assessment of their performance, we illustrate that most of the developed methods are actually complementary in the interactions they infer. Specifically, we demonstate that the low overlap in predicted interactions for different methods does not necessarily imply that predictions made by individual methods are wrong. Instead we point out, using real data examples, that depending on the choices that were made in the implementation, different tools are better suited for different types of reseach questions. In addition, the results motivate the construction of an ensemble of complementary methods to not only improve accuracy but also to extend the scope of what can be found.

1.3 Chapter-by-chapter overview

An overview of the organization of the thesis can be found in Figure 1-5. With the exception of this introductory chapter, Chapter 2 and the discussion, the content of all other chapters was derived from work that is already published, submitted or in preparation. Consequently, the contents of these chapters might be partially overlapping. This thesis consists mainly of two parts: in the first part (Chapters 2, 3 and 4) we focus on module inference, whereas in the second part (Chapters 5 and 6) we discuss methods for inference of the TRN.

An important application of gene expression compendia is to explore the information contained within these compendia in the context of a set of user-defined genes. To this end, different query-based data mining tools have been developed in the shape of gene prioritization methods and query-based biclustering algorithms. In **Chapter 2** we give an overview of such tools and we discuss their issues with respect to (1) handling input sets of genes that are heterogeneous in their expression and (2) defining a threshold on coexpression.

In **Chapter 3** we formulate an answer to these problems by developing an ensemble clustering strategy for query-based biclustering. This ensemble strategy incorporates a two-step procedure to simultaneously deal with the problem of defining a threshold on coexpression and deriving biclusters for a query-set that is heterogeneous in its expression profiles. The usefulness of such an approach is illustrated for an *Escherichia coli* ChIP-chip dataset, where a query-list of 90 ChIP-chip targets results in the identification of 17 biclusters each containing one or more of the ChIP-chip targets. This allows separating likely functional and true positive ChIP-chip targets from the remainder of the query-genes. In addition, this analysis reveals experimental consistencies and genes that were likely missed by the ChIP-chip assay. The work in this chapter has been accepted for publication [65]:

De Smet, R., Marchal, K. (2010). An ensemble method for querying gene expression compendia with experimental lists. Accepted for publication in proceedings of the *IEEE International Conference on Bioinformatics and Biomedicine (BIBM2010).*

In **Chapter 4**, using the same computational strategy as formulated in Chapter 3, we derive a functional map for *Salmonella* Typhimurium biofilm formation. In particular, we derive a condition-dependent coexpression network centered on a list of genes that were experimentally identified to be specifically involved in *Salmonella* biofilm formation. Building such a network for both multicellular (*i.e.* conditions that assess biofilm formation) as planktonic conditions reveals that at the transcriptional level these specific biofilm-genes are often involved in cellular processes, both required in multicellular as planktonic conditions. These results question the specificity of the transcriptional response in the biofilm formation process to a multicellular lifestyle. The work presented in this chapter is still on-going:

De Smet, R.*, Hermans, K.*, McClelland, M., Vanderleyden, J., De Keersmaecker, S., Marchal, K. (2010). Towards a functional map for *Salmonella* Typhimurium biofilm formation. *In preparation.*

In **Chapter 5** we introduce a network inference algorithm: stochastic LeMoNe ('Learning Module Networks'). We first illustrate how using a stochastic optimization scheme in combination with ensemble averaging can improve upon regulatory network inference, by prioritizing true interactions. Next we discuss how the assumptions that LeMoNe makes on the network inference problem results in particular parts of the *E. coli* regulatory network being highlighted by the method, whereas other parts can not be inferred. Finally, we compare the outcome of LeMoNe with that of CLR. Although both methods infer the regulatory network from gene expression data, they differ substantially both algorithmically and conceptually in how they approach the network inference problem. We illustrate that the conceptual differences between both methods results in the methods highlighting different parts of the *E. coli* regulatory network, suggesting that they are complementary in the interactions they infer. This work was done in collaboration with the Plant Systems Biology department of Ghent University and was published in the following three papers [32;66;67]:

Michoel, T., De Smet, R., Joshi, A., Van de Peer, Y., Marchal, K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Systems Biology, 3*, art.nr. 49, 49.

Michoel, T., De Smet, R., Joshi, A., Marchal, K., Van de Peer, Y. (2009). Reverse-engineering transcriptional modules from gene expression data. *Annals of the New York Academy of Sciences, 1158*, 36-43.

Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics, 25*(4), 490-496.

In **Chapter 6** we extend upon this observation made in Chapter 5 on the complementarity of network inference approaches. In this Chapter we argue that different state-of-the-art tools for network inference deal differently with the problem of underdetermination, by using assumptions and simplifications that reduce the number of possible solutions in order to make the problem solvable. The strategy adopted to deal with the inference problem determines the aspects of the transcriptional network that is highlighted and the type of research question that can be answered. The outcome of network inference therefore varies greatly between tools. In this chapter we give a comprehensive overview of existing network inference tools and illustrate how the different assumptions they make results in highlighting different parts of the transcriptional regulatory network. The work presented in this Chapter was published in the following paper [68]:

De Smet, R., Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology, 8,* 717-729.

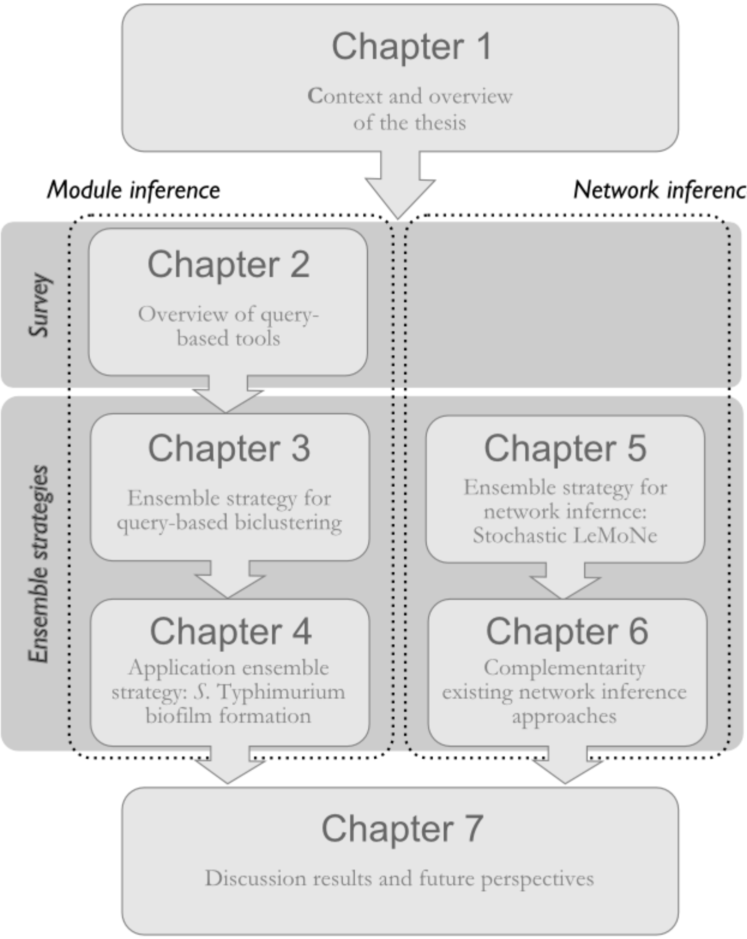

Chapter 1
Context and overview of the thesis
Module inference
Network inference
Survey
Ensemble strategies
Chapter 2
Overview of query-based tools
Chapter 3
Ensemble strategy for query-based biclustering
Chapter 4
Application ensemble strategy: S. Typhimurium biofilm formation
Chapter 5
Ensemble strategy for network infernce: Stochastic LeMoNe
Chapter 6
Complementarity existing network inference approaches
Chapter 7
Discussion results and future perspectives

Figure 1- 5 Overview structure PhD-thesis. The thesis contains an introductory chapter (Chapter 1) and a concluding Chapter (Chapter 7). The main body of the thesis consists of two separate parts, one that discusses module inference methods (Chapters 2, 3 and 4) and one that discusses network inference methods (Chapters 5 and 6). Chapter 2 gives a survey on query-based network inference methods, whereas Chapter 3, 4, 5 and 6 introduce ensemble strategies for module and network inference.

Finally, **Chapter 7** summarizes the results of this thesis and give a perspective on the future of network inference tools and ensemble methods in light of novel biological insights and current data generation technologies.

Chapter 2

Query-based exploration of gene expression compendia

2.1 Introduction

In Chapter 1 we introduced gene expression compendia as data structures that combine all available gene expression data for a certain organism. Considering the wide availability of publicly available gene expression data for model bacteria such as *Escherichia coli* and *Salmonella* Typhimurium, these compendia have high potential to study gene expression in a plethora of experimental conditions and offer to researchers the opportunity to view their own experiments in light of these data. The analysis of such compendia is however not trivial and requires the development of the appropriate data mining tools.

In this chapter we focus on query-based datamining methods. These tools treat the compendium as a database and query the compendium for genes coexpressed with a certain set of genes of interest to a researcher, hereto further referred as the *query*. This query can consist of one or multiple genes, and the query-profile is represented by the average expression profile of the query-genes in case of multiple genes or the profile of the gene itself in case a single gene is taken as input. Given a query-profile as input these methods produce as output a list of genes that shows within the expression compendium coexpression with the query. As gene expression compendia are often heterogeneous in the experimental conditions they contain it is crucial to not only select the genes coexpressed with a query, but to also select the conditions under which these genes are actively regulated. Indeed, the presence of conditions in the data set under which the transcriptional program is not active will reduce the signal-to-noise level of the data and complicates identifying sets of coexpressed genes.

Condition-dependent coexpression amongst genes generally implies functionally relatedness as genes that are coexpressed are subject to similar regulation mechanisms. Consequently, one can take advantage of the functional annotations of other genes than the query to predict a function for the query-genes or alternatively use the functional annotation of the query-genes to attribute functions to the genes that are coexpressed with the query. This strategy is also known as the 'guilt-by-association-principle'. This principle allows exploiting query-based tools to answer questions of the nature: 'Which other genes are involved in similar functions as my query?' 'What biological functions is my query involved in?' 'Under which specific conditions is biological process X activated?'

Here we distinguish between two different kinds of query-based approaches: the *prioritization* methods and the *query-based biclustering* methods. Prioritization methods rank all genes within the data set according to their similarity with the query, whereas query-based biclustering relies on module detection and outputs well-demarcated sets of genes that show condition-dependent coexpression with the query-genes. In this chapter we discuss both approaches and give examples of how they can be applied. We also discuss their shortcomings as these will be addressed in a next chapter. As in subsequent chapters we choose query-based biclustering approaches over prioritization methods we also argument this choice within this chapter.

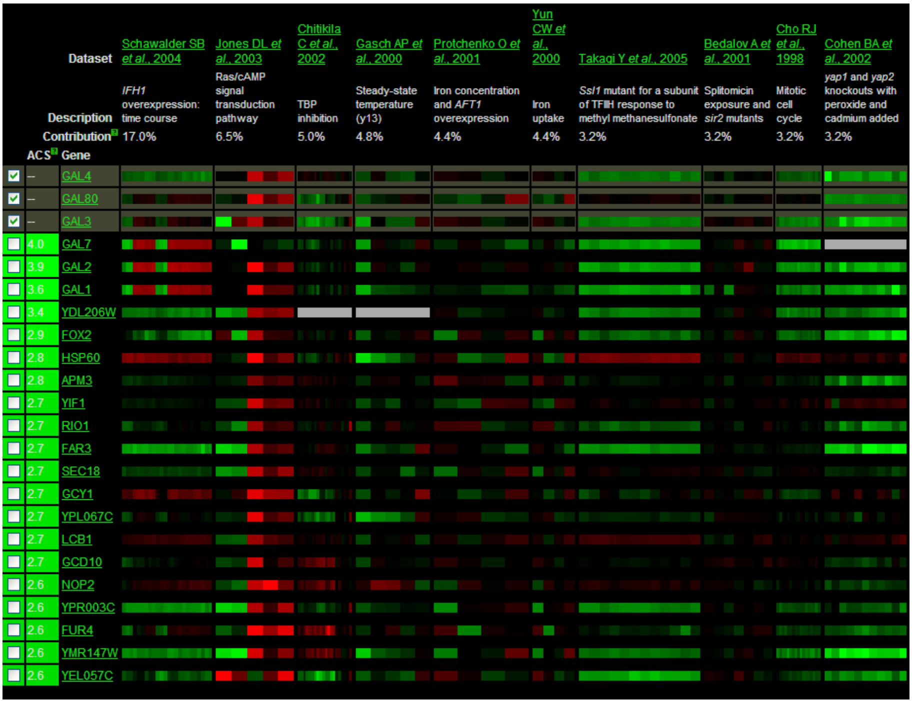
2.2 Gene prioritization methods

Gene prioritization methods are rank-based and sort all the genes in the genome based on condition-dependent similarity in expression with a given set of query-genes. The different prioritization methods differ in the criteria they use to select the relevant conditions and the way they score genes for their similarity with the query.

Most prioritization methods [69;70] follow an iterative scheme: they first calculate *condition scores* which reflect the significance of the conditions to the query: *i.e.* those conditions are chosen for which the query-genes are differentially expressed. In a second step genes are ranked according to their *gene scores* which reflect their similarity in expression to the query for the selected conditions.

Gene Recommender (GR) [69], for instance, scores conditions based on a *z*-score which measures both differential expression of the query-genes as the tightness of coexpression of the query-genes with respect to the remainder of the genes in the dataset. Conditions are selected by putting a threshold on these *z*-scores and genes within the dataset are then ranked based on their correlation with the query for the selected conditions. This procedure can be repeated for different thresholds on the condition *z*-scores and Owen *et al.* [69] propose to select as the most appropriate threshold the one that ranks the query-genes to the top. Hence the threshold for the condition scores is determined *a posteriori,* which makes the method rather computationally intensive as calculation of the gene scores needs to be repeated for a range of different possible threshold values. Owen *et al.* [69] compiled a *Caenorhabditis elegans* expression compendium containing 553 arrays, profiling gene expression in diverse set of experimental conditions. Using Gene Recommender they queried this compendium for genes coexpressed with five *C. elegans* genes involved in the retinoblastoma complex (Rb). As such two new genes could be discovered that were experimentally shown to have related functions.

The Serial Pattern of Expression Levels Locator (SPELL) [70] in contrast circumvents the need to select a condition subset relevant to the query by not putting a hard threshold on the condition scores but by using the condition scores themselves as weights to rank the genes according to their condition-dependent coexpression with the query. Specifically, SPELL groups similar conditions into 'experiments' and assesses the relevance of each experiment as the average Pearson correlation of the query-genes for this experiment. Hence, experiments



Dataset	Schawalder SB *et al.*, 2004	Jones DL *et al.*, 2003	Chitikila C *et al.*, 2002	Gasch AP *et al.*, 2000	Protchenko O *et al.*, 2001	Yun CW *et al.*, 2000	Takagi Y *et al.*, 2005	Bedalov A *et al.*, 2001	Cho RJ *et al.*, 1998	Cohen BA *et al.*, 2002
Description	*IFH1* overexpression: time course	Ras/cAMP signal transduction pathway	TBP inhibition	Steady-state temperature (y13)	Iron concentration and *AFT1* overexpression	Iron uptake	*Ssl1* mutant for a subunit of TFIIH response to methyl methanesulfonate	Splitomicin exposure and *sir2* mutants	Mitotic cell cycle	*yap1* and *yap2* knockouts with peroxide and cadmium added
Contribution?	17.0%	6.5%	5.0%	4.8%	4.4%	4.4%	3.2%	3.2%	3.2%	3.2%

	ACS?	Gene
☑	–	GAL4
☑	–	GAL80
☑	–	GAL3
☐	4.0	GAL7
☐	3.9	GAL2
☐	3.6	GAL1
☐	3.4	YDL206W
☐	2.9	FOX2
☐	2.8	HSP60
☐	2.8	APM3
☐	2.7	YIF1
☐	2.7	RIO1
☐	2.7	FAR3
☐	2.7	SEC18
☐	2.7	GCY1
☐	2.7	YPL067C
☐	2.7	LCB1
☐	2.7	GCD10
☐	2.6	NOP2
☐	2.6	YPR003C
☐	2.6	FUR4
☐	2.6	YMR147W
☐	2.6	YEL057C

Figure 2- 1 Example output for Serial Pattern of Expression Levels Locator (SPELL) [70]. Three functionally related yeast genes (*GAL4, GAL80* and *GAL3*) were taken as input. A weight (here represented as percentages with respect to the sum of all weights) is assigned to each experiment (called 'Dataset' in the figure) based on expression coherence of the query-genes for the conditions within this experiment. Next, all genes within the gene expression data set are ranked according to their weighted correlation with the query-genes. Here top-scoring genes include genes that, just as the query-genes, are involved in galactose metabolism.

in which the genes within the query have a more coherent expression profile get a higher weight (condition score). These condition scores are further used to rank the remainder of the genes based on their average weighted Pearson correlation with the query, with weights being equal to the condition scores. Consequently, genes that are coexpressed with the query in experiments for which the query-genes themselves are tightly coexpressed are prioritized by this method. An example of a SPELL-output is given in Figure 2- 1. The authors applied SPELL to a *S. cerevisiae* gene expression data set spanning 2394 conditions. Taking advantage of the extensive knowledge on the yeast gene functions, this approach could be used to attribute novel functions to different *S. cerevisiae* genes. The gene *Arp8*, for instance, was predicted and subsequently experimentally verified to be involved in cellular morphology. In addition the previously uncharacterized gene *YDL089W* was predicted to be involved in sporulation, which is in line with literature-based evidence.

Remark that both Gene Recommender and SPELL require multiple query-genes as input, as their calculation of the condition scores depends on the coherence in expression (Gene Recommender and SPELL) and differential expression (Gene Recommender) of the query-set. For Gene Recommender, for instance, the authors advice taking an input set of at least five genes as these seem to outperform random gene sets, whereas this is not the case for query-sets of smaller sizes [69]. This number depends however on the compendium used. In addition, as both methods rely on expression coherence of the query-genes to calculate the condition scores and also in the subsequent gene ranking, the genes within the query should be tightly coexpressed and consequently such an approach presumes prior knowledge on the functionality of the query-genes. Owen *et al.* [69], for instance, remarked that cases exist where a single query-set can be further divided into subgroups with distinct expression profiles. Running GR on the whole set will therefore result in a loss of specificity of the query-results and thus it is recommended to run the method on the subgroups. However, such prior knowledge on potential subgroups is often not available and therefore restricts the usability of these approaches to well-characterized systems.

To alleviate this problem of prior knowledge on the query list, Adler *et al.* [71] proposed a new approach in which it is possible to work with a single query-gene. To allow for this they depart from the classical scheme that first selects the relevant conditions before ranking the genes. Instead they construct ranked gene-lists for each experiment in the compendium, with genes being ranked according to the pairwise Pearson correlation of the genes with the query. For each gene the individual experiment-related ranks are aggregated into one score which assesses its similarity to the query-gene across all datasets. In case the query-set consists of multiple genes the average profile of these genes is used to get ranked gene lists. For each ranked gene the relevant experiments can be obtained as those for which the gene showed the lowest rank (highest Pearson correlation) w.r.t. the query. This framework is called the Multi-Experiment Matrix (MEM) [71]. Taking such an approach has several consequences. First, since gene selection precedes condition (or in this experiment) selection, the different genes that are selected as having an expression profile that is highly similar to that of the query do not necessarily have to show this similarity for the same experiments. Therefore, it is possible that these genes divide into different subgroups if the query-gene is involved in different functions. Consequently, a clustering method in post-processing might be required to obtain better insight into the query's function. Secondly, since no explicit experiment selection precedes gene ranking, spurious correlations might be detected for experiments in which the query-genes are not significantly differentially expressed. Therefore, Adler *et al.* [71] incorporated a pre-processing step which filters out all datasets for which the expression of the query is not substantially up- or down-regulated. This requires setting a threshold on the variance in expression of the query for each experiment in order to *a priori* eliminate datasets for which the query has a flat expression profile. This approach is rather *ad hoc* as the choice for the threshold depends on the user while this filtering step is rather crucial in obtaining reliable results.

Using the mouse NANOG gene as a query, the method was shown to be able to retrieve other genes that are just as NANOG known to be involved in embryonic stem cells. Likewise, using the Mini Chromosome Maintenance (MCM) protein complex as an illustrative example it was shown that additional members of the complex could be recruited if a subunit was taken as query. MEM is made publicly available as a webservice which allows querying multiple datasets for multiple organisms [71].

2.3 Query-based biclustering

An alternative to the prioritization methods mentioned above are the query-based biclustering methods. In contrast to outputting a ranked list of genes relevant to the query, these methods implement hard thresholding in their algorithmic framework to output a well-demarcated set of genes together with the corresponding conditions. We refer to these sets of genes together with their associated conditions as *modules* or *biclusters*. As these methods perform a 'query-based' search, the biclusters are algorithmically enforced to contain a set of query-genes together with genes that are coexpressed with the query. Existing query-based biclustering methods are the Signature Algorithm (SA) [12], ProBic [72] and Query-Driven Biclustering (QDB) [73]. Figure 2- 2 displays an example of a query-based biclustering output.

By thresholding the ranked gene lists obtained by prioritization methods such as SPELL and GR also well-confined sets of genes and their corresponding conditions can be obtained for a certain set of query-genes. The true distinguishing feature of query-based biclustering is, however, that here thresholding of the gene and condition scores is coupled whereas for prioritization methods this is not necessarily the case. For instance, as in both Gene Recommender and SPELL, the condition selection step precedes gene ranking, the condition content stays the same irrespective of the chosen cut-off on the gene scores. This is contra-intuitive as one would expect the condition content to change, with a changing threshold on gene coexpression. Indeed, a small set of genes is often tightly coexpressed across a larger set of conditions (*e.g.* operons), whereas with an increasing number of genes the conditions for which gene coexpression can be detected generally diminishes (*e.g.* regulons). Therefore query-based biclustering methods incorporate thresholding on coexpression in both the gene and condition direction within their algorithmic framework, in stead of defining thresholds in post-processing. As such gene selection is always coupled to condition selection and more biologically motivated modules can be obtained.

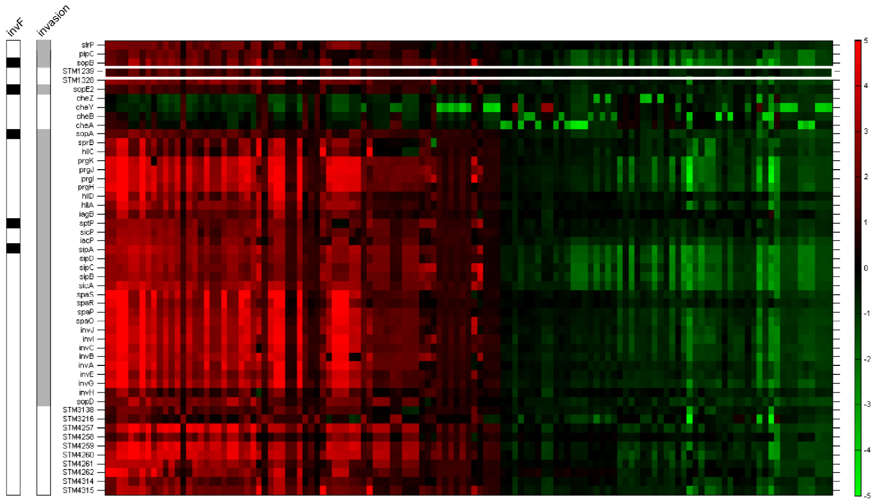


Figure 2- 2 Illustration of query-based biclustering. The figure shows the heatmap of a bicluster obtained by query-based biclustering. The query-gene profile is indicated by the white rectangle. A by ChIP-chip newly identified target 'STM1239' of the *Salmonella* Typhimurium regulator InvF [74] was taken as input of QDB [73]. The resulting bicluster seemed to contain all other known InvF targets (indicated by the black squares on the left), suggesting that STM1239 is indeed a functional target of this regulator. In addition, the bicluster is clearly enriched for invasion-related genes in *S*. Typhimurium (indicated by the grey squares on the left). This further substantiates the observation that STM1239 is a novel invF-target as invF is a known regulator of the *S*. Typhimurium invasion pathway.

By implementing hard thresholding within their algorithmic framework query-based biclustering methods have several advantages over prioritization methods:

- Query-based biclustering methods, in particular those that incorporate probabilistic frameworks (QDB and ProBic), report empty biclusters if none of the genes within the compendium could be found to be sufficiently coexpressed with the query-genes. Prioritization methods, in contrast, always output a ranked list of genes, even if there is no significant coexpression of the query-genes with any of the compendium genes.

- Each of the query-based biclustering methods described here accepts a single gene as query input. This in contrast to prioritization-based methods (with the exception of MEM), which require multiple query-genes as input and in general perform better if the query-set grows in size [69].
- Implementing hard thresholding in the algorithmic framework has as advantage that query-sets can be refined. Query-genes that drop out off the module are expected to differ in expression and also in function from the remainder of the query-genes.

In what follows we discuss several query-based biclustering methods. We focus in particular on two important features of query-based biclustering approaches: first how they retrieve biclusters centered on the query-genes. Secondly, we discuss how thresholding is defined in the gene and condition direction in order to obtain well-demarcated biclusters containing a certain subset of the compendium genes and their associated conditions. As QDB was used extensively in this thesis this method will be described into more detail than the other ones.

2.3.1 Performing a query-centered search

In Chapter 1 we mentioned that biclustering of gene expression compendia is underdetermined: there are many possible configurations of genes and conditions in biclusters that all explain the expression data equally well. Query-based biclustering constitutes a way to render the biclustering problem better-defined by taking advantage of prior knowledge on the 'location' of the bicluster in the shape of a set of user-defined query-genes. Indeed, seeding a biclustering algorithm with a set of query-genes biases biclustering towards the region of the search space one is interested in, thereby reducing the chances of the algorithm getting stuck in biological irrelevant local optima. Query-based biclustering algorithms differ in the way they exploit the expression profile of the query-genes to output biclusters containing the query. Here we discuss different methodological strategies used to obtain biclusters containing a given set of query-genes.

All three query-based biclustering methods mentioned above (Signature algorithm, QDB and ProBic) initialize the algorithm with the query and then, similarly as for SPELL and GR, a two-step procedure is followed in which first the conditions are selected that are relevant to the query-gene before genes are recruited that show similar condition-dependent expression profiles to the query-gene. In the probabilistic methods QDB and ProBic this scheme is repeated iteratively, such that the gene and condition content is continuously refined until convergence is reached. Therefore it is important for these methods to include a mechanism that prevents the bicluster to drift away from the query in consecutive iterations. Both in QDB and ProBic strong informative priors are used to enforce the bicluster to remain centered on the mean expression profile of the query. While this imposes a bicluster to contain the query, this restricts the flexibility in the biclusters that can be found. Indeed, bicluster solutions that still contain the query but that are not centered on the expression profile of this query-set, will not be retrieved by these methods.

SA, in contrast, does not rely on a Bayesian framework and assesses whether the weighted uncentered covariance of the genes with the query-profile exceeds a certain threshold in order to assign genes to the bicluster. Hence, this method can find bicluster solutions that are not constrained to be centered on the query-profile. As the SA does not rely on a probabilistic framework it is not possible to use priors to stop the bicluster from drifting away from the query, therefore gene and condition selection is only performed once and no further refinement of gene and condition content in iterative steps is performed. Later the SA was extended to the Iterative Signature Algorithm (ISA) which just as the SA is initialized by a set of genes (the query) but then iteratively performs multiple SA-steps in order to further refine genes and conditions until convergence in both gene and condition direction is reached [75]. However, as ISA does not implement a strategy to stop the bicluster from drifting away from the initial gene set, there is no guarantee that the resulting bicluster will still contain the query. Therefore ISA is generally used as a global biclustering method, aimed at identifying the more dominant patterns in the data set that do not necessarily contain the genes that were used to seed the method. The SA is then the query-based variant of the ISA algorithm.

2.3.2 Incorporating the threshold

Query-based biclustering approaches distinguish themselves from the prioritization methods in that they do not output all genes ranked according to a score, but implement hard thresholding to output well-demarcated sets of genes and conditions. For SA two different thresholds need to be defined, one to select the genes relevant to the query (*gene score threshold*) and one to select the corresponding conditions (*condition score threshold*). Alternatively, the probabilistic methods QDB and ProBic rely only on one parameter that needs to be set in order to define the bicluster boundaries: an informative prior, called the *priorvariance* (see 2.3.3.1 for a more detailed explanation on this parameter).

Both the priorvariance as the gene score threshold determine the bicluster size, *i.e.* these parameters control the expression coherence within the bicluster and concomitantly the number of genes and conditions. The more stringent these parameters are set, the more tightly genes within the bicluster are coexpressed and the smaller the number of genes belonging to the bicluster. We further refer to these parameters as the *resolution parameter* of the algorithm as it controls the biological detail of the bicluster (small, very homogeneous biclusters vs. larger, less homogeneous biclusters).

In ProBic the value for the resolution parameter is fixed *a priori*. However, choosing one fixed threshold for all queries, as is the case for ProBic, ignores the possibility that for not all biological processes the same stringency in coexpression is equally important. Indeed, depending on the query and the interest of a certain researcher a different stringency in coexpression might be desirable as is illustrated in Figure 2- 3. Therefore, SA and QDB reason that it is not *a priori* known how tightly other genes should be coexpressed with the query to be deemed biologically relevant and these methods incorporate a *resolution sweep approach* (illustrated in Figure 2- 4). In this approach a linear range of possible values for the resolution parameter are scanned and hence a whole range of biclustering solutions is obtained. As will be discussed below different ways exist to select appropriate values of the resolution parameter amongst the spectrum of solutions that results from the resolution sweep approach.

Figure 2- 3 This figure illustrates that the stringency of coexpression depends on the biological process studied. A gene known to be regulated by PurR (*purH*) and FNR (*fdnG*) were both taken as input of respectively QDB and the signature algorithm. We retrieved biclusters for different stringencies of coexpression (*x*-axis), for QDB this is controlled by the priorvariance whereas for the SA the gene score threshold controls coexpression tightness. To assess the biological relevance of the obtained biclusters for the different values of the respective resolution parameters we calculated their ability to reflect known regulon information present in RegulonDB [65]. Specifically, the F-measure (*y*-axis) balances the precision and recall with which both methods infer known targets of the regulators PurR and FNR. Maximal values for the F-measure are indicated by red (PurR) and black (FNR) circles. As can be seen from the figure, for both methods maximal F-measures correspond to different values of the resolution parameter. This suggests that depending on the biological process studied different stringencies of coexpression are required to obtain the most biologically relevant outcomes.

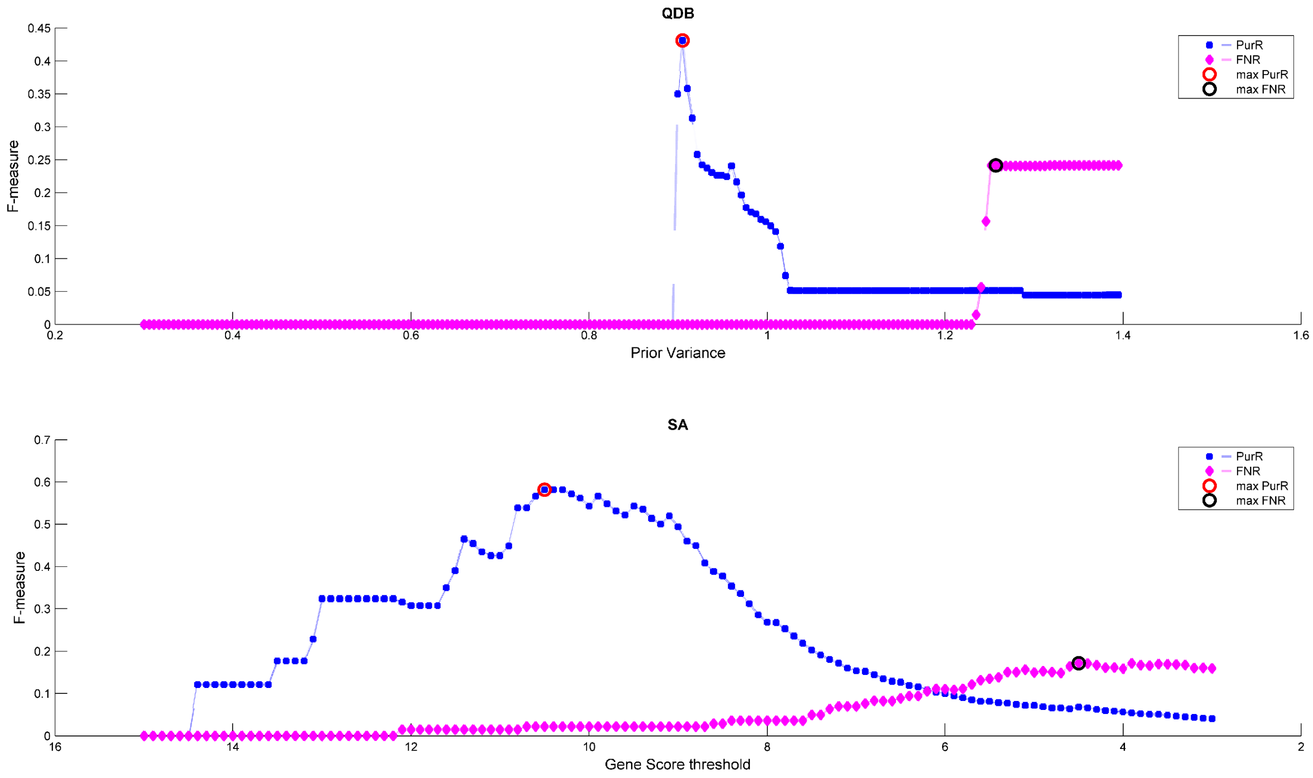


Figure 2- 3 Caption on previous page.

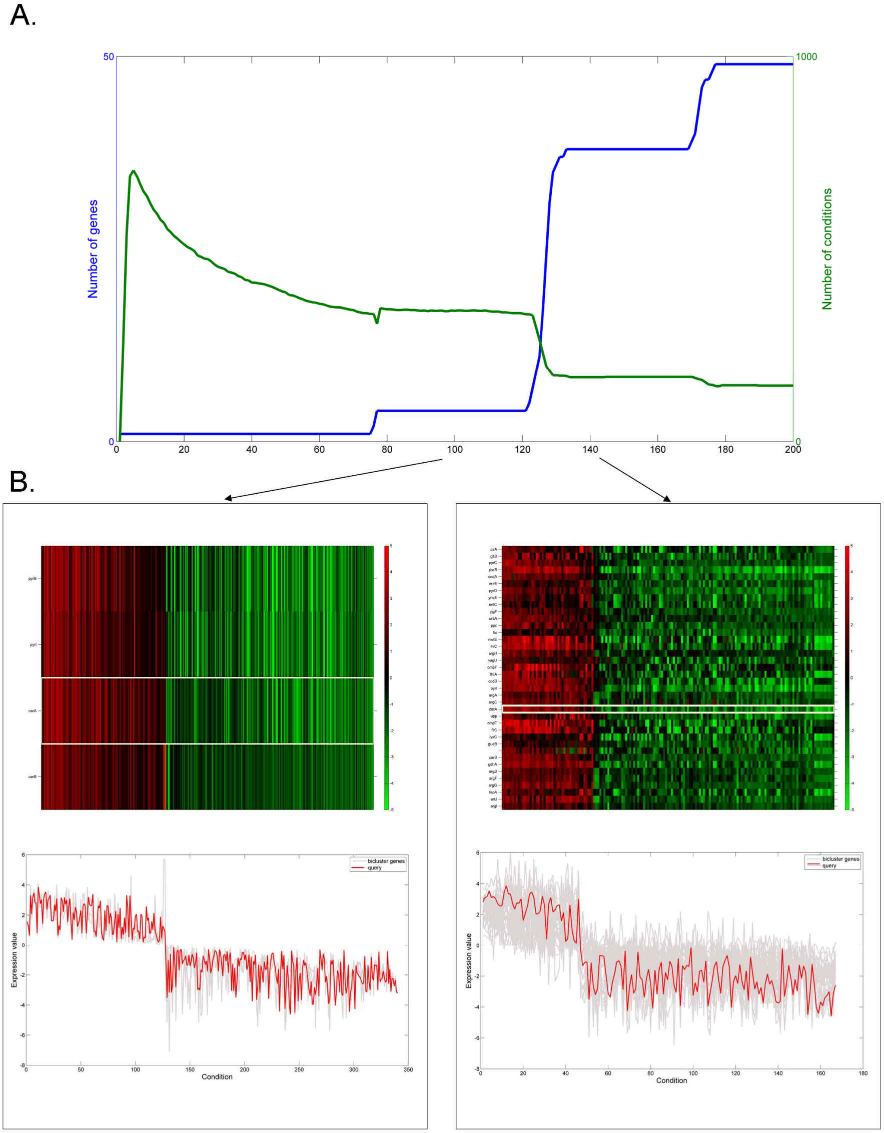


Figure 2- 4 Caption on next page.

Figure 2- 4 Illustration of the resolution sweep approach for QDB. A. Represents the evolution of the bicluster gene (blue) and condition (green) content for an increasing value of the resolution parameter in QDB (*x*-axis). With increasing values of this parameter the biclusters grow in the number of genes whereas the number of conditions they contain decreases. B. The heatmap (top) and expression profiles (bottom) are plotted for two different values of the resolution parameter. Query-genes are indicated in white rectangles in the heatmap plot, whereas in the profile plot the red profile corresponds to the expression profile of the query-genes (*x*-axis are conditions, *y*-axis the expression values). The grey profiles in the profile plot correspond to the remainder of the bicluster genes. The left panel illustrates the bicluster for a smaller value of the resolution parameter, resulting in a bicluster with only 4 genes that are all very tightly coexpressed with the query-gene (profile plot). For a higher value of the resolution parameter (right panel) a bicluster is obtained with more genes and fewer conditions. In addition, here coexpression of the bicluster genes is less tight, as indicated by the profile plot.

2.3.3 Intermezzo: a Bayesian framework for query-based biclustering

At the core of the query-driven biclustering algorithm (QDB) is a probabilistic biclustering framework developed by Sheng *et al.* [76]. The advantage of using a probabilistic framework is that it allows for the introduction of a query as a prior distribution.

2.3.3.1 Biclustering framework

Let $E = (e^{c_1} \quad e^{c_2} \quad ... \quad e^{c_m})$ be a gene expression data set, with e^{c_j} a vector that describes the expression values of the data set's genes under condition *j*. Then the expression data can be described by two Gaussian models, one for the bicluster data and one for the background data. We can not exclude the possibility that the background data also contains modules, however here we concentrate on extracting one module at a time: the one that contains the query-genes. As for the bicluster model we define Gaussian distributions for each condition (column) within the data set: $e^{c_j} \sim N\left(\mu_j^{bcl}, \left(\sigma_j^{bcl}\right)^2\right)$, $j = 1...m$ (Figure 2- 5 – left panel). A similar condition dependent Gaussian distribution is defined for the background model.

Given the bicluster and background model, for each gene and condition within the data set a loglikelihood-ratio can be calculated that a gene/condition belongs to the bicluster as opposed to the background. The loglikelihood-ratios represent respectively the gene and condition scores. Besides the Gaussian models, the framework also incorporates hidden labels for genes and conditions (Figure 2- 5 – right panel). These labels indicate whether a certain gene/condition belongs to the bicluster. Because we work in a Bayesian manner we assume that these labels are the outcome of Bernoulli distributed random variables, one for each row (gene) and column (conditions). Labels are determined both by the loglikelihood ratio as parameters which express prior knowledge on the bicluster size. We refer to Figure 2- 5 for a schematic representation of the framework. The algorithm proceeds by iteratively determining the parameters of the Gaussian models ($\mu_j^{bcl}(\mu_j^{bgd})$ and $\sigma_j^{bcl}(\sigma_j^{bgd})$) from the data, while keeping the labels for genes and conditions fixed and deriving the labels for genes ($\mathbf{g}$) and conditions ($\mathbf{c}$) while keeping the model parameters fixed. This procedure is repeated until convergence to a local optimum is reached. For this task Conditional Maximization is used [77].

2.3.3.2 Query

Within Bayesian statistics prior probabilities are often used to represent ones believe in a certain event. In this framework the query is treated as prior knowledge with respect to the location of the bicluster. Therefore, we discuss in this section the prior distributions within QDB that allow performing a query-based search. Setting the parameters of these prior distributions in an intuitive way allows directing the search for biclusters that contain genes whose expression profile resembles that of the query.

By choosing for an informative *prior on the mean* of the bicluster model μ_j^{bcl} (equation 2.1) it is guaranteed that the bicluster remains around the query-genes. Indeed, this parameter is determined by the weighted average of the mean expression of the bicluster genes $\overline{\mu}_j^{\mathrm{bcl}}$ and the prior mean φ_j^{bcl} (Figure 2- 5). The trade-off between both terms is determined by the number of pseudocounts κ^{bcl} and the number of genes in the bicluster $\|\mathbf{g}\|_1$. By setting the number of pseudocounts to infinity we make this prior very informative and the mean of the bicluster model will be mainly determined by the prior mean. To restrict the location of the bicluster around the query the prior mean φ_j^{bcl} is set equal to the average expression profile of the query-genes.

$$p(\mu_j^{\mathrm{bcl}} \mid \mathbf{g}, c_j, \sigma_j^{\mathrm{bcl}}, \varphi_j^{\mathrm{bcl}}, \kappa^{\mathrm{bcl}}) \propto N\left(\gamma_j^{\mathrm{bcl}}, \left(\lambda_j^{\mathrm{bcl}}\right)^2\right)$$

$$\begin{cases} \gamma_j^{\mathrm{bcl}} = \dfrac{\kappa^{\mathrm{bcl}} \varphi_j^{\mathrm{bcl}} + \|\mathbf{g}\|_1 \overline{\mu}_j^{\mathrm{bcl}}}{\kappa^{\mathrm{bcl}} + \|\mathbf{g}\|_1} \\ \left(\lambda_j^{\mathrm{bcl}}\right)^2 = \dfrac{\left(\sigma_j^{\mathrm{bcl}}\right)^2}{\kappa^{\mathrm{bcl}} + \|\mathbf{g}\|_1} \end{cases} \qquad (2.1)$$

As the prior on the mean determines the location of the bicluster but not its size a second informative prior is introduced on the *variance* of the bicluster model σ_j^{bcl} (equation 2.2). Similar to the prior on the mean this parameter is determined by the weighted average of the variance in expression of the biclustergenes (v^{bcl}) and the priorvariance $\left(s_j^{\mathrm{bcl}}\right)^2$, with proportionate weight of both being determined by the number of pseudocounts v^{bcl} and the number of genes in the bicluster ($\|\mathbf{g}\|_1$). Here also the prior is made informative by setting a high value for the pseudocounts and hence the variance of the bicluster model is primarily determined by the priorvariance $\left(s_j^{\mathrm{bcl}}\right)^2$. Because we do not know in advance the exact value of s_j^{bcl} for which biologically relevant biclusters can be found, a *resolution sweep* approach is used in this framework. As such the value of the priorvariance s_j^{bcl} increases linearly in the course of the algorithm. For low values of this parameter small homogeneous biclusters are obtained whereas for increasing values large, heterogeneous biclusters are retrieved. Hence, the priorvariance mainly determines the size and the compactness of the biclusters.

$$p\left(\left(\sigma_j^{\text{bcl}}\right)^2 \mid \mathbf{g}, c_j, v^{\text{bcl}}, \left(s_j^{\text{bcl}}\right)^2\right) \propto \text{Inv}-\chi^2\left(\eta_j^{\text{bcl}}, \left(\varsigma_j^{\text{bcl}}\right)^2\right)$$

$$\begin{cases} \eta_j^{\text{bcl}} = v^{\text{bcl}} + \|\mathbf{g}\|_1 \\ \left(\varsigma_j^{\text{bcl}}\right)^2 = \dfrac{v^{\text{bcl}}\left(s_j^{\text{bcl}}\right)^2 + \|\mathbf{g}\|_1\left(\bar{\sigma}_j^{\text{bcl}}\right)^2}{v^{\text{bcl}} + \|\mathbf{g}\|_1} \end{cases} \tag{2.2}$$

Besides the prior for the mean and the variance of the bicluster, priors for the other model parameters (mean background model, variance background model, gene labels and condition labels) are defined. These are however of less importance to the introduction of the query and therefore we refer to Dhollander *et al.* [73] for further discussion of these parameters.

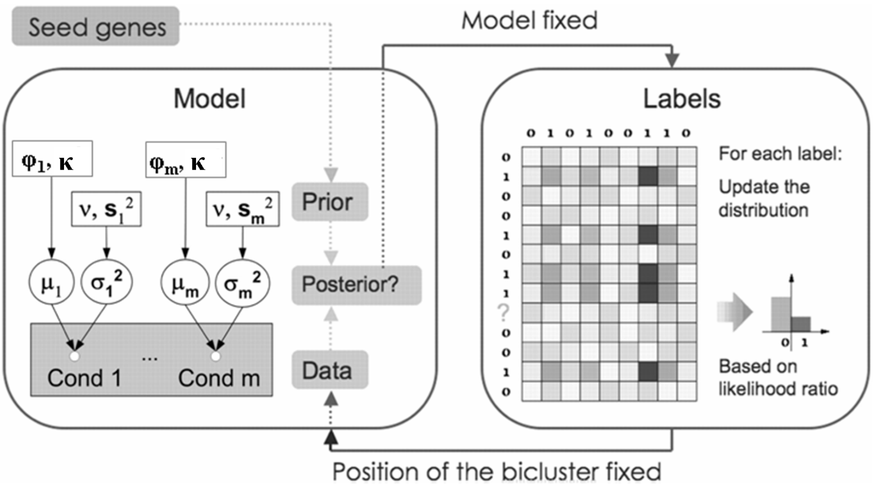


Figure 2- 5 Schematic representation of the probabilistic framework for query-driven biclustering [73]. In the left panel the columnwise normal distributions of the bicluster and the background model are represented. The mean and the variance of these distributions are indicated by circles, whereas the hyperparameters of the corresponding priordistributions are indicated by rectangles. In an iterative procedure the parameters of these statistical models are determined in case the location of the bicluster (determined by gene and condition labels) is presumed fixed. The right panel of the figure represents how the binary gene labels are determined if the model parematers and the condition labels are assumed fixed. Figure is taken from [73].

2.3.4 Bottlenecks of query-based biclustering approaches

2.3.4.1 Selecting the appropriate threshold

As was discussed in the previous section both SA and QDB incorporate a resolution sweep approach and consequently output for the same set of query-genes a large range of biclustering solutions corresponding to different thresholds on coexpression. Therefore running these algorithms for a specific query-set requires selecting the most appropriate biclustering solutions *a posteriori*. Different selection criteria for such task have been previously defined. Here we distinguish between *internal* and *external* selection criteria [78]. Internal criteria refer to selection methods that assess the output of query-based biclustering using criteria inherent to the method used, *i.e.* they use no external data to assess the bicluster quality. External criteria measure the performance by assessing the outcome relative to external information which refers to the true class labels of the genes (*e.g.* GO annotation).

Internal criteria

- As QDB incorporates a Bayesian framework the loglikelihood-score for each solution, corresponding to a certain value of the resolution parameter, can be calculated and can consequently be used to identify the statistically most relevant solutions. To accommodate for model complexity (*i.e.* the likelihood score depends on the number of conditions in the bicluster) a Bayesian model selection criterion based on the Akaike Information Criterion (AIC) [79] was proposed by Dhollander *et al.* [73]. AIC-score = $2\,l - 2k$ with l the loglikelihood and k the number of model parameters. Statistically relevant biclusters are selected as those that represent local optima of the AIC-score.
- An alternative way of selecting the solutions for as well QDB as ISA is to select those resolutions for which the

gene content changes significantly. It is indeed possible that certain values of the resolution parameter introduce larger changes in the gene content of the biclusters than others (see Figure 2- 6 for an example). Such transitions might correspond to biologically interesting patterns, *e.g.* the transition of a bicluster containing tightly coexpressed operon-members towards a bicluster containing less-tightly coexpressed regulon members. To identify those values of the resolution parameter for which the bicluster content changes drastically, the distance (*e.g.* cosine distance) in the gene score vectors obtained for biclustering outcomes corresponding to subsequent values of the resolution parameter can be calculated. Large jumps in this distance measure correspond to large changes in gene content and indicate potentially interesting bicluster solutions.

- Alternatively, to get a sense of the reliability of the identified biclusters, Ihmels *et al.* [12] introduced the *recurrence property*: a bicluster is deemed more stable if it can be retrieved from multiple input sets. However, such a property goes against the intuition of a query-based approach as the assessment measure requires multiple query-sets to be assessed such that the reliability of the module can be obtained.

- Segal *et al.* [80] introduced *expression coherence* as a measure that assesses the tightness in expression of a bicluster. Expression coherence is measured as the fraction of gene pairs within a bicluster for which the Euclidean distance between their expression profiles is less than a certain threshold (usually 5th percentile of the expression distances for all genes in the genome). To estimate the significance of the expression coherence score, the score for the bicluster is compared to that of randomly generated genesets of the same size and with the same number of conditions. To our knowledge this score has not been applied yet to assess the significance of the bicluster solutions but could potentially be applied to this end. However, as it relies on permutation testing to assess the significance of the expression coherence score it is rather computationally intensive.

- Alternative measures have been defined to assess the *expression quality* of the biclusters that do not rely on permutation testing. Cheng & Church [[81] introduced the *mean squared residue* to assess bicluster expression coherence - and this measure was later used to assess bicluster quality [82]. The mean squared residue calculates for each condition the variance of the expression values contained within the bicluster. As for a certain condition the genes within the bicluster should be all coherently up- or down-regulated this variance should be as small as possible. Hence, biclusters that are coherent in their expression profiles have a minimal mean squared residue value. This measure, however, only assesses coherence and gives no information on the differential expression of the genes within the bicluster. Consequently, it is possible to assign high mean squared residue values to biclusters that only constitute noise (*i.e.* genes that have expression patterns close to zero for all conditions). Therefore Zhao *et al.* [72] not only assessed expression coherence of the bicluster genes within all conditions, but also the variance of the gene expression profiles across conditions (*STD-across*). Biclusters that are both coherent in their expression and that contain genes that are differentially expressed have a high ratio of their STD-across to their mean-squared residue. On a similar note Reiss *et al.* [73] introduce the R*oot Mean Square Deviation (RMSD)* which accounts both for data set coverage and coexpression coherence.

External criteria

- An alternative way to choose the appropriate bicluster solutions amongst the spectrum of solutions outputted by the different algorithms is to rely on functional annotations or annotations of the regulatory interactions. Indeed, a biological relevant bicluster corresponds to one that contains functionally coherent genes. Therefore functional enrichment analysis for bicluster approaches and gene set enrichment analysis [83] for prioritization-based approaches constitute a good alternative to select biological relevant biclusters in the presence of reliable gene annotations. However, such an approach biases towards what is known and is only applicable to well-characterized biological systems and/or organisms.

Figure 2- 6 Comparison of different threshold selection methods to detect relevant values for the resolution parameter. Two different query-genes, fdnG and purH were taken as input of QDB. For both query-genes a large range of biclustering solutions was obtained corresponding to different values of the resolution parameter (X-axis). For the outputs for both queries three different selection methods were used to select the most relevant biclustering outputs: (1) the Akaike information criterion ("AIC") (internal), (2) the cosine distance ("cosine") function to assess similarity between genescore vectors (internal) and (3) the enrichment for known targets of the TFs regulating the respective query-genes (FNR and PurR) as assessed by hypergeometric test ("regulonDB") (external). The figure shows that the three different selection criteria do not agree on the most relevant solutions. The solutions indicated by AIC and the cosine distance clearly differ from those deemed biologically most relevant by the hypergeometric test. In addition each of them indicates different possible solutions, making it difficult to decide upon one of them.

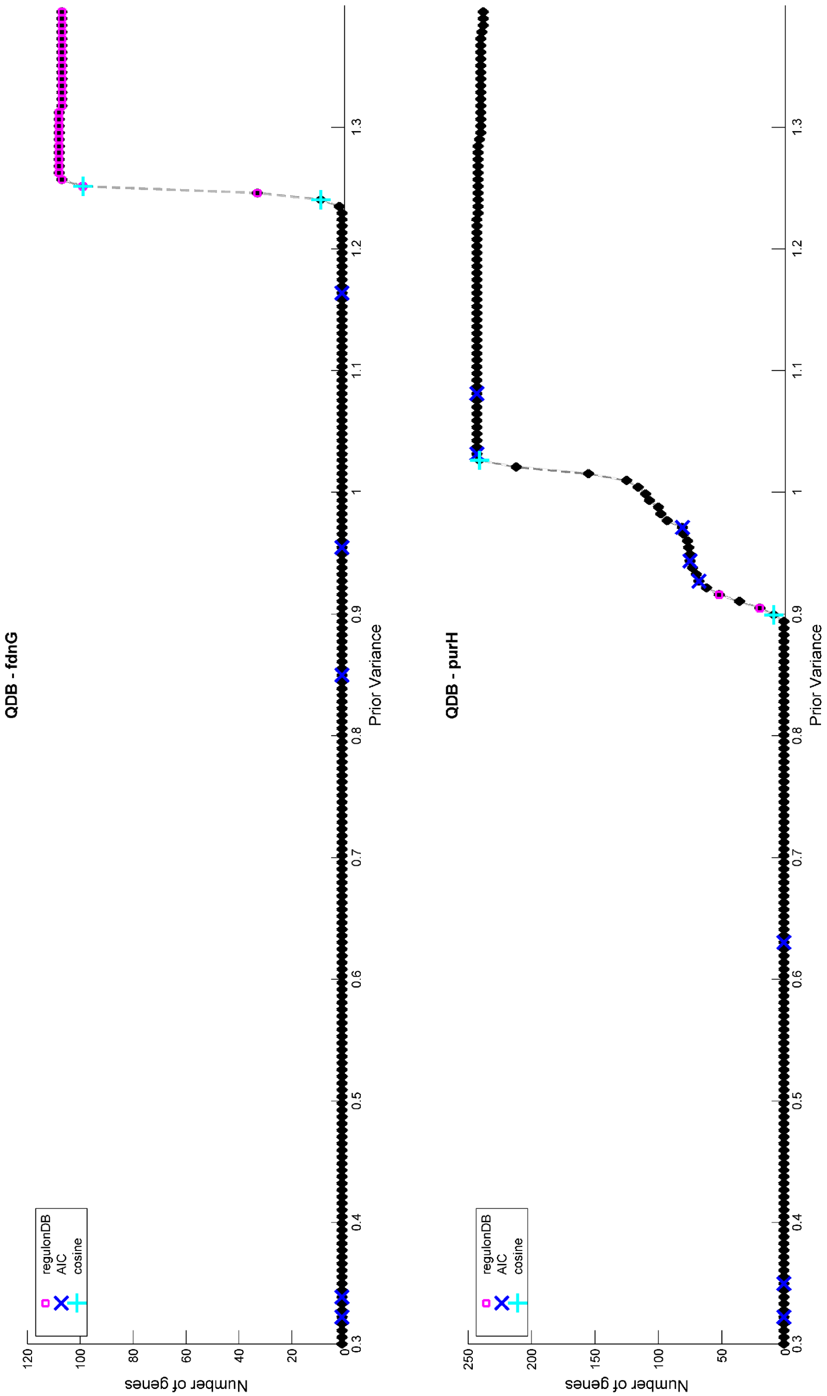


Figure 2- 6 Caption on previous page.

Hence, depending on the method used and the prior information available on the biological system studied different selection criteria might be more appropriate. The internal criteria described here each define objective selection criteria that can be used to pinpoint a few relevant biclusters amongst the whole range of bicluster solutions that is outputted for a query-set in case of a resolution sweep approach. There is however no guarantee that these solutions also correspond to the ones that are most biologically relevant. In addition different selection criteria seem not to agree on the most relevant solutions (Figure 2- 6). In case biological information on the system is well abundant (external selection criteria) this might point the user to the most informative solution, however for non-model systems such information is often not present and in addition this biases the solution strongly to what is known and therefore hides new biology. Remark that this problem of defining a threshold on coexpression is not confined to query-based biclustering and also prioritization methods suffer from this problem.

2.3.4.2 Heterogeneous query lists

The second bottleneck corresponds to the restrictions each of the methods puts on the list of query-genes. Indeed, each of the query-based biclustering methods described here performs very poorly if a list of genes is taken as input which does not contain all genes with a similar expression profile.

This failure is inherent to the way these methods treat the query-set: they all depart from the average expression profile of the query-genes. If a certain fraction of the genes within this set shows an expression profile that deviates from that of the remainder of the genes (*i.e.* outlier genes) then the query-profile will be deteriorated resulting in a loss of specificity. Hence, algorithms that use the average profile of the query-set presuppose knowledge on the expression similarity of the query-genes. The robustness of query-based biclustering methods to such outliers in the query-set can be tuned to a certain extent by for instance making the prior on the mean of the bicluster model less informative in the probabilistic frameworks (QDB and ProBic) [72]. SA, on the other hand, is inherently more robust to such outliers as it does not enforce the bicluster to be centered on the average expression profile of the query-genes. However, even in these cases, once the number of outlier genes in the query exceeds a certain threshold no longer relevant biclustering outputs can be retrieved. Ihmels *et al.* [12] for instance illustrated that the SA is able to kick out a certain set of randomly added genes from a set of related query-genes as long as the ratio random genes to related query-genes does not exceed a certain threshold. In addition, when the query-set consists of gene sets that participate into different functional groupings, these different groupings can not be retrieved as the algorithm only outputs a single bicluster solution for a certain query and coexpression threshold.

Hence, the fact that query-based biclustering methods use the average expression profile of the query-set limits their biological utility. To remediate this problem query-based biclustering methods are able to take as input single query-genes. Therefore if a query-set is heterogeneous in its expression profile, this problem can be alleviated by taking each gene of this set independently as input of the query-based biclustering method. This, however, often leaves the researcher with the interpretation of a large set of query-based search solutions, a problem that will be addressed in the next chapter.

2.4 Applications of query-based search strategies

As was discussed above the different prioritization and query-based biclustering methods implement different strategies to query gene expression compendia. Evidently, this has consequences for their practical use. Here, we give some guidance as to what specific problems can be solved using a certain tool. We cover the major possible applications of query-based biclustering methods as were retrieved from literature. It is, however, possible that other applications than the ones presented here can be thought of and therefore this list is not meant to be exhaustive.

If a researcher is interested in a gene with as of yet unknown function and wishes to get additional insight into this gene's function, then query-based biclustering methods and MEM can be used to interrogate gene expression compendia for other genes coexpressed with this query-gene. Indeed, these methods do not require prior knowledge on the functionality of the query-genes as they work on query sets consisting of single genes. SPELL and GR in contrast require a set of functionally related genes as query and therefore are less appropriate when one wants to get insight into a gene's function.

If one wants to recruit additional genes with a similar function as the query-genes, all methods described here can be used as they are all meant to expand upon the query-set. They can do this by either outputting a well-demarcated set of genes coexpressed with the query as is the case with query-based biclustering methods, or by ranking genes according to their similarity in expression with the query, as is the case for the prioritization methods.

Another possible application of query-based search methods is to refine the set of query-genes. Assume one departs from a set of genes with mostly functionally related genes, but with a few outliers. A possible application of query-based search methods would be to refine this query-set to remove these outlier genes. For this purpose primarily query-based biclustering approaches are opportune as they output well-demarcated sets of genes that not necessarily need to contain all query-genes. GR, in contrast, is not well-suited for this purpose as it chooses the threshold for the condition scores as the one that ranks the query-genes to the top. Ihmels *et al.* [12], for instance, used the Signature Algorithm (SA) to refine the gene set involved in the TCA cycle in *S. cerevisiae* with the homologs of 37 *E. coli* TCA cycle genes as query.

Remark, however, that due to the fact that all query-based search methods discussed here depart from the average query-profile in case multiple query-genes are given as input, their ability to refine the query-sets is rather limited. Indeed, Ihmels *et al.* [12] illustrated that upon adding random genes to a specific query-set the same bicluster solution is retrieved until certain critical threshold is exceeded.

2.5 Conclusion

In this chapter we introduced methods developed to interrogate gene expression compendia for genes coexpressed with a certain user-defined set of genes: the query. We distinguished between prioritization methods, which rank genes according to their condition-dependent coexpression with the query, and query-based biclustering methods, which output well-demarcated set of genes, coexpressed with the query, and their corresponding conditions.

The major distinction between prioritization-based methods and query-based biclustering methods is the fact that the latter couples gene threshold selection and condition threshold selection. This allows query-based biclustering methods to be used to explore different biclusters obtained for different thresholds on gene coexpression in a resolution sweep approach. As depending on the interest of the researcher or the biological case studied this threshold on coexpression might vary, such a resolution sweep approach allows to maximally explore the relevant information contained within a compendium for a certain query-gene by not restricting the outcome of the method to one out of many possible solutions. However, while providing extra flexibility in the biclustering-solutions that can be obtained for a query-gene, this results in difficult post-processing of the outcome. In particular, we presented different possibilities to select the most relevant biclustering solutions for a single query out of a whole range of solutions outputted for different values of the resolution parameter. However, these selection criteria do generally not agree on what the most 'interesting' biclustering outcomes are and what is most 'interesting' seems to generally depend on what aspect of the bicluster a researcher is most interested in (*e.g.* tightness in coexpression, enrichment for certain functional categories, etc.).

In addition, we also highlighted the problem query-based strategies in general struggle with if the query-set is heterogeneous in its expression profile. While query-based biclustering methods, in particular, seem to be able to cope with some outliers in the query-set [12;72], they fail to output a relevant bicluster if the proportion of outliers exceeds a certain threshold [12]. In addition, when the genes within the query-set partition into different functional groupings, these can not be readily detected as the query-based biclustering methods only output one bicluster for a certain query-set and threshold on coexpression.

In the following chapter we introduce an ensemble strategy for query-based biclustering that is able to simultaneously deal with query-sets heterogeneous in their expression profiles and the multitude of solutions generated by a resolution sweep approach.

Chapter 3

An ensemble method for querying gene expression compendia with experimental lists

3.1 Introduction

With the large body of publicly available gene expression data, compendia are being compiled that assess gene expression in a plethora of conditions and perturbations. Comparing own experimental data with these large scale gene expression compendia allows viewing own findings in a more global cellular context and pinpointing inconsistencies between public data and own experiments. In the previous chapter we introduced query-based search approaches such as prioritization-based methods [69-71] and query-based biclustering techniques [12;72;73] to query a gene expression compendium for genes that are coexpressed with a given gene or gene list. These approaches generally combine gene with condition selection to identify genes that are coexpressed with the query in a subset of the compendium conditions.

These query-based methods usually work well when the query list contains one gene only or a set of genes that are mutually tightly coexpressed as they query the expression compendium with the average expression profile of the query-set. However, when query-lists are compiled from the output of experimental assays this list will often contain genes with diverse expression profiles. For instance a query-list derived from a ChIP-chip experiment might partition into different coexpressed groups due to the existence of combinatorial regulation. Hence, when faced with a query-set that is heterogeneous in its expression, these methods will fail to output meaningful gene sets. A solution to this problem is to run these query-based methods on each gene from the query-list separately. This avoids the query-profile to be deteriorated by genes within the list that exhibit a different expression behavior, but will inevitably result in at least partially redundant bicluster solutions as mutually coexpressed genes within the query will output similar biclusters.

A second issue when using query-based biclustering concerns defining a threshold on the minimal level by which the additionally recruited genes should be coexpressed with the query. Indeed, it is often not *a priori* known how tightly a set of genes should be coexpressed to be biologically meaningful. In addition, this level of coexpression might depend on the biological process the query genes are involved in (some processes are more tightly coexpressed than others). To allow for a maximal flexibility, some query-based biclustering methods offer the possibility to use a resolution sweep in which a whole range of possible threshold values is scanned. The most relevant solutions can then be selected *a posteriori*, either based on the intuition of the user or by using other *ad hoc* defined selection criteria (such as functional overrepresentation).

The combined effect of having to run the query-based biclustering on each of the genes from the query list separately with the fact that for each of these single runs also an optional parameter sweep can be performed will result in highly redundant clustering results that all have to be filtered manually by the user.

In this chapter we present an ensemble clustering strategy to merge multiple query-based biclustering results into a few non-redundant consensus biclusters. The chapter is organized as follows: first we give a brief overview of the developed ensemble approach before providing methodological details. Next, we evaluate possible ways of constructing consensus biclusters on real data. Finally, we illustrate the usefulness of the developed approach on a biological case study.

3.2 Overview developed ensemble approach

Using query-based biclustering to interrogate gene expression compendia for gene lists heterogeneous in their expression profiles requires the method to be applied to each gene from the list separately (Figure 3- 1, panel 1), often resulting in at least partially redundant solutions (Figure 3- 2). To summarize the results of these gene-specific solutions, we developed an ensemble approach.

As here we have to merge the outcome of a biclustering instead of a clustering algorithm, we separated the task of merging the gene sets obtained for the query-genes from that of merging the condition sets. Hence, we first construct a new similarity matrix from the ensemble of biclustering results (Figure 3- 1, panel 2) and restrict the construction of this matrix to the gene direction. This similarity matrix (the consensus matrix) represents the evidence for co-clustering of a certain gene pair assessed over multiple biclusterings of the data [54]. Graph clustering [84] is then applied to this matrix to partition the genes into consensus clusters (Figure 3- 1, panel 3). Finally, for each of the obtained gene consensus clusters the corresponding conditions are retrieved from the original biclustering-outputs (Figure 3- 1, panel 4).

Central to the ensemble approach is the construction of the consensus matrix. Depending on whether the query-based biclustering method used incorporates a resolution sweep approach or not, construction of this consensus matrix might run over one (without resolution sweep) or two phases (with resolution sweep).

Query-based biclustering algorithms that incorporate a resolution sweep approach [12;73] evaluate in one run of the algorithm, for a single query-gene, different biclustering outcomes corresponding to a varying threshold on coexpression. Therefore in a first phase, for each query-gene a gene-specific consensus matrix is constructed which summarizes the co-clustering of the genes across these different biclustering outcomes. We reason that genes that co-occur in both fine-grained and coarser-grained biclusters, corresponding to a decreasing tightness of coexpression, are more likely to be truly functionally related than genes that only co-occur in coarser-grained biclusters. Therefore, genes that frequently co-occur over the results obtained with the varying resolution

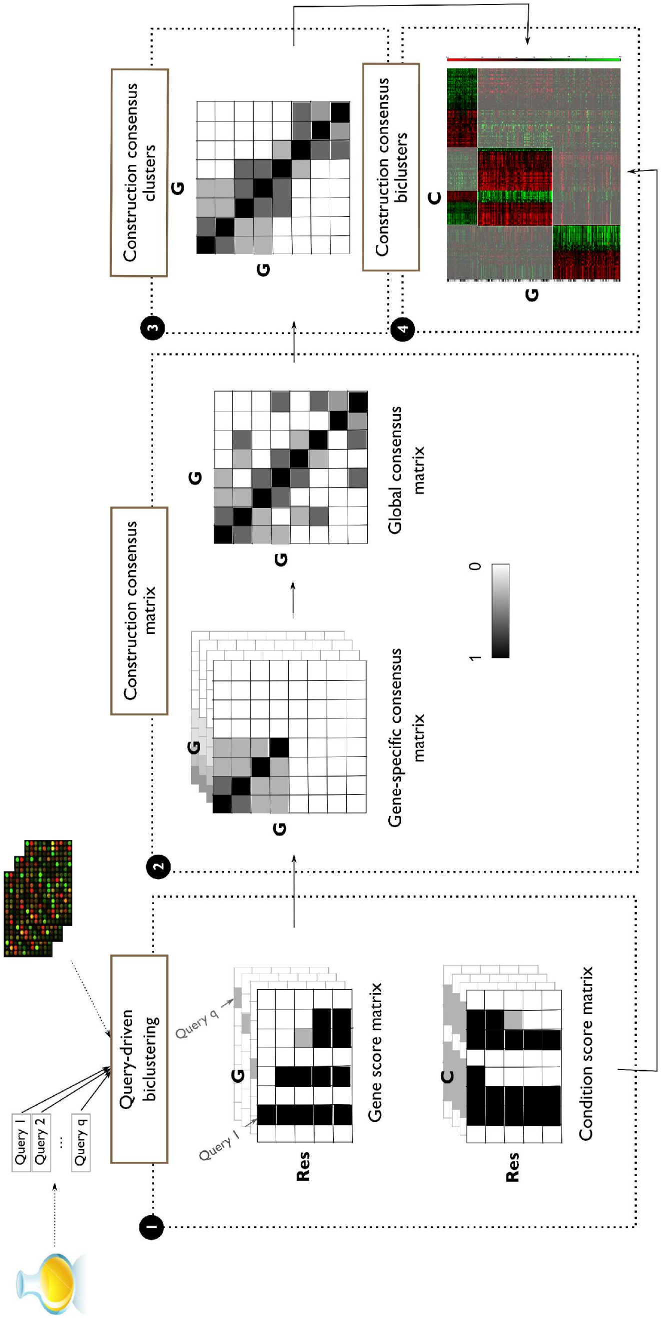

Query 1
Query 2
...
Query q
Query-driven biclustering
Query 1
Query q
G
Res
Gene score matrix
C
Res
Condition score matrix
1
2
Construction consensus matrix
G
G
Gene-specific consensus matrix
G
G
Global consensus matrix
1
0
3
Construction consensus clusters
G
G
4
Construction consensus biclusters
C
G

Figure 3- 1 Caption on next page.

Figure 3- 1 Overview of the ensemble biclustering approach. 1) Targets from a ChIP-chip analysis were each taken as input of the query-driven biclustering algorithm (QDB) [73]. Per single ChIP-chip target QDB results in a gene score and condition score matrix (G and C refer respectively to the gene and condition dimension of the matrices), containing for each value of the resolution parameter (indicated with Res) the loglikelihood score that a gene or condition belongs to the bicluster. Shades of grey are representative for the magnitude of the gene scores and condition scores. 2) Constructing the consensus matrix proceeds in two steps in which first a gene-specific consensus matrix is constructed for each query from its gene score matrix. This gene-specific consensus matrix summarizes for each query-gene the QDB-solution obtained at different values of the resolution parameter. In a second step the gene-specific consensus matrices for all genes in the ChIP-chip list are merged into a single consensus matrix, representing the frequency of co-occurrence (again indicated by shades of grey) of two genes across the different QDB-solutions in which at least one of the genes occurs. 3) Next, by applying graph clustering the consensus matrix is partitioned into consensus clusters. 4) Eventually, consensus biclusters can be obtained by retrieving for each consensus cluster the corresponding conditions from the original QDB-solutions.

parameter will obtain higher consensus scores (*i.e.* have a higher weight of belonging to the same consensus bicluster) than those that only sporadically co-occur. This first step is only needed when using a query-based biclustering algorithm that uses a resolution sweep.

In a second phase, these gene-specific consensus matrices are merged into a single consensus matrix, which summarizes the outcomes of the query-based biclustering runs across all query-genes in the list. Here, we remove the redundancy in the gene sets obtained by the different biclustering runs by assuming that genes that repeatedly co-occur in different biclustering outcomes of distinct query-genes form a single grouping. However, as not all query-genes give rise to similar gene sets, we aim not only at reducing the redundancy amongst the gene sets, but also to preserve to a maximal extent biclustering results that were

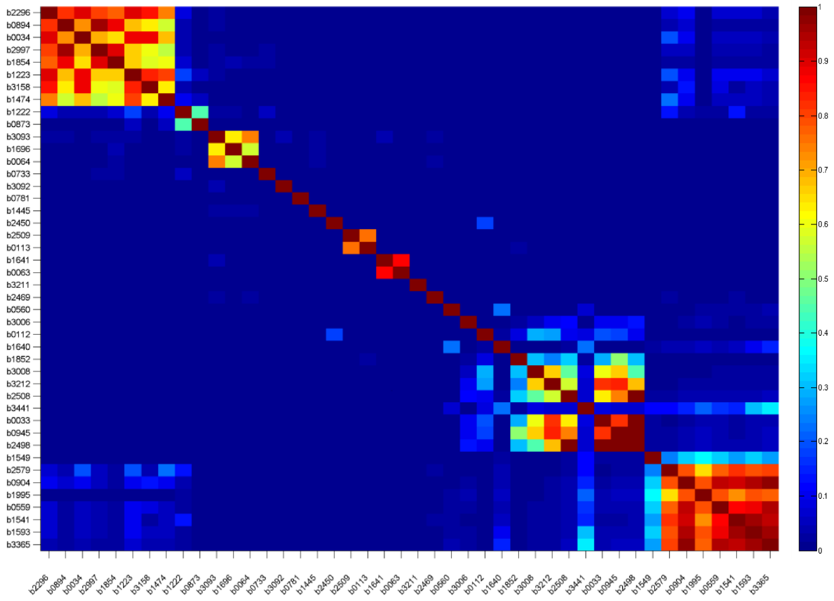


Figure 3- 2 Illustration of redundancy amongst the query-driven biclustering outcomes derived for a query-list from a ChIP-chip experiment [85]. This query-gene by query-gene matrix reflects the redundancy amongst the query-driven biclustering results. Each matrix element represents the maximal overlap in terms of the genes belonging to each query-driven biclustering solution, as assessed by geometric coefficient (3.3.2.3). A value of 1 means that for a certain value of the resolution parameter the two query-driven biclustering solutions have exactly the same gene content whereas a value of 0 indicates that for none of the values of the resolution parameter the two query-driven biclustering results have a gene in common. The matrix was clustered and clusters represent groups of query-genes with highly similar QDB-outcomes. This figure indicates that there are indeed query-genes with highly similar query-driven biclustering results and therefore redundancy in the output exists.

not repeatedly retrieved for different query-genes (*i.e.* the non-redundant gene sets). As such we stress genes that co-cluster consistently across different runs while also retaining gene sets specific to a certain query-gene in order to retain as much information as possible contained in the original biclustering outcomes.

3.3 Methods

3.3.1 Query-driven biclustering

The strategy proposed in this paper can be used in conjunction with any query-based strategy. For illustrative purposes we use here the query-driven biclustering algorithm (QDB) [73]. As QDB incorporates a resolution sweep approach, one run of the algorithm on a single query-gene outputs multiple biclustering solutions, each corresponding to a different value of the resolution parameter. We further refer to the output of one such run of the algorithm as the 'QDB-solution'.

Briefly, QDB [73] incorporates a Bayesian framework and gives as output the loglikelihood ratio of each gene and condition belonging to the bicluster versus a background model, respectively called the gene score and condition score. A prior on the mean of the bicluster distribution enforces the bicluster to be centered on the expression profile of the query-gene. A second prior, on the bicluster variance, determines the degree of coexpression within a bicluster. The algorithm uses a sweep on the prior of the bicluster variance (*i.e.* resolution sweep) to evaluate in a single run of the algorithm all possible solutions that correspond to different degrees of coexpression (*i.e.* the resolution). As such the algorithm grows a bicluster around a query, first outputting bicluster solutions with only a limited number of genes and then growing the bicluster, for increasing values of the resolution parameter, in the number of genes. With an increasing number of genes, the bicluster gets coarser-grained as coexpression amongst the genes becomes less pronounced. A single run of the algorithm outputs the results for a specific query-gene and multiple values of the resolution parameter and consists of two matrices: the gene score matrix and condition score matrix. These matrices contain for each resolution the gene and condition scores (hence for each resolution a gene score vector and condition score vector is given).

3.3.2 An ensemble approach for query-based biclustering

An overview of the computational framework is given in Figure 3- 1.

3.3.2.1 Consensus matrix construction

First, all results for a single QDB-run on a single query-gene from the list, obtained for n_{res} different values of the resolution parameter, are merged into a *gene-specific consensus matrix* according to the same principle as Monti *et al.* [54]. Matrix entries C_{ij} reflect the average gene pair-to-bicluster membership across a sweep over the resolution parameter (*i.e.* the coexpression threshold):

$$C_{ij}^{qdb} = \frac{\sum_{t=1}^{n_{res}} Gs_{i,t}.Gs_{j,t}}{n_{res}} \qquad (3.1)$$

Here, $Gs_{i,t}$ represents the genescore for gene i for the t-th value of in total n_{res} possible values for the resolution parameter.

In a second step these gene-specific consensus matrices are merged in a final *consensus matrix*, which summarizes the outcomes of the QDB-runs across all n_{qdb} query-genes in the list. For this purpose, we introduce a *distributed consensus matrix construction* approach. Here, the frequency of co-occurrence for a gene-pair (gene consensus score) is calculated as its sum over all gene-specific consensus matrices, normalized by the number of times a certain gene pair co-occurred in the gene-specific consensus matrices:

$$C_{ij}^{global} = \frac{\sum_{r=1}^{n_{qdb}} C_{ij}^{r}}{\sum_{r=1}^{n_{qdb}} O(C_i^r, C_j^r)} \qquad (3.2)$$

with $O(C_i^r, C_j^r)$ representing the co-occurrence function which is 1 if both genes belong to the same specific consensus matrix and otherwise 0. The reason for this altered normalization as compared to the gene-specific consensus matrix is that simply averaging the gene-specific matrices across all QDB-solutions would erroneously downweigh those genepairs specific to a certain QDB-run (*i.e.* those that correspond to a query-gene not coexpressed to any of the other genes in the list) and reward genepairs retrieved by multiple QDB-runs (*i.e.* those that correspond to query-genes that are mutually coexpressed).

We also tested whether the following transformations of the consensus matrix could further improve the quality of the obtained ensemble solution:

- The Topological Overlap Matrix (TOM) [86] which does not only account for pairwise gene-gene co-occurrence, but also accounts for similarity in the other genes with which both genes co-occur in the QDB-solutions.
- Pruning the consensus matrix by setting statistical insignificant consensus scores (*i.e.* low consensus scores) to zero. Statistical relevance of consensus scores is assessed by the disparity filter [87]. This method compares for each gene *i* the distribution of its consensus scores with all other genes (*i.e.* the values $C_{i\cdot}^{global}$) to a null model and sets the least significant scores to zero. We choose our significance threshold such that 90% of the total consensus score (*i.e.* the sum of all consensus scores) was retained to avoid eliminating too many elements with large consensus scores from the matrix.

3.3.2.2 Extracting consensus clusters from the consensus matrix

We aim at obtaining non-redundant gene consensus clusters from the consensus matrix, each corresponding to distinct QDB-solutions. This problem can be approached as the clustering of a weighted graph, with weighted edges representing the gene consensus scores and nodes representing the genes. Here we compared several graph clustering methods that can be applied to weighted graphs. These methods include the Newman spectral modularity algorithm [88], affinity propagation (AP) [89], Markov clustering (MCL) [90], hierarchical clustering and a recently published fuzzy spectral graph clustering method [55].

The Newman spectral modularity algorithm and the fuzzy spectral method select automatically the number of clusters. To select the optimal number of clusters for the remainder of the methods, we use for AP the default parameters, for MCL the efficiency measure [90] and for hierarchical clustering the median split silhouette coefficient [91].

Graph clustering of the consensus matrix might result in gene consensus clusters that do not contain any of the genes included in the query-list. As these clusters do not contribute to the interpretation of the query-list in terms of the expression compendium these clusters are discarded.

3.3.2.3 Obtaining consensus biclusters

To map the conditions to the gene consensus clusters, we trace back the obtained gene consensus clusters to the original QDB-solutions from which they were derived. To find the corresponding QDB-solutions we use the geometric coefficient [92] to quantify the overlap in the genes for the original QDB-solutions and the consensus clusters:

$$Overlap = \frac{\left|G_{cons} \cap G_{qdb}\right|}{\sqrt{\left|G_{cons}\right|\left|G_{qdb}\right|}} \quad (3.3)$$

with G_{cons} representing the genes in the consensus cluster and G_{qdb} the genes in the original QDB-solution. Since, each QDB-solution corresponds to different gene sets retrieved for different values of the resolution parameter, the overlap is calculated for each resolution separately. The condition score vector (*i.e.* QDB loglikelihood scores) that corresponds to the resolution for which this overlap is maximized is then retained. Next, the condition consensus scores for a particular gene consensus cluster are calculated as the weighted mean of all condition score vectors retained for this consensus cluster. The weight is chosen equal to the geometric coefficient, hence giving higher weight to condition score vectors belonging to bicluster outcomes better reflected by the gene consensus clusters. Finally conditions with a consensus score exceeding 0.75 (conditions occur in at least 75% of the condition score vectors) are retained.

3.3.3 Applying the ensemble approach

As a proof of concept we applied the proposed ensemble biclustering approach to presumed FNR-targets obtained by ChIP-chip analysis [85]. In this experiment binding of FNR under anaerobic conditions was evaluated. The authors identified 63 genomic regions at which FNR binds. These 63 genomic regions were mapped to 90 genes as the authors assigned a bound region located in the promoter region of two divergently regulated genes to both genes [85].

Each of the 90 potential FNR-targets was used separately as query in the query-driven biclustering algorithm [73]. As gene expression data set an *E. coli* gene expression compendium spanning 870 conditions was used [8]. For each of these query-genes 200 biclustering outcomes were obtained corresponding to 200 different values of the resolution parameter. For 44 out of the 90 query-genes QDB-solutions could be retrieved which contained all together 61 out of the 90 FNR ChIP-chip targets. For the remaining 29 genes no significant QDB-solutions were obtained, either because no additional genes were found to be coexpressed with the query-gene (26 cases) or because the number of conditions under which the genes were found to be coexpressed was not sufficient (here at least 10 conditions were required to be included in the bicluster).

For each of these 44 query-genes a gene-specific consensus matrix was constructed to aggregate its 200 biclustering outcomes. QDB-solutions for each of these 44 query-genes were at least partially overlapping (Figure 3- 2), therefore these 44 gene-specific consensus matrices are merged into one consensus matrix. Consensus biclusters are finally obtained by applying graph clustering to the (transformed) consensus matrix and by retrieving the matching condition set from the 44 QDB-solutions.

To analyze the gene content of these consensus biclusters gene functional GO-categories were taken from EcoCyc [93]. To verify the presence of known FNR-targets within the consensus biclusters the known *E. coli* regulatory network was taken from RegulonDB [94].

Heatmap visualizations of the consensus biclusters were made by using ViTraM [95].

3.3.4 Performance evaluation

The following quality measures for the obtained consensus biclusters were devised, each of which measures a different aspect of these consensus biclusters.

Degree to which they recapitulate the information contained within the original QDB-solutions: The *overlap measure* calculates for each consensus bicluster its maximal overlap in number of genes with the original QDB-solution by the geometric coefficient (see formula (3.3)). To obtain one score for the whole set of consensus biclusters obtained from a single consensus matrix we averaged this measure for the different consensus biclusters. This overlap measure is however asymmetric: different consensus biclusters might show maximal overlap with the same QDB-solution and hence the overlap might be high while the consensus solution is biased towards certain QDB-solutions only. Therefore we also calculate the query-gene coverage (*coverage measure*), which assesses whether the obtained consensus biclusters cover the information content of the QDB-solutions in its entirety. The query-gene coverage is calculated as the number of query-genes in the original QDB-solutions that belong to a non-trivial consensus bicluster (*i.e.* a consensus bicluster with more than 1 gene).

Degree to which they remove redundancy: The *redundancy measure* evaluates the extent to which the consensus biclusters are able to reduce the redundancy present in the original QDB-solutions. We assume that query-genes with largely overlapping (or highly redundant) QDB-solutions should belong to the same consensus bicluster. Consequently, we use Normalized Mutual Information (NMI) [96] to assess how well a clustering of the query-genes based on overlap in their QDB-solutions corresponds to the partitioning of the query-genes according to the consensus biclusters (see A.3).

Biological relevance: The *functional coherence measure* assesses the biological relevance of the set of consensus biclusters produced from the consensus matrix. For each consensus bicluster a p-value for functional enrichment is calculated using the hypergeometric test ($p < 0.01$, Bonferroni-corrected for multiple testing). As from each consensus matrix multiple consensus biclusters are obtained, we use the clustering score function [59] to aggregate all p-values obtained for all consensus biclusters derived from the same consensus matrix into a single score. Let n_s be the number of significantly enriched clusters and n_i the number of insignificant clusters for a p-value cut-off *c*, then the functional coherence of a consensus solution is defined as follows:

$$fc = 1 - \frac{\sum_{k=1}^{n_s} \min(p_k) + (n_i * c)}{(n_s + n_i) * c} \qquad (3.4)$$

Statitical quality: We also assessed the objective quality of the consensus biclusters by assessing whether consensus clusters derived from the consensus matrix have more intra-cluster edges than between-cluster edges as evaluated by the *modularity* function. Modularity Q [97] compares, given a clustering and corresponding consensus matrix C_{ij}, the fraction of the edges that falls within a given cluster minus the expected fraction if edges were distributed at random. The higher the modularity the better the cluster separation, with a maximum value of 1 for strong modular structures. Let k_i be the weighted degree of node *i*, *m* the total weighted number of edges and the δ function yields 1 if vertices *i* and *j* belong to the same cluster (otherwise the function is 0), then the modularity is given by:

$$Q = \frac{1}{2m} \sum_{ij} (C_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \qquad (3.5)$$

Each of these metrics has a maximum value of 1, which makes their interpretation and comparison rather straightforward.

3.4 Results

3.4.1 Analysis of different ensemble constructs

To develop an ensemble approach that was able to maximally remove redundancy by merging the outcome of redundant biclusters into a single consensus bicluster, while also retaining as much as possible the information contained within the original query-based biclustering results (the results obtained before applying the ensemble approach), we tested (1) different transformations of the final consensus matrix (see 3.3.2.1) and (2) different graph clustering methods (see 3.3.2.2) to extract the consensus biclusters from the consensus matrix. The final consensus matrix, before applying any of the matrix transformation methods, was obtained as described in 3.3.3.

To evaluate each of these strategies we used the following metrics that assess the quality of the final consensus biclusters: the degree to which the consensus biclusters are capable of recapitulating the original bicluster ("the overlap and coverage measure"), the degree to which they remove redundancy ("redundancy measure"), their biological relevance ("the functional coherence") and their statistical relevance ("modularity").

We separately evaluated the result of both the matrix transformation and the graph clustering step as to not confound their effects on consensus bicluster construction. First, we compared the results obtained by applying different transformations of the consensus matrix in combination with the same graph clustering method. Specifically, we used a non-transformed consensus matrix, pruned consensus matrix and a Topological Overlap Matrix (TOM) (see 3.3.2.1). Pruning of the consensus matrix (see 3.3.2.1) might improve the outcome of the ensemble biclustering approach by excluding noise from the consensus matrix, as this filtering sets low consensus scores to zero. TOM on the other hand aims at increasing the robustness of the consensus scores by modifying them such that they do not only account for pairwise co-occurrence of the genes within a biclustering solution, but by also accounting for their joint co-occurrence with other genes. Although the absolute values of the quality measures depended on the used graph clustering method we consistently observed that using the pruned consensus matrix outperformed the two other consensus matrices for all evaluation metrics (Figure A- 1). As the effect of matrix transformation and graph clustering do not seem to be confounded we show in Figure 3- 3A one representative obtained for a single graph clustering method (*i.e.* fuzzy clustering, the clustering for which the best quality measures were obtained). Although the effect was marginal, graph clustering with a pruned consensus matrix resulted in a slightly higher modularity value, biological relevance and overlap metric than using a non-transformed consensus matrix. A higher overlap value indicates that the consensus biclusters obtained with the pruned matrix are more truthful to the original QDB-solutions than those obtained with the non-transformed matrix. A higher modularity suggests that by applying the disparity filter before graph clustering more densely connected clusters can be obtained from the consensus matrix. These results suggest that it indeed makes sense to first prune the consensus matrix by filtering out irrelevant consensus scores before performing graph clustering. Applying TOM, on the other hand, seems to disturb the match between the consensus solution and the original QDB-solutions as indicated by the lower overlap and redundancy measures.

Figure 3- 3B illustrates the effect of using different graph clustering methods on the obtained quality metrics. As a representative example we show the results obtained by applying the different graph clustering methods on the pruned consensus matrix. Results obtained by using the alternative consensus matrices as input are represented in Figure A- 1. We observe that fuzzy clustering has the best trade-off between removing redundancy while still agreeing largely with the original QDB-solutions (with an average overlap of 70% with the original QDB solutions and a redundancy score of 0.84). AP performs similarly for these metrics but fuzzy clustering has a higher coverage for query-genes and outperforms AP with respect to the cluster density as assessed by the modularity measure. Figure 3- 4 indeed shows that the clustered consensus matrix obtained with fuzzy clustering shows

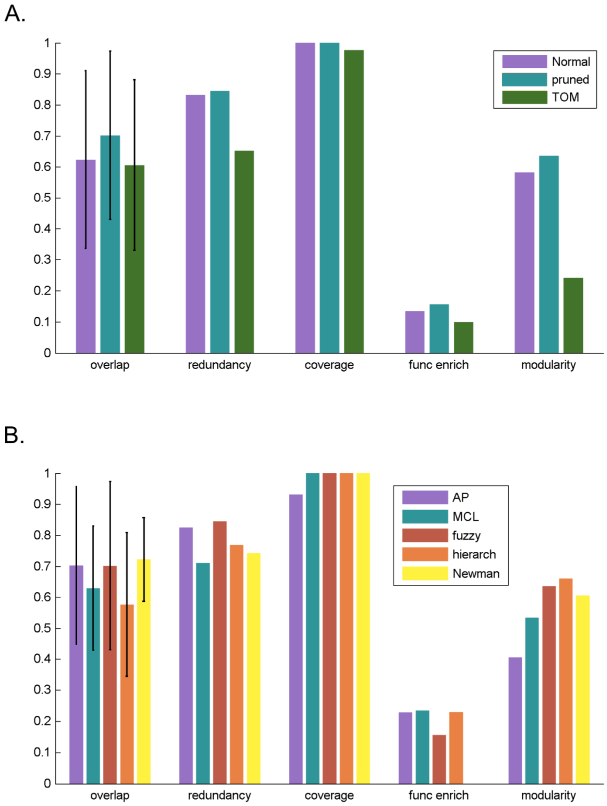


Figure 3- 3 Comparison of different ways to construct the consensus biclusters. A. Compares the influence of using different consensus matrix transformations on the quality of the final consensus biclusters assessed by respectively their overlap with the original QDB-solutions ('overlap'), the extent to which redundancy amongst the QDB-solutions is removed ('redundancy'), their coverage for query-genes ('coverage'), their functional coherence ('func enrich') and the modularity of the obtained clustered consensus matrix ('modularity') (*x*-axis). For illustrative purposes we show the assessment of the final consensus biclusters for different matrix transformations, each time used in combination with fuzzy clustering. B. Comparison of the effect of using different graph clustering methods to extract from the consensus matrix the final consensus biclusters. Same assessment criteria as in panel A were used. Missing bars reflect a value of zero for the corresponding evaluation metric. For illustrative purposes only results obtained on the pruned consensus matrix are shown.

a more consistent block-diagonal structure than that obtained with other cluster algorithms.

The results in Figure 3- 3 and Figure A- 1 further show that irrespective of the graph clustering or consensus matrix transformation method used, the obtained consensus biclusters are in good agreement with the original QDB-solutions, illustrated by high values for the overlap and coverage measure. In addition all tested strategies are capable of grouping highly redundant solutions into the same consensus bicluster as is illustrated by the redundancy evaluation metric. As the combination of the pruning transformation step with the fuzzy clustering outperformed the other methods for the used quality criteria we used this combination in the subsequent application.

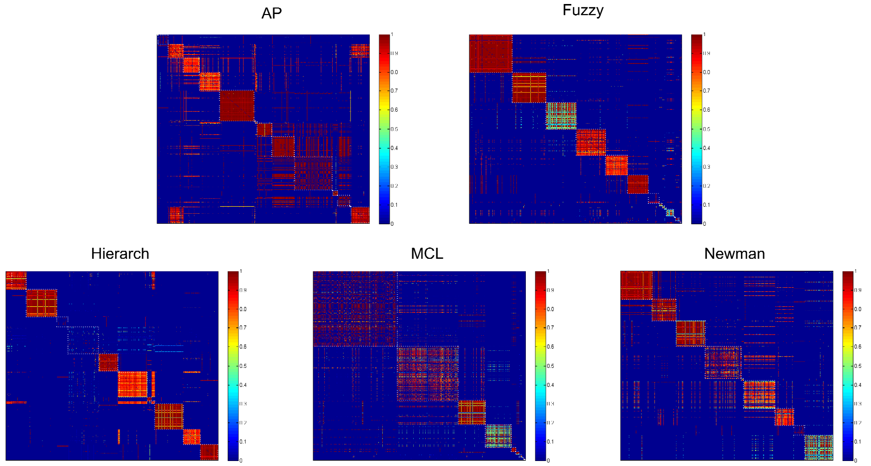


Figure 3- 4 Comparison of different graph clustering methods. The consensus matrix was sorted according to the cluster memberships obtained with the different graph clustering. A crisp clustering results in a block diagonal ordering of the values with high consensus scores.

3.4.2 A ChIP-chip case study

Chromatine immunoprecipation in combination with microarray technology (ChIP-chip) is increasingly being used to measure protein-DNA interactions *in vivo*. Being a high-throughput technology, ChIP-chip data inevitably gives rise to false positives. In addition, the technology fails to distinguish non-functional from functional binding [98]. Hence ChIP-chip experiments need to be backed up by expression data that provide information on whether the identified target genes are indeed being regulated by the bound transcrtion factor (TF).

We applied the proposed workflow to presumed FNR-targets obtained by ChIP-chip analysis (see 3.3.3) [85]. In this experiment binding of FNR under anaerobic conditions was evaluated, yielding a list of 90 query-genes containing 26 known FNR-targets [94]. For 44 out of the 90 query-genes, biclusters could be retrieved that were efficiently merged into 17 consensus biclusters (Table A- 1). These 17 biclusters cover 61 of the 90 ChIP-chip targets, amongst which 24 known FNR-targets (Table 3 - 1).

In what follows we use the results of these consensus biclusters to interpret the results of the FNR ChIP-chip experiment, *i.e.* to distinguish within the list of possible ChIP-chip targets the functional from the non-functional or false positive ones and to pinpoint likely false negative targets that were not recovered by the ChIP-chip experiment.

Table 3 - 1 Overview of ChIP-chip targets. The table (next page) contains the name and locustag of the 90 ChIP-chip targets. Black shaded boxes indicate whether a ChIP-chip target belongs to any of the consensus biclusters (‘In consensus’), whether it is a known FNR target according to RegulonDB (‘FNR target’), whether it belongs to any of the consensus biclusters enriched for ChIP-chip targets (‘In enrich’) ($p < 0.01$, Bonferroni-corrected) or whether it belongs to any of the consensus biclusters with a high coverage for known FNR-targets (‘In coverage’) (at least one third of the genes is a known target).

LocusTag	Name	In consensus	FNR target	In enrich	In coverage
b1241	adhE	■	■	■	■
b3365	nirB	■	■	■	■
b4139	aspA	■	■	■	■
b4154	frdA	■	■	■	■
b2579	yfiD	■	■	■	■
b0733	cydA	■	■	■	■
b0894	dmsA	■	■	■	■
b0904	focA	■	■	■	■
b1109	ndh	■	■	■	■
b4123	dcuB	■	■	■	■
b1474	fdnG	■	■	■	■
b2296	ackA	■	■	■	■
b4070	nrfA	■	■	■	■
b1256	ompW	■	■	■	■
b0034	caiF	■	■	■	■
b0113	pdhR	■	■	■	■
b1223	narK	■	■	■	■
b0873	hcp	■	■		■
b3092	uxaC	■	■		■
b1222	narX	■	■		■
b2498	upp	■	■		
b4238	nrdD	■	■		
b0781	moaA	■	■		
b3212	gltB	■	■		
b1854	pykA	■		■	■
b0559	ybcW	■		■	■
b1541	ydfZ	■		■	■
b3157	yhbT	■		■	■
b2997	hybO	■		■	■
b2509	xseA	■		■	■
b3158	yhbU	■		■	■
b1445	ydcX	■			■
b0033	carB	■			
b0945	pyrD	■			
b1852	zwf	■			
b2508	guaB	■			
b4069	acs	■			
b1593	ynfK	■			
b1640	ydhH	■			
b1696	ydiP	■			
b2450	yffS	■			
b3211	yhcC	■			
b3440	yhhX	■			
b4141	yjeH	■			
b0063	araB	■			
b0112	aroP	■			
b0715	abrB	■			
b3006	exbB	■			
b0560	nohB	■			
b0961	yccF	■			
b1548	nohA	■			
b3093	exuT	■			
b4140	fxsA	■			
b1995		■			
b3441	yhhY	■			
b4142	groS	■			
b0064	araC	■			
b1643	ydhI	■			
b2469	narQ	■			
b3008	metC	■			
b1549	ydfO	■			
b2499	purM		■		
b3493	pitA		■		
b3206	npr				
b1473	yddG				
b2295	yfbV				
b0514	glxK				
b1108	ycfP				
b1255	yciC				
b1333	uspE				
b1341	ydaM				
b2298	yfcC				
b2443	yffL				
b3009	yghB				
b3120	yhaB				
b3492	yhiN				
b4038	yjbI				
b1642	slyA				
b2468	aegA				
b3746	yieN				
b1343	dbpA				
b1853	yebK				
b4155	poxA				
b1342	ydaN				
b1641	slyB				
b1697	ydiQ				
b2580	ung				
b3747	trkD				
b0962	helD				
b4351	mrr				
Total		61	26	24	28

Figure 3- 5 represents the 4 most interesting consensus biclusters obtained by interrogating the *E. coli* expression compendium with the initial FNR ChIP-chip target list. Consensus biclusters 5 and 12 contain a high enrichment for ChIP-chip targets (Table A- 1), suggesting that the ChIP-chip targets within these consensus biclusters constitute functional targets. Indeed, we expect that ChIP-chip targets of the same TF are co-regulated and thus should be coexpressed. The hypothesis that the targets within these consensus biclusters are functional targets is further supported by the observations that these consensus biclusters are mainly composed of conditions that measure the effect of oxygen (Figure 3- 5A) and show a high coverage for known FNR-targets (Figure 3- 5B). In total these consensus biclusters covered 24 FNR ChIP-chip targets of which 7 novel ones, not documented in [94]. In addition they contained

Figure 3- 5 Consensus biclusters obtained by interrogating an *E. coli* expression compendium with the target list of a FNR ChIP-chip experiment. A. Heatmap representation of the consensus biclusters 5, 10, 12 and 16. Rows represent the genes, whereas columns represent the conditions. Different consensus biclusters are indicated by colored rectangles. At the top of the picture conditional categories present within the gene expression data set are shown [8]. A colored square on top of the heatmap indicates that a condition belongs to a particular conditional category. B. Overview of the content of the 4 consensus biclusters in terms of the number of ChIP-chip based ('ChIP-chip') and previously described FNR-targets ('FNR'). Ratios in the table represent the number of transcription units in the consensus bicluster belonging to a certain category (*i.e.* (1) identified by ChIP-chip ('ChIP-chip') and (2) known FNR-target ('FNR')) against the total number of ChIP-chip targets in the consensus bicluster (left – "ChIP-chip"), the total number of FNR-targets in the consensus bicluster (left – "FNR") and the total number of transcription units in the consensus bicluster (right – "Total"). 'Novel' refers to ChIP-chip targets not documented to be regulated by FNR according to RegulonDB [94]. Consensus biclusters indicated with an asterisk are significantly enriched in ChIP-chip targets. The number of known FNR-targets that correspond to a certain regulatory mode of FNR (repressor, activator, dual regulator or combinatorial regulation with NsrR) are also indicated. For each consensus bicluster, the predominant regulatory mode is indicated in red.

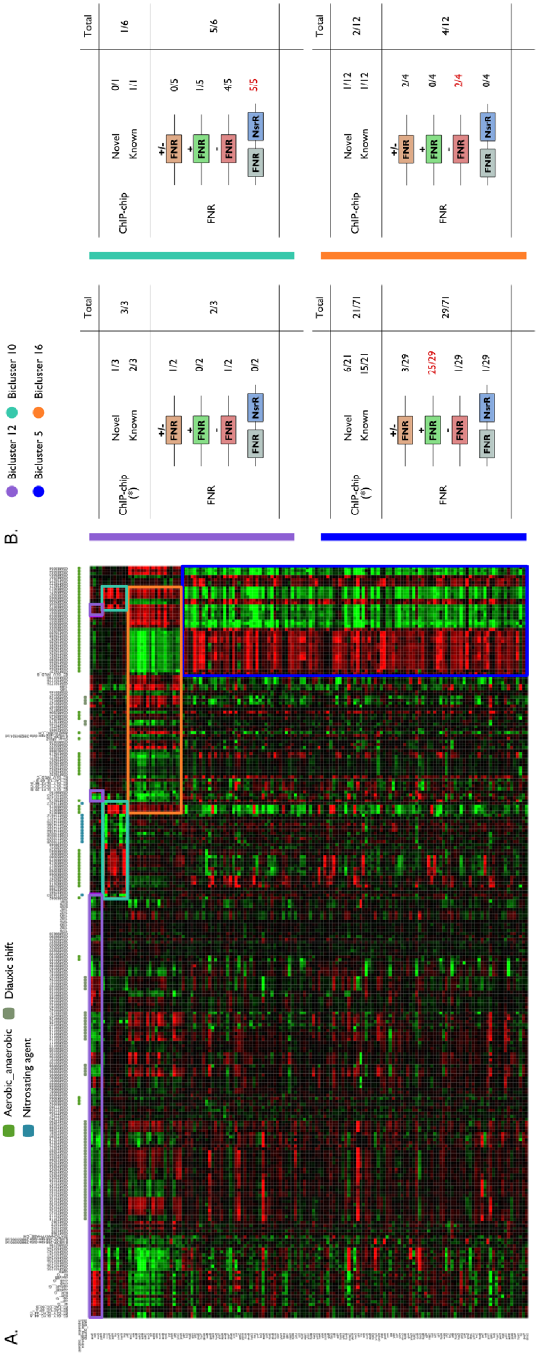


Figure 3- 5 Caption on previous page.

14 previously described FNR-targets that were missed by the ChIP-chip analysis (false negatives).

The 2 other biclusters in Figure 3- 5, consensus bicluster 10 and 16, are not enriched in the ChIP-chip targets, but show a high coverage of previously described FNR-targets (Table 3 - 1). In addition they are just like biclusters 5 and 12 enriched in oxygen related conditions (Figure 3- 5). Interestingly, the expression pattern of the genes within bicluster 16 is anti-correlated to that of the genes in the ChIP-chip enriched consensus biclusters (Figure 3- 5A). This anti-correlated behavior reflects the different mode of action of the dual regulator FNR, which acts as an activator on the targets of the ChIP-chip enriched consensus biclusters 5 and 12 and as a repressor on consensus bicluster 16 (Figure 3- 5B).

The distinct expression behavior of the genes in consensus bicluster 10 can be explained by joint regulation of the genes within this consensus bicluster by NsrR and FNR (Figure 3- 5B), which also explains the presence of nitrosating conditions within this consensus bicluster (Figure 3- 5A). Similarly to consensus bicluster 16 the genes within this consensus bicluster not retrieved by the ChIP-chip experiment are also known to be repressed by FNR. Seemingly the conditions used in the set up of Grainger *et al.* [85] were biased towards selecting positively regulated targets (bicluster 5 and 12), but missed most of the repressed targets (bicluster 10 and 16). Together these consensus biclusters 10 and 16 contained 3 ChIP-chip targets of which 1 novel one (*i.e.* not documented as an FNR-target in RegulonDB [94]) and 7 additional previously described FNR-targets not retrieved by the ChIP-chip experiment (Figure 3- 5B).

The remaining 33 ChIP-chip targets belong to biclusters not enriched with genes from the ChIP-chip experiment, nor having a high proportion of known FNR-targets. For these targets the results are less conclusive. Six of these ChIP-chip targets are known FNR-targets according to RegulonDB (Table 3 - 1). Considering that many targets of FNR perform global cellular functions, it is indeed possible that due to pleiotropic functions of these genes some FNR-targets end up in biclusters not having a high coverage for known FNR-targets. However, we can expect a large proportion of these 33 genes to correspond to false positives or non-functional targets. Not only because of the ChIP-chip procedure itself, but also because of the way the ChIP bound regions were mapped to the genes: the presence of a ChIP bound region located in the intergenic region between two divergently transcribed genes does not automatically imply that both genes are transcriptionally regulated by the bound TF [85].

3.5 Discussion

In this chapter we developed an ensemble method to be used in combination with query-based biclustering methods for the interrogation of expression compendia with a list of experimentally derived genes.

The method exploits the possibility some query-based biclustering methods offer to explore a whole range of thresholds that influence the bicluster size. Instead of having to choose the 'best bicluster with the most optimal coexpression level' based on some user-defined *ad hoc* criteria, our ensemble merges the results of the multiple runs in a single consensus cluster, whereby genes that were repeatedly retrieved at multiple biological resolutions will receive a higher weight to belong to the same consensus cluster. The ensemble method thus offers a statistically inspired way to merge the outcomes for different thresholds on coexpression.

The ensemble method is also devised to cope with the 'split and merge strategy' that is needed when using a query-list containing genes with different expression behavior as input. The ensemble procedure was used to merge the partially redundant biclustering-outcomes that were obtained by running query-based biclustering on each of the genes of the query list separately. The main goal of the consensus solution here is to remove redundant biclusters that were obtained by using query-genes that show a similar coexpression behavior. However, as in this case also the biclusters that share no overlap with any of the other biclusters are valuable (these were derived from genes in the query list that do not show any similarity in coexpression behavior with the rest of the list) the consensus solution not only needs to reduce redundancy, but at the same time should reflect to a maximal extent the distinct solutions that were present in the query-based biclustering outcomes of the individual query-genes. This application of an ensemble based strategy is inherently different from its traditional use where it is mainly meant to increase accuracy of clustering results by searching for genes that were found coexpressed in multiple runs [54;96].

The ensemble approach was validated using different evaluation metrics that assess both the agreement with the original biclustering-solutions as the quality of the consensus clusters independent of these query-based biclustering outcomes. We tested the influence of using different transformations of the consensus matrix in combination with different graph clustering methods on the quality of the consensus biclusters. While all tested combinations of matrix transformations and graph clustering methods resulted in consensus biclusters that recapitulate the original query-based biclustering solutions and reduce redundancy, using fuzzy clustering to extract consensus clusters from a pruned consensus matrix gave the overall best results.

To illustrate how query-based biclustering in combination with our ensemble approach can be used to interrogate a gene expression compendium with own experimental data, we applied it to an FNR ChIP-chip case study. By combining the ChIP-chip list with the public data we could obtain a view on its quality: not only could the analysis point out potential false positive ChIP-chip targets, but it also showed that most of the targets repressed by FNR were missing from the ChIP-chip list.

Chapter 4

Towards a functional map for *Salmonella* Typhimurium biofilm formation

4.1 Introduction

Salmonella enterica serovar Typhimurium (*S.* Typhimurium) is an important pathogen causing host-specific diseases ranging from self-limiting food-borne gastroenteritis to life-threatening systemic infections. *Salmonella* infections still constitute a serious public health burden and represent a significant cost to society in many countries. *Salmonella* is able to form microcolonies and mature biofilms on both biotic [99-102] and abiotic [103] surfaces. Biofilms, which are the predominant mode of bacterial cell growth in natural habitats [104], are structured communities of bacterial cells enclosed in a self-produced matrix, adhering to inert or living surfaces [105]. Cells within these biofilms are physiologically distinct from planktonic cells: single cells from the same organism that swim/float freely in liquid medium [104]. This biofilm-forming ability increases the pathogen's resistance to antibacterial treatments [106;107] and enhances its spread and persistence in non-host environments [108]. Because *Salmonella* infections generally occur after the ingestion of contaminated food or water, environmental *Salmonella* biofilms (in for example stables, slaughterhouses and on kitchen surfaces) are, next to the traditional routes of infection (contaminated meat, eggs and poultry), indeed a source of reappearing occurrence by this pathogen [109]. Further on, *Salmonella* biofilm formation is a strategy to induce chronic infections [110;111] and even a possible way to colonize host organisms [102;112].

As compared to their planktonic counterparts and despite the vast amount of biofilm related research during the last decades, bacterial biofilms have remained poorly understood. Inherent complexities associated with biofilm studies originating from spatial and temporal biofilm heterogeneity [113] and uncharacterized growth parameters [114] are contributing to this lack of knowledge. In the recent years, within model organisms there has been a change of focus from a simple hunt for genes involved in biofilm formation towards a more global analysis of biofilm-related genes through DNA microarray analysis (*e.g.* [115-120]). However, the first transcriptional profiling study of *Salmonella* biofilms was only recently performed [105]. These and related studies generally result in a list of genes experimentally determined to be involved in the biofilm formation process, but provide however little information on the specific functional roles of these genes within biofilms. In addition, recently there has been some debate on the existence of gene sets that are specifically involved within biofilms: when using microbial genetics to identify the causal genes for biofilm formation it is often not clear whether altered behavior of these genes within biofilms reflects a biofilm-specific pathway or whether this behavior is a consequence of changes in cell metabolism and altered growth dynamics resulting within this new, multicellular, environment [121;122].

Within this chapter we aim to investigate whether biofilm formation constitutes a specific, *i.e.* involving genes with functions that are limited to biofilm formation, rather than a global adaptive response to changing environmental conditions. To this end we first compile a core list of genes experimentally determined to be involved in *S.* Typhimurium biofilm formation. This list consists of genetic hits identified in a screening of a recently constructed targeted *Salmonella* deletion mutant library [123] to identify mutants impaired in biofilm formation. We combined this data together with data generated using a single cell approach to study *Salmonella* biofilm formation [124]. As such we obtain an extended list of genes involved in *Salmonella* biofilm formation.

Secondly, we leverage publicly available gene expression data for both planktonic and multicellular conditions to deduce whether the genes within this core list all belong to the same biofilm-specific pathway (*i.e.* are all coexpressed under conditions that assess multicellular behavior).

The work within this chapter was done in collaboration with the Salmonella and Probiotics group of CMPG (KULeuven; Ir. K. Hermans, Dr. Ir. S. De Keersmaecker and Prof. Dr. Ir. J. Vanderleyden), where *Salmonella* biofilm experimental assays were performed. Therefore, details of the experimental protocols are out of scope of this PhD thesis and we focus within this chapter mostly on the bioinformatics analysis and the biological interpretation of the results.

4.2 Results

4.2.1 Overview of the approach

In this work we aim to examine the specificity of the transcriptional response in biofilm formation as compared to planktonic conditions. An overview of the approach is given in Figure 4- 1.

We start by compiling a core list of 70 genes experimentally identified to be involved in *Salmonella* biofilm formation. Specifically, we derive the gene list from two complementary approaches in order to not bias the list towards a certain experimental approach (see 4.2.2).

In a second step we interrogate two different *Salmonella* Typhimurium gene expression compendia for genes that are coexpressed with the genes within this core list. The first compendium assesses gene expression in multicellular conditions ('multicellular compendium') (195 conditions), whereas the second one contains conditions that measure transcriptional responses under planktonic behavior ('planktonic compendium') (522 conditions). We used the ensemble approach that was described in Chapter 3, which combines query-based biclustering with an ensemble post-processing strategy, to identify biclusters of genes that are centered on the expression profile of the genes within this core list for both gene expression compendia. As such we obtained a set of ‘multicellular biclusters’ by interrogating the multicellular compendium for the core list and a set of ‘planktonic biclusters’ obtained by interrogating the planktonic compendium for the core list. As condition-dependent coexpression is usually indicative of functional relatedness this approach allows attributing functional categories to the genes within the core list by leveraging functional annotations of the genes with which they are coexpressed (guilt-by-association principle). To this end we calculated enrichment of each of these biclusters for GO-categories. In addition, joint coexpression of the genes within the core list reveals mutual functional relationships amongst these genes.

Lastly, to investigate the specificity of the transcriptional response of the core list to multicellular conditions we compared the gene content of the mutlicellular biclusters with that of the planktonic biclusters.

Figure 4- 1 Overview of the approach. First, as input data (left panel) we create a core list of experimentally derived genes assayed for their relevance to biofilm formation. Further we also create two separate gene expression compendia, one containing conditions that assess multicellular behavior and one with conditions that assess planktonic behavior. Using query-based biclustering in combination with the ensemble approach of Chapter 3, for each compendium a set of biclusters is obtained, centered on the genes from the core list (middle panel). In a last step the biclusters obtained for both compendia are compared (1.) and functional enrichment of each of the biclusters is calculated (2.) (last panel).

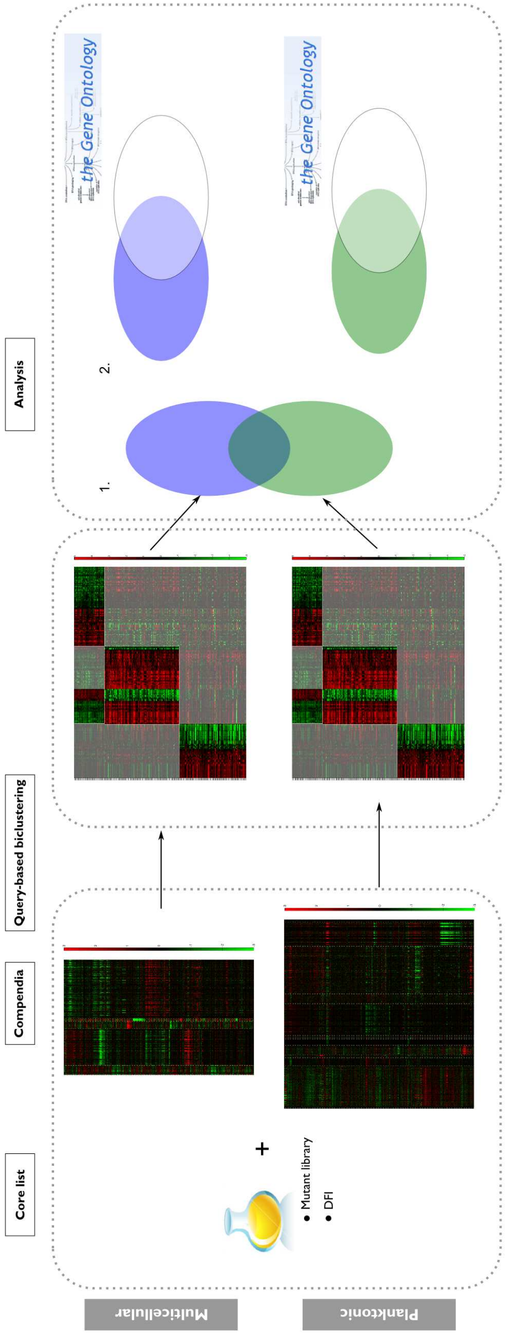


Figure 4- 1 Caption on previous page.

4.2.2 Composing a core list of *Salmonella* Typhimurium specific biofilm genes

A core list of biofilm genes was derived from two complimentary experimental approaches. As such we can compile a set of genes involed in *Salmonella* biofilm formation, not confined to a single experimental method or single cellular process within biofilm formation, yet tuned towards focusing on genes involved in biofilm formation as compared to planktonic conditions.

In the first approach a DFI single-cell enrichment method was used to study *Salmonella* biofilm formation [124]. In this method, within single *S.* Typhimurium clones transcriptional activity of genomic DNA is measured by fusing the segments with a promoterless *gfp*-gene (*i.e.* the reporter gene). In case of transcriptional activity of the fused segment fluorescent signals can be detected because of GFP (Green Fluorescent Protein) production. These fluorescence signals can be monitored by a fluorescence-activated cell sorter (FACS). In this experiment single clones were specifically filtered based on differential fluorescence during biofilm growth as compared to planktonic growth. Subsequent sequence determination of the genomic segments led to the identification of the DNA sequences that caused the increased expression of the promoterless *gfp*-gene. As such 27 genetic loci were identified showing biofilm specific increased expression. Of these, 17 could be narrowed down to promoter regions of already annotated genes of which 5 encoded proteins without known functions. The remainder 10 genetic loci coincided with putative unknown regulatory elements of known genes or in intergenic regions. The advantage of this experimental approach is that it monitors gene expression at the level of single cells in stead of at the population level (as is the case in microarrays). As such genes can be identified that play a crucial role in biofilm-related processes in only a subpopulation of cells. In addition, genes are specifically filtered for upregulation in biofilms as compared to planktonic conditions in consecutive selection rounds, each time filtering the most fluorescent cells in biofilm conditions and the least fluorescent cells in planktonic conditions. Therefore, it is expected that the identified genes play a role in biofilm-related processes. Lastly, this assay is not biased towards identifying particular gene sets since the *gfp*-reporter strains are constructed in a random fashion. This is in contrast to for instance mutant libraries (see below), where only genes non-essential for growth can be monitored. In this chapter, we used the list of the 17 annotated promoter regions as a first input to identify clusters of coexpressed genes.

A second input gene set was derived from the screening of a mutant library [123] for altered biofilm behavior (Kim Hermans (CMPG - KULeuven), personal communications). In particular, this mutant library contained around 1000 targeted *S.* Typhimurium deletion mutants primarily hit in genes without ortholog in *Escherichia coli* [125], including almost all 200 genes associated with *Salmonella* virulence such as the type III secretion systems and their effectors. Further on, this mutant library also contained mutants in nearly all genes involved in fimbrial and surface antigen regulons as well as a small subset of genes shared between *Salmonella* and *E. coli* including sRNAs (confirmed and candidates) and genes involved in motility, regulation and pathogenesis. The construction of such a mutant library with mainly *Salmonella* specific genes was motivated by recent evidence that suggests a link between horizontally acquired genes and biofilm formation (*e.g.* [126-128]). Moreover, subtle differences between closely related bacteria considering extracellular matrix production [129] further emphasizes the importance of different gene sets and/or expression patterns between related bacteria considering biofilm formation. Screening of this mutant library resulted in 55 genes for which the corresponding mutants showed significant induction or reduction of the biofilm phenotype.

Experiments within yeast have already shown that transcriptional and genetic screening approaches complement each other when investigating the specific pathways involved in certain phenotypes of interest [130-132]. Here also, we find that both lists only have 2 genes in common and that therefore they are complementary in the gene sets they find.

Together these lists contain 70 genes (Table A- 2) that can be subdivided in different groups according to their importance in biofilm formation. Firstly, genes that clearly have been attributed to *Salmonella* biofilm formation before, such as *csgD* and *csgB* [133-135]. Secondly, genes that have been been documented for mechanisms known to be important in biofilm formation, such as polyamine (*potF*) [136] and iron metabolism associated genes (*sitA*, *iroN*, *fhuA* and *fes*) [137], but were for the first time reported to be important for the specific case of *S.* Typhimurium biofilm formation, either in the DFI screening or in the mutant library screening. A third category concerns specific *Salmonella* genes with unannotated functions.

4.2.3 Core list does not correspond to a single pathway

First, we wanted to verify whether the 70 genes within the core list all belong to a single biofilm specific pathway. To this end, we used the method introduced in Chapter 3 to investigate whether the genes within the core list are all mutually coexpressed (*i.e.* belong to the same bicluster) under multicellular conditions. This method outputs one or multiple biclusters (sets of genes together with the conditions in which these genes are differentially expressed) in the gene expression data set that each contain at least one of the genes from the core gene list: *i.e.* the bicluster solutions are centered on the expression profiles of the genes in this core list. Using this approach we queried the multicellular compendium (see 4.2.1) for our core gene list consisting of 70 genes. As such, 9 multicellular biclusters (Figure 4- 2), containing in total 38 out of the 70 query-genes (Table A- 2), were obtained.

Figure 4- 2 Multicellular biclusters. A. Consensus matrix (see Chapter 3) for the Query-Driven Biclustering results obtained for the genes in the core list. Separate biclusters are color coded in the bar on the left. B. Heatmap of the 9 multicellular biclusters. Biclusters are indicated in colored rectangles (the same color code as in A is followed). Genes are represented in the rows and conditions in the column. For representational reasons only biclusters which were significantly functionally enriched for a certain GO-category are displayed.

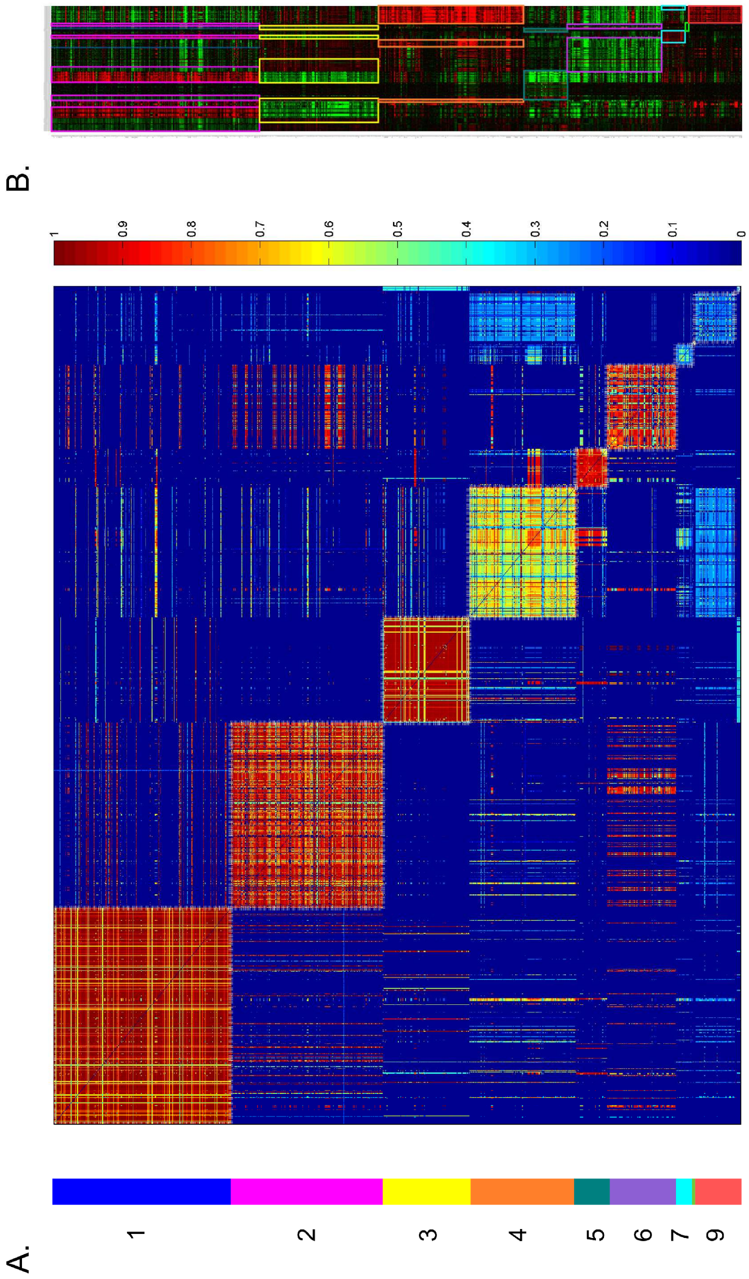


Figure 4- 2 Caption on previous page.

The fact that these 38 genes from the core list separate into different biclusters according to their expression in multicellular conditions, suggests that they belong to different pathways. It is however, possible that false positives of the experimentally assays described in 4.2.2 give also rise to multiple biclusters. Indeed, false positives of the experimental assay that are not directly linked to the biofilm process will likely constitute a different functional role within the cell and therefore will be coexpressed with other genes than the true positives of the experimental assays. We observe however, that each of the biclusters contain at least 2 genes from the experimental assay (Table 4- 1). This suggests that the genes from the core list are at least to some extent mutually coexpressed under multicellular conditions. This makes it very unlikely that these genes are false positives of the assay as it would imply that the false positives are biased towards a particular functional pathway. Therefore, our results suggest that the genes assayed to be involved in biofilms separate into different functional groupings according to their condition-dependent expression patterns within a compendium that assesses *Salmonella* multicellular behavior.

Table 4- 1 Content of the multicellular biclusters in the number of genes.

Bicluster	Number of genes (total)	Number of genes (core list)
1	319	6
2	273	8
3	155	2
4	192	2
5	56	2
6	124	6
7	30	7
8	5	2
9	70	3

4.2.4 Functionality query-genes is not limited to multicelullar behavior

As genes that are coexpressed are assumed to be functionally related we took advantage of known Gene Ontology (GO) annotations to functionally annotate the biclusters these 38 genes belonged to. Bicluster function was determined by retaining for each biclusters those GO functional classes that are statistically overrepresented within the biclusters (p-value < 0.01) (see 4.4.4 for details). This analysis reveals that the genes within the obtained biclusters are mainly involved in general cellular processes such as oxidation-reduction (P= 6.82E-08), translation (P = 1.00E-10), chemotaxis (P = 1.00E-10), transport (P= 8.61E-05) and pathogenesis (P = 1.00E-10) (Figure 4- 3A). Interestingly, bicluster 7 is significantly enriched for genes involved in lipopolysaccharide (LPS) (P = 1.12E-09) synthesis, which has previously been associated with biofilm formation [129;138].

Taken together, this suggests that biofilm formation triggers a large variety of cellular responses that are not all specific to biofilm formation. To further assess the specificity of the triggered pathways to biofilm formation, we compared the set of biclusters to one that was obtained on the planktonic compendium (assesses gene expression in free-living conditions).

Starting from the same core list of 70 biofilm specific genes, we could retrieve 10 biclusters (Figure 4- 4) containing in total 44 out of the 70 query-genes (Table A- 2). With the exception of bicluster 9 and 6, each of these 10 biclusters contained at least 2 genes from the core list. The fact that for the majority of these 70 genes within the core list biclusters could be obtained for planktonic conditions, suggests that the function of these genes is indeed not limited to biofilm formation. In addition, functional enrichment analysis of these biclusters revealed that these biclusters are enriched for similar GO terms as was the case for the multicellular biclusters: *i.e.* translation (P = 1.00E-10), chemotaxis (P = 1.00E-10), pathogenesis (P = 1.00E-10), oxidation-reduction (P = 4.58E-10) and transport (P = 2.46E-04) (Figure 4- 3B).

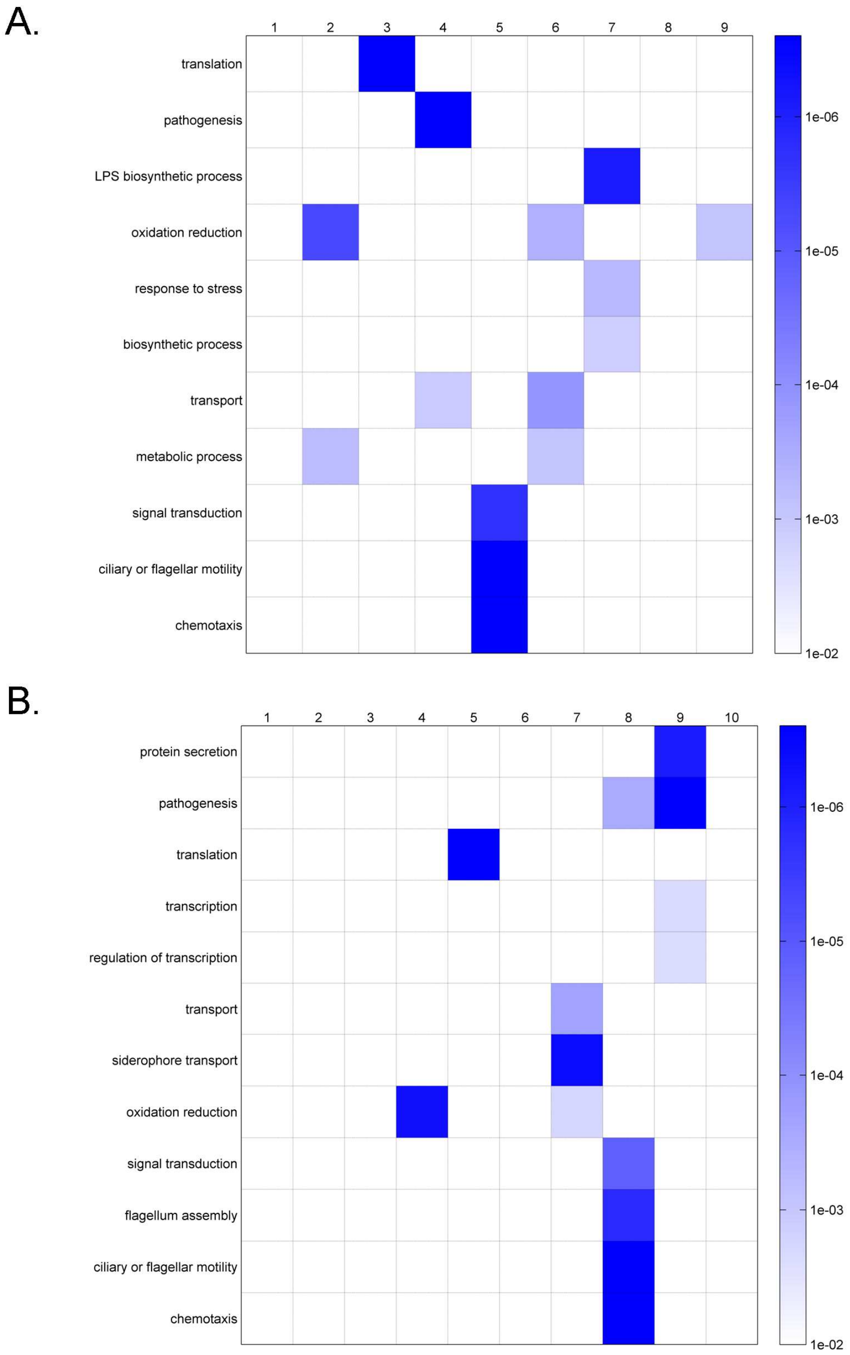


Figure 4- 3 Functional enrichment analysis of multicellular (A) and planktonic biclusters (B).

To further assess the similarity in the transcriptional response for genes in the core list for multicellular and planktonic conditions we evaluated overlap of the biclusters obtained for both data sets in terms of their gene members. Figure 4- 4 illustrates that except for bicluster 7 and 8 all biclusters derived for the multicellular compendium show significant overlap in their genes (as assessed by hypergeometric test) with at least one of the biclusters derived for the planktonic compendium. In addition, they are also enriched in similar GO-categories. This suggests that the transcriptional response for each of the genes within the core list is not specific to multicellular behavior and supports the hypothesis that the majority of the genes within the core list are involved in general cellular functions in stead of pathways specific to biofilm formation. Interestingly, bicluster 7 for multicellular conditions could not be associated with any bicluster corresponding to planktonic behavior. As mentioned above this bicluster was enriched for genes involved in LPS synthesis and coordinated expression of these genes seems to be limited to conditions that assess multicellular behavior. This suggests an important role for these genes within *Salmonella* biofilm formation.

Table 4- 2 Comparison of the multicellular and planktonic biclusters in terms of the genes they contain. Overlap significance levels (middle column) were calculated by hypergeometric test (see 4.4.4). Insignificant functional enrichments (p > 0.01) are left blank.

Multicellular bicluster	Planktonic bicluster	Overlap p-value	Function multicellular	Function planktonic
1	1	6.34E-04		
2	4	6.40E-04	Oxidation reduction	Oxidation reduction
3	5	0	Translation	Translation
4	9	0	Pathogenesis	Pathogenesis
5	8	0	Chemotaxis	Chemotaxis
6	7	4.53E-14	Transport	Siderophore transport
7		> 0.01	LPS biosynthesis	
8		> 0.01		
9	4	5.92E-13	Oxidation reduction	Oxidation reduction

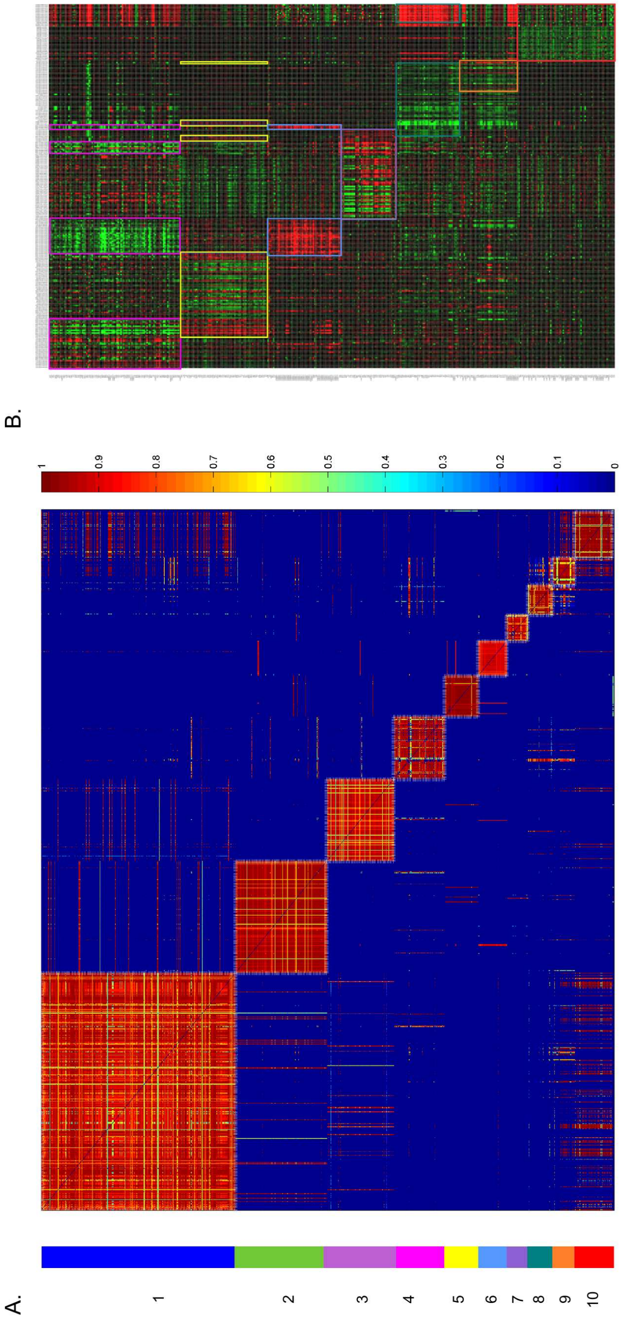


Figure 4- 4 Caption on next page.

Figure 4- 4 Planktonic biclusters. A. Consensus matrix (see Chapter 3) for the Query-Driven Biclustering results obtained for the genes in the core list. Separate biclusters are color coded in the bar on the left. B. Heatmap of the 10 planktonic biclusters. Biclusters are indicated in colored rectangles (the same color code as in A is followed). Genes are represented in the rows and conditions in the column. For representational reasons only biclusters which were significantly functionally enriched for a certain GO-category are displayed.

For multicellular bicluster 8 also no planktonic counterpart could be identified. However, as this bicluster only contains 5 genes coexpressed in only a limited subset of conditions we consider this bicluster as trivial.

4.3 Discussion

4.3.1 Functionality majority genes core list is not limited to biofilms

The production of biofilms by bacteria remains a subject of major research interest as the underlying regulatory mechanisms are still unclear. In particular there is still much debate on the specific genes and pathways that are involved in biofilm formation as even global analyses of biofilm gene expression through microarray analysis do not agree on the lists of genes differentially expressed [139].

Here we wanted to investigate whether biofilm formation constitutes a specific cellular response, involving a set of genes belonging to a biofilm specific pathway. In particular we first compiled a core list of 70 *S.* Typhimurium genes that were experimentally determined to be involved in biofilm processes. In addition, in stead of focusing on single gene expression studies, we took advantage of the large body of publicly available gene expression data to compile two different gene expression compendia: the first one contains conditions that assess multicellular behavior, whereas the second one contains conditions that assess planktonic behavior. We interrogated these gene expression compendia for biclusters centered on the expression profile of the genes from the core list to obtain biological meaningful groups containing these genes.

We find that the genes within the core list partition into different biclusters and therefore do not seem to function into a single biofilm specific pathway. This is further corroborated by the observation that the multicellular biclusters largely overlap in their genes with planktonic biclusters. These observations suggest that the genes within the core list participate in transcriptional responses that are not limited to biofilm formation. We hereby want to stress that the genes within the core list do indeed seem to have a key role within biofilm processes as they were either shown to be specifically upregulated in biofilm conditions as compared to planktonic conditions (DFI-experiment) or their respective mutants seem to show impaired biofilm formation (mutant). However, our analysis reveals that their functionality is not limited to biofilm processes only, nor that they constitute a single biofilm-specific pathway.

4.3.2 The role of the ECM in the observed response

Interestingly, for one of the multicellular biclusters no planktonic counterpart could be retrieved. This cluster was enriched for genes functioning in LPS biosynthesis which is known to be an important constituent of the Extracellular Matrix in which *Salmonella* Typhimurium cells reside in biofilms [138]. Furthermore, LPS is not only an important constituent of the matrix, its proper expression is also important for maintenance of the balance between different extracellular matrix factors (such as cellulose and curli fimbriae) since LPS mutants showed a totally altered extracellular matrix constitution [140]. This Extracellular Matrix is essential for biofilm formation [129] and apparently coordinated transcription of genes involved in this matrix formation only takes place under multicellular conditions.

These observations that the transcriptional response of this core list of 70 genes is rather ubiquitous with the exception for genes involved in construction of the Extracellular Matrix is consistent with recent work of White *et al.* [122] that studied metabolomic changes in an *S.* Typhimurium *csgD*-mutant incapable of constructing an Extracellular Matrix. By comparing the metabolome of wild type cells with this mutant, it was revealed that these mutants mainly experienced a shift in central metabolism and at the level of the osmotic stress response. In line with our analysis here also differences in gene expression for planktonic and biofilm-forming cells were attributed to quantitative differences in expression for similar gene sets in the planktonic and biofilm cells, rather than to specific gene sets. Based on these findings the authors hypothesized that the major changes in metabolism and stress response were mainly a consequence of the microenvironment created by the ECM. Our findings confirm such a hypothesis, first we identify that genes involved in biofilm formation according to experimental set-up are part of general cellular pathways that are differentially expressed in both multicellular and planktomic conditions. Secondly, genes involved in synthesis and, maybe more importantly, homeostasis of the ECM are only coexpressed in multicellular conditions. This suggests that studies aimed at unravelling biofilm regulatory mechanisms are often confounded by the unique environmental niches that cells within ECMs reside in [113], which complicates identifying the deterministic biofilm-specific regulatory pathways that lead to biofilm formation [121].

4.3.3 Do biofilm-specific pathways exist?

We do however not completely rule out the possibility that specific biofilm-related pathways exists. First, within our analysis we only focus on the transcriptional level: *i.e.* we searched for biofilm-specific pathways based on a coexpression analysis of a set of genes experimentally determined to function within biofilms. Hereby, likely regulatory mechanisms on the post-transcriptional or metabolic level are ignored. Second, our analysis did not reveal biclusters for all of the genes within the core list. For instance, for *csgD,* a gene that has been previously associated with *S.* Typhimurium biofilm formation, no bicluster could be obtained. This could be explained by the fact that this gene shows stochastic expression behavior (*i.e.* is only expressed in a subpopulation of cells at a certain time point) [113] and that therefore the expression level of this gene on the population level as measured in microarrays does not exceed a certain threshold level. In addition, previous work has shown the importance of phosphorylation for mediating biological activity of this gene [141]. For the remainder of the genes within the core list similar explanations might exist as to why we do not retrieve biclusters for these genes. Furthermore, is the gene expression compendium used not exhaustive in the possible biofilm-related and planktonic conditions that cells within their natural environment might encounter. Therefore, it is possible that for gene sets that function under very specific conditions no corresponding biclusters could be retrieved.

Although, we can not exclude the possibility that some of the genes in the core list are involved in biofilm-specific pathways we were able to narrow down the likely candidates for genes that specifically function within biofilms. In addition, our results revealed that although the experimental assays were tuned towards selecting genes that are relatively specific to biofilm formation, either by selecting genes whose expression is particularly induced in biofilm conditions as compared to planktonic conditions (DFI experiment) or by focusing on genes specific to *Salmonella* Typhimurium (mutant library), the function of most of these genes seems not to be limited to biofilm formation. Recent research has shown that this observation is not limited to our work [121;122;126] and therefore care needs to be taken in interpreting the outcome of such assays with respect to the specificity of the derived gene sets to biofilm formation.

4.4 Materials and methods

4.4.1 Composing a biofilm specific gene list

A core list of biofilm specific genes was derived from two complementary assays: a list of 17 genes was derived from a recent DFI-experiment for biofilm formation [124], whereas a second list of 55 genes was derived from screening a *S.* Typhimurium mutant library [123] for biofilm formation (Kim Hermans (CMPG – KULeuven, personal communications) for more details). As such a core list of 70 genes (there is some limited overlap between the two lists) was obtained.

4.4.2 Constructing gene expression compendia

Cross-platform *S.* Typhimurium compendia were compiled from data stored in public repositories [16-18]. Here we took advantage of experiment descriptions to built separate compendia that assess multicellular and planktonic behavior. Specifically experiments annotated with 'swarming' were considered as assessing multicellular behavior (no other experiments assessing multicelullar behavior could be found), whereas the remainder of the experiments assessed planktonic behavior. As such a multicellular compendium was obtained that contains the expression values for 4525 genes profiled under 195 conditions and a planktonic compendium compendium that profiled gene expression under 522 conditions. The data was normalized appropriately in order to allow for cross-experiment and cross-platform comparison (normalization procedures were similar as in [8]).

4.4.3 Query-based biclustering

Biclusters were obtained by running a query-based biclustering algorithm (QDB) [73] on this compendium, using the genes within the core list (4.4.1) as input (*i.e.* as query). QDB was applied to each of these genes separately, using a resolution sweep approach to evaluate in a single run of the algorithm all possible solutions that correspond to different degrees of coexpression with the query-gene. We used the ensemble approach introduced in Chapter 3 [65] to remove redundancy amongst the outcomes obtained by QDB. Heatmaps with biclusters were plotted using ViTRaM [95].

4.4.4 Enrichment analysis

GO functional categories for *S.* Tyhimurium were obtained from the Uniprot GOA proteome sets through EBI (http://www.ebi.ac.uk/GOA/proteomes.html). Functional enrichment p-values were calculated by hypergeometric test. P-values for the overlap in the number of genes for planktonic and multicellular biclusters were calculated by hypergeometric test [142]. Let *N* be the total number of genes in the gene expression data set, *M* the number of genes annotated with a certain GO functional category, *K* the number of genes in a certain bicluster. Given that *x* out of the *K* genes in the bicluster are annotated with a certain GO functional category the p-value of this occurring by chance can be calculated using the hypergeometric distribution:

$$p = 1 - \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}} \qquad (4.1)$$

Chapter 5

An ensemble strategy for module networks learning

5.1 Introduction

In the previous chapters we discussed and applied methods for module detection: algorithms that search for sets of genes that are coexpressed. As transcriptional regulation underlies gene coexpression it is interesting to also identify the regulatory proteins that regulate the expression of these genes: *i.e.* to describe gene coexpression causally. We refer to these methods that do not only infer coexpressed gene sets but also predict causal regulators as methods that infer the transcriptional regulatory network (TRN).

In this chapter we introduce such a network inference method, LeMoNe (LEarning MOdule NEtworks). This method is inspired by the pioneering module networks method of [30]. A module network is a probabilistic graphical model [143] which consists of modules of coexpressed genes and their transcriptional programs. The transcriptional program consists of a set of regulators that is assigned to the module and that best predicts the condition-dependent mean expression of the genes in a module. As the module networks method only requires gene expression data as input, which is well abundant for many organisms and easy to obtain, and requires little prior knowledge on the underlying network, this method has already been widely used [30;144-148]. Segal *et al.* [30] used a deterministic optimization algorithm that searches simultaneously for a partition of genes into modules and a regulation program for each module. This results in an output that consists of a list of regulator-module assignments which are all equally likely. LeMoNe [32] extends this module networks framework by using a stochastic optimization scheme in combination with ensemble strategies to obtain a more refined module networks.

In this chapter we used LeMoNe to characterize the *E. coli* transcriptional regulatory network. As a model organism, the TRN of *E. coli* has been under extensive study. Databases such as RegulonDB [94] and EcoCyc [93] give a comprehensive overview of our current knowledge on the *E. coli* TRN. It is, however, far from complete as for only about one third of the genes in *E. coli* some interaction information is present in RegulonDB and it is estimated that only half of the transcription factors (TF) have been characterized [42]. Extension of this network by computational methods therefore remains a problem of outstanding interest. Methods which infer transcriptional regulatory networks from gene expression data alone are particularly interesting, since their predictions are not restricted to characterized regulators, as is often the case with data-integration methods [8;31;35]. In addition, due to the abundant information on the *E. coli* regulatory network this network is considered as a reference network for reverse-engineering algorithms [34] and therefore can assist us in obtaining insight in the true power of the method in reflecting known biology.

The work presented in this chapter was performed in collaboration with the Plant Systems Biology (PSB) department of the VIB (Univeristy Ghent; Dr. T. Michoel, Dr. A. Joshi and Prof. Y. Van de Peer). As the algorithmic framework was mainly developed at PSB, a detailed algorithmic explanation is out of scope of this thesis and we restrict the methodological explanation to a concise algorithmic overview. Instead, within this chapter we focus on the validation of the ensemble approach and the comparison to related network inference methods. First, we illustrate the advantages that come with using an ensemble learning strategy for network inference. Secondly, we characterize the type of interactions that can be inferred using LeMoNe and compare it to the transcriptional regulatory network that was obtained using another popular network inference method, CLR [34]. We demonstrate that

LeMoNe is highly qualified to recapitulate existing knowledge in *E. coli* and that it is able to make novel high-confidence predictions in the form of experimentally verifiable hypotheses.

5.2 LeMone ('Learning module networks')

In contrast to the module networks method of Segal *et al.* [30] which searches for the modules and their transcriptional programs simultaneously, LeMoNe considers both as separate tasks. This algorithm improves the original module network learning algorithm of [30] at both the level of learning modules as well as assigning regulators (a detailed comparison with the results of [30] is given in [32]).

In Figure 5- 1 an overview of the algorithm is given. The first step of the algorithm consists of generating an ensemble of non-overlapping gene clusters (or modules) with condition partitions (two-way clustering). Algorithmic details for this step are presented in [55]. The generated modules consist of genes that are coexpressed across all conditions in a gene expression data set. When searching for modules, often many local optima exist with partially overlapping modules differing from each other in a few genes. Therefore LeMoNe exploits this fuzzy property of a module to increase the reliability of the predicted interactions: instead of reporting only one cluster solution (local optimum), a stochastic approach is used to generate many equiprobable, but partially redundant cluster solutions from which an ensemble averaged solution is generated. This solution consists of so-called 'tight clusters', subsets of genes which cluster together in almost all local optima. To generate the tight clusters a Gibbs sampling method for two-way clustering was developed [55]. This method iterates between the gene and condition direction to identify clusters of genes that show condition-specific coexpression patterns. Briefly, genes of the same cluster belong to the same mixture of Gaussian distributions (Figure 5- 1 B), with each mixture component corresponding to a condition partition. The mean and variance of each Gaussian component are defined as the average and the variance of the gene expression of the cluster genes for that condition partition. Hence, each condition partition qualitatively describes gene expression for the conditions within that partitition (upregulated, basal expression level down-regulated). The algorithm starts by randomly partitioning genes into clusters and within each gene cluster partitioning the conditions in condition clusters. Using Gibbs sampling gene cluster assignments and the partitioning of conditions within a gene cluster is iteratively updated until a stationary state is reached [55]. The number of clusters is decided upon automatically. As one such run of the algorithm results in a local optimum, multiple runs are performed to cover the whole space of solutions and in a subsequent step this ensemble of solutions is averaged into tight clusters.

Figure 5- 1 LeMoNe working mechanism. A. LeMoNe takes as input a gene expression compendium and a set of candidate regulators. First, modules are constructed using a model-based two-way clustering approach. Gibbs sampling is used to derive multiple equiprobable but different clustering solutions from the gene expression data set. These are combined into tight clusters of genes which co-cluster repeatedly across the different gibbs sampling runs (the centroid solution). Next, for each tight cluster a transcriptional program is predicted based on the clusters' condition partitions (see B.). Also here for each module an ensemble of possible transcriptional programs is generated in a probabilistic fashion. Hence regulator assignment scores can be calculated reflecting the frequency with which a certain regulator was assigned to a certain module. Regulator-to-module assignments can be ranked according to this score. B. Illustrates how transcriptional programs are predicted for each module. Two-way clustering is used to generate modules of coexpressed genes. Such a method does not only cluster the genes but also the conditions within a module: *i.e.* each module is represented by a mixture of Gaussians with each mixture component representing a condition partition. Consequently, a single cluster in the gene direction is associated with multiple clusterings in the condition direction. Transcriptional programs are predicted based on how well the regulator's expression profile explains the modules Gaussian mixture profile.

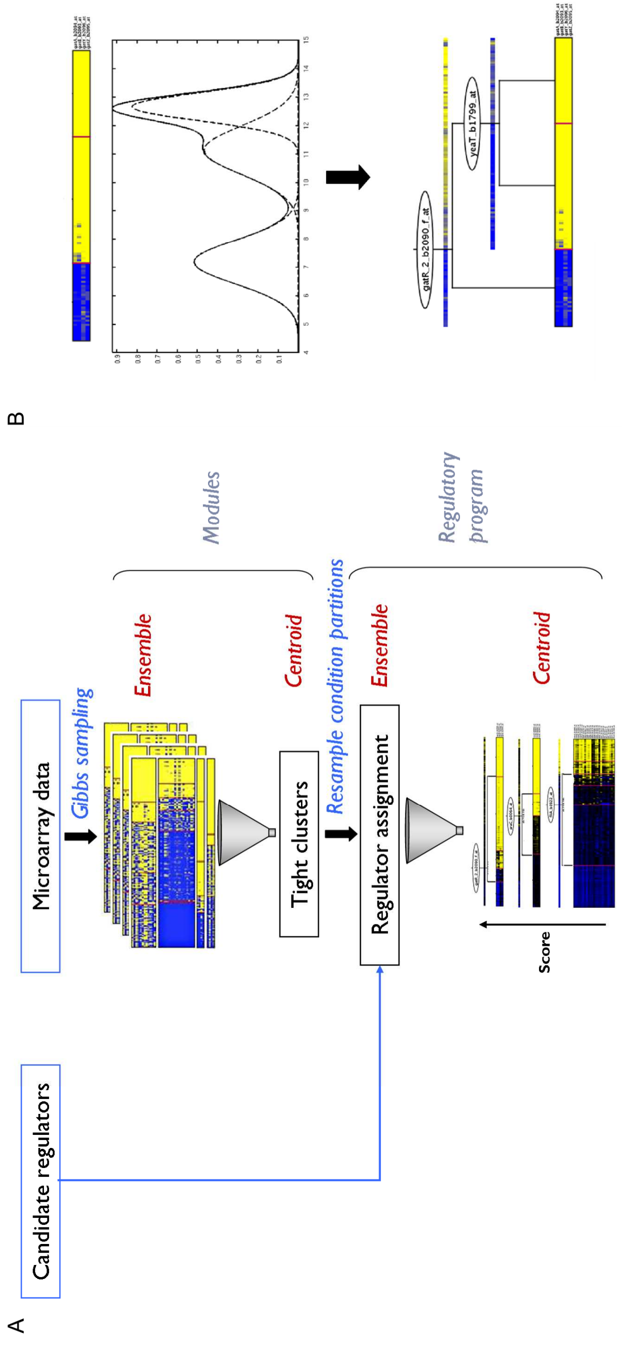


Figure 5- 1 Caption on previous page.

In the second step the condition partitions within the tight clusters are exploited to assign transcriptional programs to each cluster. Here, the algorithmic details are described in [32]. The method takes as input a list of known and predicted regulators and constructs a transcriptional program based on how well the expression pattern of the regulators explains the condition partitions in the tight clusters, as defined by the mixture of Gaussian distributions corresponding to each module. Similar as in [30] the transcriptional program has a tree structure with top-regulators explaining all condition partitions within a module and regulators at the lower tiers only explaining a subset of the condition partitions. As such the method can predict combinatorial regulation: if two regulators are assigned to the same condition partitions they are predicted to regulate the module genes combinatorially for those conditions. As for each tight cluster multiple equiprobable condition partitions can be generated (each corresponding to a local optimum), for each module an ensemble of transcriptional programs can be constructed that each correspond to a different partitioning of the conditions for the same module. This ensemble of transcriptional programs is merged into a statistical score, the 'regulator assignment score'. This score accounts for how often a regulator is assigned to a module, with what strength and at which level in the regulation tree. This score can be used to prioritize regulator-to-module predictions.

5.3 Application to public *E. coli* gene expression compendium

We applied LeMoNe to a compendium of *E. coli* Affymetrix gene expression profiles [34]. Transcriptional programs were learned from a list of 316 known or putative transcription factors in *E. coli* [42;93].

We first illustrate the advantages that come with using an ensemble method. Next we interpret the resulting module network in terms of the hierarchical topological structure of the *E. coli* TRN inferred from the static interactions present in RegulonDB [149-151]. At the top, global regulators sense and react to major environmental signals, which are further fine-tuned by local, more specific regulators in the lower tiers, and processed by modules of functionally related genes at the bottom of the hierarchy[150]. We show that methods for inferring TRNs from gene expression data are mainly useful to characterize the lower layers of the hierarchical network, with a preference for autoregulators and neighbor regulators (colocalized on the chromosome with their targets). These neighbor regulators are often acquired through horizontal gene transfer [152].

5.3.1 Illustrating the power of the ensemble strategy

Here we highlight how using an ensemble approach can improve upon regulatory network inference. First, we focus on the module detection part, as module quality is determinative to learning module networks since the learning of the transcriptional program depends on it. We observe that potential false positive targets can be filtered out by using an ensemble approach. Indeed, an example is given by the GadE-regulon (involved in pH homeostasis) (Figure 5- 2). Most of the target genes of this regulator consistently cluster together over the different gibbs sampling runs. The many local optima, generated by the Gibbs sampling approach resulted in 62 additional genes, which were clustered at least once with one of the GadE-regulon genes. With the exception of *yhiV*, *gadW* and *gadX*, none of the remaining 62 genes are known GadE-targets and therefore this list consists primarily of false positives. Co-clustering of these 62 genes with the 10 module genes was not sufficiently significant to be retained in the final module. Similar observations were made for other regulons, such as the AraC- and GalS-regulons. This illustrates how the ensemble approach guarantees an effective filtering of false positives while retaining the true positives.

With respect to the inference of transcriptional programs for each module, regulator-to-module edges can be ranked according to their score which reflects the statistical significance of the assignment of a regulator to a module. To assess whether this ranking is biologically

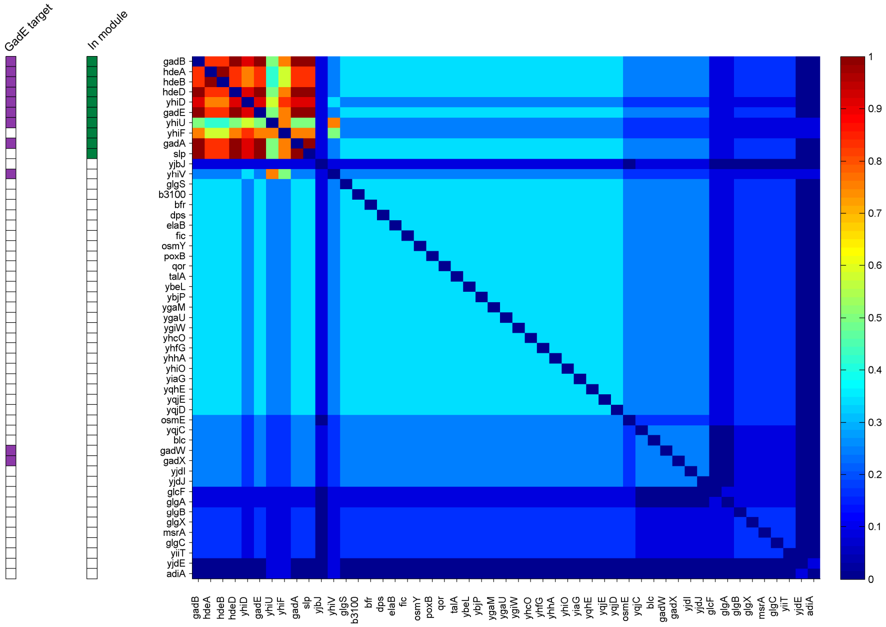


Figure 5- 2 This figure illustrates how the ensemble solution contributes to a more accurate construction of the modules. The image on the right illustrates co-clustering of GadE-targets with other genes over 12 Gibbs sampling runs. The purple bar on the left indicates that most GadE-targets consistently co-cluster across these 12 runs, whereas they only co-cluster sporadically with 62 other genes, which are mainly false positives. The final module (green bar), mainly consist of true GadE-targets (purple bar) and the false positive targets are effectively filtered out.

meaningful, true regulators are derived for each module using information from RegulonDB. For each module the 'true' regulator is considered the regulator for which the module shows the highest enrichment in targets. This criterion has the advantage of being well-defined, albeit very stringent (it allows only one true prediction per module). Figure 5- 3 shows the number of true predictions for a given number of predictions ranked by their score. It shows that the regulator score indeed prioritizes the most reliable regulator-to-module assignments. A reliable ranking of the predictions can assist in characterizing novel regulators and is thus an important advantage compared to the method of [30] which fails to rank its predictions.

5.3.2 Topological characterization of module network edges

We further analyze the *E. coli* regulatory network constructed by LeMoNe. To this end we keep all predictions with scores above a cut-off equal to a regulator assignment score of 20% of its maximum value (red lines in Figure 5- 3). This results in a bipartite regulator-to-module network with 57 regulators, 69 modules and 82 edges. The 69 modules together contain 956 genes. The network is shown in Table 5- 1.

From the 82 edges in the network 30 involve an uncharacterized regulator. For 20 edges the inferred regulator is the most enriched in known targets (Table 5- 1, column 'Target Enrichment').

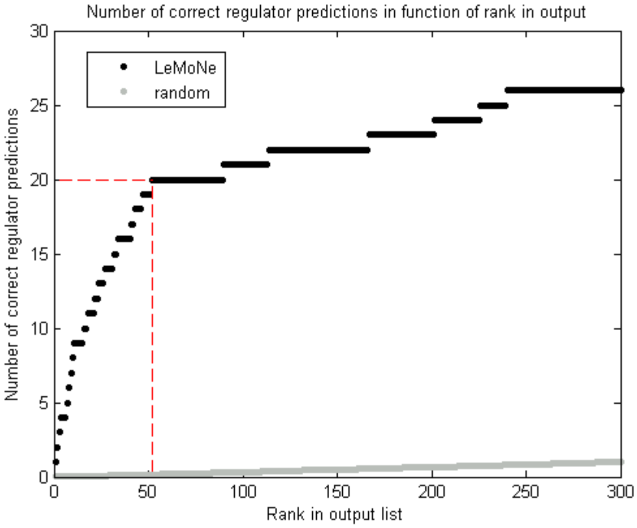


Figure 5- 3 Regulator prioritization showing the number of true predictions for a given number of predictions ranked by their score. A 'true' prediction is a regulator for which the module shows the highest enrichment in targets. The red lines indicate the position of the score threshold (20% of the maximum value). We also compared to a baseline of random predictions, obtained by repeatedly assigning regulators at random to the obtained modules.

For another 9 edges the predicted regulator has known targets in the cluster (Table 5- 1, column 'Target Enrichment"). For the remainder of the predictions 8 regulators were shown to have targets which are involved in the same functions as the module-genes, hence explaining their assignment to the module (Table 5- 1, 'Pathway'). Several newly predicted interactions could be validated by literature.

This network only represents a fraction of the network information present in RegulonDB and therefore we further investigate which parts of the TRN can be accurately captured by LeMoNe.

Table 5- 1 Biological validation of the LeMoNe network for *E. coli.* Target enrichment: (*) module is enriched in known targets of the predicted regulator, (**) module is most enriched for predicted regulator. Autoregulator: (*) regulator is an autoregulator. Pathway: (*) module is enriched in the same function(s) as the regulator. Local: (*) regulator is in the same operon as the module genes, (**) Transcription unit of regulator is adjacent to transcription units of the module genes. Function: enriched functions of the module. Regulators in italic face are putative regulators without known targets; module IDs in italic face consist only of uncharacterized genes.

Regulator	Module ID	Score	Target Enrich.	Autoreg.	Pathway	Local
gatR_2	73	1912.98	**		*	**
gadE	48	1844.50	**	*	*	**
gutM	38	1807.24	**	*	*	*
ymfN	*58*	1749.11				*
ymfN	*33*	1711.17				*
fliA	12	1510.48	**	*	*	**
rcsB	*62*	1261.72			*	*
fecI	57	1200.77		*	*	
gatR_2	42	1176.55	**		*	**
yahA	82	1171.92				
rcsA	87	1151.97	**	*	*	
lexA	20	996.62	**	*	*	*
lldR	65	976.84	**	*	*	*
fliA	45	956.70	**	*	*	
fliA	18	903.46	*	*	*	
nac	85	827.17		*	*	
yiaG	15	816.55				
ydaK	23	815.75				**
ydaK	154	805.22				

fnr	23	798.27	*	*	*	**
lrp	5	777.80		*	*	
araC	46	760.44	**	*	*	**
appY	50	748.75				
yfiE	*67*	736.50				
osmE	15	734.87				
lexA	78	726.67	**	*	*	
purR	144	708.63		*	*	
uidR	81	708.36		*		
araC	21	678.10	*	*	*	
yfeG	29	663.94				
b1450	*53*	662.16				
flhC	18	650.64	**		*	
ogrK	83	645.35				
fliA	17	637.28		*		
rpoS	14	637.13	**		*	*
pdhR	*55*	633.52		*	*	
tdcA	31	619.06	*	*	*	*
yebK	*106*	617.44				
araC	56	608.17	**	*	*	
csgD	26	599.30		*		
hycA	66	596.27				
tdcR	11	593.75				
fliA	24	593.05	*	*	*	
chbR	24	590.31		*		
hycA	29	563.45				*
galS	76	561.25	**	*	*	**
nlp	77	559.41				
yfeC	119	549.33				
b1506	36	548.33				
lrp	10	528.90	*	*	*	
cspB	37	527.86				
cusR	68	515.56	**	*	*	**
b1284	51	514.78				
nanR	9	508.87				
yohL	90	496.21				
lrp	126	493.60		*	*	
yjjQ	*179*	491.02				
yehV	*63*	483.29				
ogrK	27	481.75				
slyA	3	474.43				
ydcN	16	467.66				
cpxR	9	465.39	*	*	*	
yehV	34	451.77				
fruR	*63*	449.25				
araC	64	441.57	*	*	*	
fis	19	436.12	**	*	*	*
fadR	16	435.98	*			
purR	10	431.78	**	*	*	
cadC	37	429.32		*		
fecI	54	429.28		*		
rstA	*102*	428.94				
tdcR	*61*	428.84				
flhC	24	426.88	**		*	*

5.3.2.1 Local vs. global regulators

Different studies [149-151] find a hierarchical structure for the TRN of *E. coli.* Global regulators on top of the hierarchy regulate large numbers of genes all involved in a reaction to a major environmental signal (such as glucose starvation, absence/presence of oxygen, etc.) [153]. The targets of global regulators often perform quite distinct tasks within the global response, explained by the fact that global regulators cooperate with more specific local regulators, downstream in the hierarchy [153]. In the module network in Table 5- 1, most of the modules with known targets involve local regulators. This suggests that differences in expression between transcriptional modules are explained by the more specific, 'local' regulators. We take CRP as an example to further illustrate this. Figure 5- 4 contains a list of modules of which 50% of the genes are known to be regulated by CRP. Differences in expression between the modules can be explained by the presence of specific regulators, often in a feed-forward loop together with CRP. CRP does not show expression correlation with any of the modules, but some of the local regulators do (red edges in Figure 5- 4; six out of seven of the red edges in Figure 5- 4 are inferred in the module network). CRP acts as

Figure 5- 4 Illustrates assignment of local instead of global regulators by LeMoNe in case of CRP. A. Displays a subset of modules (square nodes, inferred from expression data) and known regulators (circular nodes, from RegulonDB). A red edge refers to regulators showing expression correlation with the module (6 out of 7 of the red edges were inferred by LeMoNe). B. Profile plots for two modules (average module expression is indicated in black), the local regulators regulating the module genes (green) and the global regulator CRP (blue). This figure illustrates that for both cases expression of the local regulators is correlated with expression of the module genes, whereas expression of CRP is not.

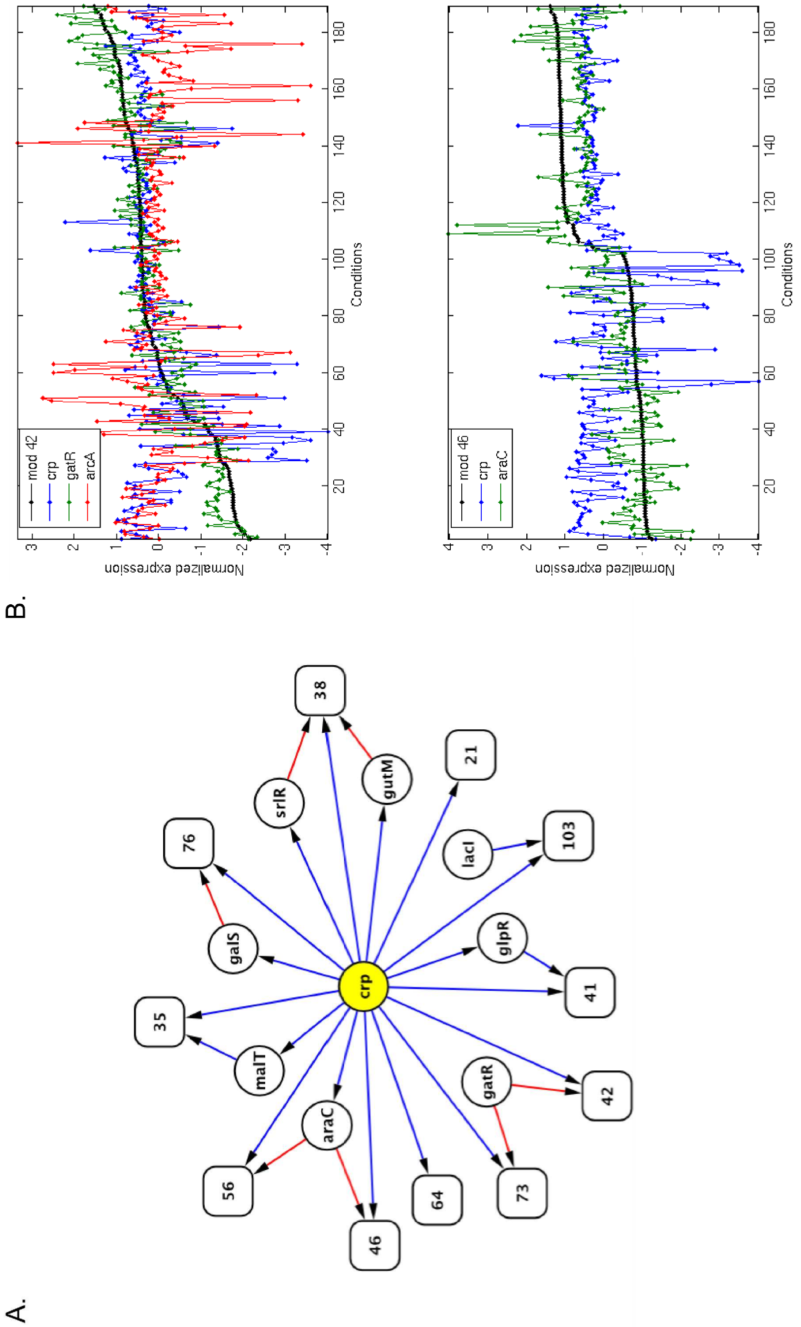


Figure 5- 4 Caption on previous page.

a global regulator, activated by glucose starvation, but depending on the presence of certain alternative carbon sources, CRP will need assistance of more specific regulators to further activate genes that are required for transport and metabolism of those alternative carbon sources (*e.g.* GalS mediates the response to galactose, LacI the response to lactose, etc.) [154]. Other global regulators, such as ArcA, form similar star-like structures, which may be overlapping, *i.e.*, some modules are regulated by multiple global regulators. And here also only the local regulators were correctly assigned to modules by LeMoNe. An exception is given by Fis which is correctly assigned to some of its targets. This can be explained by the fact that it is the only global regulator which regulates most of its genes in an independent matter [153].

5.3.2.2 Autoregulation

Feedback in the E. coli TRN is mainly manifested at the level of autoregulation [149;150;155]. In RegulonDB, 57% of all known regulators are autoregulators. In the module network, 70% (23/33) of the known regulators are autoregulators. If we limit to regulators with edges supported by RegulonDB, this increases to 80% (Table 5- 1). Autoregulators are not overrepresented in the bottom layers of the TRN (for instance 54% of the regulators which do not regulate other regulators, besides possibly themselves, are autoregulators) and we conclude that autoregulators tend to be more coexpressed with their targets.

5.3.2.3 Incoherent interaction and expression correlation sign

In LeMoNe, we get as additional information whether a predicted regulator is positively or negatively correlated with its target module. However, although theoretically possible, we could not detect biologically relevant patterns (supported by RegulonDB) of anticorrelation, in line with previous studies [156]. Even though the assumption of anticorrelation seems intuitively plausible in case of repressors, it is a too simplistic representation of reality. Indeed LeMoNe finds many targets of mainly autorepressors (*e.g.* LexA, PurR, LldR, GatR and GalS), but they all were positively instead of negatively correlated with their targets. This can be explained by the fact that the activity of such autorepressors is dependent upon the presence of corepressing signals. In the absence of the corepressing signal the repressor is active, limiting its own production as well as that of its target genes. In presence of the corepressing signal the repressors are inactive, which enables the production of both inactive repressor gene and its targets [157-159]. For example in case of GalS, correlation with its targets follows from the action of the inducer galactose/D-fucose in the absence of glucose [157]. Upon DNA-damage, LexA's DNA binding capacity is disrupted causing joint expression of LexA and its targets [159]. In the absence of the corepressors hypoxanthine and guanine PurR shows no repression activity and all PurR targets, including itself, will be expressed [158]. These examples show that repressors rarely act purely at the transcriptional level and that therefore the activator or repressor action of uncharacterized regulators can not be inferred from expression data alone.

5.3.2.4 Neighbor regulators

Another class of high-scoring regulator predictions are regulators colocalized with their targets on the chromosome (see Table 5- 1, column 'Local'). Such regulators were termed neighbor regulators in [160]. They often regulate just a few operons, which are known to be tightly coexpressed [160]. They are suggested to be involved in niche specific functions [152] and are hence the prototype of specific, local regulators. Examples of such neighbor regulators identified by our analysis are AraC, GatR, GalS and CusR (see Table 5- 1). In [152] it was suggested that neighbor regulators are often acquired together with their target genes through horizontal gene transfer (HGT), whereas global regulators mostly evolve through vertical inheritance. This, combined with the observation that our analysis mainly characterizes the bottom layers of the hierarchical TRN, is in line with the suggestion of [149] that the bottom layers of the TRN mainly evolve through addition of nodes by HGT. Many uncharacterized neighbor regulators in E. coli were also transferred through horizontal gene transfer [152]. High-confidence predictions can therefore be made for uncharacterized regulators which have a high score in the module network and also lie adjacent to the genes in the module they are predicted to regulate (*e.g.* YmfN to module 33 and 58 and YdaK to module 23).

5.4 Comparison with CLR

In general, the scientific community has mainly focused on the overall performance of newly developed methods in reconstructing the known network of certain model organisms as compared to a reference network, measuring algorithmic performance with standard measures such as recall and precision. Less attention has been paid to what extent conceptually different approaches differ in the networks they infer. Nonetheless, in order to get a better understanding of the systems studied it is also important to understand which specific problems can be tackled using a certain method, irrespective of the overall performance of the different methods.

Broadly speaking we can distinguish between two classes of methods for reverse-engineering transcriptional regulatory networks from gene expression data which differ vastly in how they approach the network inference problem. Direct methods infer individual regulator-target interactions using a pairwise correlation measure between the expression profiles of a transcription factor and its putative targets [34;161]. Module-based methods assume a modular structure of the transcriptional regulatory network [12;30;162], with genes subject to the same regulatory input being organized in coexpression modules.

While different direct methods have been compared to each other in the past [34;163;164], no systematic comparison between direct and module-based methods has been undertaken so far. In this study we perform such a comparison using a representative method from each class. The CLR (Context Likelihood of Relatedness) algorithm [34] considers all possible pairwise regulator-target interactions and scores these interactions based on the mutual information of their expression profiles as compared to an interaction specific background distribution. It has been shown to outperform other direct methods [34]. The LeMoNe (Learning Module Networks) algorithm uses probabilistic, ensemble-based optimization techniques [32;55] to infer high-quality module networks [30], where genes are first partitioned into coexpression modules and regulators are assigned to modules based on how well they explain the condition-dependent expression behavior of the module. It has been shown to outperform the original module network algorithm [32].

Here we compared both methods using a public expression compendium for *Escherichia coli* [34], an organism for which relatively large databases of known transcriptional regulatory interactions exist [165]. We first use recall versus precision curves to give a comparison of the global performance of both methods. We then show that due to the different assumptions underlying both methodologies, they infer topologically distinct networks with limited overlap, even at equal performance thresholds. Biological validation of the inferred networks cautions against over-interpreting recall and precision values computed using incomplete reference networks.

5.4.1.1 Global comparison using recall and precision

The output of LeMoNe and CLR consists of a list of respectively ranked regulator-module and ranked regulator-target interactions, scored according to their statistical significance. As a first, global, comparison, we can therefore compute recall and precision with respect to the given reference networks at different score cut-offs. For CLR we can directly compare the inferred network with the true network; for LeMoNe we draw an edge between each regulator assigned to a module and all genes in the module, thereby ignoring at this stage the extra information present in the module structure. We computed recall and precision as in [34]: if an edge is predicted between two genes present but unconnected in the reference network it is counted as a false positive.

Figure 5- 5 shows the recall versus precision curves for both algorithms. Both algorithms succesfully prioritize true positive interactions in *E. coli*: all curves go from a high precision, low recall region to a low precision, high recall region. For CLR the curves show a smooth course while for LeMoNe they are more staircase-like. CLR scores individual interactions and as a result, in the recall-precision curve interactions will be added one by one, but interactions corresponding to a certain regulator will be dispersed continuously throughout the recall-precision curve. LeMoNe on the other hand assigns a regulator to a module as a whole and all targets belonging to the same module are added at the same time in the recall-precision curve. For a stringent threshold and subsequently a low number of interactions inferred, the CLR network will cover few interactions for many regulators while the LeMoNe network will retrieve many interactions for few regulators.

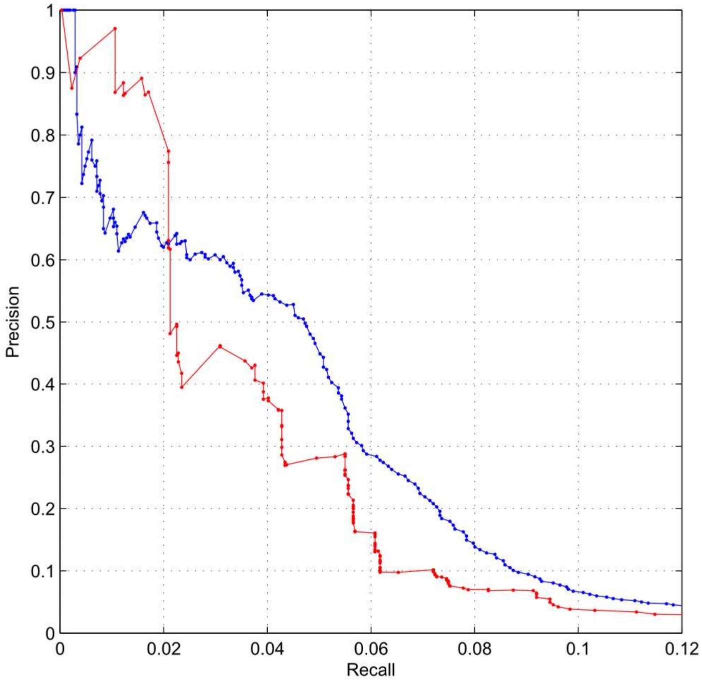


Figure 5- 5 Global comparison of LeMoNe and CLR using recall versus precision curves. Recall versus precision curves for LeMoNe (red) and CLR (blue).

A simple 'area under the curve' (AUC) measurement would suggest that CLR performs slightly better in *E. coli* (the AUC is 0.043 for CLR and 0.035 for LeMoNe). LeMoNe was however shown to outperform CLR in the more complex eukaryote *S. cerevisiae* [66]. This suggests that performance of the algorithm depends on the dataset used and the complexity of the system. In addition, as we will show below, both algorithms infer complementary information in both organisms.

5.4.1.2 Topological distinctions between inferred networks

As explained in the previous section, due to how interactions are scored, direct and module-based methods will infer different kinds of networks at stringent precision thresholds. We compared the LeMoNe and CLR networks at a 30% precision threshold where both networks have nearly equal recall and precision (see Figure 5- 5). The LeMoNe network consists of 53 regulators assigned to 62 modules for a total of 1079 predicted interactions; 594 of these interactions are between genes in RegulonDB, with a precision of 29%. The corresponding CLR network contains 1422 predicted interactions for 242 regulators; 597 of these interactions are between genes in RegulonDB, with a precision of 30%. 51 out of 53 LeMoNe regulators are also present in the CLR network, but only 277 interactions are predicted in both networks.

The networks inferred by LeMoNe and CLR are topologically very distinct. This distinction can be quantified by their in- and out-degree distributions (Figure 5- 6). The in-degree is the number of regulators assigned to a certain target gene and the in-degree distribution counts for each value *k* the number of targets with in-degree *k*. Likewise, the out-degree is the number of targets assigned to a certain regulator and the out-degree distribution counts for each value *k* the number of regulators with out-degree *k*. CLR infers for each regulator only the most significant targets. As a result, the out-degree distribution is skewed to the left, with the majority of regulators having only few targets. The in-degree distribution on the other hand has a long tail of genes assigned to many different regulators. LeMoNe infers for each module the most significant regulators, resulting in opposite characteristics of the degree

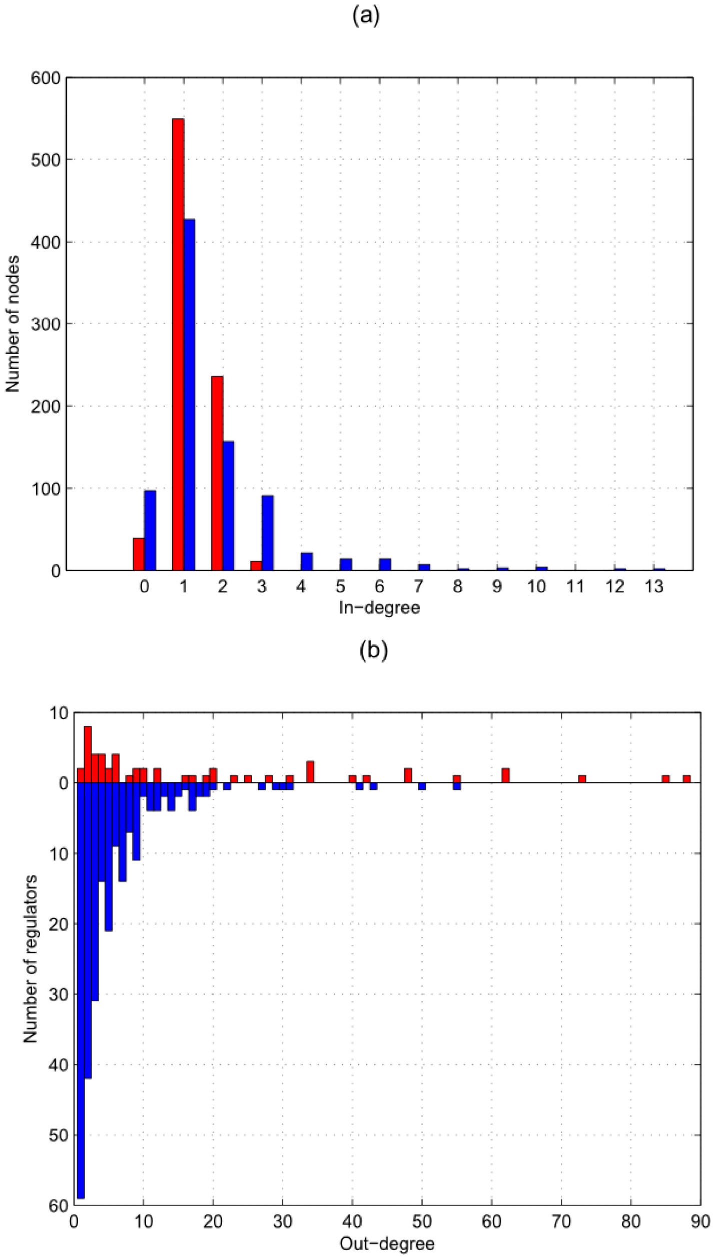


Figure 5- 6 In- and out-degree distributions of LeMoNe and CLR networks**.** (a) *E. coli* in-degree distribution for LeMoNe (red) and CLR (blue) at 30% precision threshold. (b) *E. coli* out-degree distribution for LeMoNe (red) and CLR (blue) at 30% precision threshold.

distributions. The in-degree distribution has no tail since for most modules at most 2 significant regulators are identified. The out-degree distribution on the other hand has a long tail since each regulator assignment involves a whole module of genes. For these reasons, we say that CLR is 'regulator-centric' and LeMoNe is 'target-centric'.

5.4.1.3 Regulator specific comparison

We make a further comparison of the two methods, focusing on how they differ in the type of regulators they assign. We compared again the 30% precision networks for E. coli. For both methods, a large fraction of the regulators for which known targets are inferred are autoregulators. LeMoNe and CLR have respectively 19 and 32 regulators with at least one true positive; 15/19 (79%) and 27/32 (84%) are known autoregulators, while the fraction of autoregulators in the total reference network is 95/150 (63 %). The abundance of autoregulators is not surprising since autoregulation is a simple mechanism by which the expression profile of a regulator and its targets can be correlated.

Similarly to LeMoNe, CLR fails to predict correct interactions for repressors, unless it is an autorepressor. Regulators for which the module-based and direct methods differ in performance are in line with the topological distinctions. CLR is better at inferring interactions for regulators that are known to regulate just one or a few operons (*e.g.* BetI, CsgD, DnaA, MarA, Yhhg, see Figure 5- 7). These operons are found with a relatively high rank in the CLR network since their regulators often belong themselves to the operons and are thus by definition tightly coexpressed with their targets. The clustering method employed by LeMoNe appears to be too coarse-grained to identify these operons individually, since they are mostly part of larger clusters. LeMoNe on the other hand is superior at inferring interactions for regulators that are known to regulate larger regulons, such as Fis, LexA, PurR, and RpoS, for which the level of coexpression is not as high as the one observed within a single operon (see Figure 5- 7).

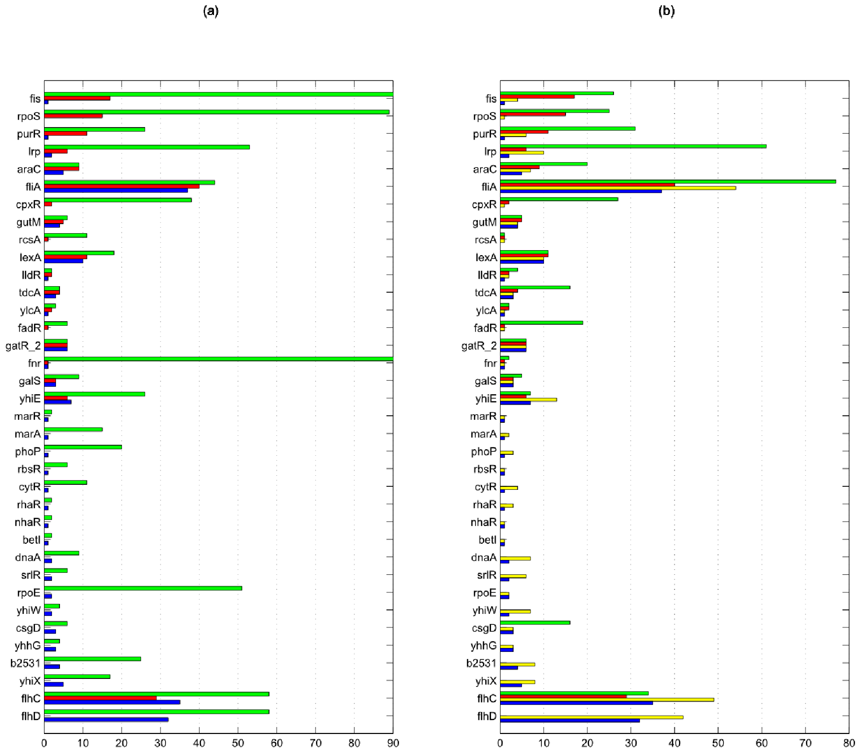


Figure 5- 7 Regulator specific comparison of LeMoNe and CLR on E. coli**.** For each regulator in *E. coli* with known interactions inferred: (a) the number of interactions in the reference network (green) and the number of true positives in LeMoNe (red) and CLR (blue); (b) the number of interactions inferred (green) and the number of true positives (red) in LeMoNe, and the number of interactions inferred (yellow) and the number of true positives (blue) in CLR. LeMoNe and CLR networks are both at 30% precision threshold. Regulators are sorted by the difference $\mathrm{TP}_{\mathrm{LeMoNe}} - \mathrm{TP}_{\mathrm{CLR}}$. The total number of true positives is 171 for LeMoNe and 180 for CLR. For clarity, the *x*-axis in (a) is truncated, the true number of targets for Fis and Fnr is respectively 111 and 173. The number of interactions inferred only counts targets that belong to the reference network.

5.4.1.4 Biological validation of inferred networks

Due to the lack of a negative gold standard, we have denoted in the previous analysis an edge as being false positive if both regulator and target are present but not connected in the reference network (the positive gold standard). Since the coverage of these reference networks is still very incomplete, it is likely that the number of false positives is overestimated. Moreover, about half of the predicted regulators in *E. coli* are not present in the reference network and their predicted interactions are thus never evaluated.

Here we have performed an in-depth biological validation of the 30% precision module network inferred by LeMoNe. To biologically validate the obtained regulator-module assignments, we calculated for all modules functional enrichment scores [93] and enrichment in targets of previously annotated regulators [165].

Table 5- 1 shows that in nearly all cases the module is enriched in known targets of the predicted regulator (column 'Target Enrichment') or at least involved in the same biological function (column 'Pathway'). In several cases the predicted regulator is the one which has the best target enrichment p-value (column 'Target Enrichment').

Nearly half of the regulators are putative regulators without any currently known targets, and these assignments cannot be validated. However, many of the correctly predicted regulators involve neighbor regulators [160] (Table 5- 1, column 'Local'), *i.e.* regulators colocalized with their targets on the genome. In 5.3.2.4 it has been suggested that many of the putative regulators in *E. coli* constitute such neighbor regulators [152]. Hence this feature of gene neighborhood can be used to attach additional significance to the high-scoring predictions for uncharacterized regulators. One of the advantages of a module-based approach is the fact that if a certain module contains several known targets of the assigned regulators, the rest of the unknown targets in this module can be considered high confidence predictions for that regulator.

Biological validation of inferred networks is tedious and does not provide an easy alternative to the automatic estimation of true and false positives using an established reference network. The results of this section do show however that many 'false positives' with respect to an incomplete network are likely true positives when additional information is taken into account and that recall versus precision plots such as in Figure 5- 5 have to be interpreted with caution.

5.5 Conclusion

We have constructed an ensemble-based module network for *E. coli* from expression data using a Gibbs sampler clustering algorithm and a method for inferring probabilistic transcriptional programs. The module network connects modules of co-regulated genes to condition-specific regulators which explain the expression profile of the module. Unlike a single optimum, ensemble averaging allows the assessment and prioritization of the statistically most reliable modules and their condition-specific regulators. As regulators are ranked according to a significance score, the method is especially useful to make initial high-quality predictions for uncharacterized regulators and unknown genes. The quality of these predictions is supported by assignments of known regulators to modules enriched in their targets, and several new predictions for known regulators could be validated by literature.

An analysis of the module network in the context of the hierarchical topology of the known *E. coli* transcriptional regulatory network shows that local regulators near the bottom of the hierarchy explain expression differences between different modules and are more often coexpressed with their targets than global regulators near the top of the hierarchy. The bottom layers of the hierarchical network have evolved mainly through addition of new regulators together with their target genes by horizontal gene transfer and consequently we find many high-scoring regulator assignments colocalized with their targets on the chromosome. These results illustrate that LeMoNe only characterizes a fraction of the *E. coli* regulatory network (*i.e.* the bottom layer) and therefore complementary methods are necessary to get a comprehensive view on this TRN.

Consequently, we compared our module-based approach (LeMoNe) with another approach for reverse-engineering transcriptional regulatory networks: the direct CLR method. We have found that CLR is 'regulator-centric', making few but highly significant predictions for a large number of regulators. LeMoNe on the other hand is 'target-centric', identifying few but highly significant regulators for a large number of genes grouped in coexpression modules.

Through a regulator specific comparison and analysis of specific biological subsystems, we have shown that at stringent significance cutoffs, the conceptual differences in statistically scoring potential regulatory interactions lead to topologically distinct inferred networks containing different kinds of regulators and biological information. Our results show that the choice of algorithm should be made primarily based on whether the biological question under study falls within the target-centric or regulator-centric viewpoint, and not on global metrics which cannot be transferred between organisms. Ideally, several network inference strategies should be combined for the best overall performance. It is an important challenge for future research to develop sound statistical methods for optimally combining the output of multiple, existing reverse-engineering algorithms.

5.6 Methods

The *E. coli* microarray data compendium [34] contains expression profiles for 4345 genes under 189 different stress conditions and genetic perturbations. We selected a subset of 1882 differentially expressed genes (standard deviation larger than 0.5) and used a list of 316 known or putative transcription factors [42;165] to reconstruct regulatory networks. LeMoNe [32] (software available at http://bioinformatics.psb.ugent.be/software/details/LeMoNe) identified 108 ensemble-averaged modules from 12 independent Gibbs sampler runs, containing 1761 genes in total. It inferred a ranked list of regulator-module edges from an ensemble of 10 transcriptional programs per module with 100 regulator samples per transcriptional program node (see [32] for more details on the meaning of these parameters). We applied CLR [34] (software available at http://gardnerlab.bu.edu/clr.html) on the data for the 2084 selected genes (the union of the 1882 differentially expressed genes and 316 candidate regulators) and kept all mutual information z-scores between the 316 transcription factors and 1882 target genes. As a reference network we used RegulonDB version 5.7 [165], a database of 4840 known transcriptional interactions in E. coli between 167 transcription factors and 1693 genes. Recall values are computed with respect to RegulonDB restricted to the subset of 2084 genes. This subnetwork contains 3110 edges between 150 transcription factors and 1053 genes.

We used EcoCyc [93] to compute functional enrichment of modules. Target and functional enrichment in Table 5- 1 were computed using a cumulative hypergeometric distribution, Bonferroni corrected for multiple testing, with confidence level 95%.

Chapter 6

Exploring complementary aspects of network inference approaches

6.1 Introduction

In bacteria regulation at the transcriptional level is pivotal to guarantee metabolic flexibility and cellular integrity [1;166]. Deciphering the coexpression or the transcriptional regulatory network (see Chapter 1) is thus crucial to understanding bacterial cellular behavior. The number of computational methods that are being developed to reconstruct TRNs from genome-wide expression data is rapidly increasing. Indeed, the examples presented in this thesis only constitute a small portion of the module and network inference methods that have been introduced in the past decade.

In Chapter 1 we introduced network inference and module inference as a problem that is computationally underdetermined: due to the large number of possible solutions (or the large search space), together with the restricted number of independent data points and the relatively low information content of the available data [43;44;167;168] different solutions are possible that all explain the data equally well. Therefore, different methods incorporate different strategies to deal with this problem of underdetermination, often resulting in different outcomes of the inference problem.

In this chapter we provide a scheme that allows classifying state-of-the-art transcriptional inference methods based on the strategies used to solve the inference problem. In contrast to previous categorizations, our classification uses a combination of criteria that directly relate to the biological interpretation of the outcome rather than being merely dataset related [169] or computationally focused [170;171]. We use representative tools of each class to show how using different strategies results in inferring different types of interactions, hereby extending the observations made in the previous chapter. In addition, we draw attention to first tentative attempts that have been made to leverage the predictions made by different network inference tools in an ensemble-based strategy.

6.2 Strategies to deal with underdetermination

The problem of underdetermination relates to the size of the search space: the larger the search space, the larger the complexity of the inference problem and the more difficult it will get to find the unique solution that approximates the biological truth. To tackle this problem of underdetermination, module and network inference methods adopt different strategies which reduce the search space and/or extend the amount of independent information (Figure 6- 1).

'Conceptualization by simplifying biological reality' is a commonly used strategy that renders network inference more tractable. Transcriptional regulatory networks have been shown to be modular in structure [150]. Module-based approaches exploit this observation and simplify the complex structure of the TRN by modeling the network as sets of overlapping modules of functionally related genes. Genes belonging to the same module all act in concert under certain environmental cues [11-13], explaining their coordinated expression behavior. Modules are identified by module detection methods that rely on clustering or biclustering [28;29]. Module-based network inference procedures, primarily designed to infer transcriptional programs, also make use of this concept of modularity: in contrast to assigning an individual program to each single gene as with direct network inference methods, they assign a transcriptional program to pre-grouped gene sets or modules. This drastically lowers the number of interactions that must be evaluated during the inference process. Another simplification relates to the definition of combinatorial regulation where multiple regulators act together to mediate specific condition-dependent responses. Inferring a transcriptional program that allows for combinatorial regulation implies that all possible combinations of regulators and their possible binding modes (*i.e.* cooperative, synergistic, etc.) must be evaluated in order to explain the observed expression behavior. As this is computationally intractable for large datasets, all large-scale network inference methods make an approximation of combinatorial regulation.

A second strategy relates to extending expression data with other available information (integrative versus expression-based methods). Integrative methods that combine the expression with complementary data describing the TRN from a different angle, such as for instance chromatine immunoprecipitation on chip (ChIP-chip) or motif data, often obtain more reliable interactions and a more complete picture of the network. Moreover, by prioritizing during the search predictions for which the different independent data sources agree allows traversing the search space more efficiently.

As a third strategy, query-based methods reduce the search space by intentionally restricting the number of interactions that needs to be evaluated: instead of searching for a global pattern as is done by global inference methods, query-based methods concentrate their search on a predefined set of core genes or on a subnetwork one is interested in and expand upon this core gene set or subnetwork given the present data. Treating the inference as a classification problem as is done with (semi-)supervised methods (fourth strategy) can be considered as a specific way of exploiting known information in a query-based way.

As each strategy implies different assumptions and poses different constraints, adopting a specific strategy or combination of strategies determines the type of interactions that can be found. This will be further illustrated with results obtained from state-of–the-art inference tools that have successfully been applied to microbial datasets.

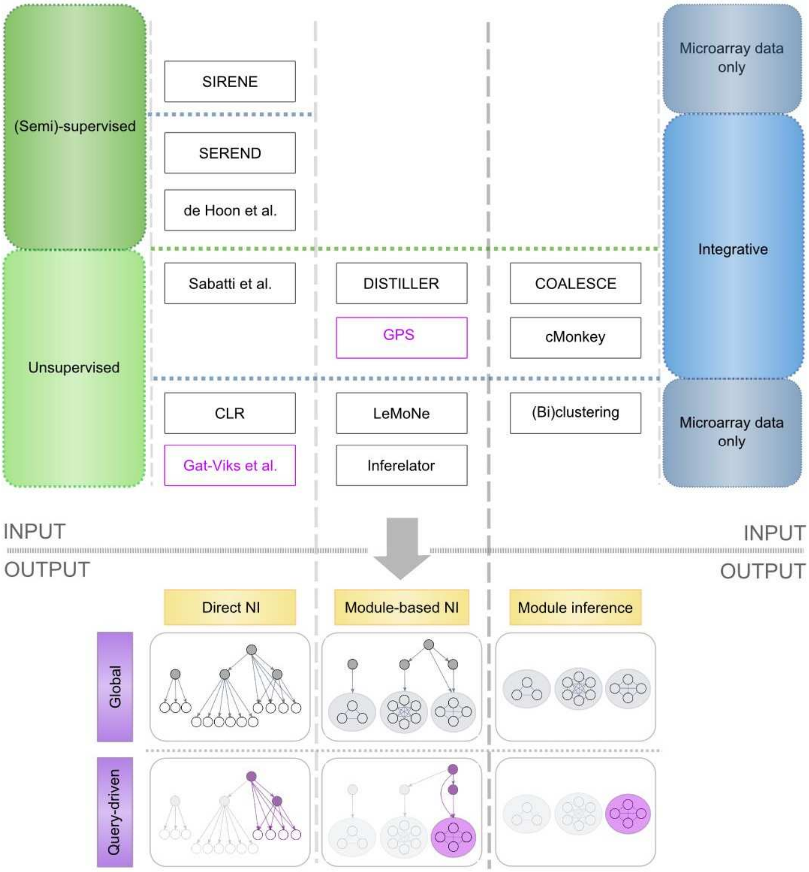


Figure 6- 1 Caption on next page.

Figure 6- 1 Categorization of different state-of-the-art methods for module and network inference. Module inference methods search for sets of coexpressed genes. Network inference methods on the other hand search for a transcriptional program that explains an observed expression behavior. According to the strategies methods use to cope with the problem of underdetermination, they can be categorized as follows: 1) integrative versus non-integrative methods (blue): methods that complement expression data with additional data sources versus methods that use expression data only, 2) module-based versus direct inference, (yellow): methods that conceptualize the network by treating sets of coexpressed genes as single entities (modules) versus methods that consider all genes on an individual basis, 3) query-based versus global (purple): methods that start from a predefined set of core genes or core pathways and expand upon those versus methods that search for global patterns in the data, 4) supervised versus unsupervised (green): methods that treat the inference problem as a classification problem versus methods that do not. Most of the methods can be either used in a query-based or global mode. The methods indicated in purple, are specifically designed to be query-based.

6.2.1 Module-based versus direct network inference

Usually module-based network inference methods use module inference based on biclustering as a first step, prior to the assignment of the transcriptional program. Exploiting the concept of modularity offers advantages from both the biological and the statistical point of view. Most module-based approaches do not only predict regulatory interactions, but also identify the experimental conditions under which the predicted interactions take place. This information can be helpful in designing the appropriate conditions under which experimental validation of the predicted interactions should optimally be performed [8;162] Assuming modularity also contributes to the statistical robustness of the inferred interactions: all coexpressed genes within a module confirm each other in providing evidence for a certain transcriptional program, while for direct methods this evidence for a particular regulator-target interaction is only based on a single gene observation.

How adopting the concept of modularity determines the interactions that can be inferred is illustrated by a comparison between the results of the direct method CLR and the module-based network inference method 'Stochastic LeMoNe' (Table 6- 1) (Figure 6- 2). By exploiting modularity, LeMoNe and related methods [30] are able to assign programs with an expression profile that is less similar to that of its target genes than is the case with CLR or similar methods [161;172]. Indeed, LeMoNe performs better than CLR in inferring transcriptional programs for genes grouped in coarse-grained modules that correspond to larger pathways (*e.g.* Fis, RpoS and PurR) and for which the genes show an overall low degree of coexpression with each other or with their

Figure 6- 2 Complementarity in the type of interactions inferred by direct and module-based inference methods. CLR and Stochastic LeMoNe, as representatives of respectively direct and module-based inference methods were applied to the same E. coli compendium [34]. The precision of the inferred interactions was calculated as described in Faith *et al.* [34] using experimentally documented interactions in RegulonDB [94] as a standard. Panel (a) compares the precision with which true interactions were inferred per regulator between both methods by calculating per regulator the difference in precision obtained with CLR and LeMoNe. Regulators were ranked according to this difference in precision. High negative value indicate a higher precision of LeMoNe than of CLR, high positive values indicate the opposite. Panel (b) shows the values of the regulator-specific precision for Stochastic LeMoNe (green) and CLR (blue). Panel (c) illustrates the size distribution of the known regulon membership according to RegulonDB for the regulators where respectively LeMoNe (upper part) and CLR (lower part) show a higher precision. Panel (a) and (b) illustrate the complementarity between both methods in retrieving interactions for different regulators. Panel (c) shows how Stochastic LeMoNe predicts on average correct targets for more global regulators (a larger regulon size), whereas CLR typically predicts targets for regulators with a smaller number of known targets. Note that predictions for regulators not documented in RegulonDB were not included in this plot.

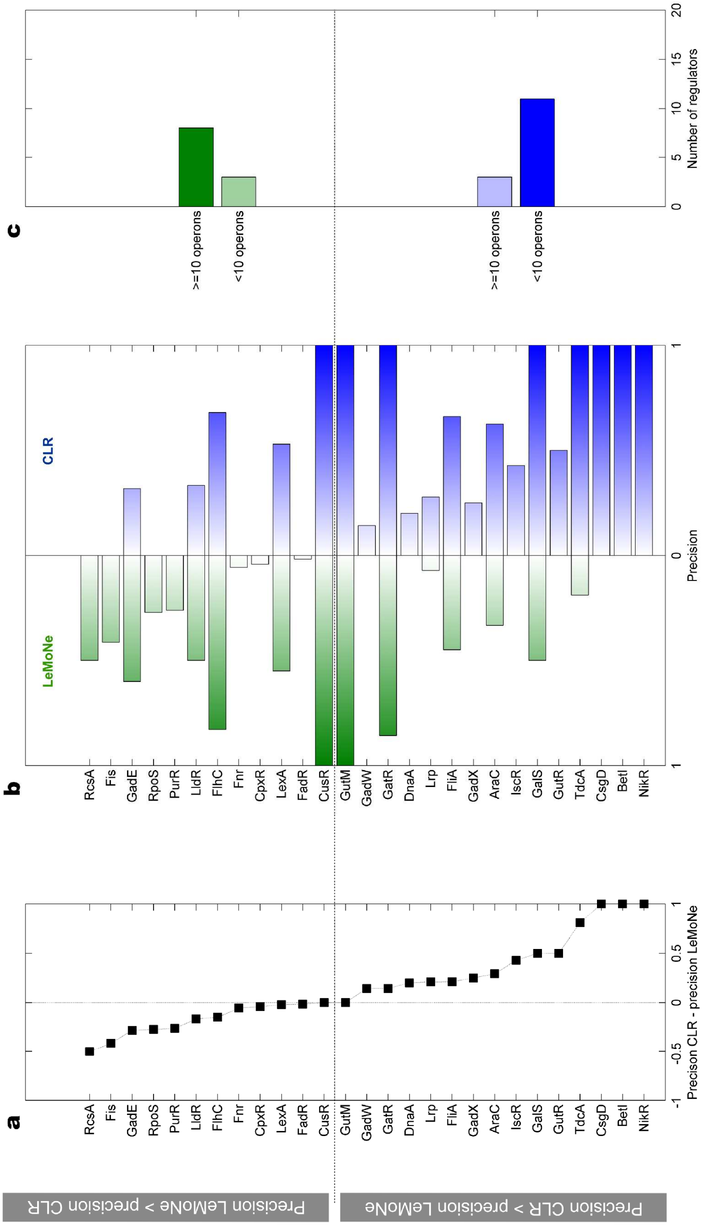


Figure 6- 2 Caption on previous page.

transcriptional program [66]. Conversely, CLR showed a higher precision than LeMoNe in identifying targets for regulators that are dedicated to one or at most a few operons as in prokaryotes such operonic regulators are tightly coexpressed with their targets (*e.g.* GutR, IscR, BetI, AraC). A direct method such as CLR also covers interactions for a larger range of regulators than a module-based method such as LeMoNe as a module–based inference method looses interactions with target genes that are not coexpressed with a sufficient number of other target genes [66].

6.2.2 Modeling combinatorial regulation

Inferring combinatorial regulation in its full complexity is also computationally intensive. Most direct methods, supervised (SEREND [31], SIRENE [33]) (see further) (Table 6- 1) as well as unsupervised (CLR [34]), simplify the problem by assigning regulators one by one to their target genes and composing the combinatorial transcriptional programs in a post-processing step as sets of regulators that belong to the same target genes. Although significantly reducing the complexity of the network inference problem, such a stepwise approach renders it impossible to distinguish between truly complex combinatorial regulation, where the signal of multiple TFs is integrated to trigger the observed gene expression pattern, or condition-dependent regulation, where different TFs act independently from each other to mediate expression under different subsets of conditions. For instance, applying CLR to *E. coli* resulted in the correct assignment of the regulators GadW, GadX and GadE to several genes involved in the acid response [34]. The true more complex relation between these regulators with GadE, which acts as the main regulator of the acid response and is under control of both GadW and GadX [173], could not be unveiled.

Module-based inference methods such as Stochastic LeMoNe [32] and DISTILLER [8] (Table 6- 1) automatically take into account the condition-dependency of the inferred transcriptional programs: regulators that are assigned to the same genes, but under different subsets of conditions are assumed to act independently from each other, while regulators that were predicted to regulate the same target genes in similar conditions presumably act combinatorially. Using DISTILLER, Lemmens *et al.* [8] detected for instance that the *E. coli* global regulator CRP interacts, depending on the conditions, with different specific regulators. Neither DISTILLER nor Stochastic LeMoNe can infer the mode of the combinatorial interactions between the assigned regulators, *i.e.* whether the assigned TFs act together in an additive or synergistic way (AND), whether in a combinatorial interaction the presence of one of the regulators is sufficient to trigger expression of the target gene (OR) or whether their binding is mutually exclusive (XOR). By combining the expression profiles of the regulators according to these different possible interactions (AND, OR or XOR) before assessing how well they explain the target's expression behavior, Inferelator [162] can infer those more complex modes of transcriptional interactions. Recently, also CLR was extended to account for synergistic relations (synergy augmented CLR), *i.e.* the latter occurs when the expression value of a third gene can be better explained by two genes together than by each of them separately [174;175]. Using this approach, the authors could uncover novel links in the original *E. coli* CLR network such as, for instance, that the expression of *fecA* depends on both *fecI* and *aceK,* with *aceK* presumably acting as an indirect inhibitor of ferric citrate transport mediated by FecA [175].

6.2.3 Integrative versus expression-based approaches

Non-integrative expression–based network inference methods extract information on regulator-target interactions from the expression data itself. Except for the supervised expression-based methods, such as SIRENE [33] that exploit the observed coexpression behavior of known targets of a particular TF (see further), most non-integrative methods assume that the regulator's expression profile is a proxy for its activity (Stochastic LeMoNe [32], CLR [34], Inferelator [162], and correlation-based methods [176;177]). The latter assumption disregards the important role of regulation mechanisms at levels other than the transcriptional one [178] and restricts the interactions that can be inferred to those of regulators that are either coexpressed or anticorrelated with their targets [156] (Figure 6- 3). As a result expression-based inference methods, such as CLR and Stochastic LeMoNe or related ones [30;161;172], are biased towards inferring interactions of auto- or operonic regulators that were shown to be tightly coexpressed with their targets [66]. Moreover, most expression-based inference methods are not able to distinguish between regulators that actually regulate the gene (direct causal effect) or regulators that are simply coexpressed with it (mere correlation). This problem can partially be alleviated by inferring networks from dynamic instead of from static data as time series inherently contain information on causal effects, if one assumes that the expression of the TF needs to be altered before it can affect its targets (in a direct way or via a regulatory cascade). Inferring networks from dynamic data requires special techniques that capture the dynamics (*e.g.* the lag in expression profiles between genes). Schmitt *et al.* [177], for instance, used time-lagged correlation analysis (Table 6- 1) to infer the regulatory network that mediates the response to alternating light conditions in the cyanobacterium *Synechosystis*, while Shaw *et al.* [176] inferred the *B. subtilis* regulatory network using the same technique. In practice, inference of networks from dynamic data is restricted by the insufficient time resolution of the available samples, which complicates distinguishing true from noisy signals and results in missing fast responses.

By complementing gene expression with additional transcriptional information (such as motif data, DNA-protein interaction data), integrative network inference methods [8;31;35;179-181] can extend the scope of their predictions beyond interactions that can be inferred from coexpression behavior and result in general also in more reliable predictions (Figure 6- 3). Sabatti *et al.* [182] propose a direct integrative approach based on hidden component analysis (Table 6- 1) that overlays a network topology derived from known and motif-based interactions with expression data. It was used to infer the transcriptionally active edges in the *E. coli* network. By exploiting known information on regulatory motifs and transcriptional interactions derived from EcoCyc [93] in a supervised way, the direct integrative method SEREND could infer novel interactions for previously characterized regulators of *E. coli* (see further).

Module-based network inference methods such as DISTILLER [8], cMonkey [183] and COALESCE [184] (Table 6- 1) rely on an integrative module detection step to derive their transcriptional program. Integrative module inference searches for genes that are not only coexpressed with each other, but also share a common regulatory binding site (identified by motif detection or ChIP-analysis). Exploiting complementary data sources to confirm expression-based module assignments reduces the assignments of false members to true modules or the detection of spurious modules. As the observed coexpression in a module now also truly implies coregulation, the module inherently contains information to infer the transcriptional program: for instance, to each module the regulator is assigned that is known to recognize the motif or binding site found to be associated with the module. Applying DISTILLER to a cross-platform *E. coli* expression compendium and motif data for 67 known regulators resulted in the prediction of 278 novel interactions for 29 different regulators [8]. Of the 11 novel interactions for the regulator FNR that were experimentally verified by ChIP-qPCR, none were retrieved by the non-integrative methods CLR [34] and Stochastic LeMoNe [66]. When using these integrative approaches in combination with *de novo* detected motif sites, the assignment of a cognate regulator will be based on additional computationally derived criteria (*e.g.* average proximity of the regulator to its targets) [25] or on a concomitant expression-based inference of the transcriptional program [162]. In the future, integration with data resulting from protocols that globally survey an organism's proteome for sequence-specific interactions with putative DNA regulatory elements will further facilitate mapping of cognate regulators to novel motifs [185;186].

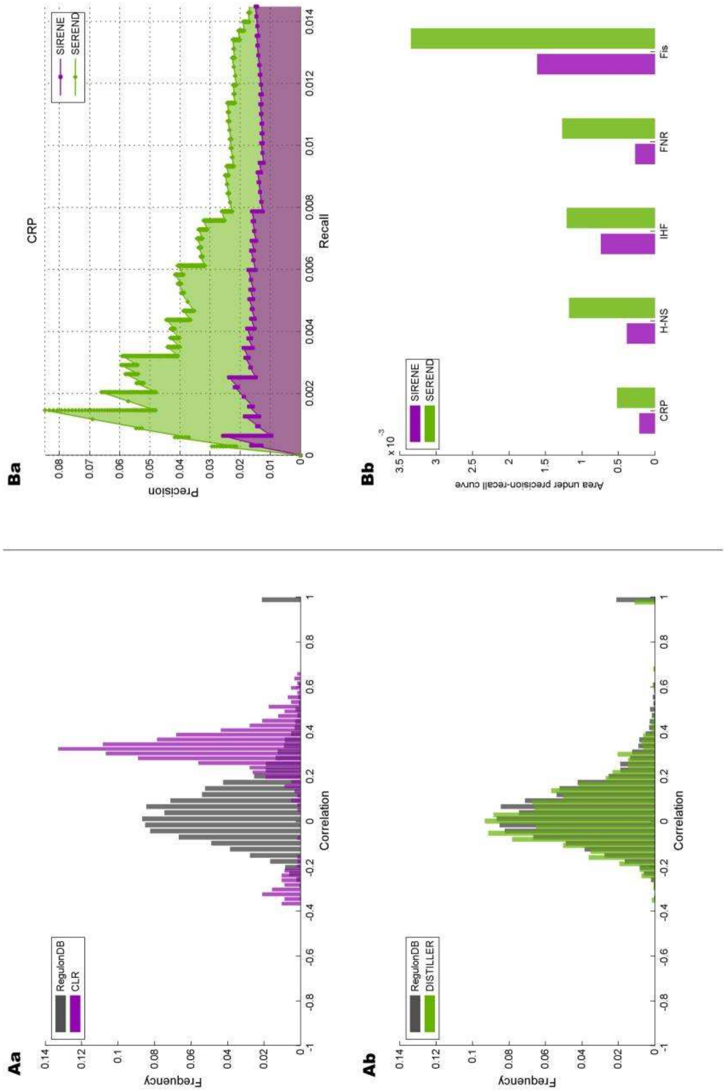


Figure 6- 3 Caption on next page.

Figure 6- 3 Illustration of the different characteristics of interactions inferred by either expression-based or integrative network inference methods. Left panel: illustrates how expression-based methods that estimate the activity levels of the regulators from their expression profiles are biased towards predicting interactions for regulators that are tightly (anti-)correlated with their targets, while for methods that infer their transcriptional program from complementary data sources, this is not the case. The expression-based methods CLR (Aa) and the integrative network inference method DISTILLER (Ab) were applied to the same E. coli expression compendium (results were taken from Lemmens *et al.* [8]). The histograms displays the number of predicted pairwise TF-target interactions as a function of their mutual coexpression. In black is shown, by means of a reference, the same distribution, but for all regulators documented in RegulonDB [94]. The peak at correlation coefficient 1 corresponds to the situation where the two considered profiles of respectively the regulator and the target gene are exactly the same, which is the case for auto-regulators. Right panel: illustrates how the integrative methods result in more reliable predictions than those obtained with an inference method that only uses expression information. The performance of an expression-based (SIRENE, purple) and an integrative network inference method (SEREND, green) were compared using ChIP-chip data as external standard. The upper figure (Ba) displays the precision-recall curve for SEREND and SIRENE for the regulator CRP. The area under the precision-recall curve, indicated by the shaded area, is used as an estimate of the overall performance. Figure Bb compares the area under the precision-recall curve for both SIRENE and SEREND of five different regulators for which ChIP-chip data [85;187;188] is available. This figure shows that the integrative SEREND method outperforms SIRENE in retrieving ChIP-chip targets for each of the regulators considered.

So, inference methods that only use expression data are useful for organisms with little additional information available. Integrative methods on the other hand provide a more complete view on the network and are more likely to predict true positive interactions. The additive value of integrative methods, however, depends both on the quality of the additional data [189] and of the used algorithm.

6.2.4 Global versus query-based inference

Global module inference methods [75;76;81;190-195] search for the modules that explain most of the data. This generally corresponds to identifying large pathways, consisting of many genes, and usually responsible for general responses upon major metabolic or condition shifts such as, for instance, flagellar synthesis, amino acids biosynthesis or DNA damage. As such, global approaches provide a general view on the active TRN and its resulting physiological state. Query-based module detection methods on the other hand [12;72;73] search for genes that are coexpressed in a condition-dependent way with a predefined set of genes (also called query-genes). These algorithms are deliberately biased towards finding a specific local solution in the search space that is of particular interest to the user. This solution is not so trivial to find in a global way as either the expression signal of the query-genes is too low to be significant or the local solution is obscured by a more global one. As an example of the latter effect searching an *E. coli* compendium for a PurR-related module using a known PurR-target as query results in a module that is indeed significantly enriched for previously known PurR-targets (p-value $< 1\text{x}10^{-15}$) while with a global approach the module that contains most PurR-related genes is under default conditions much larger and enriched for more general functions related to amino acid biosynthesis and translation. Query-based approaches are thus typically used to expand or curate a particular pathway or process by either searching for additional genes that are coexpressed with genes known to be already involved in the pathway or by filtering out genes that are not coexpressed with the majority of the so called pathway genes. Ihmels *et al.* [12], for instance, used the query-based Signature Algorithm (SA) to refine the gene set involved in the TCA cycle in *S. cerevisiae* with the homologs of 37 *E. coli* TCA cycle genes as query.

Regarding network inference methods, most of the already described global methods can be applied in a query-based setting through restricting their input sets. In some cases this can be advantageous, for instance methods such as CLR, Stochastic LeMoNe and Inferelator perform better if the transcriptional program can be inferred from a prespecified list of regulators than from a full gene list because *a priori* erroneous interactions with non-regulators will be eliminated. Algorithms specifically designed for query-based network searches focus on one or few core pathways [196;197;197;198]. By constraining their search space to only those solutions that contain the query, these methods can afford making more detailed network models than would be possible in a global setting. Gat-Viks *et al.* [198] (Table 6-1) formalized qualitative existing knowledge on the yeast osmotic response as a probabilistic model. Interrogating this model with expression data allows both refining the model by correcting erroneous interactions and extending the original network with novel targets affected by components of the original network. Alternatively, kinetic approaches for modeling the dynamics between TF and target gene from time series expression data, that are still intractable on a genomewide scale, have been successfully applied in a query-based mode to validate the outcome of a ChIP-chip experiment. So far they have only been applied in higher eukaryotes [199]. The GPS algorithm [196] (Table 6- 1) is another query-based network inference method that takes advantage of detailed promoter descriptions in combination with mutant expression data to extend the regulon of a predefined regulator. More specifically, GPS could identify four additional PhoP targets in *S. Typhimurium* that previously were thought to be only indirectly PhoP-dependent. In addition, the identified PhoP targets in *E. coli* fell apart in separate modules of coexpressed genes, one of which primarily contained genes involved in acid resistance. This allowed establishing a novel link between PhoP regulation and bacterial acid resistance [196;200].

6.2.5 Supervised versus unsupervised inference of the transcriptional program

Supervised methods treat the inference as a classification problem. They start from a set of known TF-target interactions. Based on this predefined training set, characteristic features are derived, such as TF binding sites (SEREND [31] and de Hoon *et al.* [201]) or the degree of coexpression between TF targets (SIRENE [33], SEREND [31] and de Hoon *et al.* [201]). These characteristics are subsequently used to evaluate a novel candidate gene as a potential target of a TF. Genes that share many characteristics with the known targets of the TF are classified as true targets, the others as non-targets. Such a classification strategy depends on the quality of the used training set of true positive and negative interactions. While it is for model organisms such as *E. coli* and *B. subtilis* quite straightforward to extract examples of positive interactions from curated databases, such as RegulonDB [94] (*E. coli*), EcoCyc [93] (*E. coli*) and DBTBS [202] (*B. subtilis*), the definition of true negative interactions is much less trivial. Genes not known to interact with a specific regulator, *i.e.* 'unknowns', are then often treated as the negatives. As our knowledge of TRNs is still limited, there is a good chance that such a set of 'unknowns' contains as of yet uncharacterized true positive interactions for a given TF, in which case the classification results will be deteriorated.

By extrapolating previously known information, interactions predicted with supervised methods are generally highly reliable, but are restricted to regulators with a sufficiently high number of previously known targets, such as global regulators and sigma factors from well characterized model organisms (*E. coli* [31;33] and *B. subtilis* [201]) (Figure 6- 4). SEREND was shown to be very useful in extending the repertoire of interactions for the *E. coli* global regulators IHF, H-NS, CRP, FNR and Fis [31].

To infer interactions in less-studied systems with little previously characterized information, unsupervised approaches are more suitable (*e.g.* Stochastic LeMoNe, CLR, DISTILLER, Inferelator) as they do not necessarily depend on previously known information and can also infer interactions for regulators with little or no prior information available (Figure 6- 4). Mainly the unsupervised methods that can infer their transcriptional program from only expression data such as CLR and Inferelator have been shown useful to provide a first global view on the TRNs of for instance *Salmonella Typhimurium* [203;204], *Shewanella oneidensis* [205], *Halobacterium salinarum* [206] and Cyanobacteria [207].

6.3 Choosing benchmark datasets

Benchmarking is important to understand the reliability of the reconstructed network. It is based on the calculation of the precision and recall according to a predefined external standard. By collecting all curated interactions of a particular organism and treating them as true positives, and treating all predicted interactions between a gene and a TF other than the ones it was documented to interact with as false positives, a standard is generated. Using such a standard tends to overestimate the false positive prediction rate as most genes probably interact with many more TFs than is currently documented. Moreover, all novel interactions with TFs for which no interactions are documented yet are ignored in the assessment. As a result, using an external standard rewards methods that merely reproduce current knowledge, but penalizes the ones that perform well in finding new biology. To compensate for this, most current studies combine validation based on an external standard with medium-throughput experiments to also validate novel biology [8;34;38].

Medium-throughput experiments avoid the infeasible task of testing all novel predictions by sampling a set of predicted interactions that is representative for the whole analysis. In practice this set usually consists of both high and lower confidence interactions for a subset or one of the assessed TFs. Mainly global regulators in *E. coli* were chosen, such as FNR [8] and Lrp [34;38] as for these regulators the good balance between yet to be discovered and already known interactions favors the benchmarking. Lemmens *et al.* [8], Zare *et al.* [38] and Faith *et al.* [34] for instance could show by combining performance analysis using RegulonDB with a ChIP-based medium-throughput experiment that their respective methods had a good sensitivity in detecting known interactions, but also that high-scoring novel predictions usually corresponded to true interactions.

For network inference methods that use predictive models, cross-validation can be used to validate the reliability of the inferred model: the ability of the model to predict the expression behavior of genes in experiments that were not used to build the model is assessed [32;162].

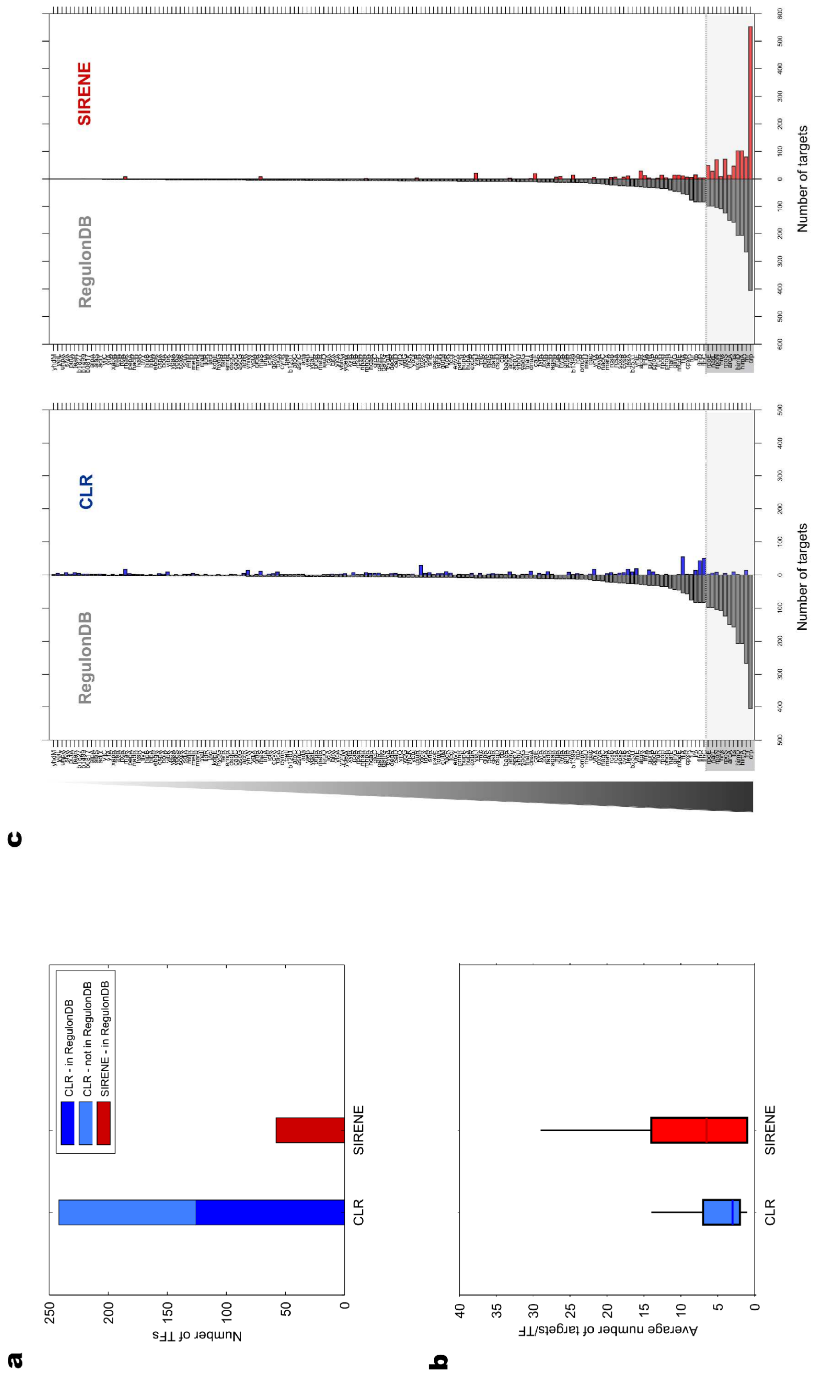


Figure 6- 4 Caption on next page.

Figure 6- 4 Complementarity in the type of interactions inferred by supervised versus unsupervised network inference methods. SIRENE (indicated by red bars) and CLR (indicated by blue bars), as representatives of respectively supervised and unsupervised network inference methods, were applied to the same *E. coli* compendium [34]. For both methods the top 1422 interactions were considered. (a) Panel a displays for each method, the number of transcription factors for which interactions could be inferred. (b) Panel b displays the average number of targets per transcription factor inferred by either method. (c) Panel c displays on the y-axis all regulators reported in RegulonDB [94], ranked according to their number of documented targets. The left hand side grey distribution displays the number of targets documented in RegulonDB per TF. The regulators for which most targets have been described so far correspond to global regulators and sigma factors (indicated by a shaded box). For either method, the number of inferred interactions per regulator (blue – CLR, red – SIRENE) is indicated at the right hand side of the plots. Panel a and c illustrate that supervised methods are biased towards predicting targets for regulators with a sufficiently high number of previously known targets (in *E. coli* corresponding to global regulators and sigma factors). Panel b shows that by exploiting known information, supervised methods are more comprehensive in predicting targets for a specific regulator than unsupervised methods.

In several studies ChIP-chip derived interactions have also been used as an alternative standard to benchmark algorithms, but as any high-throughput data source, they contain many false positive (or non-functional) and negative interactions. This explains the low performances often observed in benchmark studies with ChIP-chip data (Figure 6- 5).

Obtaining insight into the behavior of the algorithm requires a more objective validation strategy which uses perfect standards, made *in silico* by simulating data that mimick real data [208;209]. Simulated data are very useful in unveiling the qualitative properties of the algorithm under all kinds of test conditions that can never be obtained with real experimental data (*e.g.* noise robustness, sensitivity of the parameter settings, optimality of the proposed solution)[210]. Their drawback is that they can never grasp the full biological complexity of real data (such as the exact properties of the experimental noise or the multilayered aspect of gene regulation [211]). To further bridge the gap between *in* *silico* and real data, the use of synthetic gene networks has been proposed [212]. This is an engineered circuit with a well–defined network topology and interaction structure. The dynamic behavior of such circuit is fully characterized using real measurements and the resulting model is used to simulate data based on which inference methods can be tested.

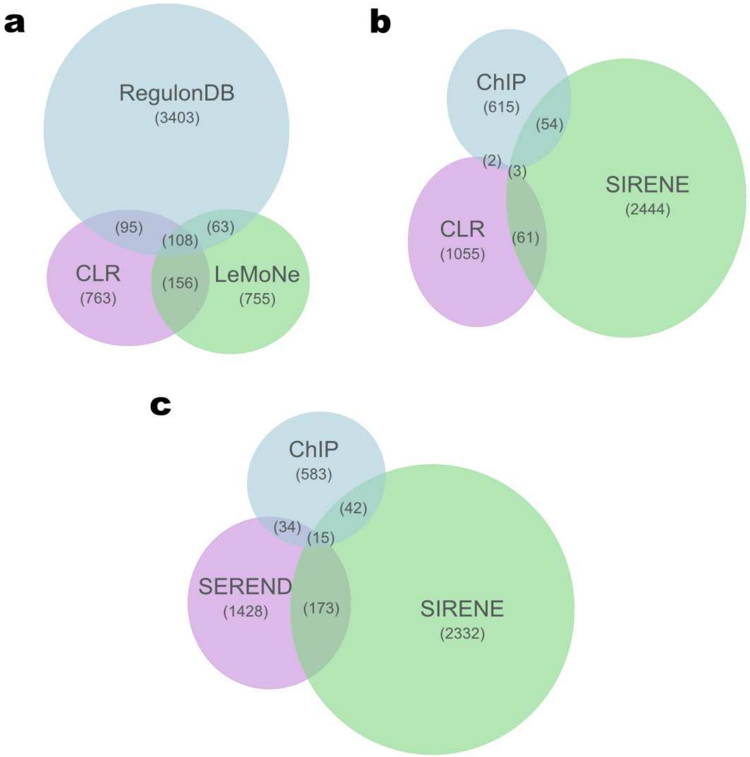


Figure 6- 5 Caption on next page.

Figure 6- 5 Illustration of the low overlap in predictions made by different network inference methods relying on different strategies. a) Mutual comparison between the results of the module-based approach Stochastic LeMoNe and the direct method CLR (both methods are non-integrative and unsupervised), (b) mutual comparison between the results of the unsupervised method CLR and the supervised method SIRENE (both methods are non-integrative and direct) and (c) mutual comparison between the results of the non-integrative method SIRENE and the integrative method SEREND which combines expression with motif data (both methods are supervised and direct). All methods were run on the same *E. coli* gene expression compendium [34]. In (a) the inferred interactions were compared to the known network in RegulonDB [94]. Since the supervised methods used in (b) and (c) use the information in RegulonDB to make their predictions, we used available ChIP-chip data for several *E. coli* regulators as an external validation standard [85;187;188]. For each pair of methods that is being compared, the proportion of the number of shared predictions on the total number of predictions ranges from 5.7 % to at most 24 %. The relative overlap with RegulonDB as external standard (number of interactions in common with the external standard/total number of predicted interactions) ranges from 15-18% and with ChIP data from 2-3% (with a very low performance of CLR on ChIP data (<1%)).

Benchmark studies are extremely useful to guide both users and developers. However, relying on a benchmark study to find out which algorithm is 'the best' is tricky as the choice of an appropriate inference tool depends on the posed research question. Fair benchmark studies should not only describe in what respect an algorithm is the best, but also where it fails. The quality of a benchmark study also largely depends on the extent to which parameter tuning is performed to guarantee that each of the applied tools optimally performs in the used setting. In this regard, the DREAM (Dialogue on Reverse Engineering Assessments and Methods) initiative [211;213] offers a platform for the unbiased assessment of network inference methods. They organize a yearly competition in which developers can participate with their own method to infer networks from blinded datasets.

6.4 Exploiting complementarity: the ensemble solution

The overlap in inferred results between methods can be very low as is illustrated in Figure 6- 5. This, together with the observation that the results of each of the tested methods shows a similar degree of overlap with an external validation standard (RegulonDB [94] or ChIP-data), indicates that it is not the failure of one of the methods in inferring biologically relevant interactions, but rather the complementarity between the methods that explains the observed discrepancy in predicted interactions.

As probably no single best method exists and different methods highlight different interaction types, aggregating the outcome of complementary methods offers a way to improve upon the breadth and the accuracy of predictions. This idea of combining the outcome of different methods has already been suggested in different contexts [61] and recently a 'reverse-engineering-by-consensus' approach has been advocated [213;214] spurred by the outcomes of the DREAM2 and DREAM3 conferences. There was shown that an ensemble of the predictions made by the best performing methods of the DREAM contest approximated much better the true interaction network than the predictions made by each method separately. For the construction of an ensemble solution that reflects an overall statistical confidence in each of the predicted interactions, inference methods are required that provide an explicit ranking of the predicted interactions according to the scoring scheme they use (such as *e.g.* Stochastic LeMoNe, CLR, DISTILLER, SEREND and SIRENE). These individual rankings can then be combined into a ranked ensemble solution that assigns a higher confidence to interactions that are repeatedly retrieved by the different methods.

As has been illustrated in the previous chapter, besides for combining the outcome of different methods, an ensemble solution can also be used to integrate different results of a single method. Because of the large search space, finding the most optimal solution to a network inference problem is non-trivial and optimization algorithms often result in suboptimal solutions that all approximate the true global optimal solution but differ slightly from each other. For methods that can capture different possible solutions, a consensus solution from interactions that are repeatedly inferred from the data [32;63] allows increasing the accuracy of the predicted interactions by better approximating the global solution.

At this stage only tentative steps have been undertaken to improve upon TRN reconstruction through ensemble methods. Much more work will be needed to assess whether and, if so, which ensemble solutions will succeed in simultaneously increasing precision and recall of the predicted interactions.

6.5 Conclusions

State-of-the-art inference tools rely on a unique combination of strategies to solve the inference problem. Because each strategy implies different assumptions, they all have different strengths and limitations and highlight complementary aspects of the network. Categorizing the tools according to their strategies, allows users to gain insights in the settings under which they can most optimally be applied. Which tool is more appropriate for a certain researcher depends on the available data and the research purpose.

The nature of the expression data generally determines whether a direct versus a module-based inference will be more appropriate. Whenever the set of available expression data gets larger and/or heterogeneous in the assessed conditions, module-based inference methods are to be preferred over direct inference methods. For less-studied microorganisms with only expression data available, expression-based network inference methods are ideal to make a first draft reconstruction. Integrating with expression data, additional high-throughput data on TF-target interactions will generally allow for a more accurate (less false positive interactions) and more complete picture of the TRN (such as the prediction of combinatorial control), but might become restrictive in inferring interactions only for those TFs for which the additional information is available. This is a disadvantage if one wants to derive global network properties. When interested in expanding knowledge on a particular part of the regulatory network rather than gaining a complete network view, query-based methods are to be chosen over global ones. When a reconstructed network is to be used as a starting point for further biological hypothesis generation, methods that provide an explicit ranking of the inferred interactions are advantageous to prioritize candidates for further experimental work. Moreover, in such case researchers benefit from using an integrative or supervised approach that exploits properties of existing interactions to infer highly reliable novel ones. However, the more the method biases towards previous knowledge, the more it will be blind towards novelty. To take full advantage of the complementarity between the different methods a 'reverse-engineering-by-consensus' approach that combines the knowledge gained from multiple inference approaches or from multiple outcomes from a single computational approach seems the ideal option [213;215].

Table 6- 1 Overview of representative algorithms, successfully applied to microbial datasets.

Name methods	**Short description of the methods**
CLR [34] (Context Likelihood of Relatedness)	CLR is an unsupervised, direct, expression-based network inference method that reconstructs an interaction between a TF and a target gene based on their consistency in expression behavior, as assessed by mutual information. The statistical significance of each inferred interaction is also evaluated.
SIRENE [33] (Supervised Inference of Regulatory Networks)	SIRENE is a supervised, expression-based, direct network inference method. The method splits the problem of network inference into multiple binary classification subproblems, one for each TF. For each particular TF an SVM-based classifier is trained based on the mutual similarities in expression profiles of respectively known target and non-target genes: genes

	known to be regulated by a particular TF are supposed to share the same coexpression behavior, while non-targets do not. This TF-specific classifier is then used to predict whether genes, other than the ones used for training are regulated by the respective TF, resulting in a ranked list of potential target genes.
Stochastic LeMoNe [32]	Stochastic LeMoNe is an unsupervised, module-based method that infers the TRN from expression data. It first uses a fuzzy two-way clustering approach to assign genes and conditions to modules and subsequently assigns a transcriptional program to these pregrouped gene sets. Each module contains the genes of which the expression profile best fits the same multivariate normal distribution that partitions the expression values of the conditions in the cluster in sets of under– or overexpression. The transcriptional program assigned to each module consists of the set of regulators for which the expression profiles best explain all or part of the condition partitions in the module.
Inferelator [162] (+cMonkey [183])	Inferelator is an inference method that can assign a transcriptional program to either individual genes or predefined modules of coexpressed genes. In the original paper these modules are obtained by the integrative module inference method cMonkey based on a multiparametric logistic regression that searches for tightly coexpressed modules, enriched for genes that make up highly connected subgraphs in metabolic and functional association networks and/or that contain statistically overrepresented *de novo* detected motifs. Inferelator itself uses standard regression with model shrinkage to build a parsimonious, predictive

	model for the modules' or gene's expression behavior using changes in environmental influences and TF expression levels as predictors. The design matrix can capture binary interactions (AND, OR or XOR) between TFs.
SEREND [31] (Semi-supervised Regulatory Network Discoverer) **De Hoon *et al.* [201]**	SEREND and De Hoon are (semi)-supervised, integrative network inference methods. Per regulator a training set of known targets (positive examples) and non-targets (negative examples) is used to determine the parameters of two separate logistic regression functions that map respectively the expression values and motif scores for the genes in the training set to their corresponding predictor variables which determine the genes' class memberships (being activated, repressed or not regulated by the TF). The basic assumption again is that targets of the same TF should share the same motif and the same expression profile. Motif and expression data are treated separately to guarantee proper balancing of the unequal number of features in either data set. A metaclassifier, also based on logistic regression, is used to combine the outcome of both the separate expression- and motif-based classifiers. Subsequently the complete classifier is used to predict the probability that genes, other than those in the trainingset belong to the same regulon.
DISTILLER [8] (Data Integration System To Identify Links in Expression Regulation)	DISTILLER is an integrative module-based network inference method. It combines expression data with interaction data (motif or ChIP-chip data) to search for coregulated modules. It uses an unsupervised strategy based on itemset mining to exhaustively

	enumerate all gene sets that are coexpressed under a subset of conditions and that share the same motifs. To identify from this exhaustive list the most significant set of non-redundant modules a probabilistic filtering step is used.
Hidden component analysis [38;182]	The method of Sabatti *et al.* [182] is a hidden component model, related to the original Network Component Analysis (NCA) [216;217] strategy that uses a linear model to decompose the measured expression data [E] in a product of a sparse connectivity matrix [A], containing the interactions between all TFs and their targets and [P], representing the hidden condition-dependent activity levels of the TFs [216]. Methods differ in the way they put constraints to achieve identifiability of [A] and [P]. Liao *et al.* [216]constrain [A] using the known network, while Sabbati *et al.* [182] use motif information as prior in a Bayesian framework to guide the reconstruction of the unobserved TF activity levels and their interactions. As these methods exploit known information to constrain their search space, they can be considered direct, integrative, unsupervised network inference methods.
COALESCE [184] (Combinatorial Algorithm for Expression and Sequence-based Cluster Extraction)	An integrative, non-supervised module inference procedure that uses a Bayesian framework to integrate sequence and expression data. *De novo* motif detection occurs concurrently with the gene and condition to bicluster assignment. Motifs, represented by probabilistic suffix trees, are assigned to a developing bicluster if their occurrence in the module is sufficiently enriched compared to their presence in the genomic background. Additional information on sequence conservation or nucleosome placement can

	be used as prior information to guide the motif and module inference.
Time-lagged correlation analysis [176;177]	Methods that explicitly use time series gene expression data to infer causal relationships. They generally consist of two different steps [176;177]. In the first step, genes with similar expression profiles across multiple time points (Pearson correlation) are grouped in a module or cluster. In a second step, causal effectors such as regulators, the modules that contain the regulators [176] or environmental inputs [177] are related to modules using time-lagged correlation, a measure that is related to the Pearson correlation coefficient, but that takes into account shifts in time between the expression of the causal effector and the target module.
Gat-Viks *et al.* [198]	The method proposed by Gat-Viks *et al.* [198] is a query-based, expression-based inference method. Qualitative knowledge on a certain pathway of interest is formalized as a Bayesian network, where the nodes represent different molecular entities (genes, proteins, metabolites) and the edges the interactions between them with their corresponding connection logics. Such a probabilistic formulation of the network allows including uncertainty on the given model. In a first model refinement step, possible model improvements (changes in topology and interaction logics) are evaluated. Refinements that result in a model that better predicts the observed expression values are withheld. In a second expansion step, transcription factor activities are predicted from the network model and a likelihood sore is used to assign additional target

	genes for which the expression can be predicted by the transcription factor activity profile. Hence the method identifies sets of genes regulated by the same set of regulators and according to a common logic (activated, repressed, feedback).
GPS [196] (Gene Promoter Scan)	Is a query-based, integrative network inference method. The method starts from a set of genes regulated by a common TF. Each gene is represented by a list of features consisting of the gene's expression profile and a detailed description of its promoter elements. Separating the set of query-genes into distinct clusters according to these features results in the input genes being grouped according to their specific regulation pattern. A fuzzy k-nearest neighbor classifier is used to extend the obtained clusters with new targets based on the similarity between the feature profile of the novel gene and that of a cluster.

Chapter 7

Conclusions and perspectives

7.1 Introduction

Recent experimental developments within the molecular biology field have enabled the unprecedented measurement of several cellular phenomena including gene expression, TF-gene binding and protein-protein interactions. Despite the fact that new data sources are becoming increasingly available, the usage of microarrays remains well-established and the number of publicly available data for model organisms runs in the hundreds of arrays. Therefore methods for module and regulatory network inference that attempt to elucidate the underlying wiring of the transcriptional regulatory network from gene expression data remain of outstanding interest. However, whereas this inundation of data provided much hope of obtaining a system's level understanding of the cell, this optimism was also tempered by the computational challenges the analysis of high-dimensional data sets, such as gene expression data, poses. Searching through high-dimensional spaces is computationally intractable, *e.g.* it is impossible to enumerate all possible cluster or bicluster configurations or to evaluate all possible transcriptional program-target combinations. Therefore, module inference and regulatory network inference must be approximated and as a consequence multiple solutions exist that each explain the data equally well. Ensemble-based strategies exploit this feature of high-dimensional inference problems by reinforcing solutions repeatedly retrieved from the same data set. Although, ensemble-based strategies have already been applied to a range of biological contexts, with a major focus on motif detection [46-49] and protein fold prediction [50;51], there only exist few cases of where they have been applied to module and network inference problems [63;213]. Therefore, in this thesis we explored such ensemble-based strategies to improve both upon existing module and regulatory network inference methods.

Key innovations introduced in this thesis include:

- Development of a generic ensemble strategy for query-based biclustering to interrogate gene expression compendia for heterogeneous gene lists, such as experimental lists. Its biological usefulness was shown on an *E. coli* and *S.* Typhimurium case study.
- Introduction of Stochastic LeMoNe, an ensemble method for TRN inference.
- Illustration of how conceptually different methods for inference of the TRN infer complementary parts of a ground truth TRN.

In this chapter we first summarize the different ensemble-based strategies that were introduced in this thesis and refer to alternative approaches that have been developed to tackle similar problems. Secondly, we give a critical note on expression-centered studies. Lastly, we give perspectives on the future of network inference in light of novel biological insights and data sources.

7.2 Discussion and conclusions

7.2.1 An ensemble-based strategy to extend the scope of module detection

First, in Chapter 3 we introduced a 'wrapper' for query-based biclustering to extend upon its practical use. In particular we developed an ensemble approach that rendered query-based biclustering applicable to query-sets that are heterogeneous in their expression. As a consequence, the scope of these methods is no longer limited to query-sets that consist of sets of genes that are functionally related, but they can also be used on query-sets that participate in different functional groupings or that contain functional outliers. Using this extension, query-based biclustering methods can be applied to interrogate gene expression compendia with experimentally-derived query-lists. As such researchers can view the outcome of their own experiments in light of all publicly available expression data.

In Chapter 3 and 4 we illustrated how this extension can be practically applied. In Chapter 3, using this ensemble strategy, we identified experimental inconsistencies in a ChIP-chip experiment by verifying its output with a gene expression compendium. In addition, we could confirm several novel targets for the regulator tested in the ChIP-chip experiment. In Chapter 4 we applied the ensemble query-based biclustering framework to investigate the role of a specific set of experimentally derived genes in *Salmonella* Typhimurium biofilm formation.

As ensemble-based strategies have up until now mainly been applied to transcend the accuracy of the predictions made by one or multiple methods, in Chapter 3 we applied it in a rather novel way: to remove redundancy amongst the outcomes. In particular, to render query-based biclustering methods applicable to query-lists that are heterogeneous in their expression we introduced a 'split-and-merge' strategy. In the 'split-step' the problem of deriving the biclusters that contain the genes in the query-list is split into different subproblems in which a query-based biclustering solution is obtained for each gene in the query-list, hence circumventing the problem that different genes from the query-list might belong to separate biclusters. This results in an ensemble of often at least partially overlapping biclusters, each containing at least one of the query-genes. In the 'merge-step' a unique solution is derived from the bicluster ensemble by removing redundancy amongst the output. This 'merge-step' is analogous to the consensus strategy often applied to cluster ensembles (*e.g.* [53;54]) in that it relies on the construction of a new similarity matrix (the consensus matrix) which summarizes co-clustering across the ensemble of clustering solutions. However, as here we do not aim to stress gene pairs that repeatedly belong to the same bicluster, but aim to stress the distinct biclusters they belong to, we introduced in

Chapter 3 an alternative normalization function for this matrix to achieve this goal.

Whereas different approaches have already been developed to query protein-protein interaction (PPI) networks with experimental results [218-222] (see 7.3.3), to our knowledge this is the first attempt to develop a method which interrogates genome-wide expression compendia for experimentally derived gene lists. Whereas methods developed to interrogate PPI networks are readily applicable to coexpression networks (*i.e.* networks which link genes for which coexpression exceeds a certain threshold), they ignore the contextual information present in gene expression compendia (*i.e.* genes are often only coexpressed under a subset of conditions). Our approach in contrast relies on a biclustering step and therefore inherently accounts for the condition-dependency of gene expression.

7.2.2 An ensemble-based strategy to improve upon network inference

In Chapter 5 we introduced Stochastic LeMoNe (Learning Module Networks). Whereas similar in spirit to the module networks method [30], LeMoNe pursues a stochastic instead of a deterministic approach to construct modules and to learn transcriptional programs for these modules. As is illustrated in Chapter 5 and further elaborated in Joshi et *al.* [32], introducing such a stochastic optimization step in combination with ensemble averaging can improve both upon module detection and inference of the transcriptional program.

Module inference is often used as a first step in regulatory network inference (Chapter 6) and therefore the quality of the produced modules is crucial in order to obtain reliable TF-target gene predictions. In particular, modules should not only consist of genes that are coexpressed, but these genes should also be regulated by a common set of TFs: the genes within a module should be co-regulated. In Chapter 5, we illustrated that through the ensemble clustering strategy incorporated in LeMoNe spurious gene-to-module assignments can be filtered out (*i.e.*

genes that are only sporadically assigned to a module). As such modules could be obtained that consist of genes attributed to one specific regulon in stead of genes belonging to multiple regulons. Similarly, in other applications such as the clustering of tumor samples [56;57] and clustering of genes in expression data [53], the power of ensemble methods to derive a more robust and more accurate partitioning from the data has been demonstrated [53;56]. Alternative approaches to improve the quality of modules of coexpressed genes with respect to network inference have been presented by [8;25;183;184;223]. All these approaches complement gene expression data with alternative data sources, such as sequence data and ChIP-chip data to refine module construction. As such modules of genes can be constructed that are not only coexpressed but that also share motifs (sequence data) or TF-binding sites (ChIP-chip data) for the same set of regulators. Especially methods which combine the detection of coexpressed modules with *de novo* motif detection are of interest [25;183;184], as they require little prior knowledge on the presence of TF-binding sites and therefore can be ready applied to organisms for which such data is only scarcely available.

In a second step LeMoNe assigns transcriptional programs to the obtained modules. Also here LeMoNe incorporates a stochastic instead of a deterministic approach, resulting in multiple equally likely transcriptional programs being proposed for the same module. Here, to derive a consensus solution for TF-module predictions a different consensus strategy is used than for module detection. In particular majority voting (*i.e.* the regulator that is assigned to the module most often is most statistically significant) is used to attribute a significance score to each regulator-to-module assignment. In Chapter 5 we illustrated that biologically correct interactions are prioritized according to this score, as compared to a ground truth reference network. As predictions from network inference methods are often backed-up by experimental validation, a method that produces a reliable ranking of the predictions is desirable since it guides the researcher to the most probable predictions amongst the abundance of predictions the different methods produce. In addition, such a ranking assists in comparing the performance of different network inference methods as this is usually assessed through precision-recall curves, which compare method performance for different significance score thresholds. While the use of majority voting in combination with an ensemble of predictions resulting from a stochastic optimization provides an objective and statistically motivated way to prioritize predictions also other approaches have been introduced that result in the predictions of a network inference method being ranked. In contrast to probabilistic methods, the outcome of deterministic methods such as itemset mining approaches [224] or the relevance networks procedures [172], which serve as a basis for respectively DISTILLER and CLR, do not assign a significance score to the predictions. Therefore, in CLR the relevance networks procedure is extended with an adaptive background correction step to filter out spurious interactions, causing the interactions to be ranked according to a significance score. In DISTILLER, the significance of the resulting modules of the itemset mining search strategy is scored by estimating the probability that the same modules and transcriptional programs would be selected by chance. In diverse studies both the probabilistic [66] as well as the deterministic approaches [8;34] discussed here were shown to successfully prioritize known interactions, supporting the practical usefulness of the different scoring-approaches. As of yet different scoring schemes have not been assessed independently from the algorithm itself and therefore the question which scoring scheme gives the best output remains open.

7.2.3 Towards a mixed ensemble for network inference

In Chapter 5 and 6 we illustrated that different network inference methods highlight different parts of the ground-truth TRN and are complementary in the interactions they infer. This can be explained by the fact that different network inference methods pose different biological constraints on the predictions that can be inferred. Hence constructing a *mixed ensemble* of predictions obtained by different NI methods provides an opportunity to not only improve upon the accuracy, but also to extend the scope of what can be found.

Marbach *et al.* [213] were the first to propose such a mixed ensemble strategy for network inference: they combined the ranked output of different network inference methods through majority voting. They showed on synthetic data that in this manner a larger area under the precision recall curve could be obtained than for each of the individual methods. However, such a strategy is expected to mainly improve upon the precision of network inference by highlighting the interactions on which different methods agree and hence fails to account for the complementary aspect of the different approaches. We believe that mixed network inference ensembles might benefit from alternative consensus construction schemes that not only prioritize predictions retrieved by multiple methods (as is the case for majority voting) but that also prioritize high-scoring predictions unique to the different methods. However, to our knowledge so far no other consensus construction schemes have been proposed in this context. A possible alternative to the method of Marbach *et al.* [213] is to weigh each prediction in the consensus construction according to the confidence level for that prediction. Within the domain of gene function prediction Hibbs *et al.* [61] for instance took advantage of prior knowledge to appropriately weigh the ensemble of predictions made by different algorithms. However, more objective, statistically motivated scoring schemes are more desirable as to not bias the output towards what is known (*e.g,* [225]). Reliability scores as provided by most methods seem to correlate well with known biology (see for instance Chapter 5), therefore ideally a consensus scheme should exploit these scores to obtain a consensus solution that balances accuracy and sensitivity.

Besides a relevant consensus aggregation scheme, a procedure is needed to decide upon which network inference methods should be included when exploiting mixed NI ensembles. Indeed, as suggested in Chapter 1, the individual interactions need to be both accurate as diverse. In particular, here we seek for methods that are diverse in the inferred interactions and that also derive these complementary interactions accurately. As was shown in Chapter 6, this seems to be the case for several well-established network inference methods as they all show similar overlap with the ground truth network, but infer complementary parts from this network (Figure 6- 5). Each of these methods was conceptually very different from the other ones. Conversely, methods that are mainly based on similar theoretical frameworks, such as CLR and ARACNE [34;161], often do not contribute to an increased diversity in the predictions made (Figure 7- 1).

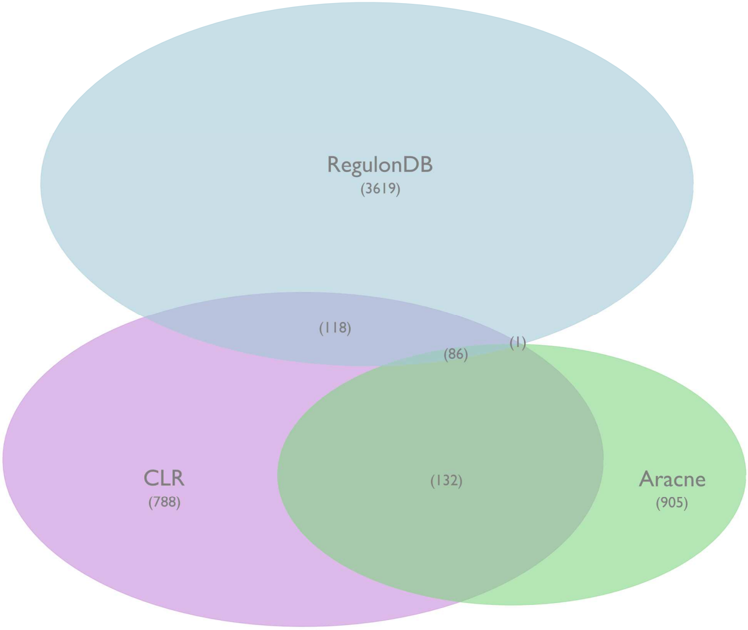


Figure 7- 1 Comparison of high-scoring interactions inferred by two similar network inference methods. CLR is built on the same principles as ARACNE but incorporates a different scoring approach. This figure shows that CLR mainly improves upon ARACNE by inferring more interactions from the ground truth network (RegulonDB). In contrast, ARACNE fails to identify predictions from RegulonDB not identified by CLR.

It needs to be further explored whether such an increased sensitivity and accuracy can also not be obtained by bootstrapping single methods. It seems intuitively plausible that a mixed ensemble leads to a larger diversity in predictions and hence can surpass single methods both in accuracy and sensitivity [46]. Recently, however, several approaches have been developed that intentionally constrain solutions obtained using a single method as to maximize the diversity amongst the outcomes [49;226].

Alternatively, first attempts have been made to combine the strengths of different network inference methods into one algorithmic framework. In [227], the authors combined the scalability of the CLR method with the power of Inferelator to infer causal relationships between TFs and potential target genes from time series data in a new computational pipeline. This mixed method was shown to outperform both CLR as Inferelator on synthetic data in the DREAM3 contest.

Finally, we note that this observation of complementarity between different computational tools aimed towards solving the same biological problem is not unique to network inference. Similar observations and different attempts towards mixed ensembles have already been made with respect to gene function prediction [61], motif detection [46;47], PPI-network based prediction of disease genes [64], protein fold prediction [50;51], etc. Hence the existence of mixed ensembles seems to become a recurring theme within the field of systems biology.

7.2.4 The limitations of expression-centered studies

The methods presented in this thesis are all expression-centered: they infer modules and regulatory networks from gene expression data as a sole data source. As gene expression data only highlights one aspect of the TRNs, *i.e.* the joint coexpression of target genes due to common regulation, such an approach comes with certain drawbacks. First, as was discussed in Chapter 2, when constructing modules of coexpressed genes an important problem concerns defining a threshold on gene coexpression. This threshold generally depends on the biological process one is interested in: *e.g.* operon-level, regulon-level or a level triggering multiple regulons responding to complex environmental changes (*e.g.* oxygen concentrations). However, which threshold on coexpression should be chosen for studying a certain biological process at a particular level of biological detail is not known in advance. This problem is in particular pressing for module-based regulatory network inference, as here obtained modules need to correspond to regulons. However, as information on joint regulation of the genes through the same set of TFs is not directly present in gene expression data, module construction on expression data alone can not guarantee that genes within the module are all co-regulated at the transcriptional level. Therefore, different methods extend the expression data with alternative data, such as sequence or motif information, to identify sets of genes that are not only coexpressed but also contain motifs for the same TFs [8;25;183;184;223].

Second, with respect to expression-based predictions of TF-target interactions we also discussed the problem of correlation vs. causation in Chapter 6: expression-based network inference methods generally predict TF-target gene interactions based on an (anti-)correlated expression pattern, however fails to explain gene expression causally. Indeed, this assumption of correlation between TF-target genes supposes that the TFs themselves are also transcriptionally regulated. TF activity is however mainly regulated at the post-transcriptional level [42]. Consequently, we observed that methods such as LeMoNe and CLR mainly predict correct interactions for auto-regulators (Chapter 5 and 6). This problem can be alleviated by relying on an integrative module detection step and assigning TFs to the modules that contain their associated motifs (*e.g.* [8]). However, as the number of TFs with well-characterized motifs is rather limited, especially for non-model organisms, these methods are restricted in scope. Alternatively, integrative module detection methods can be used that combine coexpression analysis with *de novo* motif detection [25;183;184], however as there is currently no biological motivated approach available to assign regulators to novel motifs, also expression-based methods need to be used here to predict TFs associated with the motifs [162].

These examples illustrate that integrating complementary data with existing gene expression data might reveal a more accurate and complete picture on the functioning of the TRN and this increased performance of integrative methods has also already been demonstrated in practice [68;189]. With the increasing amount of complementary data being produced data integration methods will be further challenged and issues with these methods, such as accounting for different systematic biases inherent to different experimental procedures and missing measurements [62;225] will need to be tackled. However, as gene expression data is currently still the most abundant source of information, application of data-integration methods mainly remains restricted to well-characterized systems. Indeed, the wealth of publicly available gene expression data keeps inspiring different groups to develop computational methods to explore these data for interesting biological characteristics (*e.g.* [228-231]). Especially, the dynamic aspect of expression data, *i.e.* gene expression is described under a multitude of different conditions, seems to appeal to researchers as other datasources such as PPI and ChIP-chip data are often restricted to describing biological phenomena in single conditions.

Lastly, in this work we focused on inferring the TRN of prokaryotes from gene expression data. The methods introduced here could also be potentially extended for application to higher eukaryotes. The higher complexity of these organisms (*e.g.* larger number of genes and more complex regulatory mechanisms) requires however that these methods are applicable to data sets of even higher dimensionality. In addition, due to increased complexity of gene regulation mechanisms in eukaryotic systems, network inference for these organisms will even more benefit from integrative approaches that study transcriptional regulation from different biological angles. On the other hand, recent reports suggest that within prokaryotes growth rate differences might exert global effects on gene expression [232;233]. Therefore, the possibility exists that using gene expression data to study prokaryotic transcription regulation, results might be confounded due to growth rate differences in the different gene expression experiments. As current approaches that have studied the effect of growth rate on prokaryotic gene expression have mainly been limited to theoretical and small-scale models [232;233] the more global effect of growth rate on gene expression needs to be further studied. As eukaryotic cells, however, generally grow more slowly than prokaryotic cells, here the growth rate is expected to have a less significant effect on gene expression.

7.3 Perspectives

The past decade has resulted in an unprecedented accumulation of biological data, not only on the genomic and transcriptomic level, but also on the proteomic and metabolomic level. In addition, novel high-throughput technologies have revealed a far more complex organization of the cell than was originally anticipated. These novel data types together with novel biological insights will further leverage the importance of network inference.

7.3.1 The future of network inference: accounting for regulatory complexity

The advent of novel technologies, such as tiling arrays and more recently the deep sequencing techniques [234;235], have revealed an unprecedented complexity of prokaryotic transcriptomes. Besides transcription factors, other regulatory elements such as non-coding RNAs (ncRNAs) [236] and riboswitches [237] seem to influence gene transcription levels. In addition, these data revealed that these novel regulatory elements are not cellular peculiarities, but are in stead omnipresent and involved in different cellular functions [238]. Riboswitches, for instance, are thought to regulate up to 2% of all *B. subtilis* genes [238]. sRNAs on the other hand have been shown to regulate important biological processes, such as virulence, stress response and quorum sensing [239-241]. In addition to these novel regulatory elements, alternative operon structures with multiple intra-operonic transcription sites seem to be abound within prokaryotes [36;37], allowing for increased flexibility in gene transcription within changing environmental conditions.

Although most inference methods can readily be applied to these novel types of expression data as they are insensitive to the type of technology used to generate the data, they will have to be adapted to account for the more detailed level of information that results from these novel technologies. As for instance the condition-dependent abundance of these sRNAs and riboswitches can be measured with tiling arrays or deep sequencing technologies, methods such as LeMoNe can be used to connect these regulators to potential target genes. Recently, LeMoNe has been applied to human expression data in order to predict miRNA-target gene interactions involved in cancer [242;243]. A similar strategy could be applied to connect sRNAs and riboswitches to their cognate target genes for bacteria.

Recent developments within the research for eukaryotic transcription regulation have revealed a crucial role for DNA-structure on gene expression. In particular, nucleosome occupancy, histone variants and chromatin modification seem to exert major influences on gene transcription [244;245]. Similar information in prokaryotes is much scarcer and it is currently not known to what extent chromatine structure might influence gene expression. Several reports suggest, however, that different nucleoid-associated proteins exist within bacteria [246-250] that have long range effects on gene expression, extending their control to genes for which their promoter region is not directly bound by these proteins. Therefore, as condition dependent information gets available on the effects of these nucleoid proteins on gene transcription these must be integrated into novel predictive models that attempt to explain gene expression as a factor of different possible regulatory elements, including both regulators that exert direct effects (such as sRNAs and TFs) as nucleoid proteins that exert indirect effects on gene expression.

7.3.2 The future of network inference: accounting for additional layers of gene regulation

Currently, transcriptional regulation of gene expression, and by extension regulation of protein production, attracts major attention, primarily because of the existence of mature experimental methods for transcriptomics such as microarrays, tiling arrays and ChIP-technologies. However, it is well-established that gene regulation is manifested at different levels that include besides the transcriptional level also control mechanisms at the metabolic and the protein level. *Mycobacterium pneumoniae* for instance lacks the majority of TFs and sigma factors to regulate metabolic gene transcription. Yet this bacterium shows a remarkable expression plasticity in response to changing environmental conditions, which can not be explained entirely by actions of its TRN [251]. Therefore, only focusing on the transcriptional level ignores the shear complexity of gene regulation and additional layers of regulation such as the metabolic and the phophorylation network need to be accounted for in order to obtain a comprehensive understanding on the mechanisms that underlie gene regulation. Tentative steps into this direction have already been undertaken by incorporating gene expression information in metabolic flux modelling [252;253]. However, as was illustrated in a small-scale computational model introduced by Kotte *et al.* [254] the true extent of this interplay between metabolism and the TRN still remains poorly understood. Paired datasets not only measuring gene transcription but also containing metabolite concentration provide however a step into the right direction into understanding the complex interplay between the different network layers [255]. Whereas in prokaryotes computational efforts for accounting for multiple layers of gene regulation have mainly focused on metabolism, other approaches have been developed in eukaryotes that for instance focus on post-translational modifications of TF activity (*e.g.* [256;257]). The application of similar methods to prokaryotes is however hampered by a lack of available protein-protein interaction and protein-phosphorylation data.

7.3.3 The future of network inference: towards a query-based exploration of available data

Since biological systems are very complex we often have to compromise between the system's size that we consider and the level of detail at which we model. In Chapter 6 we therefore distinguished between query-based and global network inference methods. Global network inference methods attempt to model the regulatory network in its entirety from the available data whereas query-based network inference mostly focuses on a specific part of the network and tries to model that part in more detail than is computationally possible in a global setting. As current molecular biology research often starts from a particular biological intuition or is often inspired by a certain experimental approach, we believe that query-based network inference will gain importance (*e.g.* [196;198]).

In Chapter 3 and Chapter 4 we illustrated how query-based biclustering tools can add to the information derived from experimental data sources by a query-based interrogation of gene expression compendia for experimentally derived genelists. However, as not only gene expression data is accumulating but also physical interaction data such as PPI data and ChIP-chip data are piling up, querying these physical networks for experimental output is rapidly gaining interest. Indeed different approaches have already been developed to query PPI data sets for experimental outputs, such as RNAi hits [220;221] or differentially expressed genes [219;222]. These approaches have as a goal to either filter false positives from the experimental output [220;221] or to expand upon the mechanistic insights of biological systems [219;222]. Current methods, however, generally do not account for the incompleteness of the PPI networks (or other interaction networks) and the often low accuracy of the edges in such networks. Therefore, future efforts within this direction should include strategies to account for missing links, by for instance including functional relationships (*e.g* [219;221]) or relying on statitistical properties of the networks [258] and account for weighted edges representing the accuracy of a certain interaction (*e.g.* [259]).

7.3.4 The future of network inference: constructing the genotype-phenotype map

Although to date most inference studies have focused on understanding the condition-dependent behavior of a transcriptional network in one specific model strain, the success of deep-sequencing technologies has opened up a whole new application field of 'individualized expression-centered' network inference. Expression-centered inference uses the premise that most of the mutations or changes occurring in the regulatory network at a level other than the transcriptional one will eventually lead to an altered expression profile. This assumption allows considering the expression profiles of individual strains as specific phenotypes or traits [260-265]. Additional sequence-derived genomic information can then be used to explain individually observed variations in expression behavior (similar to the identification of eQTLs in higher eukaryotes). Inference methods that generate an explicit explanatory model for the observed expression profiles (*e.g.* Inferelator, Stochastic LeMoNe) can easily be extended to this purpose [144;266-268]. Linking adaptive changes of microbial genomes [269-271] to altered expression behavior will unveil fundamental insights in microbial evolution and will identify multifactorial changes underlying industrially relevant properties of naturally occurring bacterial or yeast strains [272]. Most of the links inferred by such an expression-centered approach will only provide an indirect link between the observed genomic or epigenetic alteration and the observed strain specific expression-profiles [273]. Future inference tools should focus on completing this hidden path between a genomic and an expression alteration by exploiting information on all levels of regulation *e.g.* the (post-) transcriptional, the signaling and metabolic level [225;267;274-279].

Individualized expression-centered inference studies will not only complete, but revolutionize our understanding of bacterial regulation.

7.3.5 The role of ensemble methods in the future of NI

Paradoxically, with the increasing amount of biological data that has been generated in the past decade, confusion on the shear complexity and the inner workings of biological systems only seems to have increased. Therefore, it has become clear that there exists no single data source or no single method that can capture the full complexity of a biological system. Consequently, we believe that obtaining a systems level understanding of biological systems will increasingly rely on methods that are able to integrate biological hypothesis generated for different data sources and inferred by different computational methods that each make a different assumption on the data. Ensemble methods have been specifically designed for this task: to intelligently integrate the information obtained by multiple experts. Challenges here lie in determining the optimal combination of predictions and in finding the consensus scheme that gives the most favorable results (see 7.2.3). Much can be learned with this respect from the machine learning community where ensemble methods have already been ingrained for more than a decade now [280].

Nevertheless, the future of network inference can not entirely rely on aggregating predictions of existing tools, but novel computational approaches need to continue being developed to tackle the problems network inference tools suffer from collectively. A recent survey of Marbach *et al.* [213] on network inference tools revealed, for instance, that current tools are affected by systematic prediction errors. In particular, the majority of existing tools fails to accurately infer gene combinatorial regulation. Whereas ensemble methods can improve the accuracy and the scope of what can be found by single methods, their innovative aspect is limited: they can not output predictions beyond those captured by the single methods.

Appendix A

Supplementary materials

A.1 Comparison of different ensemble schemes for query-based biclustering

Figure A- 1 Comparison of all possible combinations of consensus matrix transformations and graph clustering to obtain consensus biclusters. A. Compares the influence of using different consensus matrix transformation methods on the quality of the final bicluster solutions assessed by respectively their overlap with the original QDB-solutions ('overlap'), their preservation of redundant relationships amongst the query-genes ('redundancy'), their coverage for query-genes ('coverage'), their functional coherences as calculated by the clustering score function ('func enrich') and the modularity of the clustering ('modularity') (*x*-axis). Each row in the figure represents the comparison of the matrix transformations for one particular graph clustering method. B. Comparison of the effect of using different graph clustering methods to extract from the consensus matrix the final biclustering solutions. Same assessment criteria as in panel A were used. Here each row represents the comparison of different graph clustering method for the same matrix transformation. Figure is represented on the next page.

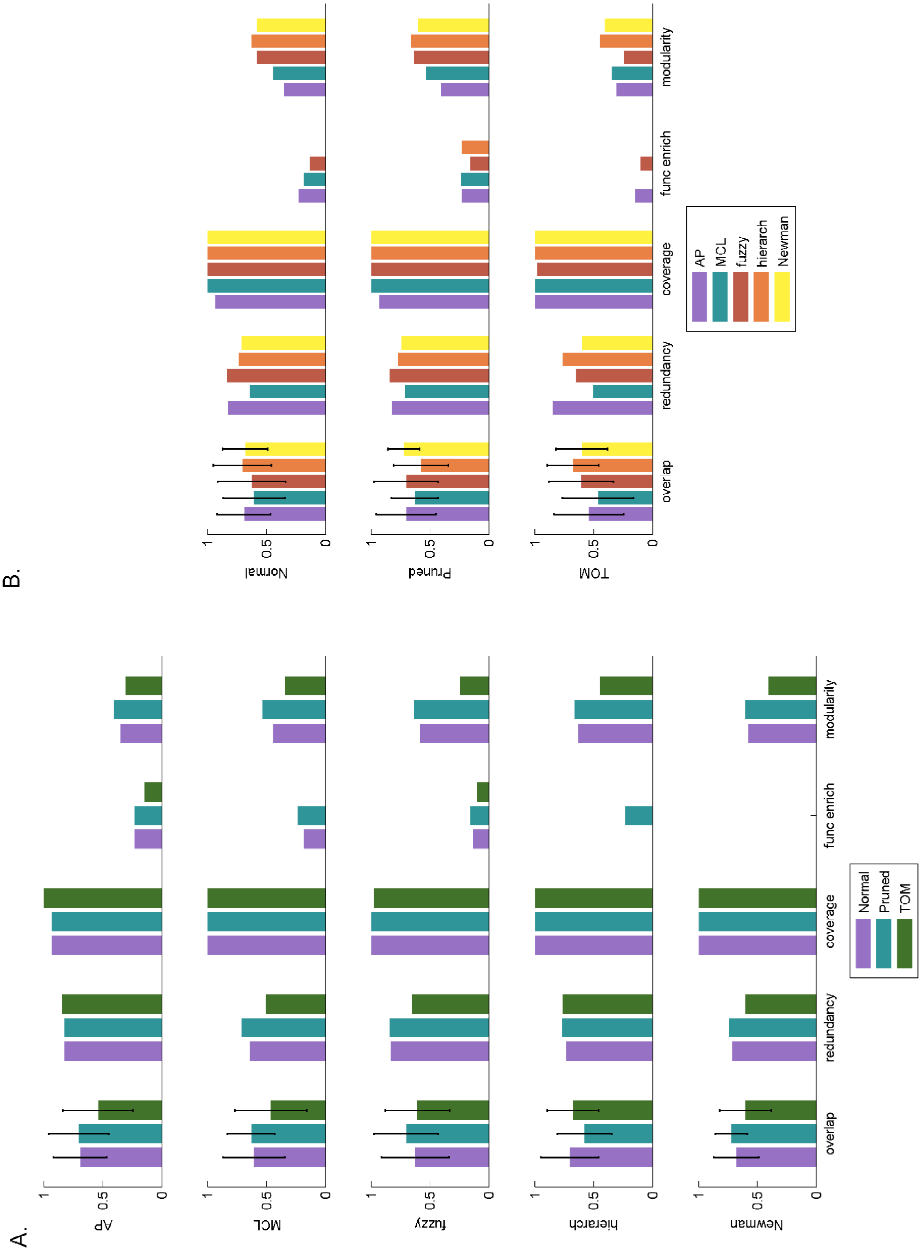


Figure A- 1 Caption on previous page.

A.2 Content consensus biclusters

Table A- 1 Overview bicluster content in number of ChIP-chip targets and FNR targets. This table summarizes the obtained consensus biclusters in terms of the number of transcription units in a consensus bicluster (second column), the enrichment of the list of ChIP-chip targets for these consensus biclusters (fifth column) and the coverage for previously known FNR targets (last column). We distinguished between ChIP-chip targets that are known FNR targets according to RegulonDB (column 4) and those that are not known to be regulated by FNR according to RegulonDB (column 5). To calculate the enrichment of the ChIP-chip list for the consensus biclusters we took the sum of both. Enrichment values were obtained by hypergeometric test ($p<0.05$, Bonferroni-corrected). Significant p-values are indicated in bold. To obtain for each consensus bicluster the coverage in FNR targets, the percentage of known FNR targets within the bicluster with respect to the total number of transcription units was calculated. For consensus biclusters with a coverage exceeding 33% the values are indicated in bold.

Bicluster	Number TU	ChIP-chip			FNR	
		Number TU		Enrichment	Number TU	% targets
		RegDB	Novel			
1	220	4	4	0.074	11	5
2	151	1	7	0.011	3	2
3	157	3	3	0.096	5	3
4	115	1	5	0.028	8	7
5	71	15	6	**<1e-15**	29	**41**
6	79	0	4	0.074	3	4
7	56	0	3	0.103	0	0
8	17	1	0	0.293	1	6
9	7	0	1	0.133	1	14
10	6	1	0	0.115	5	**83**
11	36	0	2	0.164	2	6
12	3	2	1	**7.9e-6**	2	**67**
13	1	1	0	0.02	1	**100**
14	1	1	0	0.02	1	**100**
15	2	0	1	0.4	0	0
16	12	1	1	0.023	4	**33**
17	7	0	1	0.133	0	0
Total	866	24	37		66	

A.3 Calculation of NMI as a redundancy measure

The main goal of the ensemble framework is to remove the redundancy among the QDB-solutions by assigning genes of highly redundant QDB-biclusters to the same consensus bicluster. As the query-genes form the prototype of each QDB-solution (*i.e.* the average bicluster expression profile is usually determined by the query-gene) we expect query-genes with highly redundant QDB-solutions to end up in the same consensus bicluster. Consequently we can assess how well a consensus bicluster represents the redundancy amongst the QDB-solutions by calculating the similarity in the groupings of the query-genes imposed by their redundancy in their QDB-solutions and the grouping of the query-genes imposed by the clustering of the consensus matrix. To obtain such a grouping of the query-genes according to their redundancy in the QDB-solutions we compute a query-gene by query-gene redundancy matrix (Figure 3- 2). Each element of this matrix contains the pairwise maximal overlap of the QDB-solutions of two query-genes as assessed by geometric coefficient. Next, groupings of query-genes with highly similar or redundant QDB-solutions are obtained from this matrix by clustering this matrix. Here, hierarchical clustering was used in combination with the modularity function to define a cut-off on the clustering tree. We used Normalized Mutual Information (NMI) [96] to assess the similarity in the groupings obtained from the query-gene by query-gene redundancy matrix and the grouping of the query-genes obtained from the consensus clustering. Normalized Mutual Information assesses the independency of both partitions. First, a confusion matrix is constructed in which rows represent the partitioning of the query-genes obtained from the query-gene by query-genes redundancy matrix and columns the partitioning of the query-genes according to the consensus clustering. Each matrix element N_{ij} represents the number of query-genes in common to cluster i from the redundancy matrix and cluster j from the consensus clustering. The number of communities derived from the redundancy matrix is referred to as c_{RED} and for the consensus matrix as *c*$_{CONS}$. A measure of similarity between both partitions can then be obtained by Normalized Mutual Information:

$$NMI = \frac{-2\sum_{i=1}^{C_{RED}}\sum_{j=1}^{C_{CONS}} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_{RED}} N_{i.}\log(\frac{N_{i.}}{N}) + \sum_{j=1}^{C_{CONS}} N_{.j}\log(\frac{N_{.j}}{N})}$$

A.4 Overview of *S.* Typhimurium biofilm specific gene set

Table A- 2 Overview of the list of 70 genes experimentally determined to be specifically involved in biofilm formation. The second and the third column denote if biclusters could be retrieved for these genes for the multicellular and planktonic compendium (indicated with 'x').

Gene locustag	Gene name	In multicellular	In planktonic
STM0191	fhuA	x	x
STM0473	hha	x	x
STM0557	STM0557	x	x
STM0586	fes	x	x
STM0653	ybeL	x	x
STM0731	STM0731	x	x
STM0978	aroA	x	x
STM1091	sopB	x	x
STM1140	csgF	x	x
STM1255	STM1255	x	x
STM1402	sseE	x	x
STM1583	STM1583	x	x
STM1594	srfB	x	x
STM1765	narK	x	x
STM1960	fliD	x	x
STM2065	phsA	x	x
STM2082	rfbP	x	x
STM2086	rfbU	x	x
STM2095	rfbA	x	x

STM2096	rfbD	x	x
STM2777	iroN	x	x
STM2782	mig-14	x	x
STM2805	nrdH	x	x
STM2861	sitA	x	x
STM2924	rpoS	x	x
STM3474	nirB	x	x
STM4060	cpxP	x	x
STM4423	STM4423	x	x
STM0693	fur	x	
STM1851	STM1851	x	
STM2083	rfbK	x	
STM2084	rfbM	x	
STM2085	rfbN	x	
STM2303	STM2303	x	
STM2840	STM2840	x	
STM3071	STM3071	x	
STM3502	ompR	x	
STM3713	rfaL	x	
STM3717	rfaJ	x	
STM3718	rfaI	x	
STM3721	rfaP	x	
PSLT019	pefB		x
STM0305	STM0305		x
STM0551	STM0551		x

STM0854	STM0854	x
STM0877	potF	x
STM1139	csgG	x
STM1142	csgD	x
STM1143	csgB	x
STM1144	csgA	x
STM1404	sseF	x
STM2089	rfbJ	x
STM2093	rfbI	x
STM2094	rfbC	x
STM2761	STM2761	x
STM3133	STM3133	x
STM3756	rmbA	x
STM3783	STM3783	x
STM0275	hisH	
STM0319	crl	
STM1358	aroD	
STM1432	ydhO	
STM2184	yeaH	
STM2908	STM2908	
STM2950	STM2950	
STM3342	STM2950	
STM3388	STM3388	
STM3714	rfaK	
STM3722	rfaG	

STM3981	STM3981

References

1. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318-356
2. Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269:496-512
3. Lander ES (1999) Array of hope. *Nat Genet* 21:3-4
4. Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-537
5. Wilson JW, Ott CM, Honer zu BK et al (2007) Space flight alters bacterial gene expression and virulence and reveals a role for global regulator Hfq. *Proc Natl Acad Sci U S A* 104:16299-16304
6. Natarajan K, Meyer MR, Jackson BM et al (2001) Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol* 21:4347-4368
7. Luscombe NM, Babu MM, Yu H et al (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431:308-312
8. Lemmens K, De BT, Dhollander T et al (2009) DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. *Genome Biol* 10:R27
9. Shen-Orr SS, Milo R, Mangan S et al (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 31:64-68
10. Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet* 8:450-461
11. Hartwell LH, Hopfield JJ, Leibler S et al (1999) From molecular to modular cell biology. *Nature* 402:C47-C52
12. Ihmels J, Friedlander G, Bergmann S et al (2002) Revealing modular organization in the yeast transcriptional network. *Nat Genet* 31:370-377
13. Qi Y, Ge H (2006) Modularity and dynamics of cellular networks. *PLoS Comput Biol* 2:e174

14. Holter NS, Mitra M, Maritan A et al (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci U S A* 97:8409-8414
15. Babu MM (2008) Computational approaches to study transcriptional regulation. *Biochem Soc Trans* 36:758-765
16. Sasik R, Woelk CH, Corbeil J (2004) Microarray truths and consequences. *J Mol Endocrinol* 33:1-9
17. Barrett T, Troup DB, Wilhite SE et al (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 35:D760-D765
18. Demeter J, Beauheim C, Gollub J et al (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* 35:D766-D770
19. Parkinson H, Kapushesky M, Shojatalab M et al (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35:D747-D750
20. Brazma A, Hingamp P, Quackenbush J et al (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29:365-371
21. Bammler T, Beyer RP, Bhattacharya S et al (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2:351-356
22. Irizarry RA, Warren D, Spencer F et al (2005) Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2:345-350
23. Faith JJ, Driscoll ME, Fusaro VA et al (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res* 36:D866-D870
24. Engelen, K., Fu, Q., Meysman, P., Sanchez, A., Fierro, A. C., De Smet, R., Lemmens, K., and Marchal, K. COLOMBOS: access port for cross-platform bacterial expression compendia. Submitted.
25. Fadda A, Fierro AC, Lemmens K et al (2009) Inferring the transcriptional network of Bacillus subtilis. *unpublished*
26. Eisen MB, Spellman PT, Brown PO et al (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-14868
27. Tavazoie S, Hughes JD, Campbell MJ et al (1999) Systematic determination of genetic network architecture. *Nat Genet* 22:281-285
28. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 1:24-45

29. Ihmels JH, Bergmann S (2004) Challenges and prospects in the analysis of large-scale gene expression data. *Brief Bioinform* 5:313-327
30. Segal E, Shapira M, Regev A et al (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34:166-176
31. Ernst J, Beg QK, Kay KA et al (2008) A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*. *PLoS Comput Biol* 4:e1000044
32. Joshi A, De Smet R, Marchal K et al (2009) Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics* 25:490-496
33. Mordelet F, Vert JP (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics* 24:i76-i82
34. Faith JJ, Hayete B, Thaden JT et al (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5:e8
35. Bar-Joseph Z, Gerber GK, Lee TI et al (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 21:1337-1342
36. Cho BK, Zengler K, Qiu Y et al (2009) The transcription unit architecture of the Escherichia coli genome. *Nat Biotechnol* 27:1043-1049
37. Mendoza-Vargas A, Olvera L, Olvera M et al (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One* 4:e7526
38. Zare H, Sangurdekar D, Srivastava P et al (2009) Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations. *BMC Syst Biol* 3:39
39. Kohanski MA, Dwyer DJ, Wierzbowski J et al (2008) Mistranslation of membrane proteins and two-component system activation trigger antibiotic-mediated cell death. *Cell* 135:679-690
40. Yoon H, McDermott JE, Porwollik S et al (2009) Coordinated regulation of virulence during systemic infection of Salmonella enterica serovar Typhimurium. *PLoS Pathog* 5:e1000306
41. Bonneau R, Facciotti MT, Reiss DJ et al (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131:1354-1365
42. Babu MM, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* 31:1234-1244

43. Draghici S, Khatri P, Eklund AC et al (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* 22:101-109
44. Marshall E (2004) Getting the noise out of gene arrays. *Science* 306:630-631
45. Carvalho LE, Lawrence CE (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc Natl Acad Sci U S A* 105:3209-3214
46. Hu J, Yang YD, Kihara D (2006) EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC Bioinformatics* 7:342
47. Wijaya E, Yiu SM, Son NT et al (2008) MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics* 24:2288-2295
48. Reddy TE, DeLisi C, Shakhnovich BE (2007) Binding site graphs: a new graph theoretical framework for prediction of transcription factor binding sites. *PLoS Comput Biol* 3:e90
49. Yanover C, Singh M, Zaslavsky E (2009) M are better than one: an ensemble-based motif finder and its application to regulatory element prediction. *Bioinformatics* 25:868-874
50. Ginalski K, Elofsson A, Fischer D et al (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19:1015-1018
51. Lundstrom J, Rychlewski L, Bujnicki J et al (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci* 10:2354-2362
52. Abeel T, Helleputte T, Van de Peer Y et al (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26:392-398
53. Grotkjaer T, Winther O, Regenberg B et al (2006) Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm. *Bioinformatics* 22:58-67
54. Monti S, Tamayo P, Mesirov J et al (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* 52:91-118
55. Joshi A, Van de Peer Y, Michoel T (2008) Analysis of a Gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics* 24:176-183
56. Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 19:1090-1099

57. Yu Z, Wong HS, Wang H (2007) Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics* 23:2888-2896
58. Ding Y, Chan CY, Lawrence CE (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* 11:1157-1166
59. Asur S, Ucar D, Parthasarathy S (2007) An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics* 23:i29-i40
60. Greene D, Cagney G, Krogan N et al (2008) Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics* 24:1722-1728
61. Hibbs MA, Myers CL, Huttenhower C et al (2009) Directing experimental biology: a case study in mitochondrial biogenesis. *PLoS Comput Biol* 5:e1000322
62. Ohta S, Bukowski-Wills JC, Sanchez-Pulido L et al (2010) The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* 142:810-821
63. Nachman I, Regev A (2009) BRNI: Modular analysis of transcriptional regulatory programs. *BMC Bioinformatics* 10:155
64. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26:1057-1063
65. De Smet R, Marchal K (2010) An ensemble method for querying gene expression compendia with experimental lists. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM2010).* Accepted for publication.
66. Michoel T, De SR, Joshi A et al (2009) Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst Biol* 3:49
67. Michoel T, De Smet R, Joshi A et al (2009) Reverse-engineering transcriptional modules from gene expression data. *Ann N Y Acad Sci* 1158:36-43
68. De Smet R., Marchal K (2010) Advantages and limitations of current network inference methods. *Nat Rev Microbiol* 8:717-729
69. Owen AB, Stuart J, Mach K et al (2003) A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome Res* 13:1828-1837
70. Hibbs MA, Hess DC, Myers CL et al (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 23:2692-2699

71. Adler P, Kolde R, Kull M et al (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol* 10:R139

72. Zhao, H., Cloots, L., Van den Bulcke, T., Wu, Y., De Smet, R., Storms, V., Meysman, P., Engelen, K., and Marchal, K (2010). Query-based biclustering of gene expression data using Probabilistic Relational Models. *BMC Bioinformatics*. Accepted for publication.

73. Dhollander T, Sheng Q, Lemmens K et al (2007) Query-driven module discovery in microarray data. *Bioinformatics* 23:2573-2580

74. Thijs, I. M., De Smet, R., De Coster, D., De Weerdt, A., Verhoeven, T., McClelland, M., Vanderleyden, J., Marchal, K., and De Keersmaecker, S. C. J. New target genes of Salmonella Typhimurium invasion regulator InvF. Submitted.

75. Ihmels J, Bergmann S, Barkai N (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20:1993-2003

76. Sheng Q, Moreau Y, De Moor B (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics* 19 Suppl 2:ii196-ii205

77. Gelman AB, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, USA

78. Jain AK (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31:651-666

79. Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716-723

80. Segal L, Lapidot M, Solan Z et al (2007) Nucleotide variation of regulatory motifs may lead to distinct expression patterns. *Bioinformatics* 23:i440-i449

81. Cheng Y, Church GM (2000) Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8:93-103

82. Csardi G, Kutalik Z, Bergmann S (2010) Modular analysis of gene expression data with R. *Bioinformatics* 26:1376-1377

83. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545-15550

84. Fortunato S (2010) Community detection in graphs. *Physics Reports* 486:75-174

85. Grainger DC, Aiba H, Hurd D et al (2007) Transcription factor distribution in Escherichia coli: studies with FNR protein. *Nucleic Acids Res* 35:269-278

86. Zhang B, Horvath S (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* 4:1-37
87. Serrano MA, Boguna M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci U S A* 106:6483-6488
88. Newman ME (2006) Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 103:8577-8582
89. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *Science* 315:972-976
90. Van Dongen, S. Graph Clustering by Flow Simulation. 2000. *PhD thesis.*
91. Pollard K, van der Laan M (2005) Cluster Analysis of Genomic Data. In: Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (eds) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer,
92. Goldberg DS, Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* 100:4372-4376
93. Keseler IM, Bonavides-Martinez C, Collado-Vides J et al (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res* 37:D464-D470
94. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M et al (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 36:D120-D124
95. Sun H, Lemmens K, Van den Bulcke T et al (2009) ViTraM: visualization of transcriptional modules. *Bioinformatics* 25:2450-2451
96. Strehl A, Ghosh J (2002) Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research* 3:583-617
97. Newman ME (2004) Analysis of weighted networks. *Phys Rev E* 70:056131-1-056131-9
98. Gao F, Foat BC, Bussemaker HJ (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* 5:31
99. Prouty AM, Schwesinger WH, Gunn JS (2002) Biofilm formation and interaction with the surfaces of gallstones by Salmonella spp. *Infect Immun* 70:2640-2649

100. Barak JD, Liang A, Narm KE (2008) Differential attachment to and subsequent contamination of agricultural crops by Salmonella enterica. *Appl Environ Microbiol* 74:5568-5570

101. Brandl MT, Mandrell RE (2002) Fitness of Salmonella enterica serovar Thompson in the cilantro phyllosphere. *Appl Environ Microbiol* 68:3614-3621

102. Boddicker JD, Ledeboer NA, Jagnow J et al (2002) Differential binding to and biofilm formation on, HEp-2 cells by Salmonella enterica serovar Typhimurium is dependent upon allelic variation in the fimH gene of the fim gene cluster. *Mol Microbiol* 45:1255-1265

103. Latasa C, Roux A, Toledo-Arana A et al (2005) BapA, a large secreted protein required for biofilm formation and host colonization of Salmonella enterica serovar Enteritidis. *Mol Microbiol* 58:1322-1339

104. Hall-Stoodley L, Costerton JW, Stoodley P (2004) Bacterial biofilms: from the natural environment to infectious diseases. *Nat Rev Microbiol* 2:95-108

105. Costerton JW, Stewart PS, Greenberg EP (1999) Bacterial biofilms: a common cause of persistent infections. *Science* 284:1318-1322

106. Lewis K (2008) Multidrug tolerance of biofilms and persister cells. *Curr Top Microbiol Immunol* 322:107-131

107. Fux CA, Costerton JW, Stewart PS et al (2005) Survival strategies of infectious biofilms. *Trends Microbiol* 13:34-40

108. Mouslim C, Hilbert F, Huang H et al (2002) Conflicting needs for a Salmonella hypervirulence gene in host and non-host environments. *Mol Microbiol* 45:1019-1027

109. Rasschaert G, Houf K, De ZL (2007) Impact of the slaughter line contamination on the presence of Salmonella on broiler carcasses. *J Appl Microbiol* 103:333-341

110. Costerton JW, Lewandowski Z, Caldwell DE et al (1995) Microbial biofilms. *Annu Rev Microbiol* 49:711-745

111. Davey ME, O'toole GA (2000) Microbial biofilms: from ecology to molecular genetics. *Microbiol Mol Biol Rev* 64:847-867

112. Prouty AM, Gunn JS (2003) Comparative analysis of Salmonella enterica serovar Typhimurium biofilm formation on gallstones and on glass. *Infect Immun* 71:7154-7158

113. Stewart PS, Franklin MJ (2008) Physiological heterogeneity in biofilms. *Nat Rev Microbiol* 6:199-210

114. Stoodley P, Sauer K, Davies DG et al (2002) Biofilms as complex differentiated communities. *Annu Rev Microbiol* 56:187-209

115. Beenken KE, Dunman PM, McAleese F et al (2004) Global gene expression in Staphylococcus aureus biofilms. *J Bacteriol* 186:4665-4684

116. Beloin C, Valle J, Latour-Lambert P et al (2004) Global impact of mature biofilm lifestyle on Escherichia coli K-12 gene expression. *Mol Microbiol* 51:659-674

117. Ren D, Bedzyk LA, Thomas SM et al (2004) Gene expression in Escherichia coli biofilms. *Appl Microbiol Biotechnol* 64:515-524

118. Schembri MA, Kjaergaard K, Klemm P (2003) Global gene expression in Escherichia coli biofilms. *Mol Microbiol* 48:253-267

119. Whiteley M, Bangera MG, Bumgarner RE et al (2001) Gene expression in Pseudomonas aeruginosa biofilms. *Nature* 413:860-864

120. Shemesh M, Tam A, Kott-Gutkowski M et al (2008) DNA-microarrays identification of Streptococcus mutans genes associated with biofilm thickness. *BMC Microbiol* 8:236

121. Monds RD, O'toole GA (2009) The developmental model of microbial biofilms: ten years of a paradigm up for review. *Trends Microbiol* 17:73-87

122. White AP, Weljie AM, Apel D et al (2010) A global metabolic shift is linked to Salmonella multicellular development. *PLoS One* 5:e11814

123. Santiviago CA, Reynolds MM, Porwollik S et al (2009) Analysis of pools of targeted Salmonella deletion mutants identifies novel genes affecting fitness during competitive infection in mice. *PLoS Pathog* 5:e1000477

124. Hermans, K., Anh Nguyen, T. L., Roberfroid, S., Verhoeven, T., De Coster, D., Vanderleyden, J., and De Keersmaecker, S. C. Gene expression analysis of monospecies *Salmonella* Typhimurium biofilms using a single cell approach. J Bacteriol . Submitted.

125. McClelland M, Sanderson KE, Spieth J et al (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature* 413:852-856

126. Hamilton S, Bongaerts RJ, Mulholland F et al (2009) The transcriptional programme of Salmonella enterica serovar Typhimurium reveals a key role for tryptophan metabolism in biofilms. *BMC Genomics* 10:599

127. Wells TJ, Sherlock O, Rivas L et al (2008) EhaA is a novel autotransporter protein of enterohemorrhagic Escherichia coli O157:H7 that contributes to adhesion and biofilm formation. *Environ Microbiol* 10:589-604

128. Wang S, Niu C, Shi Z et al (2010) Effects of ibeA deletion on the virulence and biofilm formation of an avian pathogenic Escherichia coli. *Infect Immun*
129. Flemming HC, Wingender J (2010) The biofilm matrix. *Nat Rev Microbiol* 8:623-633
130. Yeger-Lotem E, Riva L, Su LJ et al (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet* 41:316-323
131. Smith JJ, Sydorskyy Y, Marelli M et al (2006) Expression and functional profiling reveal distinct gene classes involved in fatty acid metabolism. *Mol Syst Biol* 2:2006
132. Haugen AC, Kelley R, Collins JB et al (2004) Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Biol* 5:R95
133. Barnhart MM, Chapman MR (2006) Curli biogenesis and function. *Annu Rev Microbiol* 60:131-147
134. Romling U (2005) Characterization of the rdar morphotype, a multicellular behaviour in Enterobacteriaceae. *Cell Mol Life Sci* 62:1234-1246
135. Ledeboer NA, Frye JG, McClelland M et al (2006) Salmonella enterica serovar Typhimurium requires the Lpf, Pef, and Tafi fimbriae for biofilm formation on HEp-2 tissue culture cells and chicken intestinal epithelium. *Infect Immun* 74:3156-3169
136. Shah P, Swiatlo E (2008) A multifaceted role for polyamines in bacterial pathogens. *Mol Microbiol* 68:4-16
137. Yang L, Barken KB, Skindersoe ME et al (2007) Effects of iron on DNA release and biofilm development by Pseudomonas aeruginosa. *Microbiology* 153:1318-1328
138. Mireles JR, Toguchi A, Harshey RM (2001) Salmonella enterica serovar typhimurium swarming mutants with altered biofilm-forming abilities: surfactin inhibits biofilm formation. *J Bacteriol* 183:5848-5854
139. Beloin C, Ghigo JM (2005) Finding gene-expression patterns in bacterial biofilms. *Trends Microbiol* 13:16-19
140. Anriany Y, Sahu SN, Wessels KR et al (2006) Alteration of the rugose phenotype in waaG and ddhC mutants of Salmonella enterica serovar Typhimurium DT104 is associated with inverse production of curli and cellulose. *Appl Environ Microbiol* 72:5002-5012
141. Zakikhany K, Harrington CR, Nimtz M et al (2010) Unphosphorylated CsgD controls biofilm formation in Salmonella enterica serovar Typhimurium. *Mol Microbiol* 77:771-786

142. Fisher RA (1932) Statistical Methods for Research Workers, 4th edition edn. Oliver and Boyd, London

143. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science* 303:799-805

144. Lee SI, Pe'er D, Dudley AM et al (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A* 103:14062-14067

145. Segal E, Sirlin CB, Ooi C et al (2007) Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol* 25:675-680

146. Zhu H, Yang H, Owen MR (2007) Combined microarray analysis uncovers self-renewal related signaling in mouse embryonic stem cells. *Syst Synth Biol* 1:171-181

147. Li J, Liu ZJ, Pan YC et al (2007) Regulatory module network of basic/helix-loop-helix transcription factors in mouse brain. *Genome Biol* 8:R244

148. Novershtern N, Itzhaki Z, Manor O et al (2008) A functional and regulatory map of asthma. *Am J Respir Cell Mol Biol* 38:324-336

149. Cosentino LM, Jona P, Bassetti B et al (2007) Hierarchy and feedback in the evolution of the Escherichia coli transcription network. *Proc Natl Acad Sci U S A* 104:5516-5520

150. Ma HW, Buer J, Zeng AP (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* 5:199

151. Yu H, Gerstein M (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A* 103:14724-14731

152. Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome Biol* 9:R4

153. Martinez-Antonio A, Collado-Vides J (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* 6:482-489

154. Stulke J, Hillen W (1999) Carbon catabolite repression in bacteria. *Curr Opin Microbiol* 2:195-201

155. Thieffry D, Huerta AM, Perez-Rueda E et al (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays* 20:433-440

156. Herrgard MJ, Covert MW, Palsson BO (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res 2003 Nov ;13(11):2423 -34 Epub 2003 Oct 14* 13:2423-2434

157. Mangan S, Itzkovitz S, Zaslaver A et al (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. *J Mol Biol* 356:1073-1081

158. Meng LM, Nygaard P (1990) Identification of hypoxanthine and guanine as the co-repressors for the purine regulon genes of Escherichia coli. *Mol Microbiol* 4:2187-2192

159. Michel B (2005) After 30 years of study, the bacterial SOS response still surprises us. *PLoS Biol* 3:e255

160. Hershberg R, Yeger-Lotem E, Margalit H (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet* 21:138-142

161. Basso K, Margolin AA, Stolovitzky G et al (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* 37:382-390

162. Bonneau R, Reiss DJ, Shannon P et al (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* 7:R36

163. Soranzo N, Bianconi G, Altafini C (2007) Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics* 23:1640-1647

164. Zampieri M, Soranzo N, Altafini C (2008) Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics* 24:1510-1515

165. Salgado H, Gama-Castro S, Peralta-Gil M et al (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34:D394-D397

166. Ptashne M, Gilbert W (1970) Genetic repressors. *Sci Am* 222:36-44

167. Marshall E (2004) Getting the noise out of gene arrays. *Science* 306:630-631

168. Johnson DS, Li W, Gordon DB et al (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res* 18:393-403

169. Bansal M, Belcastro V, Ambesi-Impiombato A et al (2007) How to infer gene networks from expression profiles. *Mol Syst Biol* 3:78

170. Bonneau R (2008) Learning biological networks: from modules to dynamics. *Nat Chem Biol* 4:658-664

171. Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol* 9:770-780

172. Margolin AA, Nemenman I, Basso K et al (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1:S7
173. Foster JW (2004) Escherichia coli acid resistance: tales of an amateur acidophile. *Nat Rev Microbiol* 2:898-907
174. Anastassiou D (2007) Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol* 3:83
175. Watkinson J, Liang KC, Wang X et al (2009) Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann N Y Acad Sci* 1158:302-313
176. Shaw OJ, Harwood C, Steggles LJ et al (2004) SARGE: a tool for creation of putative genetic networks. *Bioinformatics* 20:3638-3640
177. Schmitt WA, Jr., Raab RM, Stephanopoulos G (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res* 14:1654-1663
178. Gutierrez-Rios RM, Rosenblueth DA, Loza JA et al (2003) Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Res* 13:2435-2443
179. Lemmens K, Dhollander T, De BT et al (2006) Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol* 7:R37
180. Tanay A, Sharan R, Kupiec M et al (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* 101:2981-2986
181. Myers CL, Troyanskaya OG (2007) Context-sensitive data integration and prediction of biological networks. *Bioinformatics* 23:2322-2330
182. Sabatti C, James GM (2006) Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics* 22:739-746
183. Reiss DJ, Baliga NS, Bonneau R (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* 7:280
184. Huttenhower C, Mutungu KT, Indik N et al (2009) Detailing regulatory networks through large scale data integration. *Bioinformatics* 25:3267-3274
185. Freckleton G, Lippman SI, Broach JR et al (2009) Microarray profiling of phage-display selections for rapid mapping of transcription factor-DNA interactions. *PLoS Genet* 5:e1000449

186. Butala M, Busby SJ, Lee DJ (2009) DNA sampling: a method for probing protein binding at specific loci on bacterial chromosomes. *Nucleic Acids Res* 37:e37

187. Grainger DC, Hurd D, Goldberg MD et al (2006) Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res* 34:4642-4652

188. Grainger DC, Hurd D, Harrison M et al (2005) Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proc Natl Acad Sci U S A* 102:17693-17698

189. Lu LJ, Xia Y, Paccanaro A et al (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* 15:945-953

190. Getz G, Levine E, Domany E (2000) Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A* 97:12079-12084

191. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18 Suppl 1:S136-S144

192. Lazzeroni L, Owen A (2002) Plaid models for gene expression data. *Statistica Sinica* 2:61-86

193. Murali TM, Kasif S (2003) Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput*77-88

194. Ben-Dor A, Chor B, Karp R et al (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J Comput Biol* 10:373-384

195. Kluger Y, Basri R, Chang JT et al (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* 13:703-716

196. Zwir I, Huang H, Groisman EA (2005) Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation. *Bioinformatics* 21:4073-4083

197. Pena JM, Bjorkegren J, Tegner J (2005) Growing Bayesian network models of gene networks from seed genes. *Bioinformatics* 21 Suppl 2:ii224-ii229

198. Gat-Viks I, Shamir R (2007) Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res* 17:358-367

199. Honkela A, Girardot C, Gustafson EH et al (2010) Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A* 107:7793-7798

200. Zwir I, Shin D, Kato A et al (2005) Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica. *Proc Natl Acad Sci U S A* 102:2862-2867

201. de Hoon MJ, Makita Y, Imoto S et al (2004) Predicting gene regulation by sigma factors in Bacillus subtilis from genome-wide data. *Bioinformatics* 20 Suppl 1:i101-i108

202. Sierro N, Makita Y, de Hoon M. et al (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 36:D93-D96

203. McDermott JE, Taylor RC, Yoon H et al (2009) Bottlenecks and hubs in inferred networks are important for virulence in Salmonella typhimurium. *J Comput Biol* 16:169-180

204. Taylor RC, Singhal M, Weller J et al (2009) A network inference workflow applied to virulence-related processes in *Salmonella typhimurium*. *Ann N Y Acad Sci* 1158:143-158

205. Fredrickson JK, Romine MF, Beliaev AS et al (2008) Towards environmental systems biology of Shewanella. *Nat Rev Microbiol* 6:592-603

206. Bonneau R, Facciotti MT, Reiss DJ et al (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131:1354-1365

207. Toepel J, McDermott JE, Summerfield (2009) Transcriptional analysis of the unicellular, diazotrophic cyanobacterium *Cyanothece* sp. ATCC 51142 grown under short day/night cycles. *J Phycol* 45:610-620

208. Mendes P, Sha W, Ye K (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* 19 Suppl 2:ii122-ii129

209. Van den Bulcke T, Van Leemput K, Naudts B et al (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* 7:43

210. Van den Bulcke T, Lemmens K, Van de Peer Y et al (2006) Inferring Transcriptional Networks by Mining 'Omics' Data. *Current Bioinformatics* 1:301-331

211. Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann N Y Acad Sci* 1115:1-22

212. Cantone I, Marucci L, Iorio F et al (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* 137:172-181

213. Marbach D, Prill RJ, Schaffter T et al (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci U S A* 107:6286-6291

214. Stolovitzky G, Prill RJ, Califano A (2009) Lessons from the DREAM2 Challenges. *Ann N Y Acad Sci* 1158:159-195

215. Stolovitzky G, Prill RJ, Califano A (2009) Lessons from the DREAM2 Challenges. *Ann N Y Acad Sci* 1158:159-195

216. Liao JC, Boscolo R, Yang YL et al (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A* 100:15522-15527

217. Gardner TS, di BD, Lorenz D et al (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301:102-105

218. Ourfali O, Shlomi T, Ideker T et al (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* 23:i359-i366

219. Komurov K, White MA, Ram PT (2010) Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput Biol* 6:

220. Wang L, Tu Z, Sun F (2009) A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics* 10:220

221. Tu Z, Argmann C, Wong KK et al (2009) Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network. *Genome Res* 19:1057-1067

222. Nibbe RK, Koyuturk M, Chance MR (2010) An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol* 6:e1000639

223. Halperin Y, Linhart C, Ulitsky I et al (2009) Allegro: analyzing expression and sequence in concert to discover regulatory programs. *Nucleic Acids Res* 37:1566-1579

224. Zaki MJ, Hsiao C (2002) CHARM: An efficient algorithm for Closed Itemset Mining. In: Grossman, R., Han, J., Kumar, V., Mannila, H., and Motwani, R. (eds) Proceedings of the Second SIAM International Conference on Data Mining (SDM '02).

225. Hwang D, Smith JJ, Leslie DM et al (2005) A data integration methodology for systems biology: experimental verification. *Proc Natl Acad Sci U S A* 102:17302-17307

226. Navlakha S, Kingsford C (2010) Exploring biological network dynamics with ensembles of graph partitions. *Pac Symp Biocomput*166-177

227. Madar A, Greenfield A, Vanden Eijnden E et al (2010) DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS One* 5:e9803

228. Le HS, Oltvai ZN, Bar-Joseph Z (2010) Cross-species queries of large gene expression databases. *Bioinformatics* 26:2416-2423

229. Guan Y, Dunham M, Caudy A et al (2010) Systematic planning of genome-scale experiments in poorly studied species. *PLoS Comput Biol* 6:e1000698

230. Chen R, Sigdel TK, Li L et al (2010) Differentially expressed RNA from public microarray data identifies serum protein biomarkers for cross-organ transplant rejection and other conditions. *PLoS Comput Biol* 6:

231. Huang H, Liu CC, Zhou XJ (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc Natl Acad Sci U S A* 107:6823-6828

232. Klumpp S, Zhang Z, Hwa T (2009) Growth rate-dependent global effects on gene expression in bacteria. *Cell* 139:1366-1375

233. Scott M, Gunderson CW, Mateescu EM et al (2010) Interdependence of cell growth and gene expression: origins and consequences. *Science* 330:1099-1102

234. Sorek R, Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* 11:9-16

235. MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 7:287-296

236. Sharma CM, Vogel J (2009) Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr Opin Microbiol* 12:536-546

237. Coppins RL, Hall KB, Groisman EA (2007) The intricate world of riboswitches. *Curr Opin Microbiol* 10:176-181

238. Waters LS, Storz G (2009) Regulatory RNAs in bacteria. *Cell* 136:615-628

239. Bejerano-Sagie M, Xavier KB (2007) The role of small RNAs in quorum sensing. *Curr Opin Microbiol* 10:189-198

240. Masse E, Salvail H, Desnoyers G et al (2007) Small RNAs controlling iron metabolism. *Curr Opin Microbiol* 10:140-145

241. Toledo-Arana A, Repoila F, Cossart P (2007) Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol* 10:182-188

242. Bonnet E, Michoel T, Van de Peer Y (2010) Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data. *Bioinformatics* 26:i638-i644

243. Bonnet E, Tatari M, Joshi A et al (2010) Module network inference from a cancer gene expression data set identifies microRNA regulated modules. *PLoS One* 5:e10162

244. Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128:693-705

245. Bai L, Morozov AV (2010) Gene regulation by nucleosome positioning. *Trends Genet*

246. Vora T, Hottes AK, Tavazoie S (2009) Protein occupancy landscape of a bacterial genome. *Mol Cell* 35:247-253

247. Berger M, Farcas A, Geertz M et al (2010) Coordination of genomic structure and transcription by the main bacterial nucleoid-associated protein HU. *EMBO Rep* 11:59-64

248. Lang B, Blot N, Bouffartigues E et al (2007) High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Res* 35:6330-6337

249. Travers A, Muskhelishvili G (2005) DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat Rev Microbiol* 3:157-169

250. Cho BK, Knight EM, Barrett CL et al (2008) Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts. *Genome Res* 18:900-910

251. Yus E, Maier T, Michalodimitrakis K et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science* 326:1263-1268

252. Covert MW, Knight EM, Reed JL et al (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:92-96

253. Thiele I, Jamshidi N, Fleming RM et al (2009) Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* 5:e1000312

254. Kotte O, Zaugg JB, Heinemann M (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol Syst Biol* 6:355

255. Ishii N, Nakahigashi K, Baba T et al (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science* 316:593-597

256. Wang K, Saito M, Bisikirska BC et al (2009) Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol* 27:829-839

257. Lefebvre C, Rajbhandari P, Alvarez MJ et al (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* 6:377
258. Clauset A, Moore C, Newman ME (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98-101
259. Hsiao TL, Revelles O, Chen L et al (2010) Automatic policing of biochemical annotations using genomic correlations. *Nat Chem Biol* 6:34-40
260. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. *Nat Rev Genet* 7:862-872
261. Cookson W, Liang L, Abecasis G et al (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10:184-194
262. Cooper TF, Remold SK, Lenski RE et al (2008) Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli. *PLoS Genet* 4:e35
263. Fong SS, Joyce AR, Palsson BO (2005) Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res* 15:1365-1372
264. Mitchell A, Romano GH, Groisman B et al (2009) Adaptive prediction of environmental changes by microorganisms. *Nature* 460:220-224
265. Tagkopoulos I, Liu YC, Tavazoie S (2008) Predictive behavior within microbial genetic networks. *Science* 320:1313-1317
266. Litvin O, Causton HC, Chen BJ et al (2009) Modularity and interactions in the genetics of gene expression. *Proc Natl Acad Sci U S A* 106:6441-6446
267. Lee SI, Dudley AM, Drubin D et al (2009) Learning a prior on regulatory potential from eQTL data. *PLoS Genet* 5:e1000358
268. Gat-Viks I, Meller R, Kupiec M et al (2010) Understanding gene sequence variation in the context of transcription regulation in yeast. *PLoS Genet* 6:e1000800
269. Barrick JE, Yu DS, Yoon SH et al (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461:1243-1247
270. Conrad TM, Joyce AR, Applebee MK et al (2009) Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol* 10:R118
271. Herring CD, Raghunathan A, Honisch C et al (2006) Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* 38:1406-1412

272. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* 102:1572-1577
273. Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. *Nature* 461:218-223
274. Joshi A, Van PT, Van de Peer Y et al (2010) Characterizing regulatory path motifs in integrated networks using perturbational data. *Genome Biol* 11:R32
275. Ye C, Galbraith SJ, Liao JC et al (2009) Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS Comput Biol* 5:e1000311
276. Lee I, Date SV, Adai AT et al (2004) A probabilistic functional network of yeast genes. *Science* 306:1555-1558
277. Zhu J, Zhang B, Smith EN et al (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* 40:854-861
278. Suthram S, Beyer A, Karp RM et al (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol Syst Biol* 4:162
279. Lee E, Bussemaker HJ (2010) Identifying the genetic determinants of transcription factor activity. *Mol Syst Biol* 6:412
280. Polikar R (2006) Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6:21-45

Publication list

De Smet, R., Marchal, K. (2010). An ensemble method for querying gene expression compendia with experimental lists. Accepted for publication in proceedings of the *IEEE International Conference on Bioinformatics and Biomedicine.*

De Smet, R., Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology, 8,* 717-729.

De Smet, R., Lemmens, K., Fierro, A., Marchal, K. (2009). Systems Microbiology: Gaining Insights in Transcriptional Networks. In: Sintchenko V. (Eds.), *Infectious Disease Informatics* (pp. 93-122). New York: Springer New York.

Michoel, T., **De Smet, R.**, Joshi, A., Van de Peer, Y., Marchal, K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Systems Biology, 3*, art.nr. 49, 49.

Michoel, T., **De Smet, R.**, Joshi, A., Marchal, K., Van de Peer, Y. (2009). Reverse-engineering transcriptional modules from gene expression data. *Annals of the New York Academy of Sciences, 1158*, 36-43.

Joshi, A., **De Smet, R.**, Marchal, K., Van de Peer, Y., Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics, 25*(4), 490-496.

Zhao, H., Cloots, L., Van den Bulcke, T., Wu, Y., **De Smet, R.**, Storms, V., Meysman, P., Engelen, K. Marchal, K. (2010). Query-based biclustering of gene expression data using Probabilistic Relational Models. Accepted for publication in *BMC Bioinformatics.*

Meysman, P., Dang, T. H., Laukens, K., Wu, Y., **De Smet, R.**, Marchal, K., Engelen, K. (2010). Use of structural DNA properties for the prediction of transcription factor binding sites. Accepted for publication in *Nucleic Acid Research*.

Curriculum vitae

Riet De Smet was born in Ghent (Belgium), on July 28th, 1982. In 2000 she started her education in Bioscience Engineering at Ghent University, where she received a Candidacy degree in Bioscience Engineering in 2002 and a Masters degree in Biotechnological Engineering in 2005. In 2006 she received a Masters degree in Bioinformatics at the K.U.Leuven. Since October 2006 until January 2007 she worked as a Research Assistant in the research group ESAT-SCD, under the supervision of Prof. Bart De Moor and Prof. Kathleen Marchal. Since January 2007 she has been pursuing her PhD as a Research Assistant of the *'Agentschap voor Innovatie door Wetenschap en Technologie in Vlaanderen'* (IWT) at the CMPG research group under supervision of Prof. Kathleen Marchal and Prof. Bart De Moor.