Data-driven inference of dynamic transcriptional regulatory mechanisms in prokaryotes: a systems perspective

Aaron N. Brooks

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Caroline Harwood, PhD, Chair

Nitin S. Baliga, PhD

Ilya Shmulevich, PhD

Program Authorized to Offer Degree:
Molecular and Cellular Biology

University of Washington

Abstract

Data-driven inference of dynamic transcriptional regulatory mechanisms in prokaryotes: a systems perspective

Aaron N. Brooks

Chair of the Supervisory Committee:
Gerald and Lyn Grinstein Endowed Professor Caroline Harwood, PhD
Department of Microbiology

Microbes tailor their physiology to diverse environments despite having streamlined genomes and few regulators. Mechanisms by which microbes expand their genetic repertoire include modular reorganization of genetic expression through dynamic activity of complex gene regulatory networks (GRNs). Deciphering accurate GRNs is essential to understand how their topology contributes to cellular behavior. This dissertation develops computational methods to reverse engineer GRNs directly from genome sequence and transcriptome data. These data-driven models capture dynamic interplay of environment and genome-encoded regulatory programs for two phylogenetically distant prokaryotes: *E. coli* (a bacterium) and *H. salinarum* (an archaeon). The models reveal how distribution of *cis*-acting gene regulatory elements (GREs) and condition-specific influence of transcription factors (TFs) at each element produces environment-specific transcriptional responses. These regulatory programs partition and re-organize transcriptional regulation of genes within regulons and operons into condition-specific co-regulated modules, or *corems*. Corems capture fitness-relevant co-regulation by different transcriptional control mechanisms acting across the entire genome. Organization of genes in corems defines a system-level principle for prokaryotic gene regulatory networks that extends existing paradigms of gene regulation and helps explain how microbes negotiate environmental change.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

CMONKEY: Integrated biclustering algorithm [284] that identifies groups of genes with (1) similar patterns of differential expression, over subsets of conditions (biclusters) (2) similar de novo detected sequence motifs in their promoters and (3) related functions, inferred from functional association networks (e.g., EMBL STRING [328]).

COREM: Co-regulated module, a set of conditionally co-regulated genes discovered by applying link community algorithm [5] to backbone extracted [308] gene-gene association network (inferred by cMonkey)

EGRIN: Environment and Gene Regulatory Influence Network derived by cMonkey and Inferelator

GRE: *gene regulatory element*. 8-30nt DNA sequence. Assumed to be a binding site for a TF. Discovered by *de novo* sequence motif detection in cMonkey.

GRN: *gene regulatory network*. All $transcription factor(TF) \rightarrow gene$ interactions. Interactions typically represent physical binding of a TF to a gene promoter, determined by experimental methods (*e.g.*, ChIP-chip, ChIP-seq, Y1H, or Y2H). Can be represented as a $N \times N$ (weighted) adjacency matrix, where N is the number of genes in the genome. Weights can reflect confidence in the interaction or other information (*e.g.*, number of condition observed)

INFERELATOR: Regulatory influence inference procedure [47] that uses the regularized linear regression to find combinations of changes in TF levels and EF concentrations

that accurately model the expression changes of each cMonkeydetected bicluster.

NT: nucleotide

PROMOTER: DNA sequence upstream of the TSS or coding start site (CSS) of a gene. Most regulatory sites fall within -50 to +250 nt of the TSS of a gene, where negative values indicate downstream (3') to the TSS and positive values indicate upstream of the TSS (5').

PSSM: Position-specific scoring matrix. Representation of a GRE in terms of relative conservation of nucleotides at each position based on alignments of matching sequences

TF: *transcription factor*. A protein that binds DNA and regulates transcription at gene promoters.

ACKNOWLEDGMENTS

I have been fortunate to enjoy the company of intelligent colleagues, a loving family, and wonderful friends throughout the duration of my graduate studies. Because of these outstanding people, the six years that went into this project were not only productive, but also a whole lot of fun!

Modern biology is a highly collaborative enterprise. Without the help and guidance of many individuals, there is no way the work described in this dissertation could have been completed. As a research unit we embodied the mantra of emergence: "the whole is greater than the sum of its parts". In addition to the handful of authors on papers I co-authored, I would like to highlight the roles David Reiss and Nitin Baliga. Dave played a fundamental role in my dissertation project. I learned a great deal about computational methods from him, especially programming in R. Nitin was my graduate advisor. His keen eye for communication helped me learn how to make my science more engaging. Nitin encouraged me to "think big."; to be ambitious - sometimes even beyond my abilities. I took on several projects that I thought were too big or difficult for me to complete - but I finished them. As a result, I acquired new skills and felt constantly challenged.

There are so many people to thank:

My intelligent colleagues: Serdar Turkarslan, Adrián López García, James Eddy, Ben Heavner, Chris Plaisier, Justin Ashworth, Wei-ju Wu, Karlyn Beer, Christoper Bare, Chris Lausted, Antoine Allard, Diego Martinez Salvanha, and Cecilia Garmendia for many conversations and insights.

My previous mentor, David Bear, who taught me how to be a careful scientist. Stephen Jett and Tamara Howard for encouraging me to continue in science.

My family: Geri Myers Beel, Trent Brooks, Alan Brooks, and Lisa Brooks for constant love and support.

My friends: Peter Sudmant, Oriol Roda-Naccari Noguera, Sheila Teves, Andrew Rivers, Jairo Rodriguez Lumbiarres, Kameron Decker Harris for the adventures.

And - a wonderful woman - Teal Harbin. She was my life support at the end of graduate school. She has been so kind and generous. She taught me how to be a fulfilled person - not just a scientist.

DEDICATION

to Uncle Don

who encouraged me to be a scientist before I even knew what that was

Chapter 1

INTRODUCTION

Bacteria thrive in diverse environmental conditions that constantly change. Negotiating environmental change requires molecular systems to sense and respond to the environment cues. Given their compact genomes and relatively small number of genes, bacteria employ gene regulatory systems to reuse their genetic repertoire in different combinations to meet environmental demands. The functional outcome of a genes operation may depend on the other genes with which it is co-expressed. ATP-driven transport systems, for example, preferentially transfer different molecules depending on which periplasmic binding protein interfaces with ATP-driven pump. The way the cell organizes component genes, proteins, and other molecules into discrete functional units, we refer to as modularity. The cell is modular: it groups together components with related functions, or, in our case, related expression and regulation that lead to common function. Historically, definition of what constitutes modular organization has changed substantially, from adjacent co-transcribed genes in an operon to modern notions that consider co-regulation of genes scattered throughout the genome. The way we view biological modularity shapes our understanding of what the cell is capable of doing. In this chapter, I review the foundations of modularity in biology. I suggest that modularity emerges from activity of gene regulatory networks; I trace the development of thought concerning biological modularity over the past 50 years; and, importantly, I introduce methods one can use to infer these biological modules directly from experimental data.

Chapter Highlights

- Bacterial genomes are compact. Bacteria reuse genes to perform tasks in multiple environments. Despite few regulators, bacterial genomes are embedded in complex

regulatory networks

- Individual genes are organized in functional modules to perform specific functions
- Activity of gene regulatory networks (GRNs) generates co-regulated modules
- Understanding of what defines a co-regulated module has changed significantly over time
- Co-regulated modules can be detected directly from data

1.1 Summary

Bacteria regulate expression of their genomes using complex networks of regulatory proteins. Analyzing the topology of these networks provides insight into dynamical properties of these networks. Execution of these complex regulatory networks organizes genes into modules, or groups of co-functional, co-expressed genes.

1.2 Gene regulation in prokaryotes: a historical perspective

After the seminal discovery of DNA's structure by Watson and Crick in 1953 [357], attention rapidly shifted to decoding the contents of the genome. This has been especially true since the invention of automated DNA sequencing in 1986 [315], which provided the technology to read genomes at a staggering rate [71]. The community began to ask questions like: How is static information encoded in the form of deoxyribonucleic acids transformed by the cell into the many proteins and complex molecular machines that compose it? What effect do genomic variations have on observed cellular behaviors? In many ways, these question remain unanswered. Outside of a few examples, we still do not understand the connection between genotype and phenotype. We typically cannot predict phenotype from genotype and, when we can, the variance explained is low [236], especially in humans. Clearly, there is a complex, non-linear relationship between information encoded in the genome and the

ultimate physiology of the cell that we do not yet understand. What we do know is that complex phenotypes are somehow the result of how molecular machines read and interpret the genome in context of a dynamic, noisy chemical milieu.

An overarching goal of my work was to narrow the gap separating knowledge of genotype from understanding of phenotype. I hoped to accomplish this by developing genome-wide, data-driven quantitative models of transcriptional regulation in prokaryotes. Transcriptional regulation is critical first step in controlling how information is extracted from the genome. Prokaryotes are tractable systems with relatively small genomes and a wide array of genetic and molecular tools, making them attractive models for investigating fundamental molecular processes.

The remainder of this dissertation is dedicated to (1) tracing evolution of thought about how prokaryotes regulate their genomes, with an emphasis on modularity in biological regulatory systems, (2) describing existing approaches to reverse-engineering gene regulatory networks from high-throughput experimental data, (3) introducing new computational approaches to this problem, (4) describing the success of these new approaches, with an emphasis on how microbes leverage condition-specific TF-gene interactions to coordinate expression their genomes and what the consequences for this type or regulation are for the cell, and, finally, (5) suggesting future directions for improvements of these methods and the implications our results have for understanding the function and evolution of gene regulatory networks. We[1] were able to infer comprehensive and accurate gene regulatory networks directly from gene expression data. We used these networks to refine a central notion of modular co-regulatory organization. The result of our work is the introduction of a new term for genetic co-regulation, the co-regulated module or *corem*.Organization of genes into corems better describes how a variety functions, pathways, and regulatory mechanisms coincide to affect cellular fitness.

In this chapter, I chart a historical perspective describing how the notion of control and

[1]Projects I describe throughout the text are collaborations with many other scientists. I will use the pronoun 'we' to emphasize the collaborative nature of my dissertation work.

modular organization of the regulatory genome - both teleological and mechanistic - has changed over 50 years. This will give the reader a foundation for understanding where this work fits within the lineage of biological thought.

1.2.1 Why do prokaryotes regulate their genomes?

Microbes are complex adaptive systems that live in variable environments. Barriers separating microbes from the world around them is generally small, sometimes as little as a cell membrane. These unicellular organisms have evolved to embrace change from outside the cell, as well as from within (illustrated in Figure 1.1). Microbes deal with these changes in several ways. To begin, they possess membranes or cell walls that regulate the passage of small molecules in and out of the cell, using both active and passive transport. Microbes have also evolved regulatory systems that help control expression of genes that are useful in certain environments. A canonical example is the *lac* operon. Since it may be wasteful to produce these genes (both transporter and catabolic enzymes) when lactose is unavailable or some more readily metabolized carbon source is available, the *lac* operon is expressed only in the presence of lactose as well as the absence of glucose. This ability to turn this system on- or off- is a function of regulatory proteins that control transcription.

While the genome is mostly static with respect to the lifetime of a single individual, expression of gene products from it is dynamic. This allows the cell to adjust its physiology to changing circumstances. Depending on the environment, microbes vary which genes and pathways are expressed, tailoring their physiology to increase fitness. This process of physiological adjustment relies on sensing information about the internal or external environment, relaying that information within the cell, and ultimately responding by controlling which combination of genes are produced.

1.2.2 Model organisms to study gene regulatory systems

Many organisms have been developed to study regulation, ranging in scale and complexity from prokaryotic to mammalian systems. Notable examples of organisms used for genome-

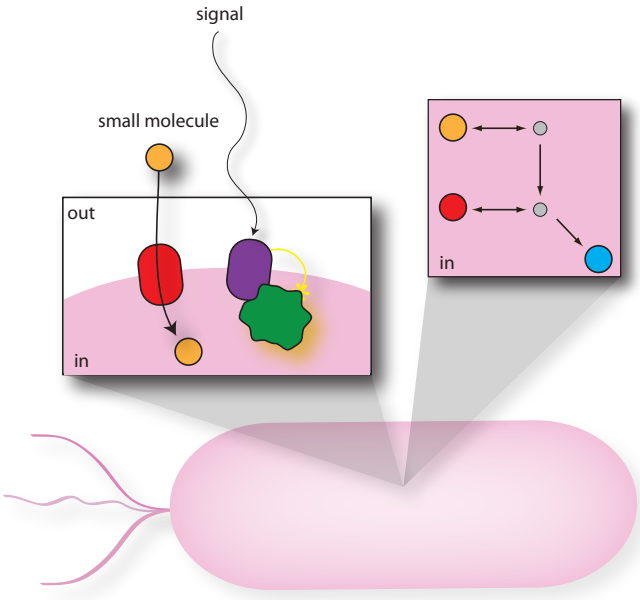


Figure 1.1: Microbes live in changing environments. Both internal and external environments constantly change. Cells sense and respond to changes in the external environment, in part, by transport and signaling. Internal cellular conditions can also be sensed. The right inset depicts several chemical transformation processes that constitute metabolism. Flux through these pathways can be monitored and adjusted by mechanisms described in Figure 1.2.

wide modeling and experimental characterizations include *S. cerevisiae* [304, 305], *B. subtilis* [264], *E. coli* [159], *H. salinarum* [50], and *M. genitalium* [180]. In this project, we utilized two of these organisms, *E. coli* and *H. salinarum*. The motivation for selection of each is described below.

Halobacterium salinarum sp. NRC-1 is a halophilic archeon. Several biological features make it an interesting organism to study: (1) *Halobacterium* has evolved to thrive in environmental conditions that would be lethal to most species, including exposure to extreme salinity (up to 5.2M NaCl), frequent desiccation-rehydration cycles, high doses of UV radiation, and regular influxes of transition metals and other environmental contaminants.

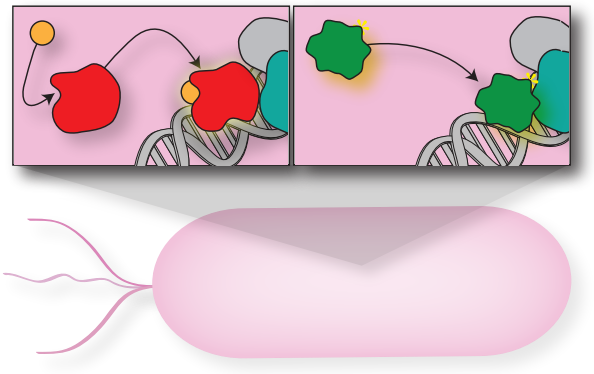


Figure 1.2: Cells relay information to regulate the genome. Continuation of Figure 1.1. Regulatory proteins called transcription factors (TFs) control expression at gene promoters. TFs contain DNA binding domains that allow them to recognize and bind to specific nucleotide sequences. In response to binding, these factors can either facilitate or inhibit recruitment of the RNA polymerase to initiate transcription. TFs can be activated by binding/unbinding of small molecules (i.e., allosteric activation/inhibition) or other post-translational modifications (e.g., phosphorylation in eukaryotes). When TFs are activated/inactivated TFs bind to/release from DNA to control the process of transcription at that location.

(2) transcriptional regulation in *Halobacterium* is a hybrid of eukaryotic and prokaryotic mechanisms. The general transcription factors (GTFs) comprise multiple orthologs of two eukaryotic GTFs: TFIIBs (TFB) and TATA binding proteins (TBPs). There are 42 possible TFB-TBP pairs, each of which may uniquely regulate expression in different conditions and contribute to niche adaptation [107, 343]. Finally, (3) *Halobacterium* possesses functional capabilities that make it a candidate for biotechnology applications. For example, the light-driven proton pump bacteriorhodopsin, which is normally used as an alternative energy source under low oxygen tension, has received interest as an element in optoelectronic devices and photochemical processes [268]. *Halobacterium* is an attractive gene regulatory model system because it contains a relatively small, fully sequenced and annotated genome (2.6 Mbp, 2400 genes) consisting of one large circular chromosome and two smaller plasmids

(pNRC100 and pNRC200). In addition, *H. salinarum* is easy to culture and manipulate genetically, making it useful for laboratory experiments. For the purposes of modeling, it is attractive because of the large amount of data and relatively little knowledge about the genetic underpinnings of its physiology, which make it ideal for model-derived biological discovery. There are currently 1,495 transcriptome profiles available for *H. salinarum* sp. NRC-1 (see 2.5.1 for additional details)

Escherichia coli K12 MG1655 is a gram-negative, facultatively anaerobic, gammaproteobacteria. Apart from human health concerns of pathogenic strains of *E. coli*, the bacteria has been used extensively in biotechnology and investigation of basic biological mechanisms since the foundational work of Lederberg and Tatum on bacterial conjugation in the 1940s [331]. *E. coli* has been subject to a wide-array of high-throughput experiments to characterize its genome [45], transcriptome [336], proteome [330], and even fitness landscape [259]. Large gene expression compendiums exist for the organism. Several computational approaches to reverse-engineer gene regulatory networks directly from gene expression data have used these public *E. coli* data sets [208, 246, 91, 251], including an organized worldwide competition [237]. There are currently 868 transcriptome profiles available for *E. coli* K-12 MG1655 (see 2.5.1 for additional details)

1.2.3 Prokaryotic gene regulatory mechanisms

Prokaryotes regulate expression of their genomes at several stages of production using multiple strategies and mechanisms. Generation of a protein product from a genetic locus can be regulated by controlling DNA structure and accessibility, transcription initiation, transcription elongation, or transcription termination; bacterial mRNAs can even be regulated post-transcriptionally through the action of small RNAs (sRNAs) This project focused on transcription initiation.

Assembly of RNA polymerase (RNAP) at gene start sites is a critical first step for transcription. In bacteria (like *E. coli*), RNAP first binds to one of the σ specificity factors,

like the housekeeping σ^{70} factor. This holoenzyme complex can then recognize and bind to specific sequences at -35 and -10 nt upstream of gene start sites. In archaea (like *H. salinarum*), the mechanism is more complicated. The archaeal RNAP resembles the eukaryotic RNA polymerase II machinery, where TBP and TFIIB (two general transcription factors) assist the RNA polymerase in locating gene start sites. *H. salinarum* encodes six *tbp* and seven *tfb* genes [23]. Regulation involving different combinations TBPs and TFIIBs has been shown to facilitate large-scale physiological changes [107] and niche adaptation [343].

Additional DNA regulatory proteins, called transcription factors (TFs), facilitate (activators) or prevent (inhibitors) recruitment of RNAP to gene start sites. TFs also recognize sequence specific sites in gene promoters. Many gene promoters contain binding sites for more than one TFs, in addition to the basal σ (or TFIIB/TBP) sites. A majority of regulatory sites occur with -250 nt to +50 nt of the gene start site in prokaryotes. TFs themselves can be categorized into two primary types: global and specific. As their name implies, global regulators regulate many genes (on the order of hundreds to thousands in prokaryotes). Typically these factors sense broad environmental shifts, like nutrient changes (e.g., CAP) or anaerobic conditions (e.g., FNR), or starvation and stationary phase (e.g., σ^{38}). Specific regulators control fewer genes, often targeting genes involved in a specific biological processes or pathway (e.g., LacI, inhibitor of lactose catabolism genes). Coordination microbial genomes results from combinations of interactions between specific and global regulators.

1.2.4 Discovery of regulatory interactions

An early goal of systems biology was to map the physical interactions between every TF and DNA. Such a physical interaction map would represent a complete gene regulatory network (GRN), which is described in detail below. A number of methods have been developed to identify TF $\rightarrow$ gene interactions, including *in vivo* methods like ChIP-chip [44] (now largely replaced by ChIP-seq, which has finer resolution [172]), yeast two-hybrid [116], and DNase I hypersensitivity [83], as well as *in vitro* methods like systematic evolution of ligand

by exponential enrichment (SELEX) [43]. Each can be used to elucidate binding sequence preferences of TFs and/or report locations throughout the genome that are physically bound by a particular TF. For those methods that report segments of the genome bound by a TF, each differs in its resolution, false positive rate, and extent to which it is a chimera (e.g., TFs for ChIP-chip are modified with functional groups for biochemical isolation),leading to contention about the trustworthiness of each approach. Knowledge generated from each of these methods, however, is limited to conditions in which the experiments were performed. Rather than providing a complete, unbiased map of TF binding locations, these approaches provide a snapshot of where TFs are bound in a particular condition.

Position-specific scoring matrices (PSSMs) quantify the sequence preference of TFs

TFs recognize DNA by sequence-specific electrostatic interactions and Van der Waals forces between evolutionarily conserved DNA-binding domains (DBDs) and DNA along the major groove. Different families of DBDs have evolved to interact with DNA in different ways. The helix-turn-helix domain, for example, consists of ∼20 amino acids that use hydrogen bonding to bind two α-helices along the major groove, whereas the leucine zipper domain forms two vertical α-helices that act as dimerization domains. Identification of putative TFs and assignment to domain families can be accomplished by locating DBDs in protein coding sequence through alignment.

A critical challenge for understanding genetic regulation by TFs is to describe what sequences they prefer to bind. This provides information about (1) how specific a TF is for DNA, and (2) where in the genome a TF may bind. Discovery of TF binding sequence preference is accomplished by alignment. Given a list of sequences putatively bound by a TF, those sequences can be aligned and a motif constructed by counting up the frequency of nucleotides at each position and scaling by its information content (Figure 1.3). These putative bound sequences can originate from observed binding (e.g., ChIP-seq) or computational identification (e.g., cMonkey see 2.2.1) There are several well-established computational algorithms for motif discovery (e.g. MEME/MAST [22, 21]), as well as searchable databases

for known motifs (e.g. STAMP [230]).

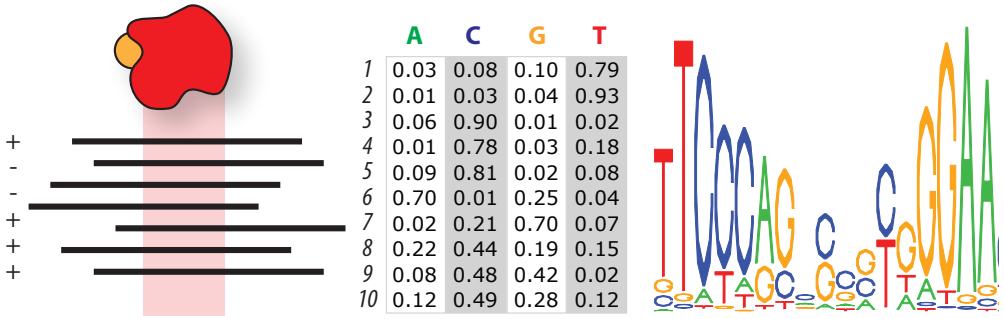


Figure 1.3: Binding sequence preferences for TFs are described by a position specific scoring matrix (PSSM) or position weight matrix (PWM). After aligning sequences putatively bound by a TF (left), the identity of each nucleotide is counted at each position to generate a relative frequency (center; PSSM or PWM). These frequencies are scaled by information content at each position for viewing as a motif logo. An example motif logo is shown at right.

1.2.5 Canonical genetic control modules

Grouping coordinated elements of biological systems into discrete modules is critical to understand their operation and evolution. Regulatory proteins, for instance, consist of multiple domains (like DNA binding domains), each with a particular structure and function. These domains include multiple amino acids that are found in similar configurations across many proteins to confer a similar activity. Likewise, single genes rarely explain the behavior (or malfunction) of biological systems. Rather, many genes act in concert to perform a particular task. We refer to discrete units of biological organization as modules. While the idea of modularity is intuitive, defining these modules turns out to be quite challenging, especially from data.

Modularity in Biology

The latter part of the 20^{th} century witnessed an explosion of scientific knowledge. Beyond genomes, countless other -omes were generated. From transcriptomes to proteomes, and even phenomes. Although such high-throughput cataloging is still commonplace, efforts to deduce how these molecular parts function together - when, why, and how - are now more widespread. Such efforts are at the heart of the nascent field of systems biology. Hartwell *et al.* wrote a perspective in 1999 [147] that anticipated an ongoing paradigm shift in biology. The authors argued that investigating molecular biology from the perspective of single genes or proteins, or even across a single molecular type would be insufficient to understand the complex organization and function of biological systems. Instead, they stressed a need for understanding how the molecular parts interact to form modules and how those modules interact to generate higher-order features.

Understanding biological systems at a genome-scale requires new tools, conceptual frameworks, and language. A first step in this process is to organize and abstract the components into modules. From the perspective of genetic regulation, modules consist of groups of genes that are either co-expressed or co-regulated. Even microbes contain thousands of genes. *E. coli*, for example, encodes 4,497 genes; the genome of *H. salinarum* contains nearly 2,400 genes. Deciphering how these genes partition into modules is non-trivial task. Furthermore, the boundaries defining modules are fuzzy. After all, co-expression in *some* environments does not imply co-expression in *all* environments.

Since the foundational work of Jacob and Monod in the 1960's, several definitions of regulatory modularity have emerged. In the following sections I define each. The following organizational paradigms attempt to group multiple genes into functional units or modules. The definitions vary in terms of the size of modules generated and with respect to assumptions regarding their basic purpose. It is important to remember that no single definition is correct; rather, each is a contrasting lens through which we can understand coordination of genetic expression.

Operon

The original co-regulatory module described by Jacob and Monod [167] is also the most simple. The *operon* simply consists of multiple, adjacent genes that are transcribed as a single, polycistronic transcript. There are many operons in bacterial genomes. *E. coli*, for example, encodes nearly 700 operons [294]. Nearly half of the genes in the *E. coli* genome are in operons. Operons, however, should not be considered dogmatically. Many operons express condition-specific isoforms. While genes of an operon may be co-transcribed in some condition, many contain internal binding and termination sites that result in production of alternative transcripts in some environments [193].

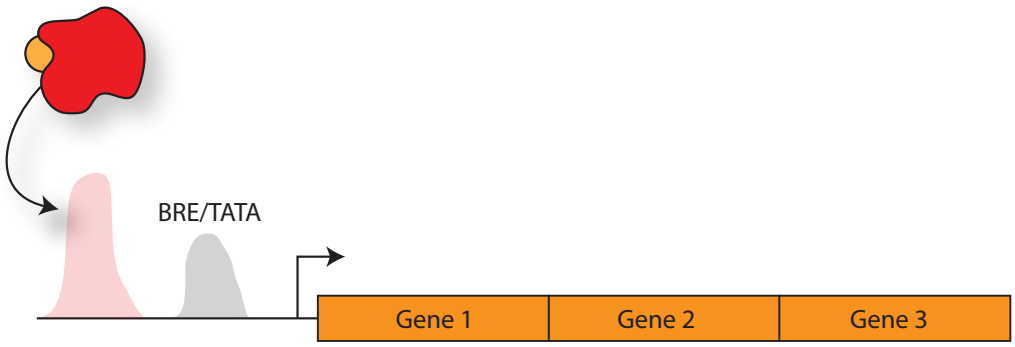


Figure 1.4: Operons are co-regulatory units wherein multiple genes are transcribed as a single polycistronic transcript. Several genes involved lactose utilization and transport were discovered to be co-transcribed as a single co-regulated unit in 1961 by Jacob and Monod [167]

Regulon

The term *regulon* was coined in 1964 to describe regulation of arginine biosynthetic genes by the repressor, ArgR [226]. Unlike operons, genes regulated by ArgR are located throughout the genome, including multiple operons like the arginine uptake system, *art*, and histidine transport genes, *hisJQMP*. A regulon consists of genes regulated by a single, common TF (like ArgR). Like operons, there are many regulons in prokaryotes - one for every TF. *E.*

coli has 83 annotated regulons [262]. The number of genes in a regulon depends on the TF. Sizes can range from several genes to many hundreds, including operons. Several databases collate information about regulons in prokaryotes [11, 263, 262]. Notably, RegRecise uses evolutionary conservation to refine regulon predictions across multiple species [262].

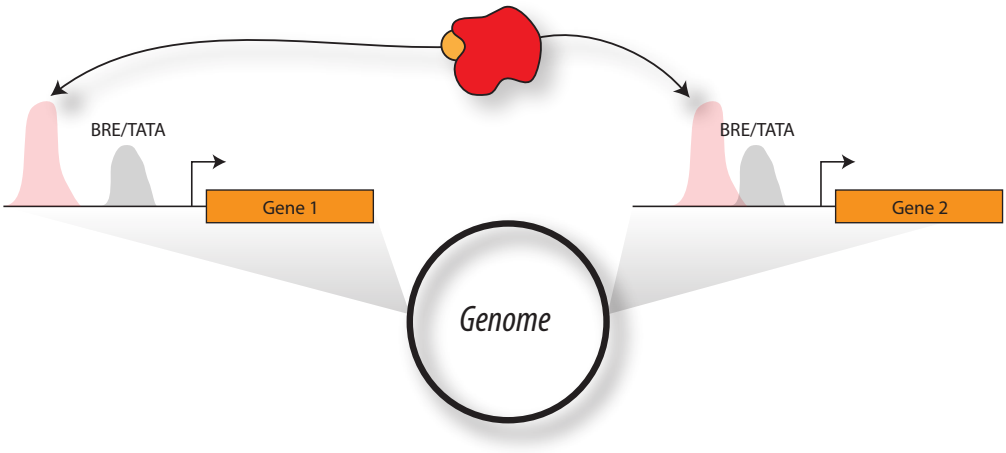


Figure 1.5: A regulon consists of multiple genes or operons located throughout the genome that are co-regulated by a common transcription factor. The ArgR regulon was first described in 1964 by Maas and Clark [226].

Modulon

A *modulon* is a collection of operons and regulons that are regulated by a common pleiotropic TF in addition to different specific TFs. Canonical examples include the CAP modulon, a nutrient-responsive modulon that affects the *lac* operon as well as the arabinose catabolism operon (*ara* operon), and the FNR modulon [138], which responds to anaerobiosis. Like regulons, all of the genes of a modulon are regulated by at least one common TF, generally a global regulator (i.e., a TF that regulates many genes, typically on account of its low sequence specificity).

Stimulon

A *stimulon* consists of operons and regulons that are regulated by the same stimulus. Well known examples include the yeast H_20_2 stimulon [133] and the *B. subtilis* heat shock stimulon [303]. Stimulons include all of the biological pathways and processes that respond to the same environmental change. As such, they can be large, sometimes including thousands of genes. Stimulon genes can increase as well as decrease in expression.

Corem

Co-regulated modules or *corems* are the condition-specific modules discovered by EGRIN 2.0. Unlike other definitions of modular genetic regulation, corems can be regulated by multiple, independent TFs. They can contain subsets of operons and regulons, reflecting condition-specific generation of multiple transcriptional isoforms from an operon. They can include genes from multiple regulons and operons. They range in size from small (3 genes) to large (100s of genes). Corems also vary in how often they are co-expressed, from rare ($<1/10$ of the observations) to common ($>2/3$ of the observations). Importantly, an individual gene can belong to *multiple* corems. A summary of corem statistics for *E. coli* and *H. salinarum* corems is available in Figure 2.6. Defining hallmarks of a corem are depicted in Figure 1.6. The generation, properties, and utility of corems will be developed throughout the text.

1.3 Networks in biology

Networks (or graphs) are mathematical representations of the relationships between objects. Networks consist of nodes (objects) and edges (relationship between the objects). Mathematically, we refer to this entity as a graph, $G = (N, E)$, where the graph, G, is an ordered pair comprised of nodes, N, and edges, E, which themselves are two element subsets of N. They can be represented using a (weighted) adjacency matrix, where each entry $N_{i,j}$ indicates a (weighted) relationship between nodes i and j.

Networks have become popular to represent biological information, especially from high-throughput screens and experiments. Regulatory influences, protein-protein interaction,

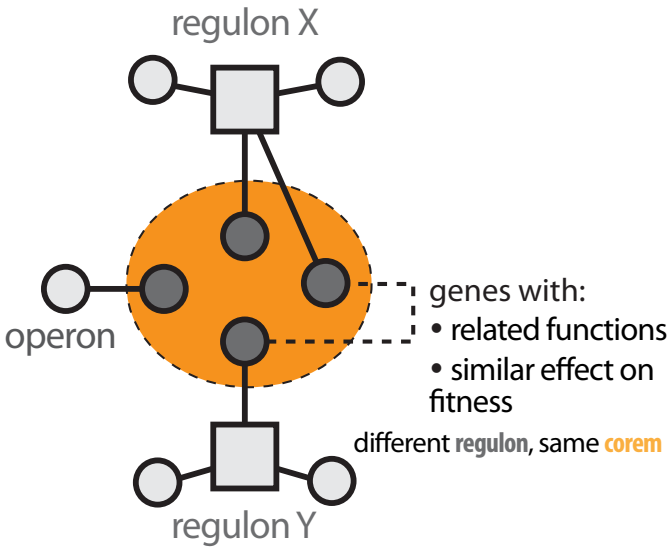


Figure 1.6: Corems are condition-specific co-regulatory modules discovered by the gene regulatory inference algorithm EGRIN 2.0. Corems can contain subsets of genes from multiple regulons and operons. Each corem has evidence for co-regulation in the form of *cis*-regulatory motifs called gene regulatory elements or GREs. GREs suggest that some corems are regulated by independent factors. Generation and analysis of corems will be described throughout the text.

and metabolic processes can be represented as networks, where nodes represent regulators, genes, proteins, metabolites or any other biological entity, and edges, which connect nodes, represent arbitrary interactions between the nodes (Figure 1.7). A variety of biological networks have been generated, including protein-protein interaction networks [304], gene regulatory networks [50], metabolic networks [123], even literature citation networks [360]. Figure 1.7 summarizes important properties of networks. Formalizing biological interactions as networks has a number of analytical advantages. Besides simplifying the representation of biological interactions, graphs have a long, well studied history. Structuring biological relationships in a graph gives access to richly developed tools of graph theory. Conveniently, many of the mathematical concepts and statistical measures developed for abstract graph structures also apply to biological networks. A graph can be analyzed mathematically to

reveal characteristics of its topology. Since the structure of interactions in a network is oftentimes directly related to dynamical properties of that network, topological analysis can give insight into the organization and function of biological systems. Important topological features of biological networks include community structure (e.g., biological modules), hierarchical organization (e.g., nearly power-law degree distributions), and overrepresented network motifs (e.g., the three gene feed-forward loop). These topological features of biological networks and their consequence for the function of biological systems will be explained in more detail throughout the text.

1.3.1 Gene regulatory networks (GRNs)

Regulatory interactions can be cataloged, visualized, and analyzed as gene regulatory networks (GRNs). GRNs can encode who regulates whom, when, where, and to what extent. Analysis of GRNs has revealed valuable insight into how cells process information [30] and control expression of their genomes.

GRN sub-networks that respond to environmental change range from simple to complex. In the simplest case, two-component relay systems directly couple environmental sensing to gene regulation through activation of a TF (e.g., osmoregulation by EnvZ/OmpR two-component system in *E. coli* [7]). Most natural environments, however, change in more complicated ways. Microbes must decipher many overlapping signals, some of which are prone to high levels of noise. To make the task even more complicated, genetic circuits are not isolated from one another. Even simple relay systems exhibit cross-regulation with other regulatory circuits and can affect other cellular components indirectly [203]. To achieve robust response to complicated environmental signals, cells have evolved mechanisms to handle errors, cross-talk, and noise. This accomplished, in part, by encoding gene regulation in combinatorial, modular regulatory circuits that confer robustness to the system.

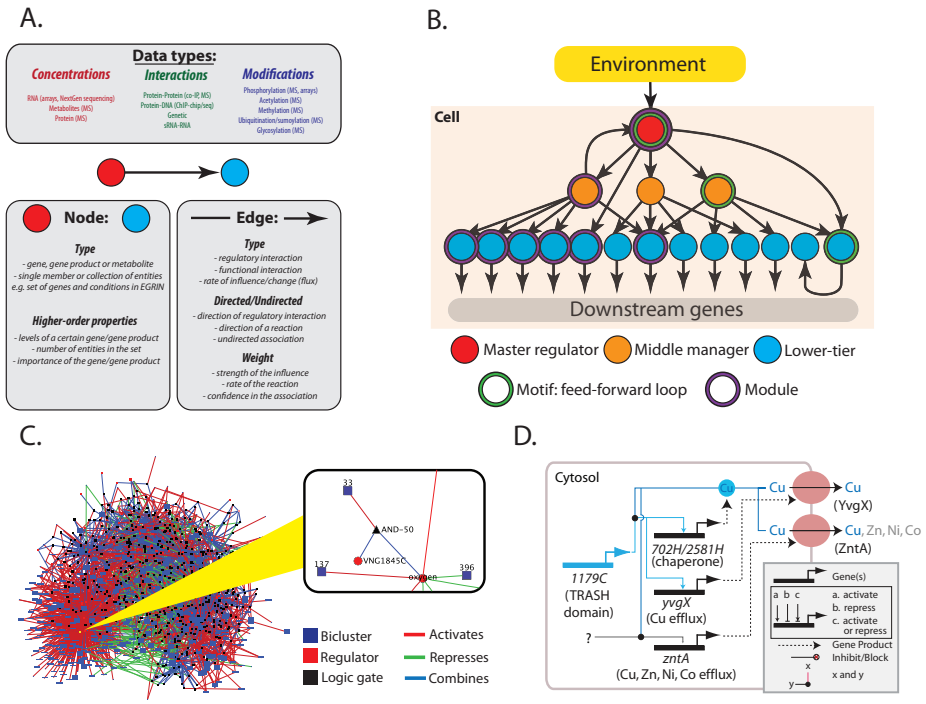


Figure 1.7: (A) The fundamental units of a network (or graph) are nodes, and edges. Three types of biological information are commonly represented as a network: transcriptional, metabolic, and protein-protein interactions. (B) Biological networks share common features. (1) hierarchy: transcriptional networks are close to scale-free in the distribution of regulatory connections and exhibit hierarchical arrangement of connections [30]. (2) modularity: biological networks aggregate pathways and functions into modules [147]. Here, we denote a transcriptional module by a purple ring surrounding the nodes in the module. (3) motifs: interesting dynamic behaviors of gene circuits are often mediated by particular wiring of the parts, which defines a network motif. The members of a particularly well-studied network motif, the feed-forward loop, are depicted in this illustration by green circles surrounding the nodes [13]. (C) Biological networks learned from experimental data, such as the Environmental and Gene Regulatory Influence Network (EGRIN) for *H. salinarum sp. NRC-1*, contain many layers of information that can be mined to aid hypothesis generation [50].

1.3.2 Discovery and Inference of GRNs

Obtaining accurate information about regulatory interactions is critical to draw conclusions about biological function from network structure. Considerable effort over the past decade has been dedicated to obtaining accurate GRNs. These efforts fall into two major categories: experimental and computational approaches. The two are not mutually exclusive; in fact, computational approaches always build from experimental observations. The two approaches do, however, vary in the amount of time and money required to obtain comprehensive and accurate GRNs.

Experimental Approaches

Experimental data is the starting point for all GRN reconstruction. An annotated genome is a minimum requirement. The units of the GRN - in this case genes - must be defined. Regulatory proteins (TFs) must also be defined. For newly sequenced genomes, this is accomplished by searching within the new genome for DNA binding domain (DBD) homologies (described above, 1.2.4) [49]. Those genes with DBDs are considered putative TFs.

After basic genome annotation, several approaches for GRN reconstruction exist. The most obvious involves direct measurement of TF binding events across the genome using ChIP-chip or ChIP-seq. Each binding event in the promoter region of a gene (observed after filtering at some confidence threshold) can be considered an edge between a TF and gene in the GRN. Other experimental methods, like *in vitro* binding of purified TF to a promoter sequence can be used to support the assignment. While attractive, such approaches suffer from two complications: (1) they require separate experiments for each TF. This, in turn, requires developing a purification strategy for each TF, typically using genetic modification (e.g., HA-tag). For bacterial genomes, this would require at least 100 separate strains and experiments for a complete GRN. (2) More troubling, is that observed TF→gene binding is condition-specific. This means that experimentally-based GRNs only reflect the conditions in which the data were collected. For interaction-based assays (like ChIP-chip or ChIP-

seq) to be comprehensive, they would need to be measured in many conditions. Designing experiments to test all pairs of TF and relevant conditions would lead to combinatorial explosion. New technologies like DNase I hypersensitivity [83] promise to overcome some of these challenges by measuring all bound sites throughout the genome in a single experiment. Currently, however, it is difficult to resolve which TF is bound to a particular fragment without *a priori* knowledge of the binding preferences of TFs. Other challenges exist for using these methods in bacterial genomes.

Several resources have been developed to share experimental knowledge of GRNs across laboratories. RegulonDB, for example, is a database of all evidence for transcriptional regulation collated for *E. coli* [293]. The collection of data in RegulonDB represents over 50 years of experimental work to characterize the GRN from a single organism. This highlights how challenging it is to reconstruct GRNs directly from experimental data. In the absence of technological innovation, GRN reconstruction in newly sequenced species through experimental approaches alone will be time and cost prohibitive. Many investigators, therefore, have considered alternative approaches to rapidly infer GRNs, especially for understudied organisms. Usually these alternatives employ computational inference methods.

Computational Approaches

Computational approaches attempt to reverse engineer GRNs directly from experimental data. While such methods are always underpowered (i.e., there are too few observations for the number of variables in the system), computational inference methods have made significant improvements, greatly increasing predictive accuracy. One of the biggest gains for these methods came from including biological priors (i.e., using biological knowledge to assist inference). Especially in understudied organisms, computational inference is the only viable alternative to costly, brute force experimentation for GRN reconstruction.

The primary source of data used for most computational approaches is gene expression. Microarray and RNA-seq allow for quantitation (absolute or relative) of every gene in the genome with a single experiment. To deduce a GRN from these measurements, inference

algorithms universally assume that these expression patterns result from some reproducible genetic "program". The aim of computational methods is to use data to figure out the program. From this perspective, biological inference is also a problem of pattern recognition (constrained by biological mechanism). Given a sufficient number of observed expression profiles (e.g., different experimental conditions), one can deduce an underlying network that produced them (at least one of many networks that is consistent with the data). The promise of these approaches is that gene expression data is relatively easy and cheap to collect, even in understudied organisms.

Full description of computational methods used to infer gene regulatory networks will be provided in Chapter 2 It should be noted that GRN inference methods vary substantially: from correlation to causal, and from statistical to information-based. De Smet *et al.* provide a comprehensive review of GRN inference methods, highlighting the advantages and shortcomings of each [90].

1.4 Systems-level view of gene regulatory organization

GRN analysis reveals valuable insight into how cells process information. A systems perspective is essential for deciphering how microbes coordinated expression of their genomes. Just as genes work in concert to mediate environmental responses, multiple TFs work together to coordinate the dynamic activity of GRNs. GRNs are impossible to characterize from the perspective of a single TF.

Early topological investigations of GRNs revealed several insights. First, not all regulators are equivalent; few transcription factors (TFs) regulate many genes, whereas the majority of TFs regulate far fewer downstream genes (Reviewed in [30]). Second, GRNs are hierarchical. "Master regulator" TFs, for example, initiate regulatory cascades. While a "master regulator" many only interact with few TFs, it can initiate a regulatory cascade by influencing activities of "middle managers" that propagate its signal to "lower tier" regulators, directly controlling specific processes [382]. Third, some patterns of interactions are statistically overrepresented in biological networks – these overrepresented subgraphs

are called motifs ([248], Reviewed in [13]). Below I elaborate on the dynamical properties of these features. Finally, abstract representations of gene regulation can yield meaningful dynamic insight into biological processes. Examining the cell-cycle regulatory network as a Boolean representation, for example, reveals the existence of biological attractor states that contribute to the robustness of the cell cycle [214].

As mentioned previously, interpretation of GRNs depends critically on their accuracy. Without correct annotation of TF→gene relationships, it is impossible to know which TF-gene pairs, for example, are in feed forward loops or which TFs are "master regulators". Network topology is critical to understand network dynamics. In the following sections, I consider the dynamical properties of GRN topological features.

1.4.1 Network Motifs

Network motifs are frequently appearing topological relationships in directed GRNs. By cataloging frequencies of three gene motifs in *E. coli*, Uri Alon and his colleagues discovered over representation of certain types of motifs in natural GRNs, including feedforward and feedback architectures [248]. These regulatory motifs possess interesting dynamic behaviors, which help cells adjust to changes in their environment. Feedforward loops, for example, buffer noise in highly deterministic biological processes, such as in development [233, 213], while feedback loops provide local sensing and specific response to chemical changes, such as product inhibition exerted in metabolic pathways [256]. Figure 1.8 provides a graphical representation of the dynamic properties of the feedforward loop. While complete enumeration of all possible three gene architectures made the study by Alon and colleagues possible [248], there are likely important higher-order motifs in biological systems as well, including global motifs consisting of many component sub-motifs (i.e., motifs of motifs, etc), that may be difficult to assess statistically because of exponential increase in possible topological permutations.

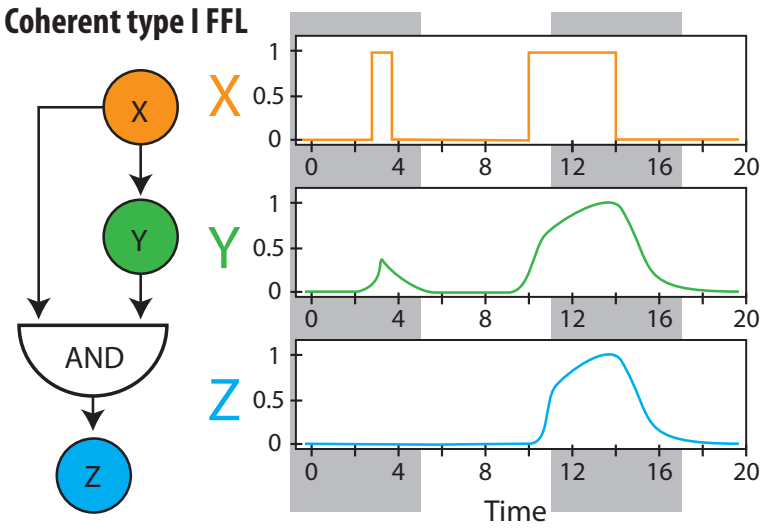


Figure 1.8: The coherent feed-forward motif is a well studied network motif that occurs commonly in biological networks. Like other network motifs, the arrangement of genes into a coherent feed-forward loop has interesting dynamical properties. Feed-forward loops have been shown to buffer noise in biological systems, as well as confer rapid response once a critical signal threshold is reached [233]. The cartoon representation depicts downstream expression of Z in response to either a transient or sustained pulse of X (adapted from [233]).

1.4.2 Regulatory Logic

The complete set of TF→gene interactions in the cell composes a complex regulatory network. This network sets upper and lower bounds on possible expression levels for genes. Dynamic execution of the network leads to coordinated expression of the genome that has evolved to match the sensed environment. Interaction topologies that make up GRNs range in complexity from simple single input motifs (e.g., SIM), to more complicated three gene motifs (e.g., feed forward loop, above), to higher-order architectures involving multiple genes and composed of multiple three-gene (and larger) component motifs (e.g., integrated feedforward loop). These interactions can even compute logic functions, like *and*/*or* gates. Figure 1.9 highlights the range of complexity in regulatory logic.

From the standpoint of systems biology, this topological framework helps us understand which genes and functional processes are coordinated in which environments. Observationally, the output of these complex regulatory circuits are co-expression patterns observed in gene expression data. The goal of this project was to group these co-expression patterns into biologically meaningful modules, determine what TFs and regulatory logics were responsible for their generation, and to understand when and why they are co-expressed.

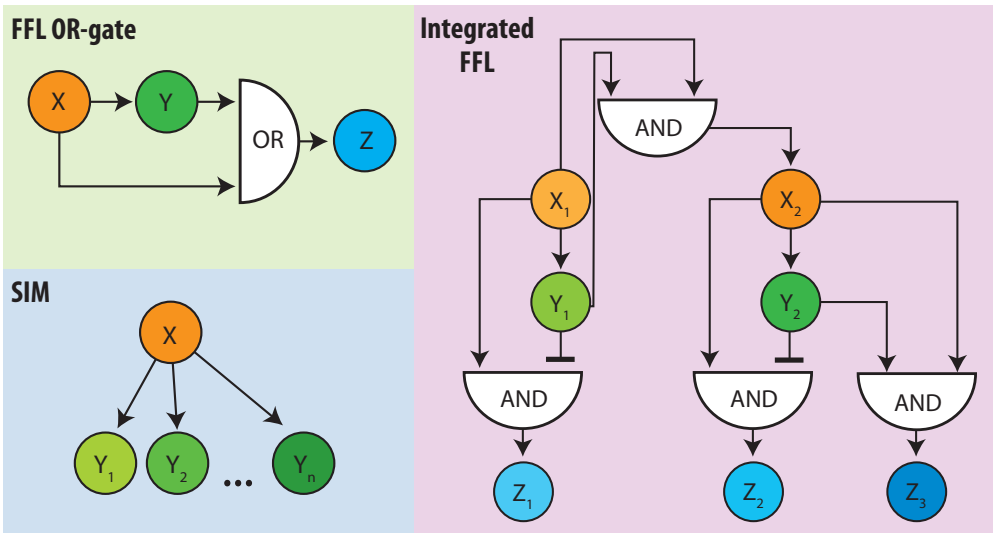


Figure 1.9: Gene regulatory logic ranges from simple to complex. Three gene regulatory circuits highlight varying degrees of complexity. Note the use of logic gates using *and*/*or* gates. Ultimately, all regulatory sub-circuits (even the simple SIM) are embedded in much more complicated regulatory networks.

1.4.3 Beyond the operon and regulon

In the 1996 edition of *EcoSal*, Neidhardt and Savageau wrote a chapter about regulation of multigene systems entitled, "Regulation Beyond the Operon" [256]. The chapter explored concepts of gene expression beyond simple co-linear arrangement of genes in an operon,

starting with the regulon and moving to more complex relationships described by modulons and stimulons (described above). While it was clear from the data that genes were co-expressed in complex, condition-specific arrangements, it has been less clear how these patterns emerge from GRN activity. Unresolved questions included: Do genes have to be regulated by a common factor to be consistently co-expressed across environments? What is the role of combinatorial regulation in prokaryotes? Are genes regulated by a common factor always co-expressed? This project attempts to quantify the relationship between gene regulatory modules, GRNs, and the nuanced condition-specific associations between TFs and genes that generate condition-specific co-regulatory modules - directly from data.

1.5 Chapter Organization

The following chapters investigate the organization, function, and evolution of GRNs in prokaryotes. I develop a comprehensive view of data-driven GRN inference, from model construction (Chapter 2) to interpretation (Chapter 3) and implications (Chapter 5). I explore how the structure and function of GRNs contributes to our understanding of how these networks evolve (Chapter 4) .

Chapter 2 discusses computational methods for gene regulatory network inference from genome sequence and large gene expression compendiums. It also describes construction of EGRIN 2.0, the network ensemble elaborated throughout the text. In addition to complete description of the algorithm, many additional details are documented, including experimental data sets used, benchmarks employed to evaluate model performance, and web-framework developed to facilitate exploration of the model's predictions. Chapter 3 focuses on biological interpretation of EGRIN 2.0. In particular, it documents the model-assisted discovery of *corems*, condition-specific co-regulatory modules that influence cellular fitness. 4 shifts focus to the evolution of GRNs. I consider how dynamics of environmental change shapes the way GRNs evolve. Finally, 5 investigates the consequences and suggests directions forward, both for improving GRN inference and in context of evolution.

Chapter 2

COMPUTATIONAL APPROACHES TO RECONSTRUCT GENE REGULATORY NETWORKS

A challenge of systems biology has been inference of comprehensive and accurate gene regulatory networks (GRNs) directly from genome sequence and transcriptome data. In this chapter, I describe computational approaches for reverse engineering accurate GRNs from gene expression data. There are diverse approaches to the problem: from information theoretic to correlational to integrated. I focus on two algorithms that play a central role in the work described here: cMonkey [284] and Inferelator [47], as well as the model derived from integrating the two, the Environment and Gene Regulatory Influence Network or EGRIN [50]. Following introduction of these components, I discuss methods developed specifically for this dissertation, an ensemble of EGRIN models. I review the history and motivation for ensemble modeling. I document all of the methods, data, and analyses used to construct models for *H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655 . Additionally, I provide validation for the model's predictions, as well as evaluation of its robustness. The details described in this chapter are the foundation from which I derive biological insight in the following chapters.

This chapter has been modified from the supplement to:

Brooks AN*, Reiss DJ*, Allard A, Wu W, Salvanha DM, Plaisier CL, Chandrasekaran S, Pan M, Kaur A, Baliga NS. (2014) A system-level model for the microbial regulatory genome. *Mol Syst Biol.* 10: 740.

* Indicates equal contribution

Chapter Highlights

- Many algorithms for GRN inference exist. Methods vary significantly in their approach.
- cMonkey integrates multiple sources of biological data, including gene expression, to identify condition-specific co-regulatory modules called biclusters
- Inferelator predicts which TFs influence condition-specific regulation of these biclusters using regression and variable selection
- Together, the two algorithms generate an Environment and Gene Regulatory Influence Network (EGRIN)
- Ensemble network inference in EGRIN 2.0 improves performance by reducing model variance and allowing detection of rare co-regulatory events
- We derived biological insight from the ensemble model by applying network-based methods for backbone extraction and link community detection
- Extensive support for model predictions from independent experimental validation data
- EGRIN 2.0 outperforms other GRN inference methods

2.1 Summary

Many methods exist for reconstructing GRNs from transcriptomic data. Integrating additional supporting data can assist network reconstruction. cMonkey is a stochastic semi-supervised machine learning algorithm for inferring GRNs that uses *cis*-regulatory motif detection and biological support in known functional networks to constrain model selection.

Inferelator predicts influence of TFs on the expression of cMonkey biclusters. Running cMonkey and Inferelator in an ensemble framework greatly improves performance. The resulting EGRIN 2.0 model is more accurate and reveals genetic mechanisms through which bacteria reuse gene modules across diverse environments to enhance fitness.

2.2 Existing computational approaches

There are many approaches to GRN inference. So many, in fact, that entire reviews have been written about them [28, 90]. There has even been a formal competition between algorithms. While full description of all approaches is outside the scope of this manuscript, I briefly cover some of the more popular methods, especially those that factor into the present study.

Methods for GRN inference vary significantly in their fundamental approach to the problem. While each tries to infer an accurate and comprehensive GRN directly from gene expression data, many of the methods take fundamentally different approaches. Some methods use regression (e.g., Inferelator described extensively below), some correlation (e.g. WGCNA); some are information based (e.g., CLR and ARACNE), whereas others are Bayesian; others still are integrated, using additional data to perform inference (e.g., cMonkey) or use other methods entirely (e.g. Genie3 [161]).Newer methods, including our own and DREAM5, aggregate across multiple runs of the same algorithm (e.g. EGRIN 2.0, the method developed here) or integrate across multiple, heterogeneous methods (e.g, DREAM5).

CLR Context likelihood of relatedness, CLR [108], is a mutual information based approach that builds on relevance networks [101, 65]. CLR (like relevance networks) calculates mutual information between each TF and every gene in the genome using mutual information (MI), where mutual information is a measure of the statistical dependence between two variables. The resulting network can be cut at some threshold MI value to produce a GRN (only TF→gene above the threshold are retained). CLR adds to relevance networks by introducing an adaptive background correction step that tries to assess how significant each

TF→gene MI value is in the context of every other possible interaction for either that TF or gene. Only those that MI values that are statistically distinguishable from background are retained. This step is designed to remove indirect dependencies in the network to retain only direct TF→gene interactions.

ARACNE ARACNE is also an information theoretic approach to reconstructing GRNs [239]. It uses MI to infer TF→gene interactions as well. Unlike CLR, it uses the data processing inequality (DPI) to remove indirect dependencies. DPI removes the lowest MI from every every gene triplet in network above a certain threshold. The basic idea is that direct interactions will have a higher MI. Thus in the following three gene scenario: $A \rightarrow B \rightarrow C$, $MI_{A,C} < MI_{B,C}$.

WGCNA Weighted gene correlation network analysis, WGCNA, computes pairwise correlation (weighted or unweighted) between every pair of genes followed by hierarchical clustering and dynamic tree cut to define gene co-expression modules [202]. Unlike ARACNE and CLR, WGCNA does not specifically predict TF→gene interaction, rather it produces gene co-expression modules (like the cMonkey component of EGRIN, described below).

DISTILLER DISTILLER, like cMonkey, produces biclusters (sets of gene co-expressed in subsets of the experiments) by integrating gene expression data with interaction data from RegulonDB [208]. DISTILLER is one of the algorithms we compare our method against (given their similarities). We obtained *E. coli* expression data from this paper.

DREAM5 Dialogue for Reverse Engineering Assessments and Methods (DREAM) is an annual competition for computational methods. In 2012, a group running the competition made a community network of all the top scoring entries for a GRN inference challenge. As described below, this is a type of ensemble model (similar in spirit to our own). The group integrated the predictions from various methods using Borda count election method. Basically, they retained the TF→gene predicted by the most algorithms.

EGRIN Environment and Gene Regulatory Influence Network (EGRIN) uses biclustering in addition to regression and variable selection to infer both gene modules (biclusters) and specific TF→gene influences [50]. The approach consists of the algorithms cMonkey [284] and Inferelator [47]. Since these algorithms are the basis for this dissertation work, they will be described in great detail below.

A comprehensive review of available methods and comparison between them is available in [90] and the supplement to [237].

2.2.1 Integrated biclustering using cMonkey

cMonkey introduction

The cMonkey integrated biclustering algorithm was described and fully benchmarked in [284]. In short, the algorithm computes putatively co-regulated modules of genes over subsets of experimental conditions from gene expression data, constrained by information provided by genome sequence (*de novo* identification of conserved *cis*-regulatory motifs in gene promoters), and functional association networks. Its defining characteristic is that it combines all three types of data (expression, sequence and networks)together into an integrated model that uses a stochastic optimization procedure to identify modules that best satisfy all three constraints, simultaneously.

The cMonkey integrated biclustering algorithm identifies groups of genes co-regulated under subsets of experimental conditions, by integrating various orthogonal pieces of information that support evidence for their co-regulation, and optimizing biclusters such that they are supported by one or more of those additional constraints. The three sources of evidence for co-regulation leveraged by cMonkey to score gene clusters are (1) tight co-expression in subsets of available gene expression measurements (similarity of expression profiles); (2) quality of *de novo* detected *cis*-regulatory motifs in gene promoters (putative co-binding of common regulators); and (3) significant connectivity in functional association or physical interaction networks (co-functionality). The algorithm served as the cornerstone

for the construction of the first global, predictive Environmental Gene Regulatory Influence Network (EGRIN) model for *H. salinarum* sp. NRC-1 [50], and has now been applied to many additional organisms (*e.g.*, [381] and unpublished).

To run cMonkey as part of an ensemble-based inference approach required significant updates to the cMonkey algorithm. These updates primarily addressed computational inefficiencies that led to long runtime. The primary algorithm modification in the new implementation is global optimization (rather than the local, individual cluster optimization utilized by the original procedure). Additional algorithm updates include changes to the individual scoring scheme for subnetwork clustering, as well as integration of the scores. All of these changes improved the procedure's runtime without significantly affecting the algorithm's performance.

Detailed cMonkey algorithm description

The cMonkey algorithm initiates by seeding k biclusters, typically using the simple, widely-used and effective k-means clustering on the input expression data set. cMonkey itself performs a global optimization, in many ways similar to the k-means clustering algorithm, which we used as a model. After beginning with an initial assignment of each gene into k clusters and a chosen distance metric, the basic$k$-means algorithm iterates between two steps until convergence: (1) (re-)assign each gene to the cluster with the closest centroid and (2) update the centroids of each modified cluster. The updated cMonkey algorithm performs an analogous set of moves with four primary distinctions: (1) the distance of each gene to the "centroid" of each cluster is computed using a measure that combines condition-specific expression profile similarity, *cis*-regulatory motif similarity, and connectedness in one or more gene association networks; (2) each gene can be (re-)assigned to more than one cluster (default 2); (3) at each step, conditions (in addition to genes) are moved among biclusters to improve their cohesiveness; and (4) at each step, genes and conditions are not always assigned to the most appropriate clusters. We now elaborate upon these four details.

cMonkey begins each iteration with a set of bicluster memberships $\{m_i\}$ for each element

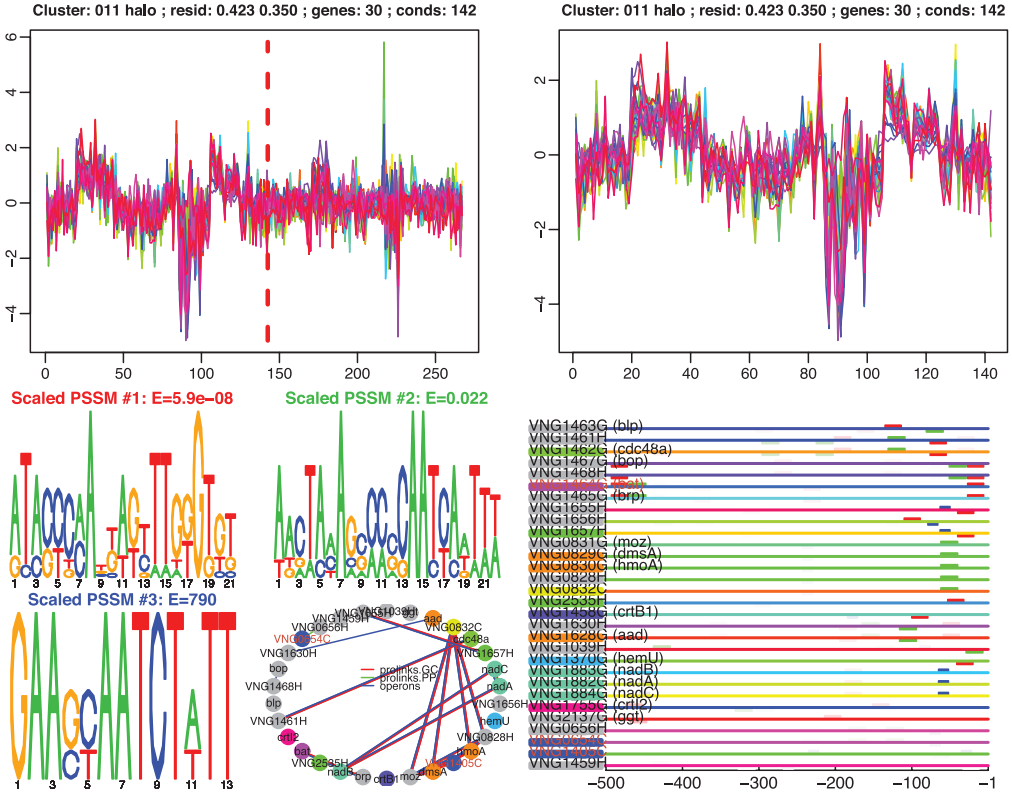


Figure 2.1: cMonkey produces a set of biclusters (∼300), each of which consists of a group of genes that are co-regulated in some (but not all) of the experiments. Typical cMonkey bicluster is shown here. Top-left: Standardized gene expression values for each gene in the bicluster. Vertical red line separates conditions that have been excluded from the bicluster. Top-right contains an expanded view of only the conditions in which the genes are co-expressed. cMonkey also uses *de novo* motif detection to support bicluster assignment. Upstream regions of bicluster genes are searched for shared *cis*-regulatory motif by MEME [22]. Similarity of the genes in annotated functional networks (e.g. STRING [328]) is also scored. Detected motif and string network shown at bottom-left. Location of the motif in gene promoters shown bottom-right.

(gene or condition) i, where by default $\|m_i\| = 2$ for genes and $\|m_i\| = N_c/2$ for conditions (N_c is the number of conditions, or measurements, in the expression data set; note that for

standard k-means clustering, $\|m_i\| = 1$ for genes and $\|m_i\| = N_c$ for conditions). cMonkey then computes score matrices (log-likelihoods, in practice) $\mathbf{R}_{ij}$, $\mathbf{S}_{ij}$, and $\mathbf{T}_{ij}$, for membership of each element i in each bicluster j, based upon, respectively, co-expression with the current gene members ($\mathbf{R}$), similarity of motifs in gene promoters ($\mathbf{S}$), and connectivity of genes in networks ($\mathbf{T}$). For the network scores ($\mathbf{T}$), the originally published procedure [284] computed a p-value for enrichment of network edges among genes in each bicluster using the cumulative hypergeometric distribution. This computation was inefficient, and moreover could not account for weighted edges in the input networks, so we replaced it with a more standard weighted network clustering coefficient [358], evaluated only over the genes within each bicluster.

Following computation of the individual component scores, cMonkey computes a score matrix $\mathbf{M}_{ij}$ containing the integrated score (a weighted sum of log-likelihoods) supporting the inclusion of gene i in bicluster j. At this stage the updated version of cMonkey then computes a cumulative density distribution from each bicluster's $\mathbf{M}_{.j}$ to obtain a posterior probability distribution p_{ij}, that each element i should be in each cluster j, which is used to classify cluster members based upon these scores. The width of the density distribution kernel is set dynamically to be larger for smaller (fewer gene) biclusters, so as to increase the tendency to add genes to small biclusters, rather than to remove them. In the updated procedure, we then add a small amount of normally-distributed random "noise:" to the scores $\mathbf{M}_{ij}$, in order to achieve a similar type of stochasticity to the original version of the algorithm (which was originally obtained using sampling, and which helps prevent the algorithm from falling into local minima; this noise decreases during the run to zero at the final iteration). The result of this noise is that at the beginning of a cMonkey run, biclusters are rather poorly defined (co-expression, for example, is weak), but during the course of a full set of 2,000 iterations, as this noise is decreased, the biclusters settle into minima.

Finally, at the end of each iteration, cMonkey chooses a random subset of elements (genes or conditions) i, and moves i into bicluster j if, for any biclusters j' which it is already a member, $p_{ij} > p_{ij'}, \forall j'$, and out of the corresponding worse bicluster j' for which

$p_{ij} > p_{ij'}$. Thus, as with the k-means clustering algorithm, the updated cMonkey performs a global optimization of all biclusters by moving elements among biclusters to improve each element's membership scores.

cMonkey software availability

The cMonkey software is available as an open-source R package [163]. With this package the algorithm can be easily applied to nearly any sequenced microbial species (given user-supplied expression data). The package automatically downloads and integrates genome and annotation data from various external sources, including RSA-tools [346]; Microbes Online [11]; and EMBL STRING [328]. Additionally, the package can generate interactive web-based and Cytoscape output [309], allowing users to explore the resulting modules and motifs in the context of external data, software, and databases via the Gaggle [310]. Examples of automatically generated output are available at the cMonkey web site. Supplementary R packages with example expression data for organisms including *H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655 are also available from the cMonkey website.

2.2.2 Assignment of putative regulators and influence using Inferelator

Inferelator introduction

The Inferelator algorithm is a method for deriving genome-wide transcriptional regulatory interactions from mRNA expression levels [47]. Inferelator is a direct inference procedure [246]. It models transcriptional regulation as a kinetic process, incorporating time information, when available, and a user-defined time constant. Inferelator uses standard regression and variable selection to identify transcriptional influences on genes or biclusters based on their mean expression levels. These influences include expression levels of TFs, environmental factors, and interactions between the two. The procedure simultaneously models equilibrium and time course expression levels. Thus both kinetic and equilibrium expression levels may be predicted by the resulting models. Through explicit inclusion of time and gene knockout information, the method is capable of learning causal relationships. The

inferred network is a predictive model comprised of linear combinations of expression profiles of various transcriptional regulators, that can predict global expression under novel perturbations with predictive power similar to that seen over training data [47].

Detailed Inferelator algorithm description

Given an input list of p putative transcriptional influences $\mathbf{X} = x_1, x_2, \cdots, x_p$ and the mean expression levels y_i of a bicluster k (over the conditions i included in the bicluster), we model the relationship between y_i and the influences $\mathbf{X}$ by the kinetic equation:

$$\tau \frac{dy_i}{dt} = -y_i + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{2.1}$$

In the steady state scenario, $dy/dt = 0$ and Eq. 2.1 simplifies to

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij},$$

and for time series measurements, we approximate Eq.2.1 as:

$$\tau \frac{y_{i+1} - y_i}{t_{i+1} - t_i} + y_i = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Clearly not all p putative influences $\mathbf{X}$ influence a given bicluster, so we use the elastic-net [389] for variable selection. This involves performing the minimization:

$$\vec{\beta} = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} \boldsymbol{w}_i (\lambda_1 |\beta_j| + \lambda_2 \beta_j^2) \right\} \tag{2.2}$$

subject to a constraint which is a tuneable combination of the L_1 (LASSO) and L_2 (Ridge) regression constraints:

$$\sum_{j=1}^{p} |\beta_j| \leq \lambda_1 |\beta_{\text{ols}}| \qquad (L_1 \text{ constraint}),$$

$$\sum_{j=1}^{p} \beta_j^2 \leq \lambda_2 \beta_{\text{ols}}^2 \qquad (L_2 \text{ constraint}).$$

The w_i in Eq. 2.2 allow different variables (β's, in this case) to be selectively constrained. For this work, we set all $w_i = 1$, *i.e.*, no differential constraints. Redefining the constraint, such that:

$$\vec{\beta} = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} \boldsymbol{w}_i \lambda \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2/2 \right) \right\} \tag{2.3}$$

defines $0 \leq \alpha \leq 1$ as a tuning parameter between the ridge (L_2; $\alpha = 0$) and LASSO (L_1; $\alpha = 1$) solutions, and λ is the single complexity parameter, which is chosen to minimize the cross-validation error (we use 10-fold cross-validation), exactly as in [47]. Substituting Eq. 2.1 into Eq. 2.3, we obtain the complete equation describing the minimization performed by Inferelator:

$$\vec{\beta} = \arg\min \left\{ \sum_{i=1}^{N} \left(\tau \frac{y_{i+1} - y_i}{t_{i+1} - t_i} + y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \sum_{j=1}^{p} \boldsymbol{w}_i \lambda \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2/2 \right) \right\}. \tag{2.4}$$

For the current implementation, we set $\tau = 10$ minutes for all TFs, and $\alpha = 0.8$ for all biclusters. In the future, we could choose τ and/or α by cross-validation as well. When $\alpha = 0$, there is no constraint, and we get the ordinary least-squares (OLS) solution with all βs non-zero. With $\alpha = 1$, we select the null model. The optimal solution is somewhere in-between, and this is usually the selected solution for each bicluster, usually ~ 6 TFs, on average; although the null model (no solution) is selected for a number of biclusters.

2.2.3 Ensemble methods to decrease model variance

Ensemble methods have become increasingly popular in statistical machine learning. The primary reason for their popularity is the observation that combining predictions reduces model variance, i.e. by combining predictions across multiple models, one can produce an aggregate model that is usually more accurate than the best of its components [307]. For most data-driven problems, there are many possible computational approaches. Ahead of

time one may not know (or have some way to determine) the method that will perform best on the data. By combing predictions into an ensemble, one can guarantee that the aggregate model will perform better than most individual methods. While it may not always outperform the best method, the aggregate model will perform better than the worst approach. This is especially useful when convenient benchmarks do not exist to evaluate model performance (to, for example, identify which methods is performs worst).

Ensemble methods gained popularity during the Netflix Prize. Netflix offered 1,000,000 dollars to anyone who could improve their internal movie recommendation system by 10%. In the end, best performance was obtained by weighing predictions from the top 30 methods [307]. A similar approach was recently taken to infer gene regulatory networks. Multiple investigators submitted algorithms to a competition called DREAM5. The competition organizers collated these into a "wisdom of the crowds" community network that performed better than most (but not all) methods on multiple data sets [237].

Some examples of well-known ensemble approaches for single algorithms include bagging [57], random forests [58], AdaBoost [120], and GradientBoosting [121]. We note that methods for single algorithm ensembles is different than heterogeneous ensembles, although the motivations are similar. The approach we took for EGRIN 2.0 is similar to bagging, i.e. bootstrap aggregation [57]. In the following sections, we detail construction of the EGRIN 2.0 ensemble.

2.3 EGRIN 2.0

2.3.1 Background and motivation

The procedure to infer a single global Environment and Gene Regulatory Influence Network (EGRIN) model from genome-wide data was described previously [50, 47, 284]. In short, the two-step procedure involves running cMonkey once to obtain a single set of ~ 300 biclusters of genes. Genes in these biclusters have tight co-expression over a subset of the measured conditions (usually about half), are supported by common putative *cis*-regulatory motif(s) in their promoters (gene regulatory elements, GREs), and are often substantiated by high

connectivity in functional association networks. Next, given a set of "predictors" (mRNA expression levels of transcription factors and/or quantitative values for environmental factors; *e.g.*, concentrations, growth media, etc.), and the mean expression levels of genes in each bicluster, Inferelator is run to choose a parsimonious subset of those predictors that can accurately predict the expression levels of that bicluster (*i.e.*., those with non-zero β [Eq 2.4]) . Predictors are selected independently for each bicluster. The combined set of TF→bicluster interactions and their associated weights (βs) give the degree of activation (or repression) predicted.

The EGRIN 2.0 modeling procedure updates this process by applying updated cMonkey and Inferelator algorithms (described above) repeatedly to subsets of the available expression data. The end result is an ensemble of EGRIN models, each model containing biclusters and their predicted regulators, tuned to a relatively small subset of the overall input expression compendium. The experimental subsets were selected semi-randomly, with available biological information constraining the selection procedure (*i.e.*., including whole groups of related experiments when one was randomly selected). For *H. salinarum*, we used manually curated metadata about each experiment to group related experiments. Since we did not have sufficient metadata from the public *E. coli* data set, we grouped the conditions based upon individual experiments instead (*e.g.*, time series).

The EGRIN 2.0 inference methodology is an ensemble learning approach, more specifically a form of bootstrap aggregation [57], or sub-bagging. Advantages of sub-bagging include simplicity (*i.e.*, basic model averaging), reduced model variance compared to individual runs [66], and avoidance of overfitting [199]. The power of ensemble learning approaches stems from their ability to average out errors in individual models. For EGRIN models, this feature helps overcome artifacts due to both experimental and algorithmic noise. Incorrect classification in a single model that are not the result of systematic error will re-occur infrequently in subsequent runs. Similarly, overfitting is mitigated by training each individual model on a small subset of the available data. Only consistently re-discovered relationships are considered significant.

Sub-bagging of experimental conditions further allows the model to effectively up-weight a restricted set of conditions for each individual EGRIN model in the ensemble. This forces each EGRIN to model regulatory behaviors present within a more narrow range of conditions. As a result, the individual EGRIN models have the opportunity to discover features that may distinguish highly related responses or occur in a very limited number of conditions in the data set (*e.g.*, conditions, genes, GREs).

To quantify this assumption, we constructed a separate ensemble of 30 EGRIN models trained on the complete *H. salinarum* data set (*i.e.*, 1,495 conditions; no sub-setting performed). We asked how often we would discover a GRE corresponding to the well-characterized anoxic *H. salinarum* TF, Bat. Given frequent detection of the Bat GRE in our full ensemble, we expected to detect ~ 20 instances of the Bat GRE in the new ensemble (*i.e.*, motifs similar to GRE #22; Figure 2.3 [26]). Surprisingly, we did not detect a single GRE matching Bat when all conditions were used for training (data not shown). This is likely because the anoxic conditions in which Bat is active represents only a small portion of the entire data set.

Ensemble-based approaches are being used more frequently in biological data analyses, including random forests (*i.e.*, bags of decision trees) [57], and the recently-published DREAM5 community ensemble of regulatory network predictions [237], which we used as a benchmark in this manuscript to evaluate EGRIN 2.0 predictions for *E. coli* K-12 MG1655 . Moreover, in principle, our approach is similar to the stochastic LeMoNe algorithm [173], which uses Gibbs sampling to learn ensembles of regulatory modules from gene expression data. EGRIN 2.0 is distinguished from LeMoNe and similar algorithms by its ability to predict transcriptional control mechanisms (*i.e.*, GREs) and the conditions in which they operate, both globally and within individual gene promoters.

To construct and mine the EGRIN 2.0 ensemble we utilized additional model aggregation and compilation procedures, including (1) motif clustering [344] and scanning [21] (Section 2.3.4); (2) gene co-regulation network construction and backbone extraction [308] (Section 2.3.6); and (3) network community detection [5] (Section 2.3.6). These methods were used

to identify GREs and their genome-wide locations, gene-gene co-regulatory associations, and corems, respectively. Each of these procedures is described in more detail below. A comprehensive workflow is provided in Figure 2.2.

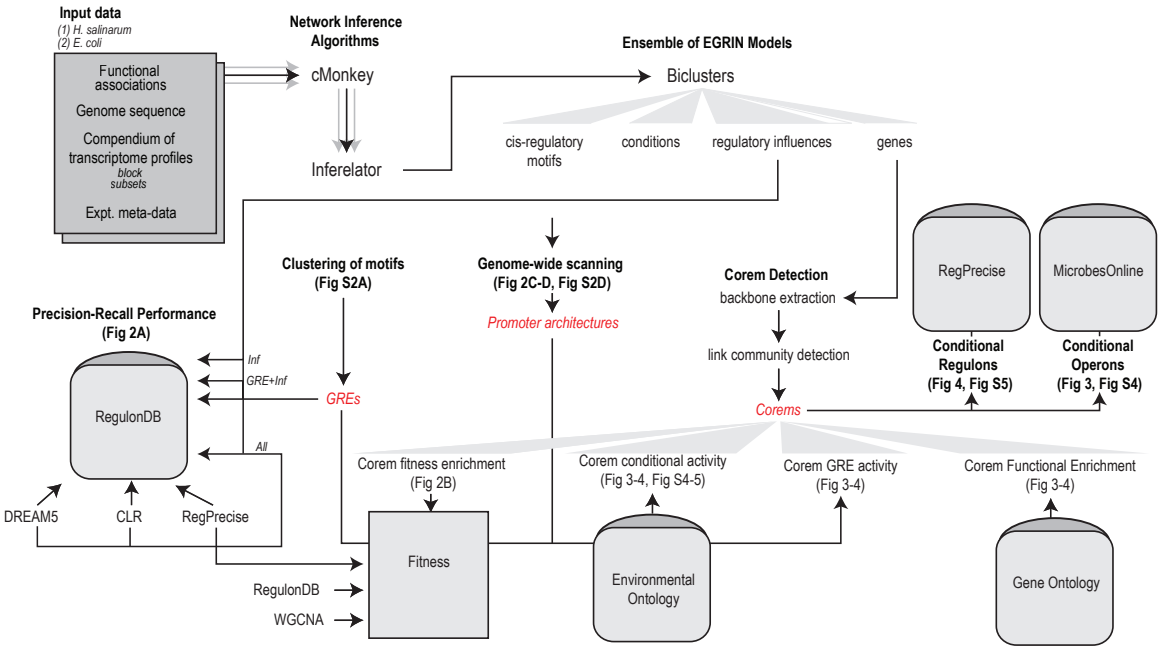


Figure 2.2: Detailed workflow for EGRIN 2.0 inference procedure. Data input, processing and analysis to construct EGRIN 2.0 model for *H. salinarum* and *E. coli*, and predictions generated. Predictions highlighted in individual figures are noted

2.3.2 Experimental Data

H. salinarum *sp. NRC-1 compendium*

A compendium of 1495 transcriptome profiles were collated from a wide array of experiments conducted by our lab over the past decade that cover dynamic transcriptional responses to varied growth (1159 arrays), nutritional (161 arrays), and stress conditions (1102 arrays), including variation in temperature (256 arrays), oxygen (285 arrays), light (786 arrays), salinity (20 arrays), metal ions (274 arrays), and genetic perturbations (643 arrays). We

Statistic	***H. salinarum***	***E. coli***
Arrays analyzed	1495	868
Genes/transcripts	2400	4213
Number of EGRIN models constructed	572	106
Number of cMonkey biclusters	142600	46520
Fraction of "good" cMonkey biclusters	61%	43%
Residual cutoff defining "good" biclusters	0.40	0.55
Number of genes/transcripts in $\leq$ 1000 "good" biclusters (Hal)	2104	
Number of genes/transcripts in $\leq$ 200 "good" biclusters (Eco)		4201
Average number of biclusters per gene	1210	212
Number of motifs (total)	269770	86167
Number of motifs (E-value $\leq$ 1)	118546	15588
Number of motifs (E-value $\leq$ 1e-6)	32739	3506
Number of motif clusters	162	402
Number of motifs contained in motif clusters	37713	13519
Number of unique GREs	135	337
Number of motifs in unique GREs	27991	12773
Number of RegulonDB GREs detected (p $\leq$ 0.01)		53
RegulonDB TFs with $\geq$ 3 experimentally characterized binding sites		86
Corems	679	590
Genes modeled by corems	1363	1572
Genes per corem: min(max)	3(377)	3(153)
Active conditions per corem: min(max)	21(1279)	69(598)
GREs per corem: min(max)	0(9)	0(12)
Co-regulatory associations: prior to backbone extraction	1573836	3094954
Co-regulatory associations: after backbone extraction	141850	170723
Co-regulatory associations: corems	56738	25976

Table 2.1: Global properties of *H. salinarum* and *E. coli* ensembles

categorized the experiments using extensive metadata collected at the time of the experiment. We used this metadata to construct a GO-like ontology of the relationships between all experiments (discussed in detail below). The annotation counts above are derived from this resource (note that a single array can receive more than one annotation). A full list of the metadata, annotations, and ontology is available on the web service. 1159 of the arrays are published ([25, 27, 50, 106, 107, 184, 185, 298, 299, 300, 366, 367]. 336 of the arrays are new for this study. Experimental protocols are identical to [50]. These data, including expression levels ($\log_2$ ratios vs. reference samples) and experimental metadata, are available online as a tab-delimited spreadsheet.

Each array in the *H. salinarum* compendium was collected using the same platform, using the same reference, and processed and normalized using the same protocol. More specifically, each RNA sample was hybridized along with a *H. salinarum NRC-1* reference RNA prepared under standard conditions (mid-logarithmic phase batch cultures grown at 37°C in CM, OD = 0.5). Samples were hybridized to a 70-mer oligonucleotide array containing the 2400 non-redundant open reading frames (ORFs) of the *H. salinarum NRC-1* genome as described in [25]. Each ORF was spotted on each array in quadruplicate and dye flipping was conducted (to rule out bias in dye incorporation) for all samples, yielding eight technical replicates per gene per sample. At least two independent biological replicates exist for all experimental conditions for a total of 16 replicates per gene per condition. Direct RNA labeling, slide hybridization, and washing protocols were performed as described by [107, 299]. Raw intensity signals from each slide were processed by the SBEAMS-microarray pipeline [241] (www.SBEAMS.org/microarray), in which the data were median normalized and subjected to significant analysis of microarrays (SAM) and variability and error estimates analysis (VERA). Each data point was assigned a significance statistic, λ, using maximum likelihood [162].

E. coli *K-12 MG1655 expression compendia*

DISTILLER expression compendium for model training A total of 868 *E. coli* K-12 MG1655 transcriptome profiles were compiled by [208] for use with their DISTILLER algorithm. These data were collated from publicly available microarray databases: 44 arrays from Stanford Microarray Database [93], 617 from Gene Expression Omnibus [33] and 36 from ArrayExpress [274], as well as 181 arrays from supplementary data in literature (for four different experiments). The experiments cover a range of conditions, including varying carbon sources (136 arrays), pH (46 arrays), oxygen (284 arrays), metals (27 arrays) and temperature (23 arrays). Overall, the compendium consists of measurements from single channel (407 arrays; including 298 Affymetrix, and 109 P33) and dual channel (460 arrays; including 337 DNA/cDNA and 126 oligonucleotide) platforms.

These microarray measurements were normalized by the authors [208], as follows: "If possible, raw intensities were preferred as data source over normalized data provided by the public repository. Dual-channel data were loess fitted to remove nonlinear, dye-related discrepancies. No background correction procedures were performed to avoid an increase in expression logratio variance for lower, less reliable intensity levels. Whenever raw data were available, single-channel data were first normalized per experiment with RMA. Logratios were then created for the single-channel data in order to combine them with the dual channel measurements. For each single-channel array, expression logratios were computed by comparing the normalized values against an artificial reference array. This artificial reference array was constructed on a per experiment basis by taking the median expression of each gene across all arrays in the corresponding experiment. When deemed necessary (e.g. experiments normalized by MAS5.0 for which the raw data was not available), a loess fit was performed on these logratios. To ensure that the artificial reference was not altered by this intensity dependent non-linear rescaling, the artificial reference expression levels were chosen for the average log intensity (instead of the mean expression levels of the respective array and the artificial reference). To ensure comparability between arrays with a different reference, gene expression profiles were median centered across arrays that share the same

reference. An additional variance rescaling of the gene expression profiles was performed to render genes with differing magnitudes of expression changes more comparable."

The authors further note that, "the array composition of the modules generated by DISTILLER is not biased towards arrays from any specific platform, indicating a correct preprocessing of the microarray compendium." [208] It is for this reason that we chose this normalized *E. coli* microarray compendium for EGRIN 2.0 analysis.

DREAM5 expression compendium for model validation To ascertain the generalizability of EGRIN 2.0 models across data sets, we inferred a second *E. coli* EGRIN 2.0 model on an independent *E. coli* gene expression compendium. By comparing this model to the original model we inferred using the DISTILLER data set, we were able (1) to understand what, if any, systematic biases exist due to normalization procedures, and (2) to cross-validate EGRIN 2.0 predictions across two data set.

We obtained the de-anonymized *E. coli* microarray compendium from the DREAM5 competition website [237]. According to the authors, these data were "compiled for *E. coli*, where all chips are the same Affymetrix platform, the *E. coli* Antisense Genome Array. Chips were downloaded from GEO (Platform ID: GPL199). In total, 805 chips with available raw data Affymetrix files (.CEL files) were compiled." Additionally, "Microarray normalization was done using Robust Multichip Averaging (RMA) 9 through the software RMAExpress. All 160 chips were uploaded into RMAExpress and normalization was done as one batch. All arrays were background adjusted, quantile normalized, and probesets were summarized using median polish. Normalized data was exported as log-transformed expression values. Mapping of Affymetrix probeset ids to gene ids was done using the library files made available from Affymetrix. Control probesets and probesets that did not map unambiguously to one gene were removed, specifically probeset ids ending in _x, _s, _i were removed. Lastly, if multiple probesets mapped to a single gene, then expression values were averaged within each chip."

Compared to the DISTILLER [208] data set, the DREAM5 [237] compendium contained

a different subset of the available *E. coli* transcriptome measurements from a different combination of platforms. While one might expect a number of arrays to be common between the two compendia, we discovered that the two data sets differed substantially in their statistical properties. The maximum Pearson correlation between arrays across the two data sets, for example, was ~ 0.63. Interestingly, the correlation among expression profiles of genes within predicted operons [279] was higher in the DREAM5 compendium (mean ~ 0.83) than the DISTILLER compendium (mean ~ 0.32). This is likely due to a combination of differences in the experiments/platforms included and normalization procedures.

Additional Data

Genome sequence data and annotations for cMonkey analysis We used genome sequences and gene annotations (coding regions) collated in RSA-tools [346] for both organisms in this study (*H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655). These data were themselves collated to annotate regulatory sequences of all sequenced genomes in RefSeq. Rather than using the RSA-tools-annotated promoter regions, we computed them ourselves as regions (-250 nt to +50 nt) surrounding the annotated translation start site of each gene/operon (see below for operon annotations).

In all cases where probe identifiers in the mRNA expression compendia used for this analysis could not be directly matched to gene annotations (or operon predictions or functional associations; see below), we used the RSA-tools "feature_names.tab"table of identifier synonyms to perform the match. In cases where the match was still not possible, we excluded the probe/ annotation/ association from analysis.

Operon membership predictions used for cMonkey analysis We used operon predictions for both *H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655 predicted by [279] from the Microbes Online database [11]. These predictions are updated regularly. The predictions are based upon genomic proximity and co-expression in publicly-available microarray data compendia. We used the versions downloaded from the website as of March, 2009. These

included predicted operon memberships for 826 genes in *H. salinarum* sp. NRC-1 and for 2,639 genes in *E. coli* K-12 MG1655 .

Predicted transcriptional regulators used for Inferelator analysis

***H. salinarum* sp. NRC-1** For *H. salinarum* sp. NRC-1 , we used the same set of putative transcription factors (TFs) as [47, 50]. This list of 124 regulators was selected from among the 2,400 *H. salinarum* sp. NRC-1 genes which are annotated as known or putative TFs based upon sequence or predicted structural homology [49].

***E. coli* K-12 MG1655** To enable direct comparison of our results to DREAM5, we used the list of 296 putative *E. coli* K-12 MG1655 transcriptional regulators collated by [237]. Their list was obtained by combining the list of TFs defined by RegulonDB [125] with TFs identified using Gene Ontology (GO) terms: *biological process* terms related to transcription (`GO:0009299;mRNA transcription` or `GO:0006351;transcription, DNA dependent`) and GO *molecular function* `GO:0003677;DNA binding` or any child terms.

Functional association networks integrated into cMonkey analysis We used EMBL STRING [328] v9.0 database of predicted functional associations between genes for both organisms (*H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655) to constrain module construction in cMonkey, as described below. The confidence scores estimated by [328] were incorporated into the cMonkey constraints. These networks included 151,826 associations among 2,559 genes in *H. salinarum* sp. NRC-1 , and 878,972 associations among 4,136 genes in *E. coli* K-12 MG1655 .

2.3.3 EGRIN 2.0: an ensemble of EGRINs, generation and statistical mining

EGRIN 2.0 model construction and analysis was performed using primarily the R statistical analysis environment, with add-on packages data.table and filehash for off-line storage (maintaining all information in memory was impossible for our large ensembles). Once the full set

of cMonkey and Inferelator runs were completed and stored, a round of post-processing was performed to agglomerate all results into a single ad-hoc database for storage and query. The following relationships could be queried to identify significant associations between biological entities described in the model:

Entity_1	Entity_2	Relationship	Associated info.
Bicluster	Gene	Contains	-
Bicluster	Condition	Contains	-
Bicluster	Motif	Contains	Associated genes
Regulator	Bicluster	Regulates	Weight
Motif	Motif	Similar	FDR q–value
Motif	Genomic coordinate	Overlaps	p-value

These relationships could then be extended to second-degree relationships, including (these relationships below are by no means all-inclusive; for brevity we denote g, g_1, and g_2 as separate genes, b as a bicluster, m as a motif, r as a regulator, and c as an experimental condition):

1. g_1 is co-regulated with g_2 if they occur in the same b.

2. g_1 is co-regulated with g_2 under condition c if g_1, g_2, and c occur in the same b.

3. m regulates g if m and g are both observed in the same b.

4. m regulates g under condition c if m, g, and c are all observed in the same b.

5. r putatively regulates gene g via m if r is predicted to regulate b which contains both g and m.

The frequency with which any of these relationships occurs throughout the entire ensemble of EGRIN models could subsequently be counted by querying the database, and a p-value

describing the significance of the frequency computed via the cumulative hypergeometric distribution. p-values were then converted to false discovery rate q-values using the BenjaminiHochberg procedure. We use this basic procedure to identify conditions associated with GRE influence, and GREs associated with gene co-regulation, as we describe below.

2.3.4 Clustering of cis-regulatory motifs to identify GREs

Each cMonkey bicluster contains at least one *de novo* MEME- detected [21] *cis*-regulatory motif. These motifs are used by cMonkey to guide bicluster optimization (in addition to other scoring metrics). There were 86,167 and 269,770 motifs detected across the entire ensemble for *E. coli* and *H. salinarum*, respectively. Each motif was represented in the model as a position-specific scoring matrix (PSSM). To determine which of these motifs represented *bona fide* GREs (as opposed to false positives), we computed pairwise similarities between all motifs using Tomtom [139] (Euclidean distance metric; minimum overlap of 6 nt) and clustered the most highly similar PSSM pairs using mcl [344].

The Tomtom motif similarity p-value threshold and the mcl inflation parameter (I) were selected to (1) maximize the density (unweighted) of edges between PSSMs inside clusters relative to the edges between clusters, and (2) ensure that the mcl "jury pruning synopsis" was at least 80 (out of 100). Criterion (1) aims to find a clustering that is as inclusive as possible, while minimizing over-clustering, while (2) is a built-in mcl metric that evaluates the quality of the clusters resulting from the user-selected pruning strategy (I). More specifically for criterion (1), we chose the clustering parameters (mcl inflation parameter I, Tomtom p-value cutoff p_c) which maximize:

$$(I, p_c) = \arg\max \left\{ \sum_{I=1}^{N} \sum_{i=1}^{n_I} \frac{\sum_{j=1}^{n_I} \delta_{ij}}{\sum_{J=1}^{N} \sum_{k=1}^{n_J} \delta_{ik}} \right\}, \tag{2.5}$$

where N is the total number of motif clusters for a given set of parameters, δ_{ij} indicates a significant similarity (subject to the given p-value threshold) the between PSSMs i and j within motif cluster I (which contains a total of n_I PSSMs), and δ_{ij} indicates a significant similarity between PSSM i in motif cluster I and PSSM j in motif cluster J. The final

parameters that maximized expression 2.5 and resulted in an mcl "jury pruning synopsis" of at least 80 were different for the two EGRIN 2.0 models: $p_c = 10^{-6}$ and mcl $I = 4.5$ for the *H. salinarum* ensemble and $p_c = 10^{-5}$ and mcl $I = 1.5$ for the *E. coli* ensemble.

We did not filter the motifs by *E*-value or other intrinsic motif quality metrics; rather, we enforced a cluster size threshold to ensure that GREs were re-detected consistently. Clusters containing at least 10 PSSMs were considered GREs. This criterion resulted in 135 GREs for *H. salinarum* (representing 27,991 PSSMs, Table S2[1]) and 337 for *E. coli* (representing 12,773 PSSMs, Table S3). Finally, we computed a "combined PSSM" for each GRE as the unweighted mean of aligned PSSMs within each cluster. This combined PSSM could be visualized as a motif logo identically to standard motif PSSMs.

The motif clustering procedure is summarized in Figure 2.3.

2.3.5 Genome-wide scanning of motifs to obtain GRE locations

We used motif scanning to discover GRE locations that were missed by the rigid definition of a promoter in cMonkey (typically -250 to +50 nucleotides surrounding the translation start site). This procedure was critical for discovering GREs in non-canonical locations, such as internal to operons. We computed how well each PSSM (described above) matched every position in the genome using MAST [21], and recorded significant matches at each genomic location subject to a position *p*-value threshold of 10^{-5}. This *p*-value cutoff corresponds to an expectation of discovering ~ 20 sites at random across the genome. For each GRE, we summed the number of significant matches to each of the GREs PSSMs at each genomic position. These counts were used to represent GRE composition in promoters (Figure 3.4 and Figure 3.5). In addition, we used these scanned locations to identify GREs located predominantly inside coding regions. Since these GREs may be spurious (*e.g.*, protein sequence motifs or trinucleotide patterns) they were flagged, although they were not removed from our global analysis.

We compared the genome-wide distribution of GRE locations to annotated start sites in

[1] All tables refer to Tables available in [59] and online

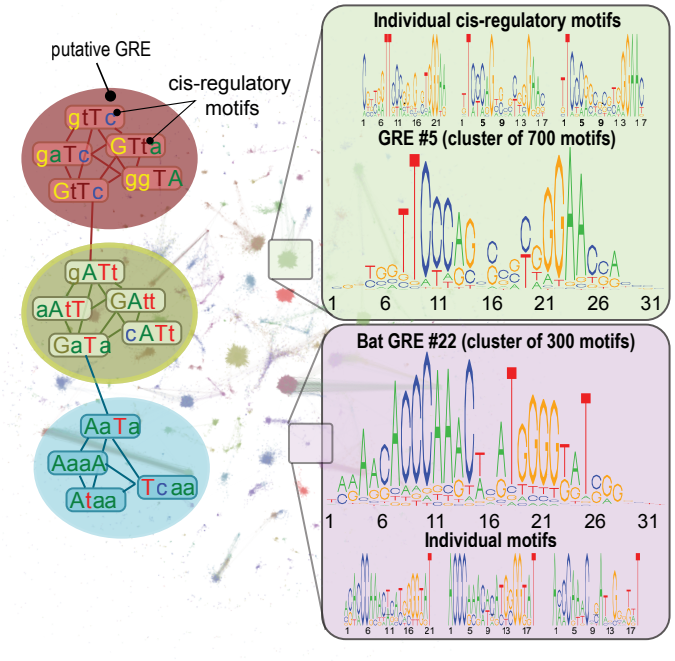


Figure 2.3: Motif clustering and GRE identification. (Left) A schematic of the approach used to align and cluster individually detected motifs to define GREs. In this example, similar motifs were aligned and clustered into three GREs using Tomtom and mcl (Details in Methods and Supplementary Methods). (Center) The *H. salinarum* network of aligned and clustered motifs. (Right) Two *H. salinarum* GREs discovered by this method. The motif logo of each GRE was generated by summing PSSMs of the individual aligned motifs in the cluster, as illustrated by three examples of individual motifs (prior to alignment) for each of the two GREs. Note that relative to the individual motifs, the averaged GRE motif is more palindromic - a hallmark of binding sites for dimeric TFs.

H. salinarum. We discovered that most GREs occur in consistent locations with respect to gene start sites. The global position of all GREs and select GREs relative to experimentally determined gene start sites is depicted in Figure 2.4.

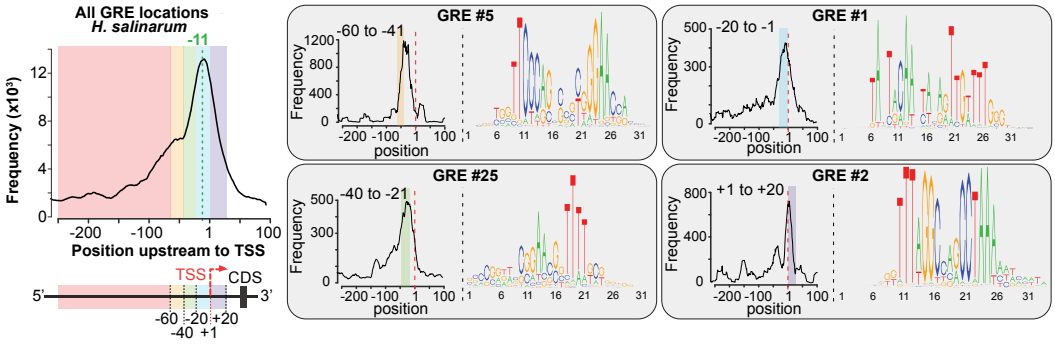


Figure 2.4: Genome-wide distribution of GREs relative to experimentally mapped transcriptional start sites in *H. salinarum*. (Left) Predicted positions for all GREs in gene promoters upstream of experimentally mapped transcription start sites (TSSs; [193]) in and (Right) four example elements. Distribution peaks for most GREs occur at characteristic locations. For instance, the location of TATA box-like elements (GRE #25) between -21 to -40 nt upstream to TSSs in *H. salinarum* is consistent with the characterized location of basal elements in archaeal promoters (-25 to 30 nt upstream to TSS). GRE location enables prediction of putative roles for the cognate TF (*e.g.*repressor, activator or a basal factor).

2.3.6 Detection of co-regulated modules (corems) by community detection

Construction of gene-gene co-occurrence network

We post-processed the EGRIN 2.0 ensemble to refine the underlying network structure and discover functionally meaningful gene co-regulatory modules present in the model. To do so, we transformed the ensemble of biclusters into a weighted gene-gene association graph G, where the nodes of G are genes and the weight of edges between the nodes is proportional to their frequency of co-occurrence in biclusters:

$$w_{ij} = \frac{|B_i \cap B_j|}{\min(B_i, B_j)}, \tag{2.6}$$

where w_{ij} is the weight of the edge between genes i and j, B_i is the set of all biclusters containing gene i. The weights were normalized by the minimum number of biclusters

containing either gene, rather than by the more typically applied union (which would make the score identical to the Jaccard Index) to avoid penalizing genes that occur infrequently in biclusters. The sum of edge weights for each gene was normalized to one. This gene-gene co-occurrence network represents how often cMonkey discovers co-regulation between every pair of genes in the genome. We note that since this network is derived from biclusters, it is also a reflection of conditional co-expression and predicted *cis*-regulatory motifs.

Network backbone extraction

After transforming the ensemble into a normalized graph, we removed edges that were statistically indistinguishable by multiscale backbone extraction (null hypothesis of uniform edge weight distribution given a node of degree k) [308]. We retained all edges satisfying the following relation:

$$\alpha_{ij} = 1 - (k-1)\int_0^{w_{ij}} (1-x)^{k-2} dx \leq 0.05, \tag{2.7}$$

where α_{ij} is the probability that the normalized weight w_{ij} between genes i and j is compatible with the null hypothesis, and k is the degree of gene i. For *H. salinarum* sp. NRC-1 , backbone extraction reduced the number of regulatory edges from 1,576,643 to 141,667; in *E. coli* K-12 MG1655 the number of edges was reduced from 3,094,954 to 170,723.

Network link-community detection

Following backbone extraction, we detected corems by application of a recently described link-community detection algorithm [5]. For this algorithm to work on our data set we modified it to accept input of a weighted graph [177]. We implemented it in C++ for efficiency. The algorithm computes a similarity score between all pairs of edges sharing a common keystone node, k, according to the Tanimoto coefficient, T:

$$T(e_{ik}, e_{kj}) = \frac{a_i \cdot a_j}{|a_i|^2 + |a_i|^2 + a_i \cdot a_j}, \tag{2.8}$$

where

$$a_i = w_{ij} + \frac{\delta_{ij}}{k_i} \sum_{l \in n(i)} w_{il}. \tag{2.9}$$

Here, e_{ik} is the edge between gene i and the keystone gene k, and δ_{ij} is the Kroenecker delta. The score reflects the similarity of gene neighborhoods adjacent to two edges sharing a gene, with the score increasing in value as the number and weight of overlapping adjacent edges increases. To transform the Tanimoto coefficient into a distance metric, we compute $1 - T$.

Following scoring, the edges were aggregated by standard hierarchical clustering. The resulting tree is cut at many thresholds to optimize the local weighted density D of the resulting clusters:

$$D = \frac{1}{M\langle w \rangle} \sum_{c \in C} m_c \langle w \rangle_c \left(\frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \right), \tag{2.10}$$

where M is the total number of edges in the entire network, $\langle w \rangle$ is the average weight of edges in the entire network, C is the set of all link communities at a given threshold, m_c is the number of edges in community c, $\langle w \rangle_c$ is the average weight of edges in community c, and n_c is the number of genes in community c. The density scoring metric D had a clear optimum corresponding exactly to the cutoff that would have been chosen had we used the unweighted scoring metric originally described (Figure 2.5). Only communities with more than two genes were retained.

Since the communities produced by this algorithm are comprised of sets of edges, we defined a corem to include all genes incident to the edges in a community. Because of this definition, each gene can be a member of multiple different corems. In *H. salinarum*, this procedure generated 679 corems ranging in size from 3 to 377 genes, covering 1,363 of the 2,400 genes in the genome, and comprising 56,738 co-regulatory associations. In *E. coli*, we discovered 590 corems, ranging in size from 3 to 153 genes, covering 1,572 of 4,213 genes and 25,976 regulatory edges. See Table2.1 and Figure 2.5 for additional statistics. Gene-to-

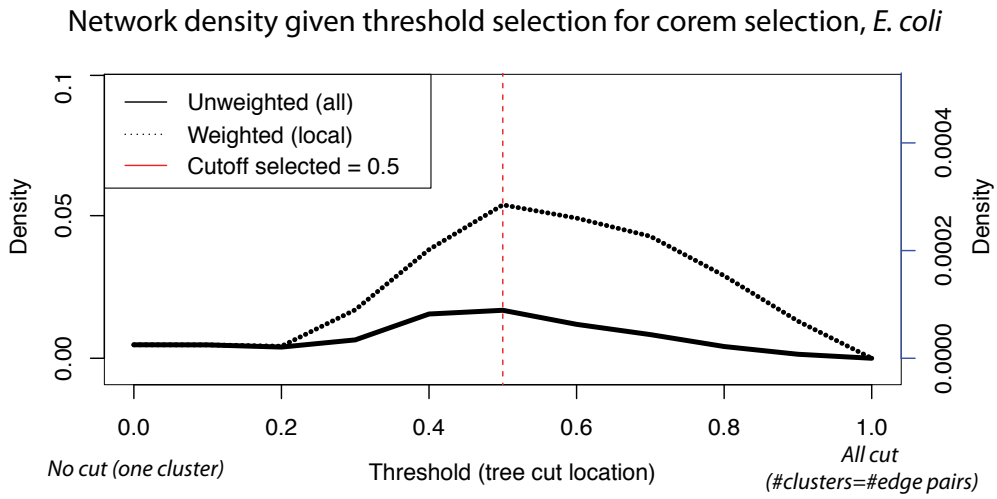


Figure 2.5: Corem density as a function of clustering cutoff threshold. Hierarchical clustering cut threshold chosen to maximize the density of resulting clusters. The cutoff chosen with modified weighted density metric is identical to unweighted density metric.

corem and corem-to-gene mappings for the *H. salinarum* and *E. coli* models are available online.

2.4 Model Evaluation

In this section we evaluate the performance of the EGRIN 2.0 model as a function of several important parameters. We focus in particular on how the performance of the model changes as a function of the number of runs included. From these evaluations, we conclude that (1) the model performs well in its final form, (2) the model has reached a stable-state wherein inclusion of additional runs does not significantly increase model performance, and (3) the model is not over-fit to particular experiments within a data set or to any data set as a whole.

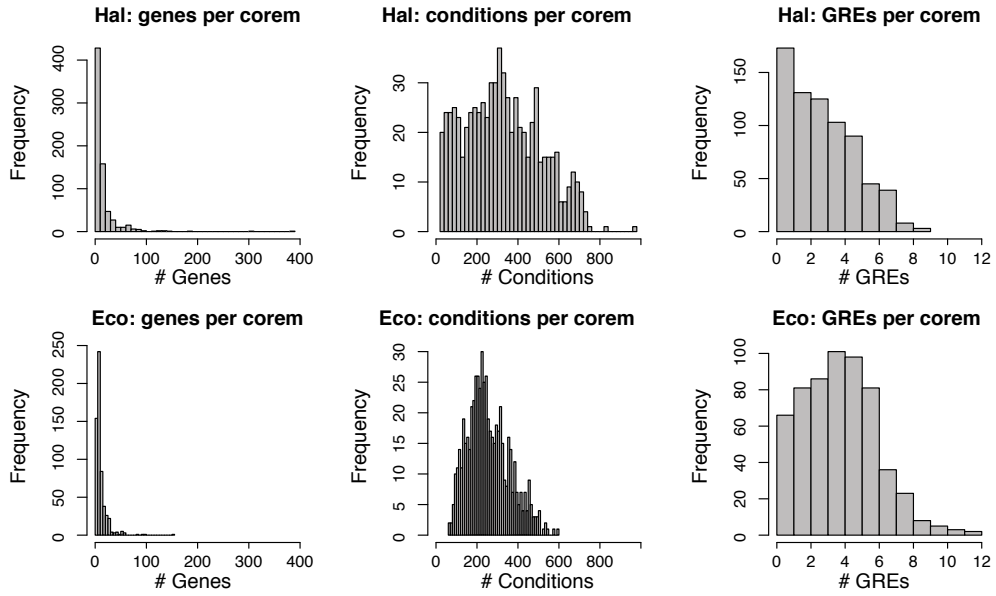


Figure 2.6: Corem statistics. Number of genes, conditions, and GREs per corem for *E. coli* and *H. salinarum* EGRIN 2.0 models.

2.4.1 Comparison with other module detection algorithms

We compared the number of RegulonDB TFs detected in the EGRIN 2.0 model to individual cMonkey runs as well as to several other module detection/clustering algorithms that were computed on subsets of the experimental data (similar to the EGRIN 2.0 ensemble; Figure 2.7). We evaluated: (a) k-means clustering, (b) WGCNA [202], and (c) DISTILLER [208]. For (a) and (b), we computed modules 100 times on random subsets of the *E. coli* expression data set (using 200-250 randomly chosen experiments per run; selection criteria were identical to *E. coli* EGRIN 2.0). We then predicted *de novo cis*-regulatory GREs in the promoter regions of genes in each module using MEME (MEME parameters were also identical to EGRIN 2.0). For (c), we performed the comparison using the original modules generated by [208]. Rather than alter module composition by re-detection, we instead varied MEME parameters applied to the modules 100 times (again, within the same ranges as those used

for EGRIN 2.0). TF-GRE matches were assigned by comparing GREs to RegulonDB TF binding sites, as previously described (Section 2.5.3).

We found that individual cMonkey runs discovered a greater number of RegulonDB binding sites, on average, than the other methods (an average of 41 for cMonkey, compared to averages of 30, 25, and 29 for k-means, WGCNA, and DISTILLER, respectively), which is consistent with previous findings [284] (Figure 2.7). Integration of all cMonkey biclusters into the complete EGRIN 2.0 ensemble outperformed all individual cMonkey runs (53 total, as described in the Manuscript). This result is typical of ensemble-based inference approaches, and supports the value of ensemble integration as part of the EGRIN 2.0 model.

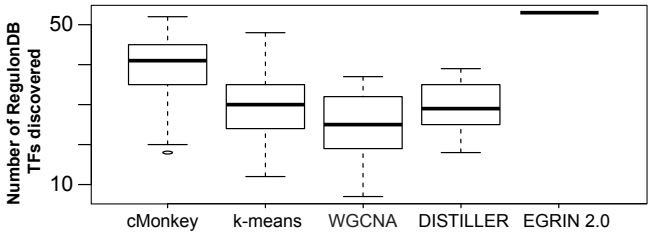


Figure 2.7: Number of TFs in RegulonDB re-discovered by various regulatory module detection methods. Comparison of EGRIN 2.0 (solid line, far right) to individual cMonkey runs, as well as multiple runs of k-means, WGCNA, and DISTILLER on subsets of the expression data. Evaluation made with respect to re-discovery of binding sites for 88 TFs with ≥ 3 unique sites in RegulonDB based on genome-wide binding site locations (FDR ≤ 0.05).

2.4.2 Convergence and stability of inferred GRNs

To evaluate the stability of the inferred EGRIN 2.0 network, we quantified how the model changes as individual cMonkey runs are excluded from the ensemble. Since the sub-bagging, as performed for the EGRIN 2.0 model inference, reduce model over-fitting, we used this evaluation understand whether the model is over-fit to particular experiments in the data set. For this task, we computed the number of individual EGRIN runs required to converge

on a consistent gene-gene co-occurrence network (see Section 2.3.6). We computed gene-gene co-occurrence networks based upon randomly selected subsets of the 106 available *E. coli* K-12 MG1655 cMonkey runs, and varied the percentage selected between 1%-99% of the 106 runs. 5 replicate samples were computed for each. To compare the networks, we computed the Pearson correlation between the two matrices (sub-sampled gene-gene co-occurrence versus the final EGRIN 2.0 gene-gene co-occurrence network). Note that since the gene-gene co-occurrence network is a weighted adjacency matrix, the correlation reflects the weighted discovery rate for every pair of genes (rather than simple presence/absence). In Figure 2.8 we demonstrate that the underlying networks converge rapidly to the final solution. By the time $\sim$ 50% of the runs have been included ($\sim$ 50 runs), the inferred network is nearly identical to the final network ($\sim$ 100 runs; cor $>$ 0.9). The backbone extracted network takes a slightly longer time to converge, likely because it requires more observations of gene-gene pairs to retain them in the final network. Since corem detection is deterministic and strictly based on the underlying gene-gene co-occurrence matrix, this convergence means that the inferred corems would be nearly identical even if up to half of the runs were excluded.

2.4.3 Confirmation of corems in an independent data set

To determine whether EGRIN 2.0 model predictions are over-fit to the DISTILLER expression compendium (or are the result of biases in that data set), we tested whether support for corems existed in an independent *E. coli* expression data set. Such evidence would suggest that corems are *bona fide* gene regulatory modules that can be re-discovered in independent data, and that their degree of condition-specificity is not biased due to normalization differences in any given data set. For this test, we used the DREAM5 gene expression compendium. As described above (Section 2.3.2), this data set is comprised of different conditions, array platforms, and, most important, was normalized by different methods, than the DISTILLER data set used for model training. We determined the condition-specific activity of corems in the DREAM5 data set using the methods described in Section 2.5.2. If a

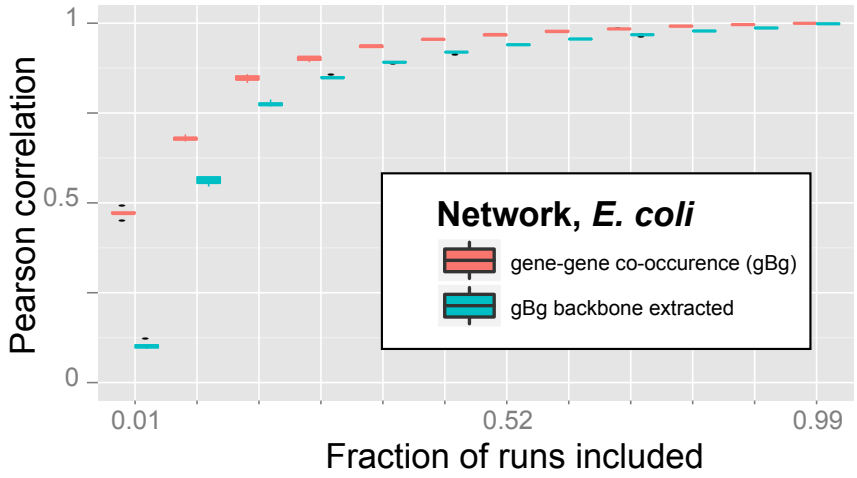


Figure 2.8: Convergence of EGRIN 2.0 co-occurrence networks. The co-regulation of genes predicted by the *E. coli* K-12 MG1655 EGRIN 2.0 model converges rapidly to a stable network. Shown is the similarity of the gene-gene co-occurrence matrix (and the backbone extraction of this matrix) to the final EGRIN 2.0 *E. coli* K-12 MG1655 network, computed when varying fractions of the cMonkey runs were excluded (Pearson correlation vs. the complete model). Each point contains a box plot representing 5 replicate sub-samples.

corem was significantly co-expressed (p-value ≤ 0.05) in at least one condition, we classified it 'supported'. To our surprise, we not only discovered support for $\sim 99\%$ of the predicted corems, we also discovered that their conditionality was very similar across both data sets – *i.e.*, corems discovered to be co-expressed in few conditions in the DISTILLER data set are also co-expressed in few conditions in the DREAM5 data set (same for corems regulated in many conditions), and similarly for corems co-expressed in a large number of conditions (Figure 2.9). Even after we removed the intrinsic relationship between the number of genes in a corem and the number of conditions in which it is co-expressed, we still observed a significant partial correlation of 0.49 (p-value $< 10^{-6}$) between the number of conditions in corems as defined from the two data sets.

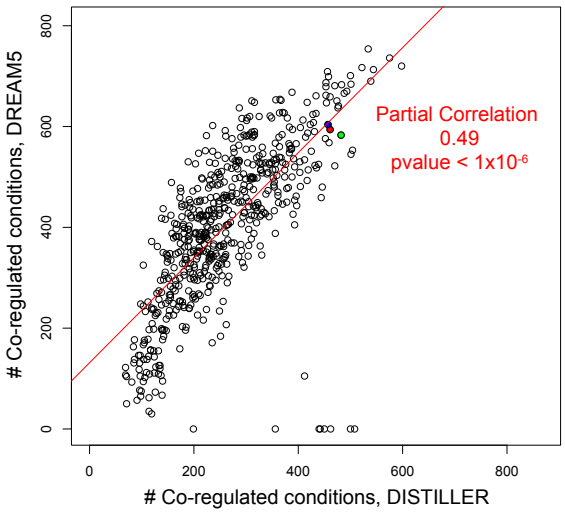


Figure 2.9: Reproducibility of corems across data sets. Number of co-expressed conditions for corems in the DISTILLER and DREAM5 expression compendia. Conditions were selected as in Section 2.5.2. Significant partial correlation of 0.49 is observed after removing the affect of gene set size (log) on the number of conditions co-expressed (p-value $< 10^{-6}$). The three corems detailed in the main manuscript are identified with their respective colors (ec512157, ec516034, ec516031)

2.5 Model Validation

2.5.1 Data For Model Validation

H. salinarum *sp. NRC-1*

Tiling array transcriptome measurements We generated *H. salinarum NRC-1* high-resolution (12 nt) tiling array transcriptome measurements over 12 points along the growth curve in rich media. These were analyzed and published in a separate study [193]. Locations of putative transcription breaks in these data were identified in using multivariate recursive partitioning, including signals from both relative changes in expression along the growth curve, as well as raw RNA hybridization signal.

ChIP-chip transcription factor binding measurements for global regulators Global binding of eight general transcription factors (seven TFBs [TFBa, TFBb, TFBc, TFBd, TFBe, TFBf, and TFBg] and one TBP [TbpB]) and three specific TFs (Trh3, Trh4, and VNG1451C) in *H. salinarum* were collected in our lab by ChIP-chip. A detailed protocol is described in [107]. Briefly, ChIP-enriched and amplified DNA for eleven regulators was hybridized to a low-resolution (500 nt resolution) custom PCR-product array spotted in-house. The resulting intensities were analyzed using MeDiChI [285] to obtain binding site locations with an average precision of 50 nt. Local false discovery rates (LFDRs) were quantified by simulation.

***kdp* promoter serial truncation measurements** *H. salinarum NRC-1 kdpFABC* truncation data were obtained from [191]. Briefly, the authors measured relative induction of a transcriptional reporter after serial truncation of the *H. salinarum* R1 *kdpFABC* promoter. The authors measured β-Galactosidase activities from truncated transcriptional fusions of the *kdpFABC* promoter to *bgaH*. β-Galactosidase activities were measured in triplicate from cultures grown in inducing (3 mM K^+) and non-inducing (100 mM K^+) conditions. We obtained data corresponding to Figure 3.4, in which the authors quantify the fractional β-Galactosidase activity (non-induced/induced) among the serial truncations (private communication). We overlaid motif predictions from EGRIN 2.0 on this data set to reach our conclusions.

E. coli *K-12 MG1655*

Tiling array transcriptome measurements We measured *E. coli* K-12 MG1655 tiling array transcriptome profiles at nine different time points during growth in rich media (LB). Growth phases spanned lag-phase (OD600 = 0.05) to late stationary-phase (OD600 = 7.3). RNA samples were prepared by hot phenol-chloroform extraction [189]. RNA was directly labeled and hybridized to custom Agilent tiling arrays containing 60mer probes tiled across both strands of the *E. coli* K-12 MG1655 genome using a sliding window of 23 nt (GEO

Platform GPL18392), as in [193]. Expression measurements were quantile-normalized as in [380] and analyzed for condition-specific transcriptional isoforms following the segmentation protocol described in [193]. Data is available on GEO (GSE55879).

PurR/ΔPurR expression data and ChIP-chip transcription factor binding sites *E. coli* PurR/ΔPurR expression data and ChIP-chip transcription factor binding measurements collected in the presence of adenine were taken from [73]. ChIP-chip relative intensities were re-analyzed using MeDiChI [285] to obtain binding site locations with an average precision of ∼25 nt.

Fitness measurements *E. coli* fitness measurements across 324 conditions were generated by [259]. In short, the authors quantitated growth rates for 3979 single gene deletions in each of 324 environments with variable stress, drug, and environmental challenges. *E. coli* mutant colony sizes were quantified on agar plates. Fitness correlations were obtained directly from the authors: http://ecoliwiki.net/tools/chemgen/. Each correlation value represents the Pearson correlation of fitness (*i.e.*, relative growth rate) for pairs of single gene deletion mutants measured across all 324 conditions that are also present in our analysis. Relative fitness scores were also obtained directly from the authors.

Effector molecule measurements *E. coli* effector molecule measurements were taken from [166]. The authors measured metabolite levels using capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS) in *E. coli* K-12 MG1655 , as well as several other biomolecules (*e.g.*., RNA and protein). *E. coli* was grown in a chemostat at several different dilution rates (0.1, 0.2, 0.4, 0.5, and 0.7 hours^{1}). We obtained the metabolite levels from the authors and computed Pearson correlation between metabolites assigned to regulate TFs by RegPrecise [263].

Experimentally mapped *E. coli* transcription factor binding sites We compared genome-wide locations of GREs in the *E. coli* EGRIN 2.0 model with experimentally-

mapped binding sites from the RegulonDBdatabase [125]. To maintain consistency with our comparisons against the DREAM5 community networks [237], we used version 6.8 of the database. For binding sites, we used the BindingSiteSet table, filtered for only interactions with experimental evidence, and used only TFs with ≥ 3 unique binding sites – a total of 88 TFs.

Experimentally measured *E. coli* transcription factor regulatory targets For the *E. coli* gold standard network, we used the same network as that used by [237] for validation of the DREAM5 *E. coli* community predicted regulatory networks. This gold standard is based upon version 6.8 of the RegulonDB database [125], and only interactions with at least one strong evidence were included, for a total of 2,066 interactions. We mapped the *aaaX*-style gene names in the DREAM5 gold standard to the *b*1234 in cMonkey using a translation table compiled in the EcoGene database, version 3.0 [386]. We were able to map a total of 4,273 gene names. The final gold standard consisted of 2,064 interactions between 141 TFs and 997 target genes. The final, complete gold standard network used for all analyses is available online.

2.5.2 Computational Methods for Model Validation

Functional enrichment estimates for genes in corems

We computed functional enrichment for genes organized into corems using DAVID [94] and the DAVIDQuery [88] R-package. Enrichments for each corem are available on the web site.

Conditional co-regulation of genes organized in corems

We defined the conditions in which genes in a corem were co-regulated as the set of experiments in which the genes of a corem are more tightly co-expressed than one would expect at chance. We statistically evaluated tight co-expression using relative standard deviation (RSD $= |\sigma/\mu|$) by resampling. We chose RSD (rather than, for example, standard deviation, σ) to avoid over-weighting conditions in which the mean relative expression is close

to zero. The significance of an RSD value for a given condition relative to each corem was estimated by resampling: for a corem with k gene members, and for each condition, c, we computed at least 20,000 RSD values for k randomly sampled expression measurements in c, to determine the likelihood that the observed co-expression has lower RSD than expected by chance (p-value < 0.01). The resampling procedure resulted in condition sets for corems that contained from 1.4% to 85.5% of the conditions in *H. salinarum* sp. NRC-1 and 7.9% to 66.6% conditions in *E. coli* K-12 MG1655 (Figure 2.6).

Conditionality of GRE influence

The upstream promoter regions of most genes contain multiple EGRIN 2.0-predicted GREs (*e.g.*, *carA* in Figure 3.5). A key insight of our model is that not all of these sites are equally important for controlling gene expression in all experimental conditions. We refer to changes in the relative influence of GREs across conditions as "conditional activity" of GRE elements. Although, to be clear, we do not imply that the transcriptional activity at a GRE is attributable to the DNA sequence itself, but rather the TF that binds to that sequence in particular environments. We leveraged the GREs discovered in genes grouped into corems and the conditional co-expression of those groups of genes to predict conditionally active GREs in EGRIN 2.0.

To identify the active GREs for each corem we combined predictions from (1) genome-wide motif scans (Section 2.3.5 above) that predict the GRE locations in an expanded region around each genes promoter in the corem using all of the ensemble predictions (1,000 nt window: -875 nt upstream to 125 nt downstream), and (2) the conditions discovered in biclusters that are most representative of the corem (*i.e.*, containing the largest fraction of genes from the corem, top decile). GREs that occurred frequently in these biclusters were considered putatively responsible for co-regulating the set of genes in the condition-specific context of the corem (q-value ≤ 0.05). Finally, we computed the average distances of all GREs to the start codons of each gene in the list (collapsing sites if they occurred within 25 nt of one another). The precise locations of all GREs for the *H. salinarum dpp* operon-

related corems (Figure 3.10) are available in Table S8 of the corresponding paper, while the locations of GREs involved in conditional modulation of the PurR regulon (3.17) are provided in Table S9.

We represented the active GREs upstream of a gene or within a corem as a pie chart, showing the normalized frequency with which the GREs computed above occurred in biclusters containing that gene. For example, if GREs 1, 2, and 3 occurred in 25, 50, and 200 biclusters containing gene A, the pie chart for gene A would have sectors of area 0.09, 0.18, and 0.73 respectively. For corems, we computed the normalized frequency of GREs for all genes of the corem. For example, if GREs 1, 2, and 3 occurred in promoters of 10, 10, and 20 of the genes of the corem, their areas would be 0.25, 0.25, and 0.5 respectively.

Detection of conditional operons

Condition-specific transcriptional isoforms of operons were predicted through corem membership. If any of the genes in an operon were found in a corem that did not contain all the other genes of the operon, we predicted that the operon had conditional isoforms. Operon annotations for both *H. salinarum* and *E. coli* were derived from MicrobesOnline [11]. All predicted conditional operons, including the specific break sites and transcriptional isoforms is available on the website. The full list of validated predictions is provided in Table S7.

Environmental ontology construction and usage

We recorded a rich set of meta-data for all 1,495 experiments conducted with *H. salinarum* and used for construction of the *H. salinarum* sp. NRC-1 EGRIN 2.0 model. The meta-data includes a detailed description of each experiment, including, for example: media composition, genetic background, concentration of perturbant, internal reference batch id, person who conducted the experiment, etc. We used this meta-information to classify experiments in an ontological framework, where two experiments can share specific meta-descriptions (*e.g.*, 10^{-3} mol/L EDTA), or inherit more general relationships from the ontological structure (*e.g.*, chemical perturbation). We used OBO-edit [89] to construct the ontology. The

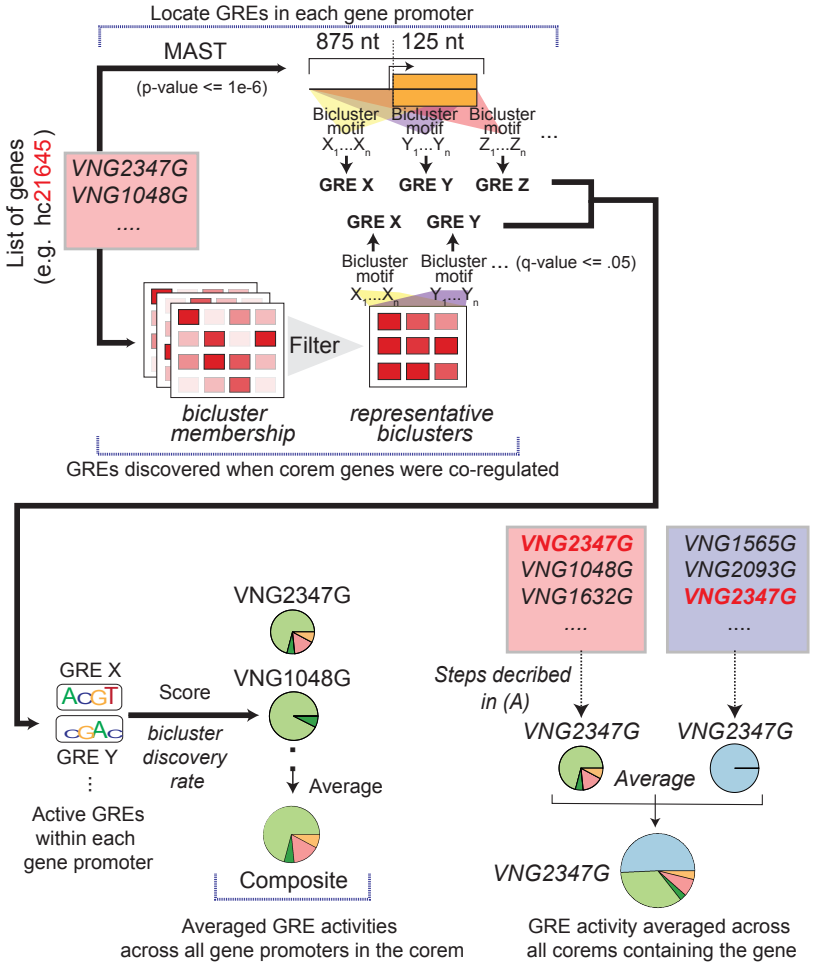


Figure 2.10: Deciphering GREs responsible for regulating corems. A GRE is implicated in regulation of a corem when it is both (1) located within an expanded region (-875nt to +125nt) around the translation start site of any gene in the corem; and (2) present in biclusters containing a large fraction of corem genes (top decile). Relative GRE influence is computed as the frequency with which each GRE was discovered in these representative biclusters (see Supplementary Methods for more details). Influence scores are illustrated as pie charts and reported for each gene individually (*e.g.*, *VNG2347G*); and as a composite by averaging across all genes in a corem. The width of each sector in the pie charts is proportional to the frequency of GRE discovery.

ontology contained 198 terms organized across three primary branches (environmental state, experimental state, and genetic state). The ontology flat file is available for download and meta-data annotations for every array in the dataset are available online.

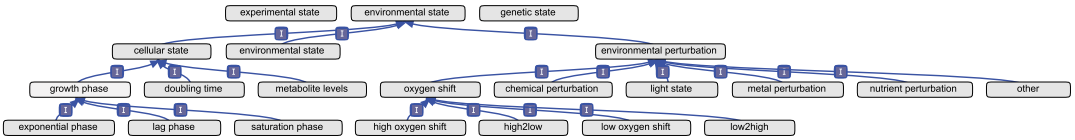


Figure 2.11: Environmental ontology hierarchically organizes relationships between experimental conditions from metadata collected across 1495 experiments in *H. salinarum*. Subset of the environmental ontology constructed for *H. salinarum* demonstrates many is-a (boxed I) relationships that organize similarities between descriptor terms descending from one of three root nodes (i.e., generic categorical descriptions). In this case a generic ontological term called 'environmental state' gives rise to much more specific terms (e.g., exponential phase or high oxygen shift) that inherit (at the highest level) a relationship through their being related to the environmental state of cells in the experiment. Each condition in the compendium is annotated with the most specific descriptors relevant to the experiment given metadata. The full environmental ontology is available for download from http://egrin2.systemsbiology.net.

We used the ontology to classify enriched environmental features for GREs and corems. For corems, we used the set of conditions in which genes in the corem are significantly co-expressed (see Section 2.5.2 above) to compute term enrichment using the ontoCAT [200] R-package. Term enrichment was assessed statistically and reported as q-values using the hypergeometric test with Benjamini-Hochberg correction for multiple hypothesis testing.

2.5.3 Global validation of GREs predicted by EGRIN 2.0

We compared the genome-wide locations of predicted GREs in the *E. coli* EGRIN 2.0 model to experimentally mapped TF binding sites from RegulonDB (BindingSiteSet table, filtered for experimental evidence and TFs with $\geq$ 3 unique binding sites; a total of 88 TFs).

We considered a GRE to be a significant match to a TF if a significant fraction (q-value ≤ 0.05) of its predicted non-coding locations overlapped with the known binding locations for a particular TF (hypergeometric p-value ≤ 0.01; see GRE definition in Section 2.3.4). In cases where a GRE significantly matched multiple TFs, only the most significant was reported.

We observed several instances where more than one GRE significantly matched the same TF. We were unable to determine whether this was the result of incomplete GRE clustering, ambiguities related to GRE scanning, limitations of the experimental data itself, or a reflection of subtle context-dependent variations in the binding preferences of these TFs. Since we did not observe clustering of GREs that map to the same TF upon re-clustering, we hypothesize that the observations may have biological origins, *i.e.*, reflect condition-dependent variations in TF binding preferences that are the result, for example, of co-activator/repressor interaction or small molecule binding. It is interesting to note that TFs with the largest fraction of GRE matches include transcriptional dual regulators, such as FlhDC and UlaR (*i.e.*, TFs with the ability to act as both activators and repressors). This is consistent with the observation that these TFs have context-dependent binding preferences. The complete set of validations, for both TFs and σ-factors, is listed in Table S4.

2.5.4 Global validation of regulatory interactions predicted by EGRIN 2.0

We assessed the ability of the EGRIN 2.0 model to correctly infer known regulatory interactions using the RegulonDB database as a standard metric for comparison. Comparison to the RegulonDB gold-standard is common practice for evaluating model performance [237]. We performed our evaluation with the version of RegulonDB used by the DREAM5 ensemble (based on RegulonDB release 6.8 [237]) so that we could directly compare our results. The authors [237] restricted the gold-standard to well-established interactions, annotated in RegulonDB with the 'strong evidence' classification. In all cases, networks were integrated from predictions among the ensemble using an approach similar to that of [237], with subtle variations noted in each section, below. To facilitate a direct comparison, we reconstructed

a new *E. coli* EGRIN 2.0 model using the same DREAM5 expression consortium as was used for the original DREAM5 competition (Section 2.3.2). The predictions of this model were used *solely* for global validation and direct comparison with the DREAM5 community network, as described in this subsection.

We performed two global evaluations of the *E. coli* EGRIN 2.0: (1) a comparison of the GREs detected in the model with experimentally mapped TF binding sites in RegulonDB (Section 2.5.3), and (2) a comparison of the predicted (TF $\rightarrow$ gene) regulation in EGRIN 2.0 with the gene regulatory network from [237]. For (2), we computed predicted regulatory networks from EGRIN 2.0 in two ways: (a) direct (TF $\rightarrow$ target) predictions from Inferelator (Section 2.5.4, and (b) a gene regulatory network derived from predicted GREs that were matched to TFs in RegulonDB (Section 2.5.4). Construction of each of these networks is described in detail below (Section 2.5.4 and Section 2.5.4). The methods for, and results of the comparisons are described in Section 2.5.4.

Conversion of EGRIN 2.0 Inferelator influence predictions into a GRN

We computed a direct (TF $\rightarrow$ gene) inferred *E. coli* gene regulatory network (GRN) from the Inferelator predictions in the EGRIN 2.0 ensemble. As with the original EGRIN model [50], Inferelator influence predictions were originally made between the 296 putative *E. coli* TFs (Section 2.3.2) and each of the $\sim 40,000$ biclusters in the ensemble. We then used a weighted average of the predicted influences among all networks in the ensemble, as follows. If Inferelator predicted a (TF $\rightarrow$ bicluster) influence with weight β then we added β to a regulatory interaction between that TF and all genes in that bicluster. Weights β were summed for each recurrence of the same (TF $\rightarrow$ gene) interaction. Note, we did not use $|\beta|$ in the individual sums, since we considered contradicting evidence to be cancelling rather than reinforcing. Finally, all (TF $\rightarrow$ gene) interactions in the final network were ranked by absolute total weight (here we *did* use $|\beta|$). As with the DREAM5 competition networks, the top 100,000 rankings were retained in the final network. The final EGRIN 2.0 Inferelator influence network is available online.

Conversion of EGRIN 2.0 GRE detections into a predicted GRN

We computed a separate inferred *E. coli* gene regulatory network from predicted GREs in EGRIN 2.0 that were matched to TFs as described in Section 2.5.3. We would like to stress that this inference relies upon (in this case, for *E. coli*) annotated binding sites for regulators, which could be statistically linked to predicted GREs through significant overlaps in their genomic locations. This enables inference of (TF $\rightarrow$ gene) direct influence predictions through the indirect relationship:

$$\text{TF} \overset{\text{anno.}}{\rightarrow} \text{GRE} \overset{\text{pred.}}{\rightarrow} \text{gene}. \tag{2.11}$$

Thus for an understudied organism, such as *H. salinarum*, such a network of (TF $\rightarrow$ gene) influences could *not* be inferred; rather a (GRE $\rightarrow$ gene) interaction network would be the final product. Such a network still contains predictions which could be validated and acted upon, for example, for engineering purposes. A future direction of our research will be to statistically link TFs to predicted GREs, for example using direct GRN predictions such as those described above (*e.g.* Section 2.5.4, or [237]).

(GRE $\rightarrow$ gene) predictions (in Eq. 2.11) were extracted from the EGRIN 2.0 model directly using the MEME predictions for motif instances in the promoters of genes in each of the $\sim$40,000 cMonkey biclusters. We then used an unweighted average of the predictions among all bicluster in the ensemble, as follows. A (TF $\rightarrow$ gene) edge with a weight of 1 was added to the predicted network if the annotated binding sites for that TF could be matched with locations of a motif (Section 2.5.3), which was detected by MEME in a bicluster in the promoter of the gene. Edge weights (1) were added for each additional prediction, in the ensemble of biclusters, of the same (TF $\rightarrow$ gene) interaction. As with the Inferelator influence network (Section 2.5.4), the top 100,000 rankings were retained in the final network. The final EGRIN 2.0 GRE-based network is available online.

Integration of predicted EGRIN 2.0 Inferelator- and GRE-based GRNs

Prior to integration of the two different predicted GRNs described above (Sections 2.5.4 and 2.5.4), we ensured that they were both equally represented in the integrated GRN by re-scaling their weights so that their sums would be equal. The GRNs were then combined into a single, integrated predicted EGRIN 2.0 GRN by simply summing the re-scaled weights for any edge predicted in both networks. Thus, this final network integration was a form of weighted average of the two (GRE and Inferelator) networks. This is *not* identical to the weighted rank average method described by [237], as it does not use a posteriori assessments of each network to assign their relative weights; rather the weights are simply adjust so that each network contributes equally to the predictions.

Network comparisons and global performance assessments

To compare EGRIN 2.0 performance to the DREAM5 ensemble, we computed standard precision-recall statistics for each network using the previously described DREAM5 gold standard GRN. We computed area-under-the-precision-recall (AUPR) statistics to summarize the predictive performance. AUPR statistics were compared directly with the DREAM5 community ensemble network. By extension, the EGRIN 2.0 AUPR performance can be compared to the individual best performers in DREAM5 as well (Figure 3.2 in [237]). The results of these analyses are summarized in Figure 3.2 in the main text. We have made all network predictions available online. Complete precision-recall curves are shown in Figure 2.12. The curves are also available in tabular form online.

We further investigated the convergence of the AUPR statistics for each of the EGRIN 2.0-predicted regulatory networks as additional individual EGRIN models are added to the ensemble. This assessment helps to address the question of whether the approach utilized for ensemble integration has the desired property of performing better than most (if not all) of the individual models. Additionally, it can address the question of how many individual EGRIN models are necessary to achieve a given performance level. We observed that this is indeed the case for the Inferelator-based predictions extracted from the EGRIN 2.0 model

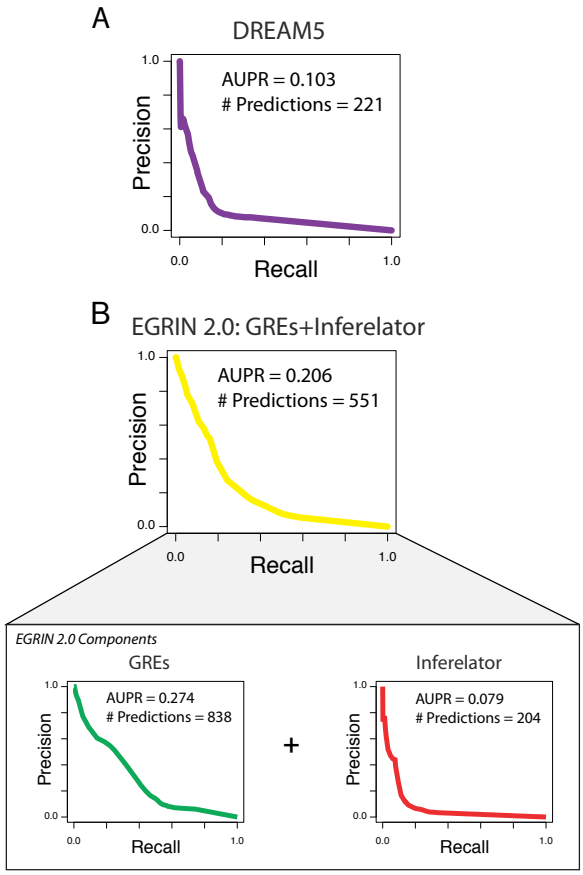
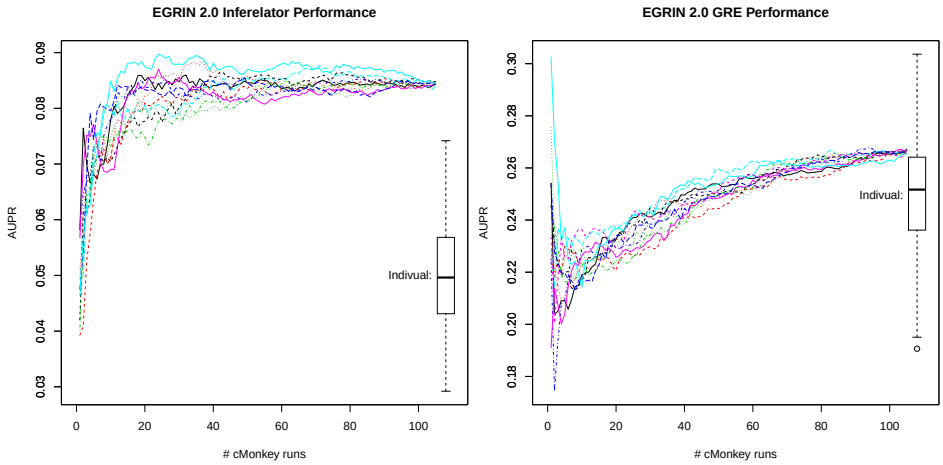


Figure 2.12: Precision-recall performance for *E. coli* networks. Comparison of precision-recall performance on *E. coli* RegulonDB gold-standard (Section 2.5.1), for the DREAM5 ensemble network (A), compared to EGRIN 2.0(B). We compare the GRE-based and Inferelator-based networks (bottom)to the integrated EGRIN 2.0 network (top). The integrated EGRIN 2.0 network consists of an equal weighting of the GRE-based and Inferelator-based networks. The EGRIN 2.0 networks were inferred using the DREAM5 mRNA expression compendium (Section 2.3.2). Area under the curve (AUPR) and the number of true-positive predictions at a precision of 25% are listed for each curve.

(Figure 2.13a), whose final AUPR of 8.5% far exceeds the rather poor performance of all 106 individual component EGRIN models (with an average AUPR of 5.0% and a maximum of 7.4%). The performance of the ensemble for this measure converges rather quickly to

the final measure, after roughly 50 of the 106 EGRIN models are integrated (taking into account the variance in models observed with integrating the models in different orders). For the EGRIN 2.0 GRE-based predicted network (Figure 2.13b), ensemble surpasses 84 (79%) of the 106 individual component EGRIN models. This measure continues to improve until $\sim$ 80 of the 106 models are integrated, suggesting that for this data set (the DREAM5 *E. coli* expression compendium), $\sim$ 100 EGRIN models was a reasonable number to use in construction of the EGRIN 2.0 ensemble.



(a) (b)

Figure 2.13: Ensemble performance of individual GRN predictions. EGRIN 2.0-inferred *E. coli* regulatory network predictive performance (AUPR vs. *E. coli* DREAM5 [237] gold standard) for Inferelator-based predictions (a) and GRE-based predictions (b) from EGRIN 2.0. Shown for both networks is the cumulative AUPR as each of the 106 individual model components is integrated in to the ensemble (as described in Section 2.5.4). Lines showing the cumulative AUPR for randomized orderings of the components' integration into the ensemble reveal the slight variations in performance that could be observed, and that these converge prior to integration of the final (106^{th}) component. Also included for comparison is a box-whisker plot which shows the distribution of corresponding AUPR scores for the 106 individual EGRIN models.

Figure 2.14 shows the inferred networks for two genes regulated by PurR and ArgR (com-

paring predictions from EGRIN 2.0, CLR, DREAM5, and RegPrecise to the annotations in RegulonDB). The result demonstrates that GRE-based approaches can discover interactions that are not predicted using direct approaches (See Section 2.5.4).

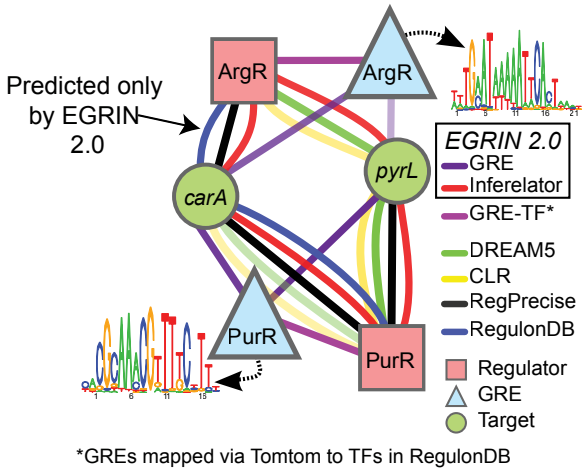


Figure 2.14: Integration of GRE discovery and Inferelator predictions yields comprehensive and detailed gene regulatory networks. EGRIN 2.0-inferred *E. coli* regulatory subnetwork for two genes (green circles) in the PurR/ArgR regulon: *carA* (*b0032*) and *pyrL* (*b4246*). The EGRIN 2.0 predictions are divided into GRE-based (dark violet) and Inferelator-based (red), and compared to predictions (or annotations) from other algorithms/databases (yellow: CLR; green: DREAM5 ensemble; black: RegPrecise; blue: RegulonDB). In two cases (ArgR→carA and ArgR→pyrL), EGRIN 2.0 discovers regulatory interactions that were missed by either hand-curated databases or expression-based inference procedures.

2.5.5 Validation of condition-specific operon isoforms by tiling array transcriptome measurements

We validated the prevalence of multiple, condition-specific transcriptional isoforms from operons in *E. coli* K-12 MG1655 by measuring changes in the transcriptome across growth, from lag-phase (OD600 = 0.05) to late stationary phase (OD600 = 7.3). The experimental platform and other experimental details are described in Section 2.5.1. We used multivariate

recursive partitioning, including signals from both relative changes in expression along the growth curve, as well as raw RNA hybridization signal to call putative transcription breaks as previously described [193]. To determine the significance of our finding, we computed a p-value describing the significance of the overlap between our predictions (see Section 2.5.2) and the experimental observations using the cumulative hypergeometric distribution.

Figures 2.15, 2.16, and 2.17 below depict several operons annotated with condition-specific transcriptional isoforms. We have integrated GRE elements discovered near break sites with the transcriptional measurements.

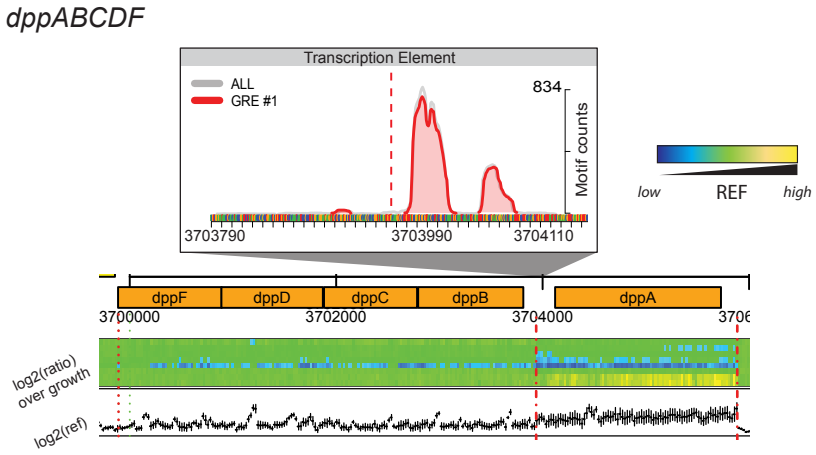


Figure 2.15: GREs regulate multiple transcript isoforms from operons in *E. coli*, *dppABCDF*. GREs coincide with experimentally measured break sites. Three examples of experimentally determined transcription break sites (red dashed lines) in operons predicted by corems to be conditionally segmented. Expression levels of these regions were profiled across growth in rich media (heatmap). Inset contains region immediately surrounding a transcriptional break site, including counts of GREs discovered at these locations.

2.5.6 Gene-gene co-fitness correlations in corems

To assess the phenotypic consequences of co-regulation in corems, we assessed whether genes grouped into corems had significantly similar fitness consequences in many environments

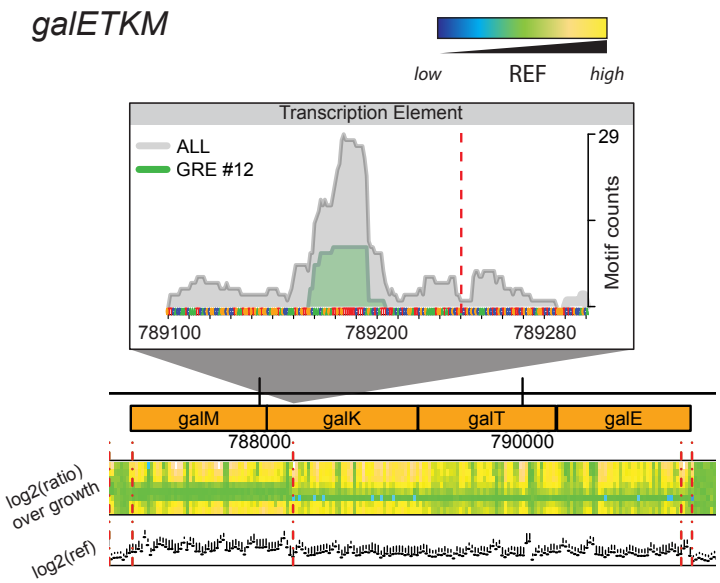


Figure 2.16: GREs regulate multiple transcript isoforms from operons in *E. coli*, *galETKM*. Caption details included in Figure 2.15

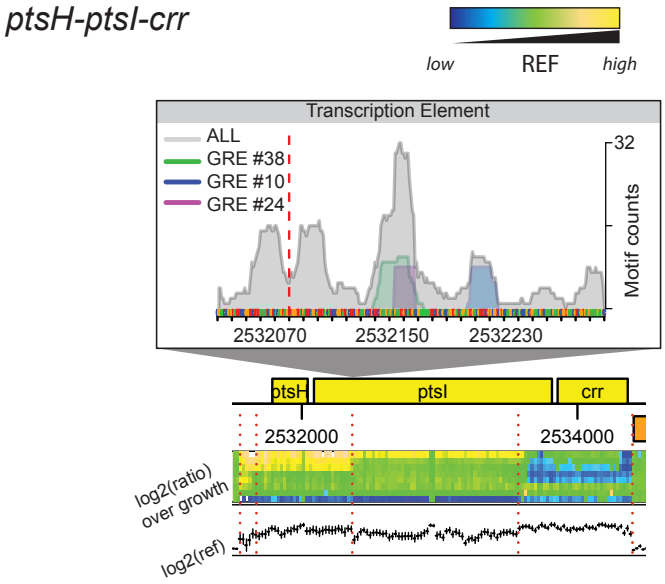


Figure 2.17: GREs regulate multiple transcript isoforms from operons in *E. coli*, *ptsH-ptsI-crr*. Caption details included in Figure 2.15.

(*i.e.*, the effect of deleting one gene is highly similar to the effect of deleting the other across many environments). We used the high-throughput fitness screen described in Section 2.5.1 to quantify these relationships.

We compared the enrichment for high co-fitness relationships in corems to other ways of assigning co-regulatory modules, including regulons (RegPrecise, RegulonDB), operons, and WGCNA. The gene modules for regulons (annotated in RegulonDB or RegPrecise [262]) consisted of genes annotated to a common TF. For WGCNA, we assigned modules using the same community detection procedures that we used to define corems from the EGRIN 2.0 ensemble (See 2.3.6). The gene co-expression modules were computed from the weighted WGCNA adjacency matrix.

For the results presented in Figure 3.3, we compared the distributions of Pearson correlations between relative changes in fitness across pairs of genes within each module, using the one-tailed Kolmogorov-Smirnov test (KS-test). We report the KS D-statistic. The precision/recall characteristics for each model are contained in Table S5.

We extended this analysis by investigating whether the enriched high co-fitness gene-gene relationships in corems consist of relationships that could be described fully by regulons or operons. To answer this question, we removed all gene pairs from corems that are also present in operons or regulons and computed the KS-test again (Figure 2.18). We still observe a significant number of high co-fitness relationships, suggesting that corems capture physiologically meaningful co-regulatory relationships between genes that cannot be explained by existing paradigms.

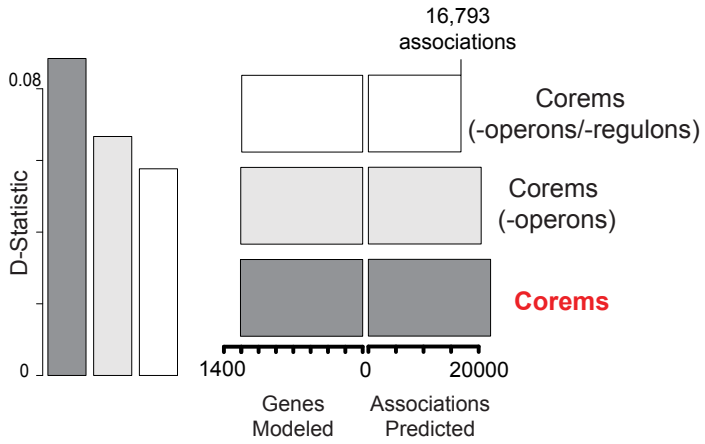


Figure 2.18: EGRIN 2.0 models highly correlated co-fitness relationships that cannot be explained by operons or regulons. (Left) Enrichment for highly correlated, pairwise fitness measurements in gene knock outs across 324 conditions before and after removing gene associations annotated by operons (Microbes Online) and regulons (RegulonDB and RegPrecise) (KS-test, D-statistic). Two-thirds of gene-pairs with most highly correlated fitness within corems are not annotated by operons or regulons. (Right) Number of genes and associations predicted.

Chapter 3

A SYSTEMS-LEVEL MODEL OF THE MICROBIAL REGULATORY GENOME

Microbes can tailor transcriptional responses to diverse environmental challenges despite having streamlined genomes and a limited number of regulators. Here, we present data-driven models that capture the dynamic interplay of the environment and genome-encoded regulatory programs of two types of prokaryotes: *E. coli* K-12 MG1655 (a bacterium) and *H. salinarum* sp. NRC-1 (an archaeon). The models reveal how the genome-wide distributions of *cis*-acting gene regulatory elements and the conditional influences of transcription factors at each of those elements encode programs for eliciting a wide array of environment-specific responses. We demonstrate how these programs partition transcriptional regulation of genes within regulons and operons to re-organize gene-gene functional associations in each environment. The models capture fitness-relevant co-regulation by different transcriptional control mechanisms acting across the entire genome, to define a generalized, system-level organizing principle for prokaryotic gene regulatory networks that goes well beyond existing paradigms of gene regulation.

This chapter has been modified from:

Brooks AN*, Reiss DJ*, Allard A, Wu W, Salvanha DM, Plaisier CL, Chandrasekaran S, Pan M, Kaur A, Baliga NS. (2014) A system-level model for the microbial regulatory genome. *Mol Syst Biol.* 10: 740.

* Indicates equal contribution

Chapter Highlights

- Method to infer a genome-wide map of gene regulatory elements (GREs) and their condition-specific activities directly from genome sequence and transcriptome profiles
- Novel co-regulatory structure, the **corem**, describes condition-specific partitioning and reorganization of operons and regulons by combinatorial and other nuanced regulatory mechanisms
- Corems group together functionally related genes that have tight co-expression in some but not all environments
- Corems associate genes from different operons and regulons that have highly similar fitness consequences

3.1 Summary

Genome-scale reconstruction of gene regulatory networks for two diverse microbial species using genome sequence and transcriptional profiles reveals complex, condition-dependent co-regulated modules (corems) and *cis*-regulatory mechanisms that generate them.

3.2 Introduction

Deciphering how microbes colonize dynamically changing environmental niches with few regulators and streamlined genomes will require mechanistic and system-level characterization of their gene regulatory networks (GRNs). Even a streamlined microbial genome encodes an intricate network of regulatory and signaling systems that sense and process extracellular and intracellular information to regulate gene expression at multiple levels (transcriptional, post-transcriptional, translational, allosteric, etc.). A significant fraction of these environmental signals are relayed by transcription factors (TFs) that modulate

transcriptional activity when they bind DNA. TFs typically bind conserved, ∼6-20 nucleotide DNA sequences located in intergenic regions immediately adjacent to transcription initiation sites. These TF binding sites are referred to as gene regulatory elements (GREs).

A goal of systems biology has been to map the complete set of TFs, GREs, and their interactions, using high throughput techniques including ChIP-chip [44], yeast two-hybrid [116], DNase I hypersensitivity [83], or more modern variants using sequencing [172]. In parallel, attempts have been made to infer GRNs directly from gene expression data [50, 90, 108, 306]. Such high throughput approaches are attractive because they would accelerate discovery in understudied organisms by circumventing significant labor and cost.

Inference of systems-scale GRNs that are both predictive and mechanistically accurate, however, has proven difficult for a number of reasons, including: (1) the statistical challenge of confidently discovering GREs across the genome, *de novo*; (2) the consequences of non-linear gene regulatory dynamics, including combinatorial molecular interactions at gene promoters; and (3) the often non-canonical locations of GREs throughout the genome (including internal to operons and within coding sequences). A remaining challenge, therefore, is to produce an unbiased map of TF-binding site locations throughout the genome, including information about what binds to those sequences, in what contexts they are bound, and, importantly, how TF-binding throughout the genome ultimately influences cellular physiology.

We previously constructed an Environment and Gene Regulatory Influence Network (EGRIN) for *H. salinarum* sp. NRC-1 [50]. This model was constructed in two steps. First, modular organization of gene regulation was deciphered through semi-supervised biclustering of gene expression, guided by biologically informative priors and *de novo cis*-regulatory GRE detection for module assignment (cMonkey; [284]). Second, using a regression-based approach transcriptional changes of genes within each bicluster were modeled as a linear combination of influences of TFs and environmental factors (Inferelator; [47]).

The EGRIN networks learned by cMonkey and Inferelator accurately predicted transcriptional changes in new environments, a feat that has subsequently been replicated by

other network inference strategies [108, 208, 237]; yet, these network models have failed to capture detailed regulatory mechanisms that operate only in specific environments, at non-canonical genomic locations, or in complex combinatorial schemes.

Here, we report significant advancement to inference of GRNs that overcomes many of these challenges. We have developed a methodology applicable to any sequenced microbe in culture to infer EGRIN 2.0 models for two representative organisms from the primary branches of prokaryotic life - bacteria and archaea: (1) *E. coli* K-12 MG1655 , a bacterium with a wealth of information about transcriptional regulatory mechanisms and related experimental data [293]; and (2) *H. salinarum* sp. NRC-1 , an archaeon with few examples of regulatory mechanisms that have been characterized in detail, but extensive experimental data from recently conducted systems biology studies [50, 193]. The wide range of prior knowledge for these organisms proved invaluable for testing our model. In addition, we have also conducted new experiments that validate EGRIN 2.0 predicted complex modulation of the *E. coli* K-12 MG1655 transcriptome structure during varying stages of growth in rich media.

EGRIN 2.0 models the organization of GREs within every promoter, their distributions across the entire genome even in non-canonical locations and links the contexts in which they act to conditional co-regulation of genes. These features are formalized in EGRIN 2.0 by condition-specific, co-regulated modules or corems. Corems are overlapping sets of co-regulated genes that, in some cases, group together genes from different regulons and, in other cases, subdivide genes of the same regulon, or even the same operon. EGRIN 2.0 formalizes how the genome-wide coordination of previously characterized and newly discovered regulatory mechanisms dynamically associates genes into corems, bringing together functionally-related genes from different operons and regulons whose deletions have similar impact on cellular fitness. Our results show how prokaryotes, much like eukaryotes, can produce complex gene expression patterns with a relatively small number of regulatory components.

3.3 Results

3.3.1 Construction of EGRIN 2.0

We developed an ensemble-framework that models the condition-specific global transcriptional state of the cell as a function of combinations of transient TF-based control mechanisms acting at intergenic and intragenic promoters across the entire genome. Specifically, for each of the two orgainsms, *H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655 , we aggregated associations across genes, GREs, and environments from many individual EGRIN models, each trained on a subset of the gene expression data, to (1) quantify confidence in each model-predicted association; (2) reveal context-dependent regulatory mechanisms that occur infrequently in the data; and (3) discover non-canonical regulatory mechanisms. We refer to the aggregated, post-processed ensemble of EGRIN models as EGRIN 2.0, and conditionally co-regulated modules as corems (details provided in Chapter 2, Figure 3.1; ensemble statistics available in Table 2.1).

3.3.2 EGRIN 2.0 discovers experimentally characterized regulatory mechanisms

A high quality GRN has to be both comprehensive (high recall) and accurate (high precision). To evaluate the quality of EGRIN 2.0, we compared its predictions on *E. coli* K-12 MG1655 to RegulonDB [125], an extensive, manually-curated, gold-standard of experimentally validated TF-gene interactions. We compared the genome-wide distribution of each *de novo* discovered GRE in EGRIN 2.0 to experimentally characterized binding locations of every TF in RegulonDB. This comparison showed that EGRIN 2.0 had accurately located binding sites for 60% of experimentally characterized TFs in RegulonDB (53 out of 88 at FDR ≤ 0.05 for all TFs with $\geq$3 unique sites; see Chapter 2). At a standard precision cutoff of 25%, EGRIN 2.0 recovered 577 "strong evidence" TF-gene interactions, which is 2.7X as many validated interactions as algorithms that exclusively use expression data, i.e. without genomic sequence information (Figure 3.2, Figure 2.12, Figure 2.13, Figure 2.14, Table S4, 2; [108, 237]. As expected, the ensemble network had greater precision and recall than indi-

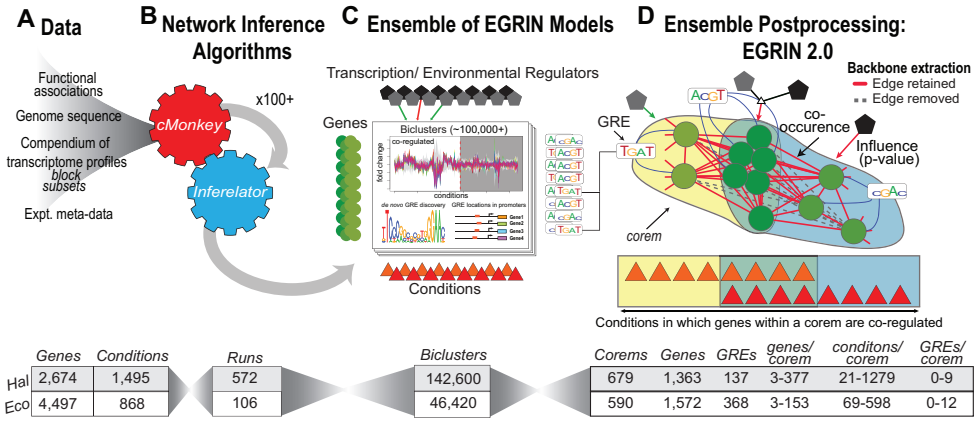


Figure 3.1: EGRIN 2.0 Model Construction. Workflow summary for EGRIN 2.0. Tables below each panel contain detailed statistics for the *H. salinarum* sp. NRC-1 and *E. coli* K-12 MG1655 models. (A) and (B). The cMonkey and Inferelator algorithms were applied many times to subsets of gene expression data from large compendiums of transcriptome profiles to construct many individual EGRIN models.(C) Individual EGRIN models were integrated into an ensemble for filtering, querying, and ranking relationships among genes (circles), regulators (hexagons), motifs (sequence logos), and the conditions (triangles) in which these relationships were discovered.(D) The library of relationships was mined using algorithms for motif clustering, backbone extraction, and community detection to construct the final EGRIN 2.0 model. In EGRIN 2.0, overlapping co-regulated sets of genes (corems, shaded regions of the graph) are statistically associated with specific gene regulatory elements (GREs, sequence logos, blue edges), regulatory influences (pentagons, green or red depending on direction), and environments in which they are co-regulated (triangles). Each node represents a gene in the model. Genes are connected via co-regulation edges, with weights that reflect the number of occurrences in the ensemble. Dashed edges were removed from the model by backbone extraction.

vidual cMonkey runs. Furthermore, integration of Inferelator-predicted TF influences with GRE-based predictions increased overall algorithm performance. These results show that integrating complementary methods, such as regression-based inference of TF regulation,

biclustering-based inference of network modularity, and *de novo* GRE detection, improves the accuracy and coverage of the inferred GRN.

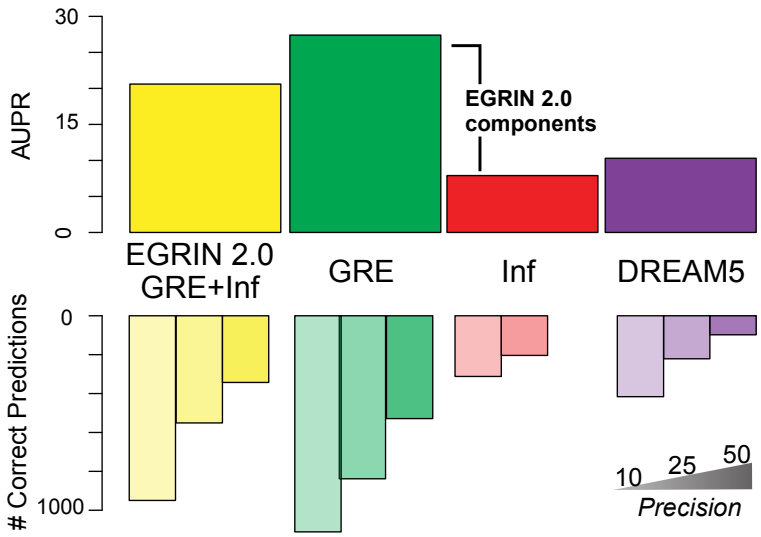


Figure 3.2: EGRIN 2.0 performance on experimentally-validated gold-standard network. Comparison of EGRIN 2.0 model components (GRE: GRE-only; Inf: Inferelator-only) to the DREAM5 community ensemble network, against RegulonDB (strong evidence code). (Top) Area under the precision-recall curve (AUPR) and (Bottom) number of correct predictions at 10, 25, and 50% precision.

Since few GREs have been characterized in *H. salinarum* sp. NRC-1 , we performed a global assessment and discovered that GREs in EGRIN 2.0 occur at consistent locations across many gene promoters throughout the genome (Figure 2.4). We could even assign putative roles for some GREs based on their location relative to transcription start sites (TSSs). For instance, the location of TATA box-like elements (GRE #25) between -21 to -40 nucleotides upstream of TSSs in *H. salinarum* sp. NRC-1 is consistent with the characterized location of basal elements in archaeal promoters (TFB/TBP complex recognition sites)

[128]. Similarly, other elements occurred either consistently downstream of the TATA box (putative repressors, e.g. GRE #1 and #2) or upstream of these basal elements (putative activators, e.g. GRE #5). Thus, even in organisms where genome-wide TF binding data are scarce, EGRIN 2.0 can be used to infer and predict putative roles for *de novo* discovered GREs.

3.3.3 Corems model genes with similar effects on organismal fitness

We investigated whether the model goes beyond simple co-expression to group together genes that have similar phenotypic contributions. We did this because previous studies have reported weak correlation between gene expression and fitness [278]. For all genes in each corem, we computed pairwise correlations of fitness effects in a dataset generated from a survey of relative growth rates for 3,902 single gene deletion strains of *E. coli* K-12 MG1655 subjected to a chemical genomics screen spanning 324 different environmental conditions [259]. We discovered that more than one-third of gene-pairs with the most similar fitness effects across environments (Pearson correlation > 0.75) were grouped together in corems. We evaluated significance of this result by performing similar analysis using modules based-on co-expression (WGCNA; [202]), and regulons (RegPrecise and RegulonDB), where a regulon is defined as a set of genes regulated by the same TF. While WGCNA and regulons also grouped significant numbers of high fitness-correlated gene-pairs (one-sided KS-test < 0.05), corems were more enriched for highly similar fitness associations (higher KS D-statistic) and in general provided greater precision and coverage (Figure 3.3). As an example, corems group together 5X as many gene-pairs with highly correlated fitness effects as RegPrecise, RegulonDB, or WGCNA (134 out of 185 gene-pairs with Pearson correlation 0.9 are discovered in corems, Table S5). Most important, corems retained a high degree of enrichment for gene-pairs with highly correlated fitness effects after removing all associations attributable to operon and regulon memberships, and even combinatorial control (Figure 2.18, Table S5[1]). This suggested that corems model regulatory associations among genes

[1] All tables available in [59] and online

that cannot be explained within the existing paradigms of regulons and operons.

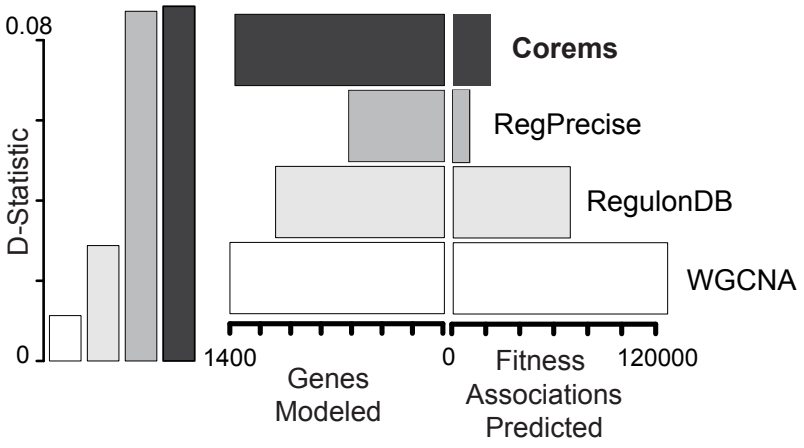


Figure 3.3: Enrichment of similar fitness effects within gene modules. (Left) Magnitude of enrichment for gene pairs with similar fitness consequences, assessed by one-tailed KS-test (KS D-statistic). (Right) Number of genes and gene-pairs predicted by each method. Comparison methods include EGRIN 2.0 corems, co-expression modules from WGCNA, and regulons from databases (RegPrecise and RegulonDB).

In other words, corems group together genes that are regulated by distinct TFs. For example, the ArgR-regulated acetylglutamate kinase, *argB*, and *ilvC*, an IlvY-regulated ketol-acid reductoisomerase have fitness correlation of 0.95 (Pearson coefficient), which suggests an important coupling between branched-chain amino acid biosynthesis and arginine metabolism (Table 3.1). Although these genes are regulated by distinct TFs (ArgR and IlvY, respectively), the high similarity of their expression changes across multiple environments brings them together into the same corem (ec512157). There are 319 highly correlated (Pearson correlation ≥ 0.75) fitness associations among genes from different regulons that are modeled by corems each of which suggests an important physiological coupling that results from the coordinated activity of TFs (Table E6). These examples illustrate how the

organizing principle of corems captures fitness-relevant associations within a GRN that are overlooked by current definitions for gene-gene co-regulation, such as regulon and operon.

3.3.4 EGRIN 2.0 predicts detailed organization and context-specific importance of GREs in gene promoters

We next investigated accuracy of EGRIN 2.0 predicted spatial organization of GREs, and their context-specific roles in mediating transcriptional regulation from specific promoters. We did this analysis in context of one of the best studied *H. salinarum* sp. NRC-1 promoters: *kdpFABC*, with data not used for model training. The *kdp* operon encodes an ATP-dependent potassium transporter that counterbalances extremely high salinity in the extracellular environment. EGRIN 2.0 predicts that at least three GREs are putatively responsible for mediating transcriptional regulation of this operon: GRE #1, GRE #148, and GRE #106 (Figure 3.4). The locations of these GREs align to regions that were experimentally characterized in an independent study as Operator and BRE-TATA elements, respectively. This demonstrates that EGRIN 2.0 is able to accurately predict the organization of GREs in gene promoters at nucleotide-resolution.

Since these sites also had characterized transcriptional roles [determined by promoter truncation experiments [191]], we asked whether EGRIN 2.0 would have been able to predict these roles given the context in which the GREs were discovered. Strikingly, we find that GRE #1 (aligned to the Operator) was discovered in environments, including low salt (hypergeometric FDR = 6.9×10^{-12}), where the transcript is repressed (one-sided t-test pval = 0.048), while GRE #106, which aligns to the BRE-TATA region, was discovered in environments, including low oxygen (hypergeometric FDR = 1.8×10^{-9}), where transcript levels are elevated (one-sided t-test pval = 1.2×10^{-3}; 2). Here onwards, we will refer to a GRE as "active" when it is predicted to be important for transcriptional regulation at a specific promoter (see Figure 2.10 for details). The environmental contexts in which the three GREs in the *kdp* promoter are predicted to be active are especially interesting because perturbations to external potassium levels and energy-producing mechanisms have

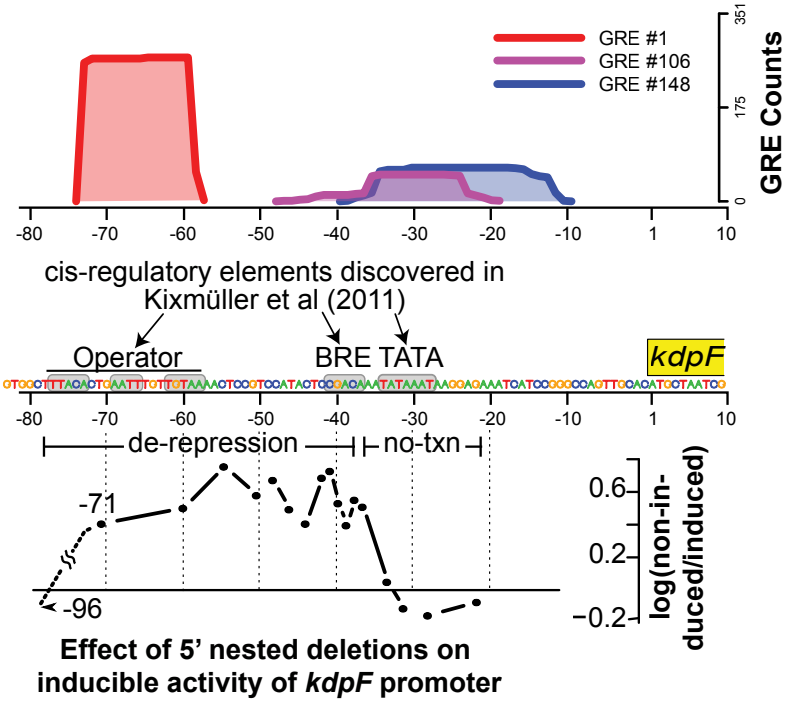


Figure 3.4: Promoter architecture of the *H. salinarum* sp. NRC-1 *kdpFABC* promoter predicted by the EGRIN 2.0 model. (Top) Frequency of GRE alignment to each position in the *kdpFABC* promoter. GREs are indicated by shaded lines. (Middle) Genome sequence marked with putative functions by [191]. (Bottom) Transcriptional activity measurements from truncated promoters used by authors to validate these sites.

been shown to significantly influence expression of this operon [377]. Thus, EGRIN 2.0 had accurately predicted that a trade-off in relative influence of GRE #1 (repressing) versus GRE#106 (activating) controls expression levels of this operon in a condition-specific manner, exactly as was characterized by independently performed experiments. This is powerful because it shows that using EGRIN 2.0 we can predict when (context) and how (activate or repress) a specific GRE(s) within a promoter might act, even though we might not know the precise regulatory mechanism (e.g., TF binding/unbinding, allosteric activation, co-factor

interaction, etc).

3.3.5 Conditionally active GREs within each promoter reorganize gene memberships within corems

We investigated whether EGRIN 2.0 accurately links the same GRE at different promoter locations, the environments in which it is predicted to be active within each of those promoters, and conditional co-regulation of the associated genes (see 2). We did this analysis with genes of nucleotide biosynthesis in *E. coli* K-12 MG1655 , including key branch-point enzymes *carA* (*b0032*) and *pyrL* (*b4246*), since they are canonical, extremely well studied pathways that are critical for survival. Regulation of *carA*, which catalyzes synthesis of an important metabolic intermediate in several amino acid and nucleotide metabolism pathways (carbamoyl phosphate), is known to be sensitive to purine and pyrimidine pools, as well as arginine [256]. EGRIN 2.0 discovered several previously characterized and new mechanisms for regulation of carA, including two GREs (GRE #4 and GRE #12) that match to consensus sequence motifs for PurR and ArgR, respectively [277] (Figure 3.5). Remarkably, EGRIN 2.0 discovered novel overlapping organization of GRE #4 and GRE #12 in the *pyrL* promoter that was not previously reported in RegulonDB (Figure 3.6). This promoter organization was verified upon mapping overlapping binding sites for ArgR and PurR precisely at the predicted locations in ChIP-chip data that were not used in model training [72, 73].

We were most interested, however, to understand the consequences of conditional regulation at ArgR and PurR-associated GREs on variable expression of *carA* in different environments. Indeed, EGRIN 2.0 predicts three condition-specific states of the *carA* promoter with respect to when PurR- and ArgR-matched GREs are conditionally active: (1) high PurR and high ArgR, (2) low PurR and high ArgR, and (3) high PurR and low ArgR (Figure 3.5). Interestingly, two of these promoter states correspond to co-regulation of *carA* with a different combination of genes (i.e., different corems), functionally separating pyrimidine from purine biosynthesis (Figure 3.18), while the third state is not associated with co-regulation of *carA* with the genes of any corem. Thus, the context in which GREs

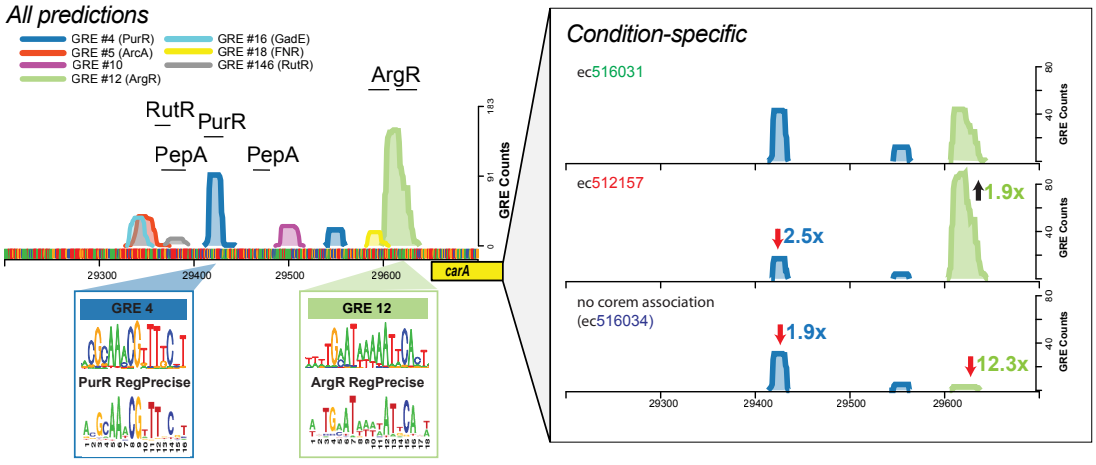


Figure 3.5: Predicted architecture of the *E. coli* K-12 MG1655 *carA* promoter across all ensemble predictions (as in Figure 3.4). Horizontal bars above peaks mark experimentally characterized TF binding sites (RegulonDB). Significant GRE matches to characterized *E. coli* K-12 MG1655 binding sites in RegulonDB are indicated in parentheses. (Right) Condition-specific states of the *carA* promoter. Variation in conditional discovery of GREs (counts and fold-change relative to ec516031, top) suggests when they are active across three different subsets of experimental conditions in the carA promoter. (Bottom). Condition subsets correspond to co-regulation of *carA* with genes in the nucleotide and pyrimidine corems (ec516031, ec512157) or environments where *carA* is not co-regulated with genes in any corem (ec516034). Motif logos for GRE #4 (PurR) and GRE #12 (ArgR) from the EGRIN 2.0 predictions compared to logos from RegPrecise.

are active accurately explains when and how genes are co-regulated in different overlapping combinations to perform distinct functions.

3.3.6 Conditionally active GREs within operons generate multiple, overlapping, and differentially regulated transcript isoforms

Some of the GREs discovered in EGRIN 2.0 occur in non-canonical locations and lead to unexpected transcriptional behaviors, such as the subdivision of operons into multiple tran-

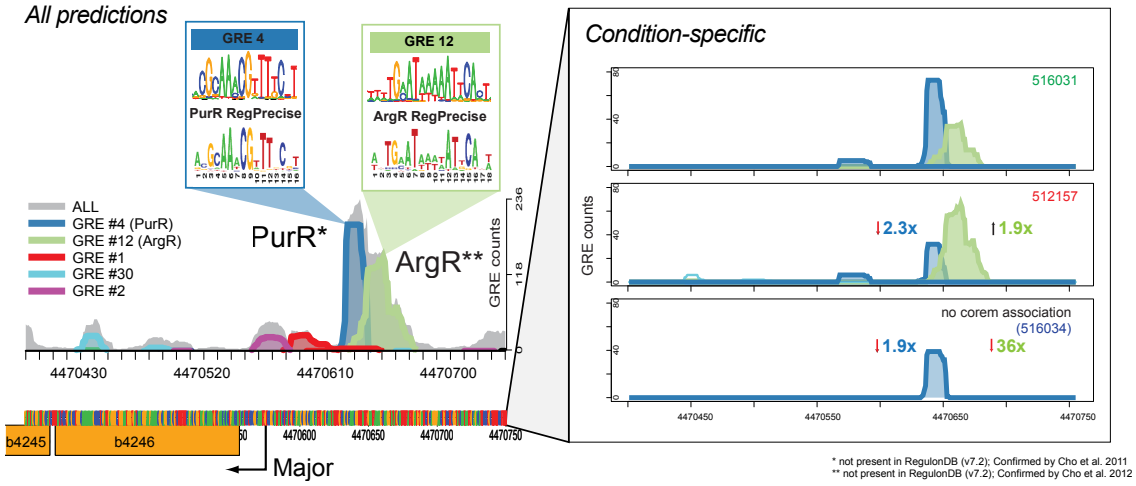


Figure 3.6: Differential GRE activity in *pyrL* promoter, *E. coli*. (Left) Predicted promoter architecture for *E. coli pyrL* (b4246). Overlapping GREs matching to PurR (GRE #4) and ArgR (GRE #12) were detected upstream of *pyrL*. These sites were not annotated in RegulonDB, but were validated in independent ChIP-chip experiments [73, 72]. Transcription start site indicated with arrow. (Right) Condition-specific promoter architectures for *E. coli pyrL* (as in Figure 3.5). Variation in predicted GRE activity across three different subsets of experimental conditions (counts and fold-change) for two GREs in the *pyrL* promoter. Experimental subsets correspond to conditions under which at least one of three nucleotide biosynthetic corems is regulated (denoted by colored names at top-right of each plot)

scriptional units. Previously, we reported pervasive modulation of the *H. salinarum* sp. NRC-1 transcriptome structure by transcriptional elements that are located within operons and coding regions [193]. EGRIN 2.0 recapitulated this phenomenon by sub-dividing operon genes into different corems. In all, the model predicted that nearly one-third of all *H. salinarum* sp. NRC-1 operons generate multiple transcript isoforms (Figure 3.7, Figure 3.8,Figure 3.9, Chapter 2 for details). Nearly half of these predictions of conditional operon structures were corroborated by experimentally mapped transcriptional breaks (hypergeometric pval $= 4.2 \times 10^{-3}$; Table S7; Koide et al., 2009). Often, these transcript boundaries

were adjacent to GREs that coincide with experimentally determined TFB-binding sites ([107]; Figure 3.10), reinforcing the accuracy of EGRIN 2.0 predictions.

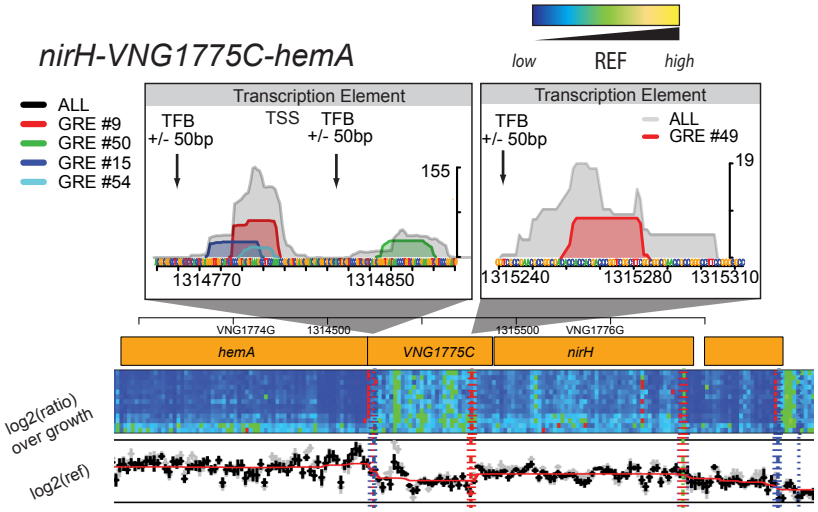


Figure 3.7: GREs regulate multiple transcript isoforms from operons in *H. salinarum, nirH-VNG1775C-hemA*. GREs located inside operons coincide with experimentally measured transcriptional break sites. Experimentally determined tran- scription break sites (red dashed lines) above expression profiles of these regions across growth (heatmap [193] and ChIP-chip TFBs [107], vertical arrows) support the role of GREs in regulating segmentation of the operon in certain conditions. Insets contain regions immediately surrounding transcriptional break sites, including counts of GREs discovered at these locations.

We further investigated whether EGRIN 2.0 provides insight into downstream consequences of differentially regulating multiple transcript isoforms from the same operon. The *dppAB1C2-oppD2-ykfD-VNG2342H* operon (hereafter the *dpp* operon) in *H. salinarum* sp. NRC-1 encodes an ATP-dependent dipeptide transporter. Some periplasmic binding proteins (like *dppA*) have the reported ability to function in conjunction with different ABC transport systems, giving support to the hypothesis that *dppA* can be regulated independently [153]. Despite high co-expression of the entire operon in the training data (mean

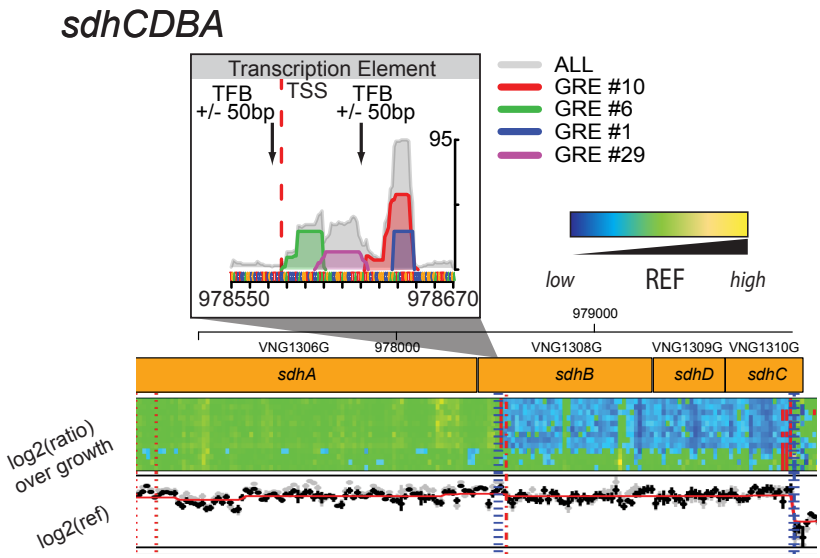


Figure 3.8: GREs regulate multiple transcript isoforms from operons in *H. salinarum, sdhCDBA*. Caption details included in Figure 3.7.

R^2 = 0.6 across 1495 conditions), EGRIN 2.0 predicted that the genes of this operon are transcribed as three different isoforms, each co-regulated with genes of a different corem: (1) the entire operon (hc21645 *dpp* corem), (2) the entire operon except the leader gene, *dppA* (hc21279 permease corem), and (3) just *dppA* (hc6326 leader corem). These predicted isoforms were verified by experimentally mapped transcript boundaries (Figure 3.11). Each of these corems contains a different dpp isoform and is enriched for a different biological function, including vitamin biosynthesis, porphyrin metabolism, and purine biosynthesis, respectively (Figure 3.11). Predicted differential regulation of the core permease (*dppB1C2-oppD2-ykfD-VNG2342H*) with porphyrin metabolism genes in the permease corem is consistent with the reported capability of this transporter system to uptake heme when it functions with a different solute binding protein (i.e., without *dppA*; [223]). Overall, EGRIN

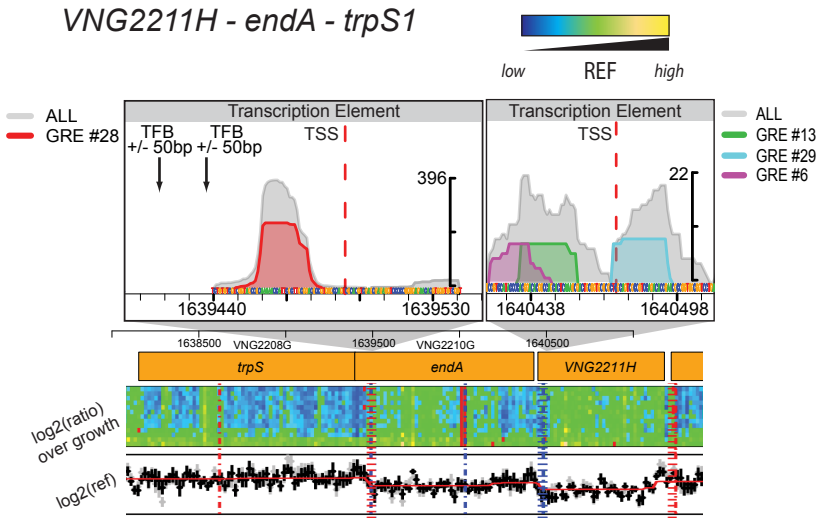


Figure 3.9: GREs regulate multiple transcript isoforms from operons in *H. salinarum, VNG2211H-endA-trpS1*. Caption details included in Figure 3.7.

2.0 provided insight into the distinct environment-dependent functional associations of each transcript isoform.

Further, EGRIN 2.0 revealed that segmentation of the *dpp* operon into multiple corems is mediated by conditionally active GREs located both upstream and internal to the operon. For example, EGRIN 2.0 predicted that GRE #6 was responsible for disassociating *dppA* transcription from the remainder of the operon. Interestingly, GRE #6 was also discovered in the promoters of nearly all of the other genes in the leader corem (Figure 3.10, Figure 3.13, Table S8). Similarly, GRE #1 was implicated in co-regulating the permease-encoding transcript with other genes in the permease corem, and GRE #17 for co-regulating the entire operon with other genes in the *dpp* corem. EGRIN 2.0 also predicted specific segmentation pattern of the dpp-operon during lag growth phase. This prediction was verified upon observing that a transcript break appears downstream to *dppB1* precisely when a batch culture transitions from lag to log phase of growth (indicated by arrow in Figure

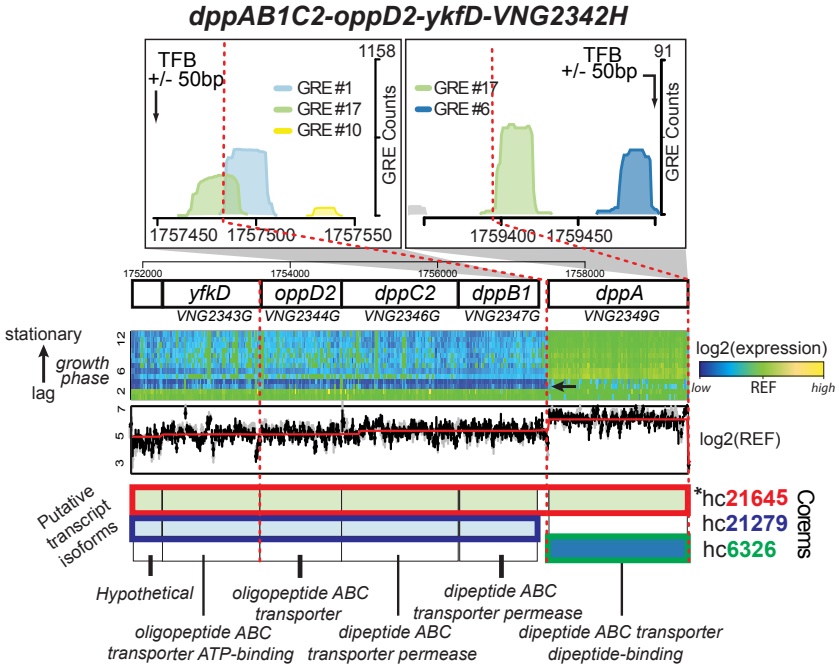


Figure 3.10: (Top) Predicted GREs located within (left) and upstream of (right) the *H. salinarum* sp. NRC-1 *dpp* operon. Locations of experimentally mapped TFB binding sites (vertical arrows; [107]), and experimentally mapped transcription break sites (vertical red dashed lines, see (B); [193]) are indicated. Locations of predicted GREs relative to coding segments of the *dpp* operon. (Midlle) Expression changes during growth in the genomic region covering the *dpp* operon measured by high-resolution tiling microarray. Raw RNA hybridization signal from mid-log growth phase indicated below. (Bottom) Three predicted transcripts from the *dpp* operon. Internal colors correspond to the GREs putatively responsible for regulating each transcript (shown at top, derived from corem membership in Figure 3.11). Boxed colors indicate corem membership for each transcript (described in Figure 3.11). Red dashed lines indicate experimentally measured transcription break sites. Transcriptional break at lag phase highlighted by an arrow. Functional annotation for each gene located at bottom.

3.10 heatmap). This is just one of 98 operons with experimentally validated conditional isoforms in *H. salinarum* sp. NRC-1 . For each instance, a similar correspondence between mechanism, context, and function could be demonstrated (Figure 3.7, Figure 3.8,Figure 3.9, and online). Interestingly, even in *E. coli* K-12 MG1655 , where previous studies report a single transcript for the *dpp* operon [2], EGRIN 2.0 discovered that it is actually transcribed as multiple, condition-specific transcript isoforms, each of which participates in a different physiological process (Figure E5A).

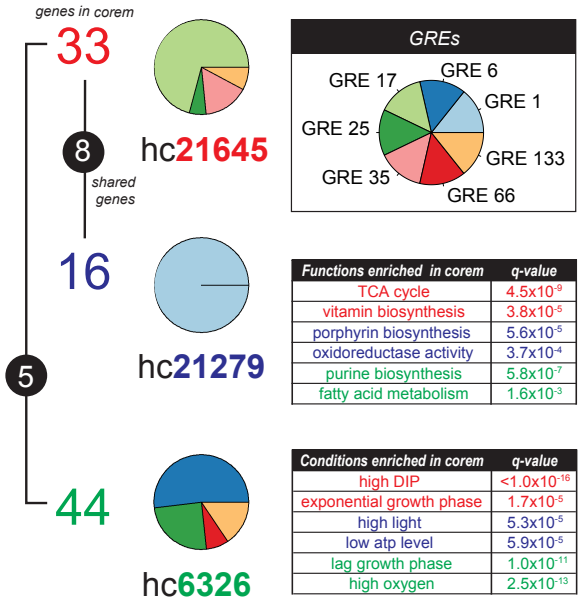


Figure 3.11: (Left) Three *H. salinarum* sp. NRC-1 corems model differential regulation of *dpp* operon isoforms: (1) the entire operon (hc21645 *dpp* corem; top); (2) five tail genes, excluding *dppA* (hc21279 permease corem; center); and (3) the leader gene, *dppA* (hc6326 leader corem; bottom). Colored numbers denote quantity of genes in each corem; numbers in black shaded circles indicate the number of genes shared between corems. Pie charts represent average predicted influence of GREs on regulation of genes in each corem (see Figure E3B for detail). (Top-Right) Pie chart key indicates GRE identity. (Bottom-Right) Tables list enriched gene functions [94] and environmental conditions for each of the corems (computed using the environmental ontology; see Chapter 2)

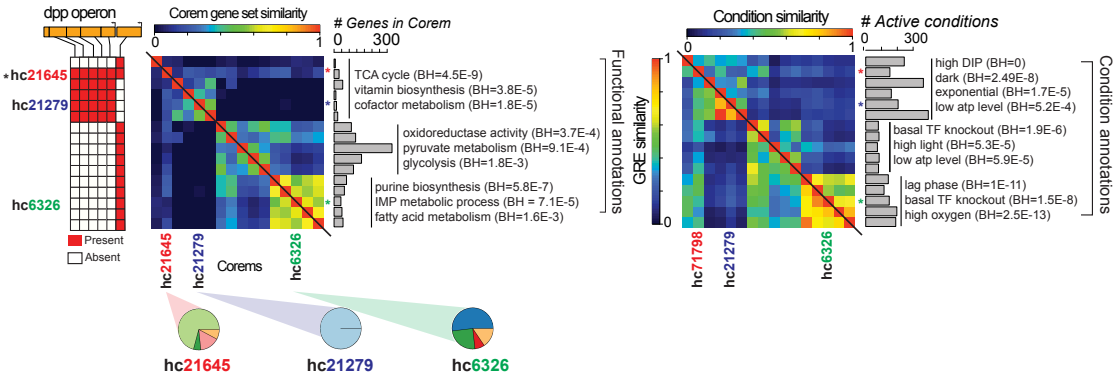


Figure 3.12: Corems group together functionally related sets of genes that are co-regulated in similar environments by similar factors (Left) Presence/absence of *dpp* operon genes in corems. Three classes of corems exist for the dpp operon: (1) the entire operon (e.g.hc21645), (2) the leader gene dppA (e.g.hc6326), and (3) five "tail" genes excluding *dppA* (hc21279). (Middle) Gene similarity between corems (heatmap, Jaccard index). Functional annotations of genes in three highly similar clusters of corems to right. GRE composition for three corems shown below (pie chart, see Figure 2.10). (Right). Similarity of conditions regulated (heatmap, upper triangle, Jaccard index) and GREs (heatmap, lower triangle, Jaccard index) among corems. Ordering is identical to (Middle). Environmental Ontology term enrichment (see Chapter 2) for three clusters depicted to right.

While we were aware of extensive transcriptional heterogeneity within operons in *H. salinarum* sp. NRC-1 , we were surprised that EGRIN 2.0 predicted that the same phenomenon also occurred extensively in *E. coli* K-12 MG1655 . To see if this were true, we mapped the *E. coli* K-12 MG1655 global transcriptome structure across varying phases of growth in rich media using a densely tiled microarray (see Chapter 2). We used this new gene expression data set to identify the corems in which different combinations of operon genes (i.e., transcript isoforms) were co-regulated in some or all phases of growth, and to characterize transcriptional breaks using previously developed methodologies [193]. We observed transcriptional breaks in nearly 20 percent of operons (including the *E. coli* K-12 MG1655 *dpp* operon) just over this 9-time point growth study, validating EGRIN 2.0 predic-

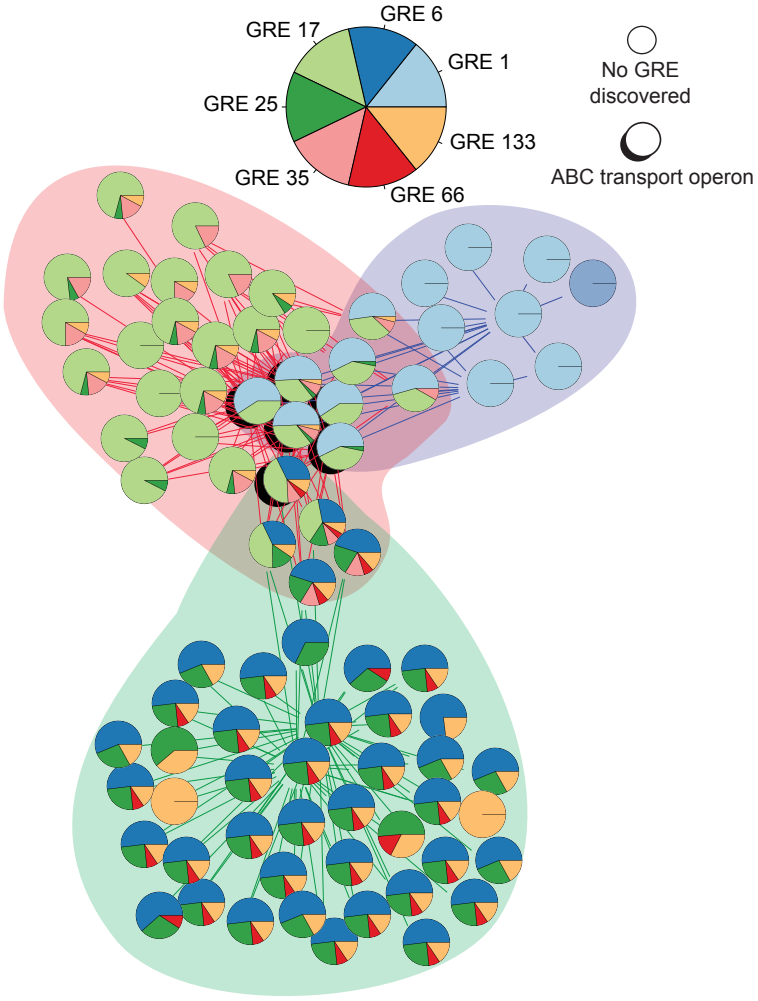


Figure 3.13: Network representation of transcriptional isoforms for the *dpp* operon predicted by corems. Genes represented by circles. Edge colors and colored region behind the network indicate corem membership. Pie charts reflect GRE composition of each gene (see Figure 2.10). Key for pie charts at top. Shading behind nodes (center of network) indicates dpp operon genes.

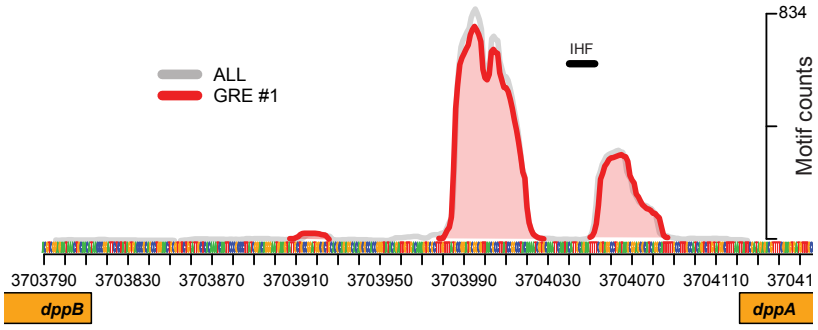


Figure 3.14: EGRIN 2.0 predicts conditional modulation of *dpp* operon in *E. coli* as well. Promoter architecture within intergenic space between *dppA* and *dppB* suggested locations for TF binding internal to the operon (as in Figure 3.10). GRE binding sites are proximal to an experimentally characterized IHF binding site (black horizontal bar; RegulonDB).

tion that nearly one-quarter of all *E. coli* K-12 MG1655 operons have conditional isoforms during varying stages of growth (hypergeometric pval = 1.07×10^{-5}, Figure 2.15, Figure 2.16, Figure 2.17, Table S7). Experimental validation of this enormous transcriptional heterogeneity among operons in *E. coli* K-12 MG1655 demonstrates the power of EGRIN 2.0 to distinguish nuanced patterns in complex data, and provide both mechanistic explanation and context for when and why the novel phenomena might occur.

3.3.7 Some TFs act similarly across certain environments to co-regulate functionally-related subsets of genes across their respective regulons

We investigated whether EGRIN 2.0 provides insights into context-dependent differential regulation of branched metabolic pathways even those that have been meticulously studied for decades, such as *de novo* biosynthesis of nucleotides in *E. coli* K-12 MG1655 [256]. At least seven GREs were implicated in partitioning (purine biosynthesis: ec516034 purine corem; pyrimidine biosynthesis: ec512157 pyrimidine corem) or co-regulating (ec516031 nucleotide corem) nucleotide biosynthesis into multiple overlapping corems (Figure 3.17,

Figure 3.15, Figure 3.16). The genome-wide locations for four of these GREs significantly overlapped with known binding locations for PurR, ArgR, MetJ, and IclR. Partitioning and co-regulation of purine and pyrimidine biosynthesis can be attributed to the location of these GREs in promoters of pathway genes, including *carA*, and the environments in which they are predicted to be active (Figure 3.5). EGRIN 2.0 predicts, for example, that MetJ (GREs #19, #87) acts in conjunction with PurR (GRE #4) to differentially regulate genes specific to the pyrimidine biosynthetic branch (pyrimidine corem), while (yet to be identified) TFs that bind GREs #2 and #206 function with PurR (GRE #4) to regulate genes in the purine branch (purine corem) (Figure 3.18). The organization of these GREs within and across promoters, and the environments in which they act to mediate regulation by specific TFs, generates complex co-expression patterns among different combinations of genes in the three corems of this highly canalized pathway (filled violin plots, Figure 3.19). These conditional co-expression patterns predict that in certain environments the two branches are differentially regulated, while in others they are co-regulated as one unit. Consistent with this observation, fitness consequences of deleting genes in these corems vary across conditions (Figure 3.20, Figure 3.21, Figure 3.22). For instance, knockouts of genes in all three corems have similar consequences on fitness in the presence of glucose. By contrast, in the presence of the toxic ionophore carbonyl cyanide m-chlorophenyl hydrazone (CCCP), only knockouts of genes in the nucleotide corem significantly alter fitness.

This example highlights two important features of EGRIN 2.0 and corems. First, EGRIN 2.0 can distinguish co-regulation by independent, similarly-acting TFs, even though their targets are co-expressed. Further, corems group together genes that are functionally-related even though their co-regulation is mediated by different mechanisms, demonstrating how conditional TF-influences in a GRN coordinate transcription of genes from different regulons whose deletions have highly correlated fitness consequences (Table 3.1). Genes of the pyrimidine corem, for example, are co-regulated by as many as five TFs. Even though promoters of each of the genes in this corem contain distinct compositions of GREs (Figure 3.16, Tables S9-S10), their expression is highly coordinated across a broad range of

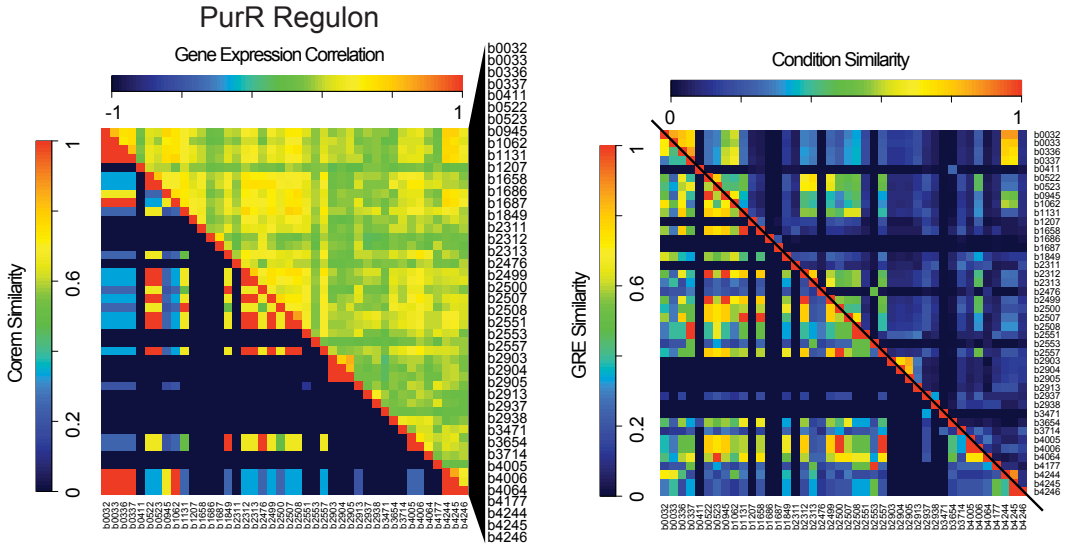


Figure 3.15: (Left) Corems identify the most highly correlated subgroupings of genes in PurR regulon. Gene expression correlation across all experiments (upper triangle) compared to similarity of corem membership (lower-triangle, Jaccard index) for genes of the PurR regulon (gene identifiers expanded to right). (Right) Similarity of regulated conditions (upper triangle, Jaccard index) and GREs com- position for these genes (bottom triangle, Jaccard index). Consistent patterns of conditional-activity and GRE composition in their promoter regions further supports subdivision of PurR genes into separate corems. Gene order is same as left.

conditions. Interestingly and counter to our expectation, transcript level changes of the similarly acting TFs are not highly correlated. Instead, we discovered correlated changes in the concentrations of effector molecules, which allosterically regulate the activities of these TFs, suggesting that coordinate regulation of genes in the pyrimidine corem is a direct consequence of metabolic dynamics (Figure 3.23; [166]).

Second, EGRIN 2.0 predicts that not all locations that match to the same GRE are functionally equivalent in all environments. Accordingly, using corems we can discern and explain why genes regulated by the same TF exhibit different expression patterns in certain

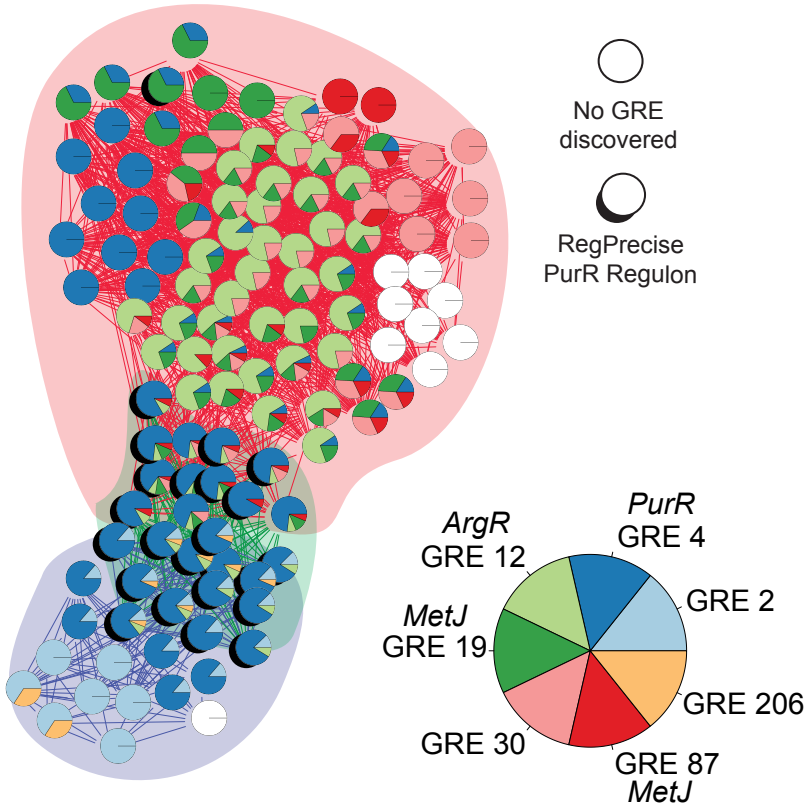


Figure 3.16: Network representation for three corems described in Figure 3.17. Genes are represented by circles. Edge colors and colored region behind the network indicate corem membership. Pie charts reflect GRE composition of each gene (see Figure 2.10). Key for pie charts at bottom. GRE-TF matches are indicated. Shading behind nodes denotes PurR regulon genes. At least 7 different mechanisms regulate the expression of these genes.

environments. For example, out of the 42 PurR-regulated genes (assigned by RegPrecise), expression changes of the 14 that are grouped into the purine corem are better correlated with each other and genes of this corem than they are to the portion of the PurR regulon that was left out (t-test, pval $< 2.2 \times 10^{-16}$, Figure 3.24). Consistent with this observation, PurR is predicted to play a variable role in regulation of genes across the three corems

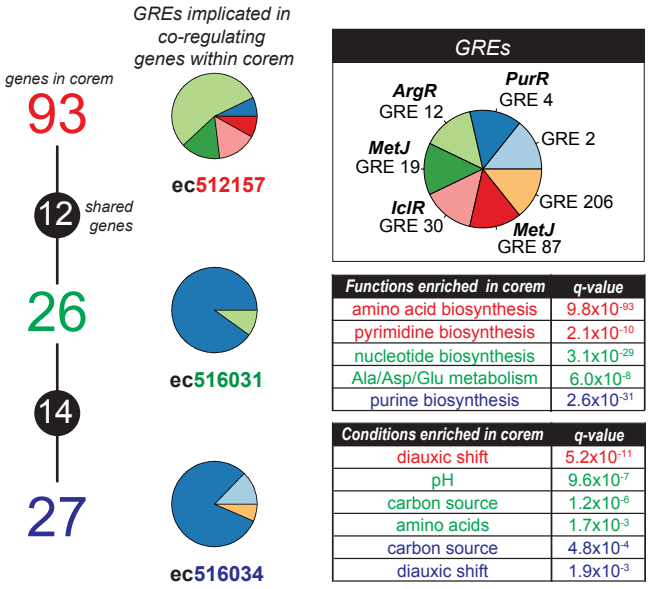


Figure 3.17: Genes of nucleotide biosynthesis are distributed in overlapping combinations across three *E. coli* K-12 MG1655 corems: purine (ec516034 purine corem), pyrimidine (ec512157pyrimidine corem), or both pathways (ec516031 nucleotide corem). (Left) Gene membership and overlap for the three corems as in Figure 3.11. Pie charts indicate average GRE composition across all gene promoters in each corem (see Figure 2.10 for detail). (Top-Right Inset) GRE key for pie charts. Matches to TFs in RegulonDB noted above the GRE name. (Bottom-Right) Tables list enriched gene functions [94] and environmental conditions for each of the corems (see Chapter 2).

(from being highly important for the nucleotide corem, to being marginally important for the pyrimidine corem, Figure 3.17). We hypothesized that the degree to which PurR is implicated in regulating genes within each corem is a good predictor of target-specific expression consequences of knocking out this TF. To test this hypothesis we analyzed global transcriptional changes in both wild type (WT) and ΔpurR deletion strains of *E. coli* K-12 MG1655 grown in the presence of adenine [73]. These data were obtained from experiments that were not included in construction of the EGRIN 2.0 model. Specifically, we calculated the relative standard deviation (a measure of co-regulation) for every PurR-associated

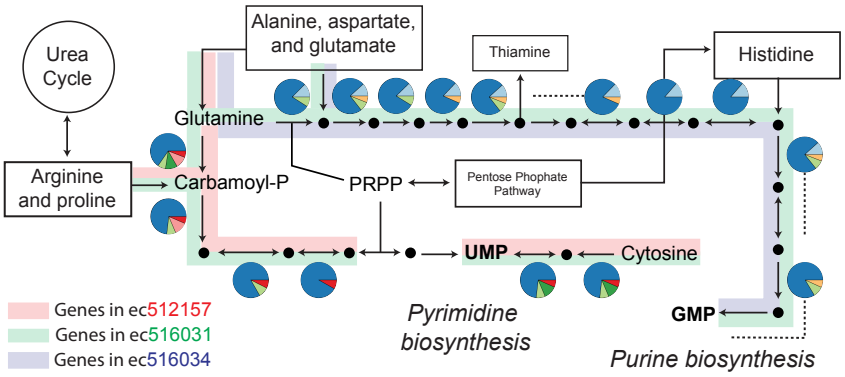


Figure 3.18: A portion of the nucleotide biosynthetic pathways, near the branch point dividing purine (top) and pyrimidine (bottom) biosynthesis. Pie charts represent GRE composition in each gene promoter (Key in Figure 3.17). Operons denoted by dashed lines, with only the leader genes promoter architecture shown.

corem in each of the two strains. As expected, genes in all three corems described above were co-regulated in the WT strain (FDR < 0.05, Figure 3.25). Strikingly consistent with EGRIN 2.0 predictions, the degree of dysregulation of genes within each of the three corems in the ΔpurR strain was proportional to the predicted magnitude of PurR influence. Maximal dysregulation of genes in the nucleotide corem and the purine corem, for example, was consistent with the predicted role of PurR as the primary regulator of genes in these corems (Figure fig:egrin2:5:C). Notably, the degree of disruption observed in these two corems surpasses that of the entire PurR regulon, suggesting that in the presence of adenine, PurR regulates only a subset of its target genes. These results illustrate how the concept of a corem captures the context in which TF binding to a GRE is functional, not just that the potential for TF-GRE interaction exists, which is how a regulon is defined.

3.4 Discussion

EGRIN 2.0 explains how microbes tailor transcriptional responses to varied environments by

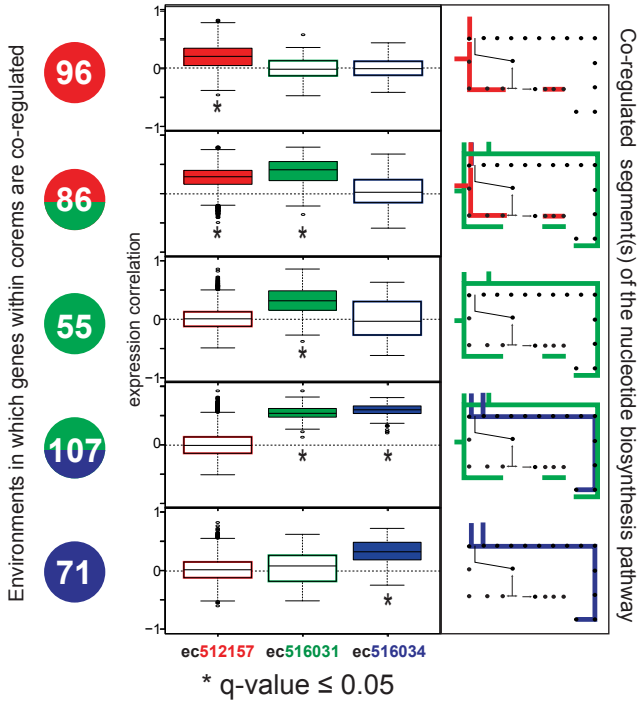


Figure 3.19: Condition-specific co-expression of genes across the three corems. (Right) The active segments of nucleotide biosynthesis (as in Figure 3.18) are color-matched to corems. (Center) Box plots plots show distributions of expression correlations between genes within each corem in relevant environmental conditions, when they are predicted to be co-regulated. Color fill and asterisks indicate corems with significantly low relative standard deviation (RSD; —/—; FDR 0.05). (Left) Colored circles indicate when genes within which corem(s) are predicted to be co-regulated (color) under how many conditions (number).

linking the genome-wide distribution of GREs to their organization and conditional activities within each promoter. The integrative model reveals the mechanisms by which microbes reuse genes in varying combinations to operationally link disparate processes and regulate flux through metabolic pathways. We have provided extensive validations for predictions made by EGRIN 2.0 for a bacterium and an archaeon (Table 2). In addition, we also performed new experiments to validate a model prediction that widespread transcriptional

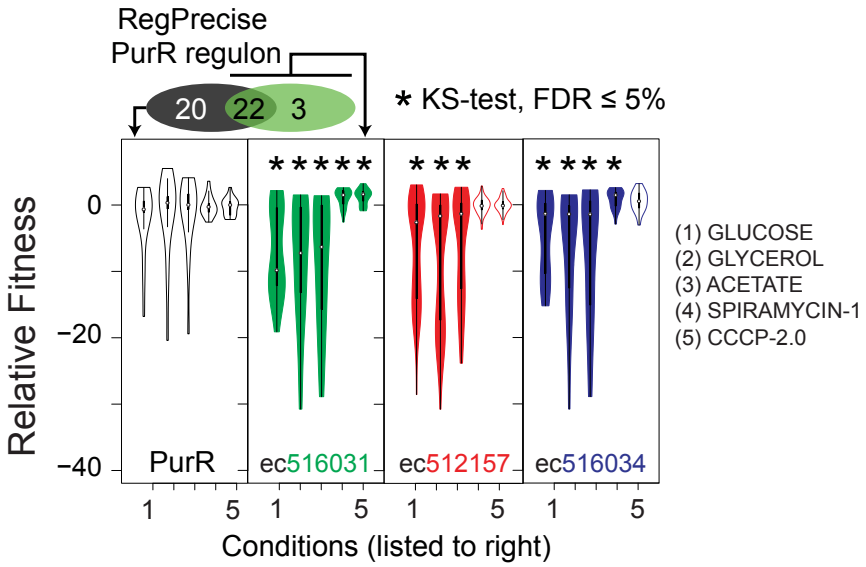


Figure 3.20: Distributions of relative fitness values for gene deletions in the three corems, as well as 20 of the 42 PurR regulon genes not modeled by ec516031 (black) across 5 representative conditions (condition identifiers listed to right, additional conditions in Figure 3.22). Asterisks denote conditions in which the distribution of fitness values is statistically significant (relative to the distribution of fitness values for all genes in that condition).

activity at non-canonical locations within genes and operons was partly responsible for complex modulation of the *E. coli* K-12 MG1655 transcriptome during growth in rich media.

Corems represent a fundamental organizing principle of GRNs that captures fitness-relevant associations among genes, forging a link between the environment-dependent dynamics of transcriptional control and phenotype. The conditional associations among genes across corems reflect the underlying structure of coupled changes in environmental factors, such as correlated changes in effector molecules. Comparative analyses of EGRIN 2.0 models, therefore, could reveal the corems associated with unique and shared environmental structures that distinguish ecotypes of the same species.

EGRIN 2.0 will provide context-relevant engineering strategies for synthetic biology be-

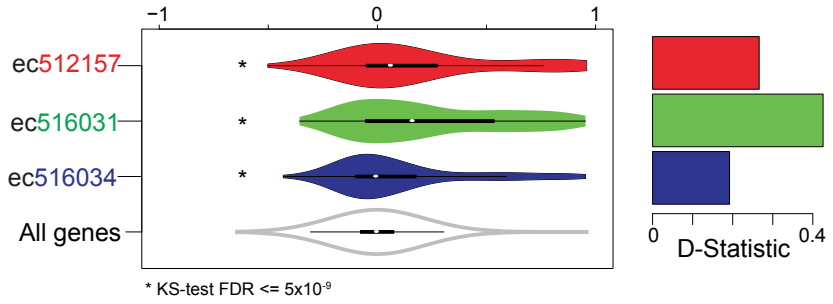


Figure 3.21: (Left) Violin plot shows distribution of all fitness correlations for genes in three nucleotide biosynthesis-associated corems compared to all genes in the data set. (Right) KS D-Statistic relates to enrichment for highly correlated gene-gene fitness associations in the corems. All three corems enrich for similar fitness effects (KS FDR $< 5\ \ 10^{9}$)

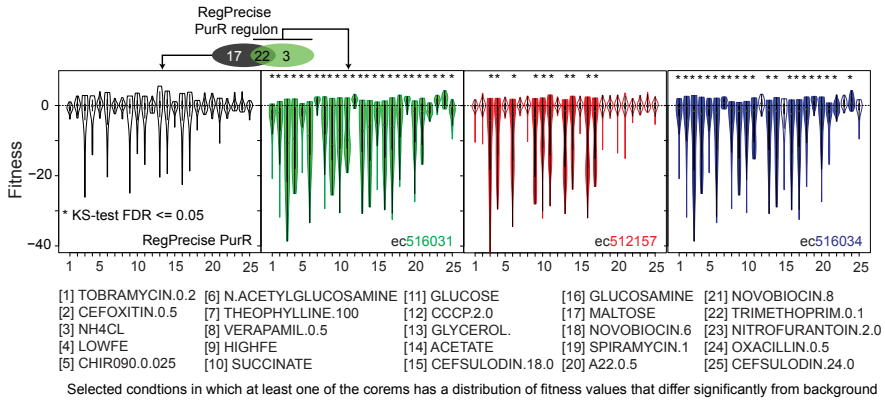


Figure 3.22: Violin plots show distribution of relative fitness among corems across conditions (negative values indicate lower fitness relative to WT). Brief condition descriptions are displayed below. Shading within the violin plot indicates that the distribution of fitness values is significantly in that condition (KS-test FDR ≤ 0.05). Fitness values for the subset of genes from the PurR regulon that do not occur in ec516031 are displayed to the left. These genes do not have significant fitness effects in any of the environments tested.

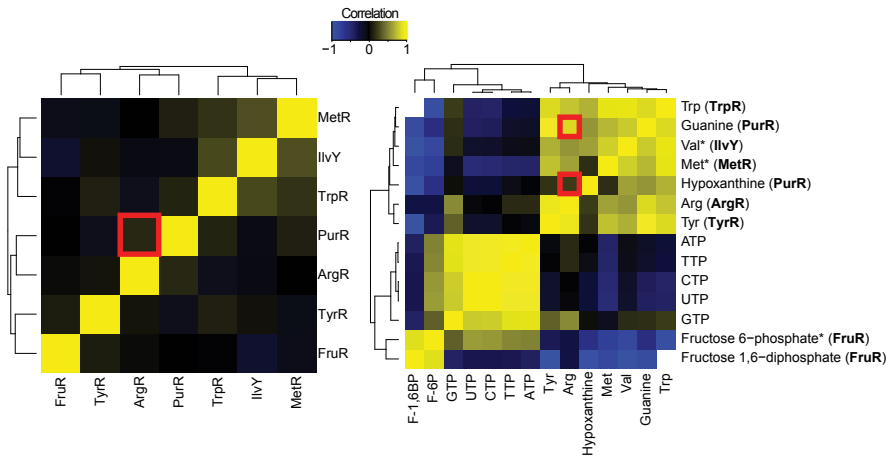


Figure 3.23: (Left) Expression correlation for TFs associated with three corems described in the text (ec516031,ec512157,ec516034). (Right) Correlation allosteric regulators for these TFs. TF regulated by each biomolecule listed in parentheses. Red boxes indicate PurR-ArgR and their corresponding effector molecules. Data from [166].

cause it models environment-dependent coordination of diverse regulatory mechanisms operating across the entire genome, including non-canonical locations. By teasing apart regulatory mechanisms that have indistinguishable outputs in certain environments, EGRIN 2.0 offers multiple strategies for introducing new genes into the GRN. For instance, there are at least five distinct mechanisms responsible for co-regulating nearly 100 genes in the pyrimidine corem in *E. coli* K-12 MG1655 . This corem coordinates genes from various segments of amino acid biosynthesis pathways, including arginine biosynthesis, as well as the pentose phosphate pathway to synchronize inputs into nucleotide biosynthesis. The conditional grouping of genes into the pyrimidine corem explains the previous observation that genes of arginine biosynthesis are repressed upon adenine addition [73]. EGRIN 2.0 predicts that this coordination of nucleotide and arginine biosynthesis is accomplished by an equivalency of PurR and ArgR activities under these conditions (possibly due to correlated changes in effector molecules), rather than by direct regulation of arginine biosynthesis genes

Table 3.1: Corems group together genes from different regulons with highly correlated fitness effects

Gene 1	Gene 2	Fitness correlation	Regulon Gene 1	Regulon Gene 2	Corems
b3774	*b3959*	0.959012	IlvY	ArgR	512157
b2913	*b3829*	0.938764	PurR	MetR	512157
b3829	*b3959*	0.934393	MetR	ArgR	512157;554056
b2913	*b3941*	0.932025	PurR	MetR	512157
b3957	*b3941*	0.931565	ArgR	MetR	512157;554056
b3172	*b3829*	0.930382	ArgR	MetR	512157;554056
b2913	*b3774*	0.927776	PurR	IlvY	512157;512477
b3941	*b3774*	0.927251	MetR	IlvY	512157
b3960	*b3941*	0.921375	ArgR	MetR	512157;554056
b3941	*b3959*	0.921282	MetR	ArgR	512157;554056

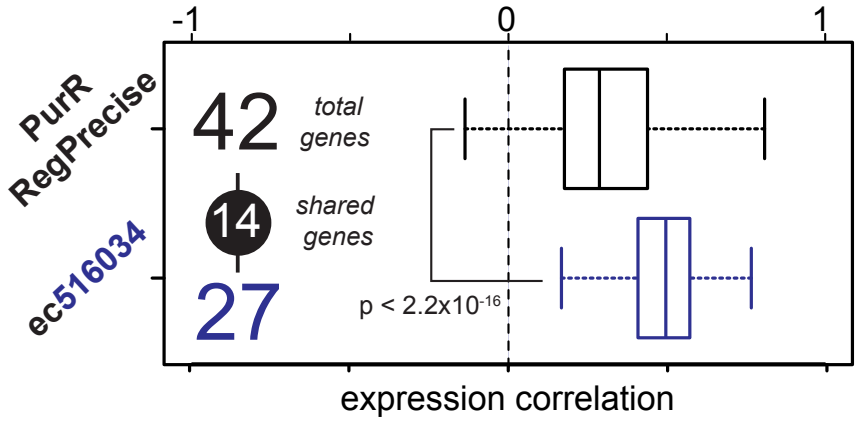


Figure 3.24: Distributions of pairwise expression correlations among all genes in the PurR regulon (RegPrecise) compared to a subset of the regulon within corem ec516034, across all environmental conditions. Also shown are the total number of genes in each group, and the number of shared genes. The two distributions are significantly different (Welch Two Sample t-test, $p < 2.2^{-16}$).

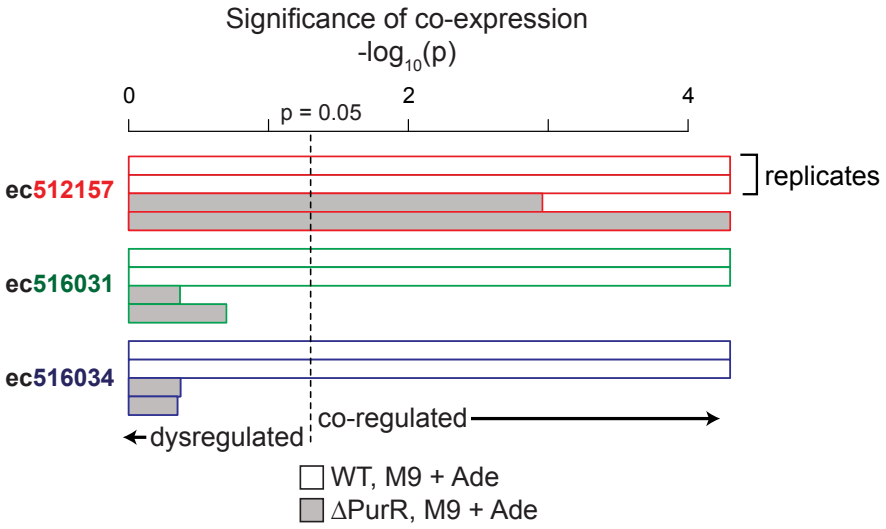


Figure 3.25: RSD of transcript level changes (resampled -log10(pval)) for the three corems in Figure 3.17 in WT and ΔpurR strains of *E. coli* K-12 MG1655 (both grown with adenine). The dashed line delineates significant co-expression (p = 0.05).

by PurR. Not surprisingly, subsets of genes within this corem belong to alternate regulatory programs (corems) under different environmental contexts. Thus, depending on the objective, we can select a reengineering strategy from a library of mechanisms that already exist within the GRN of an organism. Future work to translate the EGRIN 2.0 model into the language of synthetic biology will enable systems-level reengineering of an organism.

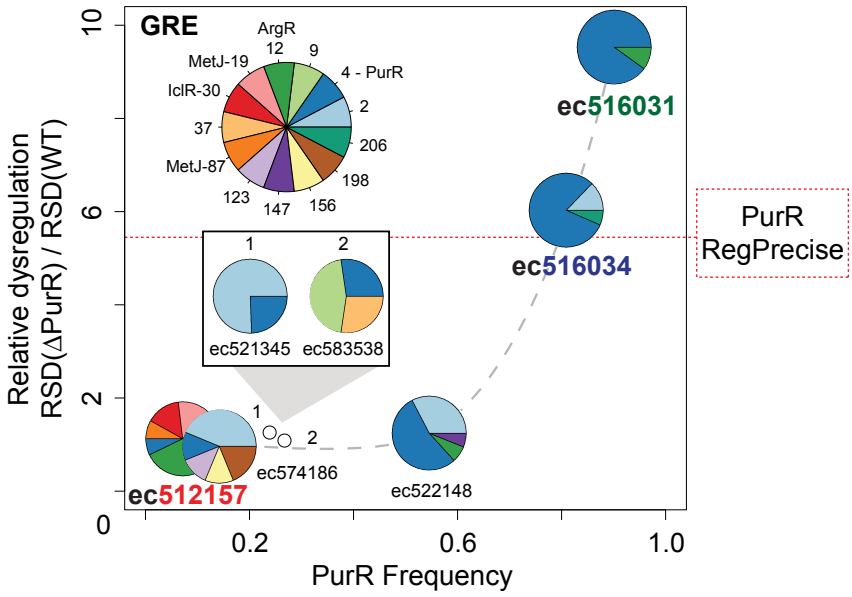


Figure 3.26: Relative RSD (ΔpurR/WT) for all seven GRE #4-associated corems plotted as a function of the frequency with which GRE #4 (PurR) is discovered within these corems. Composition of GREs discovered within each corem are shown as pie charts (as in Figure 3.17), with key in inset, top-right. Relative RSD of the RegPrecise PurR regulon is shown for reference (dotted horizontal line).

Chapter 4

EVOLUTION OF GENE REGULATORY NETWORKS IN PROKARYOTES

The evolutionary success of an organism is a testament to its inherent capacity to keep pace with environmental conditions that change over short and long periods. Mechanisms underlying adaptive processes are being investigated with renewed interest and excitement. This revival is partly fueled by powerful technologies that can probe molecular phenomena at a systems scale. Such studies provide spectacular insight into the mechanisms of adaptation, including rewiring of regulatory networks via natural selection of horizontal gene transfers, gene duplication, deletion, readjustment of kinetic parameters, and myriad other genetic reorganizational events. In this chapter, I discuss advances in our understanding of principles that shape evolution of gene regulatory networks (GRNs) for dynamic adaptation to environmental change.

This chapter has been adapted from:

Brooks AN, Turkarslan S, Beer KD, Lo FY, Baliga NS. Adaptation of cells to new environments. (2011) *Wiley Interdiscip Rev Syst Biol Med.* 3(5): 544561

Chapter Highlights

- Microbes have adapted to diverse environmental conditions
- Mechanism of adaptation depends on how fast and frequently the environment changes
- GRN evolution contributes to environmental adaptation in microbes

- Evolutionary processes in microbes can be studied in the lab

4.1 Summary

GRNs can be rewired through several genetic reorganizational mechanisms, including gene duplication, deletion, mutation, and horizontal transfer. Temporal properties of environmental variability influences what types of molecular changes are permitted and/or preferred. Some microbes may have evolved an ability to anticipate coupled environmental changes.

4.2 Introduction

Microorganisms experience myriad environmental factors over their evolutionary history, including those that remain essentially constant over long periods (e.g. geological epochs), change slowly (e.g. general increase in annual temperatures), fluctuate periodically (e.g. day-night cycles and seasonal variations), or change frequently and somewhat randomly (e.g. unpredictable nutrient loading). These changes occur over diverse timescales, ranging from the lifetime of an individual cell to multiple generations. Accordingly, microbes have evolved unique strategies to deal with the peculiarities of their environment [124]. Characteristic examples include adaptation to extreme environmental niches [188, 201], entrainment of phototrophs to day-night cycles [368, 270], and the physiological adjustment of *E. coli* K-12 MG1655 as it passes through the human intestine [249]. Organisms respond to short term environmental changes by reversibly adjusting their physiology to maximize resource utilization while maintaining structural and genetic integrity by repairing and minimizing damage to cellular infrastructure [160, 282] thereby balancing innovation with robustness. Naturally, physiological response networks emerge as products of evolution by natural selection where they can lend reproductive or fitness advantage, particularly when the environmental change is recurrent [155].

Environmental adaptation of biological systems can be considered from three evolutionary perspectives: (i) acclimation of existing cellular machinery to operate optimally in

a new environmental niche; (ii) acquisition of entirely new capabilities through horizontal gene transfer or neofunctionalization of gene duplications and (iii) reorganization of network dynamics to appropriately adjust existing physiological processes to match dynamic environmental changes. The first type of adaptation can arise through two types of events that differ dramatically in duration. Simple mutations can greatly increase fitness over very short time frames (within one or few generations). Prominent examples of short-term adaptive events include resistance to drugs [301, 321] and altered nutrient conditions [117]. Alternatively, complex mutations in multiple loci may accumulate over very long time frames, such as the evolution of acidic protein surfaces in halophilic archaea [188, 137, 319]. While the initial transfer of adaptive genes by HGT occurs rapidly [19], full integration of laterally transferred component(s) typically occurs over longer time frames (10s of millions of years), where HGT events often require regulatory rewiring to function optimally in the context of existing cellular networks [212]. Finally, physiological readjustment occurs both because of genetic and physiological robustness to withstand stress that accumulates over many generations and latent genetic variance that is revealed after environmental perturbation [292].

Here, we focus on the evolution of adaptive mechanisms for acclimation to recurrent yet transient environmental changes. When transient changes are recurrent they select for genetic traits that confer fitness by improving the ability of an organism to rapidly and reversibly adjust physiology to match current environmental conditions. These traits manifest at varying hierarchies of genetic information processing, from receptors for sensing environmental factors, to signal relay, transcriptional, post-transcriptional, translational and post-translational control mechanisms, and also at the metabolic level through modulation of enzyme function (affinities, kinetics etc.). Such adaptive changes occur over intermediate time frames (upwards of 100s of generations in *E. coli* K-12 MG1655 ; [329]) and, surprisingly, they arise repeatedly and in some cases with some regularity in distinct lineages [70] Fitness, or the number of surviving offspring after one generation [35], is a complex property that emerges from the integration of changes at all these levels. A holistic systems approach,

therefore, is necessary to fully appreciate how these varied mechanisms work together when an organism adapts to a new environment.

For the purpose of our discussion, we define environmental conditions to include both abiotic physical variables (such as light and temperature) and biotic components (such as other co-inhabiting organisms). Additionally, we restrict our analysis to asexual prokaryotic systems (including both bacteria and archaea) in which the dynamics and mechanisms of genetic evolution lack the pervasive variation that recombination by sexual reproduction promotes. Sexual populations also may not experience signatures of selection prevalent in asexual populations, such as classic sweeps associated with unconditionally advantageous mutations [64]. The reader should note, however, that the physiology of archaea and bacteria diverge substantially, with archaea sharing startling similarity to eukaryotes. Infrequently, we may explicitly refer to findings in eukaryotes that reflect mechanisms that likely also occur in prokaryotic systems.

This chapter will bridge the conceptual gap between adaptation, which by definition requires heritable genetic change [35], and physiological readjustment, which is a product of adaptation that equips organisms to attune their physiology to dynamic changes in their environments. We will suggest how systems-level methodologies and insights can be applied to better understand the strategies living systems employ to withstand and in some cases take advantage of change in their environments.

The fields of microbiology, molecular evolution, and systems biology are expansive it would be impossible to cover all adaptive mechanisms and scenarios that may influence the evolution of natural microbial populations. Instead, we will highlight major themes and new insights in microbial evolution while demonstrating how the principles of systems biology can be leveraged to develop a more comprehensive, integrative understanding of cellular adaptation to new environments. Throughout our analysis we will point the reader to other outstanding reviews that complement our discussion.

4.3 Types of adaptations and associated mechanisms

4.3.1 What is adaptation?

When we say that an organism has adapted to its habitat, we imply that it has evolved molecular mechanisms that allow it to grow optimally under the spatiotemporally varying physicochemical conditions of its environment. Evolution, however, is an unfinished process. Organisms chase fitness optima in constantly changing environments. Subtle fitness differences between individuals (due to genomic plasticity or metabolic flexibility) and phylogenetic complexity (i.e. numbers and diversity of species within a community) can lead to the diversification of species or the extinction of less fit genotypes over time [182]. Although many adaptive mutations are lost by random chance (as a result of genetic drift), mutations that confer significant selective advantage have a greater propensity to become fixed within the population, especially in large populations [269]. If selection imposed by the environment is particularly strong, fitness-enhancing genotypes will rapidly rise in frequency in the population, often carrying associated, possibly detrimental, genes along with them (i.e. selective sweep) [273] and interfering with one another via clonal interference [92].

4.3.2 Adaptation to linked environmental changes: general stress response or anticipatory behavior?

Conditions within natural environments can change continually over long periods, periodically, or transiently and unpredictably. In the context of evolution, these environmental changes occur simultaneously albeit on different timescales and exert selective pressure to enrich genotypes well matched to particular ecosystems. Not surprisingly, the repertoire of analogous solutions that characterizes success in a given environment can be similar across diverse organisms occupying similar habitats, suggesting convergent evolution of adaptive solutions that were discovered independently in divergent lineages [204, 151]. While a significant fraction of these responses are condition-specific, most organisms have also evolved robust generalized mechanisms to deal with shared aspects of stress resulting from diverse

kinds of environmental changes. In yeast, for example, a set of 900 genes responds similarly to a diverse array of environmental stresses, sharing common regulatory themes mediated by Yap1p, Msn2p, and Msn4p [127]. This generalized stress response typically includes activation of heat shock proteins, phage shock proteins, and oxidative stress response proteins [69, 149], although there are notable examples in *Candida albicans* and *H. salinarum* sp. NRC-1 where the central role of general stress response has been called into question [103, 185]. The alternate perspective offered by these studies is that changes in environmental variables are physically linked and do not occur in isolation. Elevated temperatures, for example, induce a number of associated changes in environmental conditions, including decrease in oxygen solubility [367]. Theoretical and experimental work in diverse species suggests that organisms can learn to take advantage of this natural co-variation between environmental parameters (e.g. temperature and oxygen), thereby displaying 'anticipatory behavior' [329, 367, 370]. B. subtilis, for example, retains short-term and long-term memories to inform sporulation dynamics and degradative enzyme synthesis [370].

Coupled environmental variables have important evolutionary implications that can be assessed by assaying the fitness consequences upon artificially decoupling such associations in the lab. Relative to laboratory populations evolved in an inverted environment, wild-type *E. coli* K-12 MG1655 has a fitness disadvantage when naturally linked parameters, such as temperature and oxygen, are decoupled artificially. Likewise, cyanobacteria mutants with non-functional circadian clocks are less fit and out-competed by their wild-type counterparts [171]. The enhanced fitness conferred by diurnal entrainment can be attributed to anticipation of predictable, associated stresses such as damaging UV radiation [368]. Adaptation to stress may also prepare cells to better respond to future stresses. *Vibrio parahaemolyticus*, for example, has a greater tolerance to acidity and temperature stresses following growth in media containing 3% NaCl versus 1% NaCl concentrations [365]. On the other hand, adaptation to a subset of environmental factors could come at a cost of sacrificing tolerance to others. Propionate adaptation of *Salmonella enteritidis*, for instance, leads to enhanced resistance to stresses experienced inside the host but overall decrease in infectivity [67].

For all these reasons, we need to appreciate the high degree of connectedness within environmental networks to interpret causes and functional consequences of complex biological responses.

4.4 Role of the environment in shaping GRN evolution: time matters

4.4.1 Long-term adaptation

Some environmental conditions seldom change. We define long-term adaptation as the response to environmental conditions that remain relatively constant throughout the lifetime of a species. For organisms, this represents the most fundamental level of adaptation. Over many thousands to millions of years or longer, the internal architecture of the cell changes to accurately match the general features of the habitat it occupies. Typically, the features that reflect long-term environmental adaptation are deeply ingrained in the structure of the genome, regulatory schemas, or even the molecules of the cell themselves. These molecular artifacts reflect the organism's uphill struggle towards increased fitness in an environment that changes (being directly modified by biotic factors and even the organism itself) yet remains mostly constant over long timescales. Complete reversion of these features would be difficult, if not impossible, for an organism to achieve over short time scales, suggesting that cells are poorly adapted to large variations in these environmental parameters. *H. salinarum* sp. NRC-1 , for example, thrives in high salt conditions. While *H. salinarum* sp. NRC-1 can withstand lower salt concentrations (as low as 2.5M), anything below this concentration is lethal to cells. Adaptation to high salt conditions requires a number of significant structural changes in the cell, including global alterations at the level of DNA, RNA and protein composition [86]. In high salt, water activity is low; this has profound consequences for enzymatic activity and the structural integrity of the cell membrane and genome. Adaptation to high-salt conditions has required the evolution of a highly-acidic proteome, high genomic GC content, and increased intracellular concentration of potassium cation [267]. Other factors including gene redundancy have been reported to enhance survival in fluctuating salt conditions. *Salinibacter ruber*, for instance, possesses

two or more copies of several essential genes. It has been proposed that slight differences in the amino acid composition of two versions of the same protein (ecoparalogs) might allow *Salinibacter* to survive broader fluctuations in salt concentration [295]. These strategies for thriving in hypersaline environments are generally utilized by diverse halophillic archaea. Halophilic bacteria and eukaryotes, on the other hand, have independently evolved alternate mechanisms such as overproduction of organic osmolytes (sugar and amino acid-derivatives) to live in high salt environments [267].

Adaptation to temperature is another well-studied example of molecular response to long-term environmental pressures. While all microbial species are adapted to some range of permissive temperatures, interesting mechanistic examples of thermal adaption have been reported in both extreme warm and cold environments. Psychrophillic organisms occupy extremely cold ecosystems (permanent temperatures below 5C) located deep in the ocean, and in polar and alpine regions. The habitats colonized by psychrophiles are of particular interest because they constitute more than three-quarters of the Earth's surface [112]. Similarly, hyperthermophiles (optimal growth temperature >80C) occupy high temperature habitats. These organisms are so well adapted to their environments that, under some conditions, their doubling times approach that of *E. coli* K-12 MG1655 grown at 37C [112]. In both cases, the microorganisms that occupy these habitats have evolved enzymes and metabolic strategies that allow them to survive and proliferate at temperatures that would be restrictive or lethal to mesophilic organisms. To withstand these extreme temperatures, microbial species have evolved molecular mechanisms that regulate membrane fluidity and conformational flexibility of proteins, and produce thermo-stable protein variants, all of which are significantly challenged at extreme temperatures [113, 350]. While the specific mechanisms employed to combat extreme temperatures vary across species, common mechanisms, such as increased protein stability through ion pairs, hydrogen bonding, hydrophobic interactions, disulfide bridges, packing, and intersubunit interactions at high temperatures and increased protein flexibility and membrane fluidity and exopolysaccharide production at low temperatures, are common aspects of cellular adaption to extreme environments

[168]. Like the halophilic organisms described earlier, thermophiles and psychrophiles occupy environments that impose a number of additional harsh constraints on living systems (including metal ion concentrations, nutrient limitation, and sometimes increased pressure), making them useful models to understand adaptation to stressful conditions.

4.4.2 Short-term adaptation

Environmental variables can fluctuate in unanticipated ways. Such variations are often stressful and, depending on severity, might drive selection of genotypes that are better suited to readjust physiology to manage and mitigate the consequences of stress. Genotypes can be selected on the basis of either genetic or non-genetic components. Asymmetric cell division, in which mother and daughter cells receive disproportionate numbers of molecules, alters the dynamic behavior of cells in response to environmental change and may lead to the selection of genotypes that do not necessarily contain heritable allelic alterations, but rather exhibit a temporary, dynamic state compatible with the altered environment. Sporulation in *B. subtilis*, for instance, couples asymmetric morphological changes with differential gene expression between the forespore and mother cell (Reviewed in [31]), leading to divergent cell fate. Due to asymmetric cell division, daughter cells may receive more or fewer ribosomes, transcription factors, or other cellular components, each of which might contribute to their success (or failure) in new environments. Phase variation (or phenotype switching) is another mechanism (non-genetic, though heritable alteration) by which microbial populations leverage stochastic variations in cell components to respond to uncertainty in environmental fluctuations [347]. Genetic alterations, on the other hand, can occur anywhere within the cellular hierarchy; advantageous mutations may be located within proteins, affecting the stability or kinetic parameters of the protein, or within regulatory sequences, affecting when, where, and how much of a biological product is made. A common feature of these alterations, however, is that they tend to be simple, i.e. they are the product of one or a few adaptive mutations that spontaneously occur in response to the new environment or preexisting neutral or buffered mutations whose consequence is revealed by perturbation.In

this sense, these adaptations are flexible, being easily gained and subsequently lost by genetic drift. The inherent plasticity of these mutations makes them especially important for physiological adaptation, as slight changes can drastically alter the dynamic response of an organism to stress. Short-term adaptation is tightly coupled to the longer-term adaptive mechanisms previously described. Temporary changes in long-term environmental trends may elicit genetic alterations that are short-lived, only becoming fixed within the population and canalized into regulatory programs if the stress that has elicited the advantageous mutation surpasses a temporal threshold, becoming a regular feature of the environment. Numerous studies have investigated mechanisms of adaptation to altered growth temperature [39, 40], nutrient composition [80, 46] and population structure [388]. Universally, these studies find that fitness to a new environment increases rapidly over the first several thousand generations [211, 34]. Surprisingly, adaptive mutations discovered in the lab typically occur in one or a few genes, reflecting the earliest events in the adaptive process. Only four mutations in *E. coli* K-12 MG1655 , for example, are responsible for gaining growth advantage in stationary phase: one in the stationary-phase sigma factor , one in the leucine-responsive regulatory protein, two genomic rearrangements of IS5 transposon insertion sequences, and a mutation in the *sgaC* gene [388].

Short-term, stressful environmental perturbations may also specifically induce increased rates of mutation, potentially facilitating adaptive evolution by rapid exploration of a broader genotypic space. In response to nutrient starvation, for example, the DNA damage and cell-cycle checkpoint control response (SOS response pathway) is induced in *E. coli* K-12 MG1655 . Following induction of this pathway, cells experience higher rates of mutation due to inhibition of mismatch repair and recombinational break repair, and induction of a mutator DNA polymerase [290]. In addition, some genetic loci exhibit a disproportionatly high frequency of mutation during hypermutation. These mutational hot spots may reflect regions of the genome that are more readily modified and thereby adaptive. Taken together, these observations suggest that organisms might possess the capacity to introduce a bias in the distribution of mutational variation along its chromosome(s).

4.5 Adaptation through rewiring GRNs

Systems-level coordination of cellular functions is accomplished by gene regulatory networks (GRNs) [242] The general form and features of biological networks are illustrated in Figure 1.7. Central to all GRNs is the interaction of TFs and their cognate DNA binding sites. Remarkably, the origin of DNA-binding domains (DBDs) in all present day TFs can be traced to a few ancestral classes, such as winged-helix and zinc ribbon domains [20]. This raises important questions regarding the evolutionary process(es) that led to the diversification of these few DBDs to create a vast array of distinct DNA binding specificities. It could be argued that the common origins of DBDs within TFs and their division into related classes (protein families) should help characterize GRNs in one organism and suggest projections of that information onto orthologous systems in phylogenetically related species. Functions of even structurally similar transcription factors (TFs), however, can diverge substantially through alterations in regulatory domains of either the transcription factor itself or the cis-regulatory sequences of downstream target genes. The divergence of DNA binding specificities and allosteric domains of two FNR (fumarate and nitrate reduction) family TFs in E. coli and B. subtilis are case in point. As result of subtle DNA binding differences, FNR in E. coli controls 135 genes whereas its counterpart in B. subtilis regulates only 8 genes [219]. Importantly, this result demonstrates that species-specific coevolution of interacting partners (in this case FNR and the target promoters) prohibits simple projection of TF-binding orthologies across species, even when the genes involved share recent ancestry. On the other hand, this finding underscores the flexibility of GRNs malleability of regulatory network topology can promote variation in gene expression that acclimates a species to the nuances of its environment [218].

Gene regulatory networks evolve through a number of molecular mechanisms that vary in frequency and magnitude of effect. Figure 4.1 depicts several common mechanisms of GRN evolution and their consequences for network topology. While rewiring of GRNs is an efficient mechanism to acquire new features or functions, it is important to remember that preexisting network topology constrains the space of viable and visible phenotypic outcomes

that can result from alterations to its structure [352], especially over short time scales. Historical contingency guides evolution [136], where organisms cannot liberate themselves from their past or innovate beyond the constraints of their current genetic makeup. A related, unresolved problem is how changes in genotype map to phenotype (the representation problem). From this perspective, organisms balance two opposed evolutionary characteristics: evolvability (change) and robustness (resistance to change). Recent experimental studies, for example, have found organisms to be remarkably robust at the level of gene expression, even when challenged by potentially catastrophic rewiring of regulatory components [165] or variations in gene network dosage [3]. By contrast, other (primarily theoretical) studies highlight the tendency towards increased evolvability, i.e. the capacity of organisms to generate diversity, in complex, adaptive systems (Reviewed in [353]). We suspect that the two counterpoints are a result of disparities in the time dimensions over which these two properties are assessed. Evolvability, by definition, is a property that manifests over a very long period of time whereas robustness is typically assessed in laboratory experiments which are conducted over time frames that are too short to resolve fractional fitness differences. Indeed, evolvablity itself may be a selectable trait in biological systems. Simulations of protein evolution, for example, suggest that fluctuating environments elicit large-scale genetic changes that correlate with the frequency and severity of the environmental change [98]. From an engineering standpoint, measures of the topology of biological networks may suggest regions of the network or specific genes that may be more plastic (i.e. evolvable) compared to other regions of the genome, which may guide rational reengineering.

4.6 Evolution of Gene Regulatory Networks (GRNs)

If an organism is challenged by an environmental stress, individuals within the population harboring rewired GRNs that better negotiate the environmental change may enjoy a selective advantage. Here we describe common mechanisms by which GRNs become rewired during evolution, each of which is depicted in Figure 4.1.

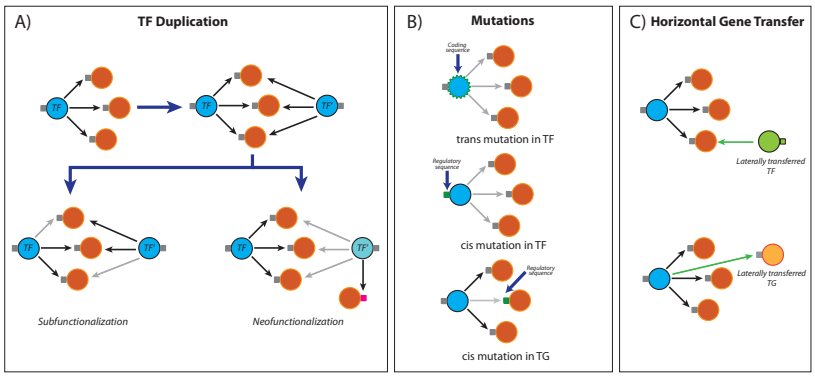


Figure 4.1: Several molecular mechanisms mediate topological changes within GRNs. (A) Duplication of transcription factor (TF) and/or target genes (TG). Duplicated copies of a TF or downstream gene initially share the same interactions as its ancestor. Following duplication, however, either copy can subfunctionalize to contain a subset of those ancestral connections or neofunctionalize by gaining new interactions. Sub- and neofunctionalization generally occur via random mutations. (B) Mutations can occur in the coding or *cis*-regulatory sequences of either TFs or TGs. Mutations in the cis-regulatory sequences of a TG only affect interaction with that particular target, while mutations in coding and cis regions of TF may affect all downstream interactions. (C) Microbial genomes can be extensively modified by horizontal gene transfers. Genomes can horizontally inherit new TF (green circle), TG (yellow circle) or both TF and its target simultaneously (not shown). Transcription factors and target genes are depicted in blue and orange circles respectively. *Cis*-regulatory regions are denoted by gray boxes attached to the circles.

Mutation . Mutational events in transcription factors (TFs) can modify the specificity or affinity of TF DNA binding domains, such as mutations in the base contacting residues of DNA binding proteins that lead to recognition of multiple target DNA sequences [220], or affect protein-protein interaction domains that confer combinatorial specificity to gene regulatory programs [23]. In *Halobacterium*, expansion of the general transcription factors allows for many possible TF-interactions [113], each of which may uniquely control cellular physiology [107]. Rewiring may also occur by mutation in downstream target genes. In yeast, the rapid loss of *cis*-regulatory motifs from multiple genes enables cells to grow

rapidly under anaerobic conditions [164].

Gene Duplication . Duplication events are common in microbial genomes. Duplicated genes constitute a genetic toolbox that cells can harness to innovate and expand their phenotypic repertoire [333, 130]. Nearly half of the regulatory interactions in *E. coli* and yeast appear to evolve through this process [333]. Functional divergence of duplicated genes (neofunctionalization and subfunctionalization) can contribute to the development of new cellular functions [215] or specialization in a condition-specific manner [355]. Whole genome duplications, though rare, can also contribute to evolution of GRNs [186].

Horizontal gene transfer (HGT) . In prokaryotes, a large proportion of genes have been acquired laterally from different microbial species [197] or even viruses [288]. Eukaryotic-derived aminoacyl tRNA synthetases [369], antibiotic-resistance genes [87], and numerous stress-response genes [258] are acquired in diverse lineages through HGT. While entire functional modules may be captured in this way, foreign DNA segments must often be integrated under the control framework of the new host, which can take tens of millions of years [212]

Chapter 5

CONCLUSIONS, PERSPECTIVES, AND FUTURE DIRECTIONS

GRN inference remains a challenge. As I have shown in the other Chapters, integrating multiple data sources and applying modern statistical learning approaches like ensemble modeling can greatly improve performance - but there is still need for improvement. We are far from obtaining comprehensive and accurate GRNs from data, even in microbes. Even our greatly improved approach only recovers ∼25% of the known regulatory interactions in *E. coli*[1]. Why is that? What are we missing? And how can we discover it? In this chapter I highlight specific areas for improvement. I outline new computational approaches to increase GRN accuracy, I highlight experimental methodologies that will help us gain deeper insight into the mechanisms of GRN adaptation, and I describe ways to make the results of GRN reconstruction more accessible to biologists, accelerating discovery.
Parts of this chapter have been adapted from:

Brooks AN, Turkarslan S, Beer KD, Lo FY, Baliga NS. Adaptation of cells to new environments. (2011) *Wiley Interdiscip Rev Syst Biol Med.* 3(5): 544561

and

Westerhoff H*, Brooks AN*, Simeonidis E*, Garca-Contreras R*, Boogerd F, He F, Jackson VJ, Goncharuk V, Kolodkin A. (2014) Macromolecular networks and intelligence in microorganisms. *Front.Microbiol.* 5:379.

---

[1]At 25% precision cutoff. This value in 2.7X greater than algorithms in DREAM5.

* Denotes equal contribution

Chapter Highlights

- New experimental strategies to study evolutionary mechanisms in the lab
- Improved GRN inference by aggregating models with varied data types
- Web-based visualization to aid hypothesis generation and sharing results
- Co-regulation is dynamic. Corems express co-regulatory associations emerging from temporal dynamics
- Beyond the GRN: integration with metabolome and proteome dynamics will enable biological insight

5.1 Summary

New experimental and computational approaches will help reconstruct GRNs with increased accuracy. These tools will provide insight into how GRNs operate dynamically and suggest how they can evolve in natural populations. The methods will be enhanced by web-based technologies that allow biologists to visualize, explore, and share their discoveries.

5.2 Systems approaches to investigate GRN evolution

Adaptational events produce molecular signatures at every level of the cellular hierarchy. Some are deeply ingrained in the molecular structures of the cell, whereas others are simpler and therefore rapidly modified. A combination of comparative genomics and systems biology is ideally suited to infer the molecular mechanisms of adaptation from diverse data types collected across taxa. Substantial progress has been made toward this goal. Here we consider advances that have contributed to our understanding of common adaptive mechanisms and highlight emerging technologies and methodologies that will deepen our exploration of evolution in microbial populations.

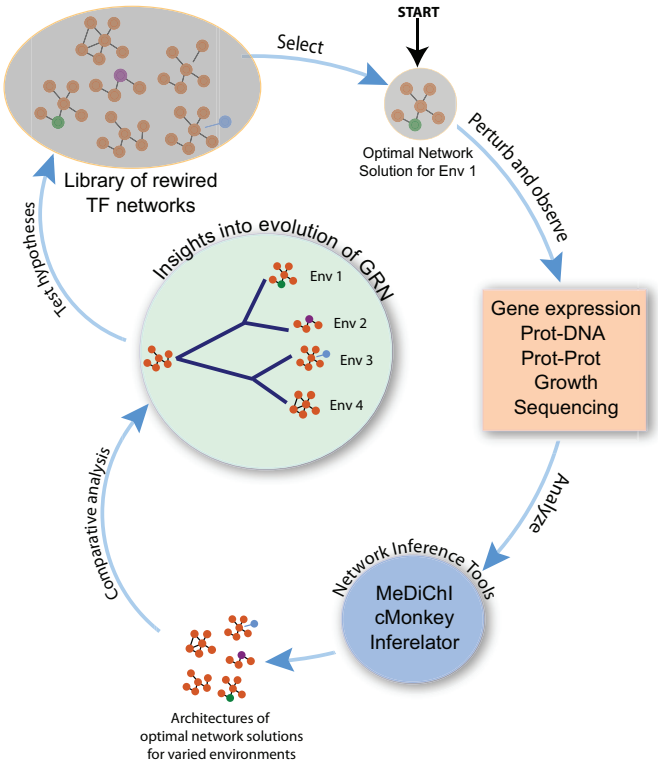


Figure 5.1: The first step in understanding GRN evolution starts with a naturally evolved GRN that has increased fitness in a given environment. Initially, the architecture of the network is unknown (indicated by gray shading). To delineate the connectivity of the network, we perturb it via genetic or environmental alterations and observe the phenotypic response across multiple data types (orange box). The data is analyzed and integrated using network inference tools to build a comprehensive view of the changes to identify architectures of optimal network solutions (Blue circle). Many iterations of this cycle for different environments yield to library of optimal network solutions which can be compared to derive evolutionary insights (Green circle) and further formulate hypotheses. These hypotheses are tested by construction of rewired GRN by introducing variations at specific components (Gray ellipse). Rewired GRN are selected at different environments and characterized in the next iteration.

5.2.1 Current limitations

Challenges of using comparative genomics alone

Since the completion of the first bacterial genome-sequencing project in 1995 [118], the number of completely sequenced organisms has increased rapidly. While initial studies confined themselves to a narrow spectrum of phylogenetic diversity, newer sequencing efforts have leveraged known 16s rRNA associations to suggest additional organisms and lineages whose gene sets may be under sampled [376]. Beyond sequencing the genome of a single species, recent metagenomic studies attempt to capture the genetic diversity present in complex environmental samples, characterizing novel genes from a number of microbial species that are uncultured under standard laboratory conditions [144, 326]. Whole genome sequences from organisms spanning diverse evolutionary domains have enabled comparative analysis between closely [100] and distantly related species and lineages [198, 231]. Orthology between genes in separate lineages can suggest similarity in function [332], participation in common pathways [179], and shared regulatory motifs [187], although the precise mechanisms behind co-regulation of similar genes is often markedly different between species [355]. A challenge of comparative approaches, however, is to separate adaptive events from exaptive events; that is, to isolate those events that are either random (i.e. the result of genetic drift) or structural consequences of other acquired traits [135]. Exaptations can be prevalent even between closely related species. For instance, small changes in gene regulatory circuits have resulted in substantial diversification of closely related species. In the fungi *Kluyveromyces lactis* and *Candida albicans*, the combinatorial circuits regulated by Mcm1 have diverged significantly [341] through gain or loss of Mcm1 binding sites and combinatorial associations with other transcriptional cofactors. After 300 million years of divergence between *S. cerevisiae* and these two fungal species, only 15% of the direct Mcm1-target gene interactions remain intact. Determining which of the myriad genetic changes results in increased fitness in new environments is a challenge for laboratory experimental evolution studies. In the subsequent sections, we consider experimental methodologies to characterize molecular

mechanisms that drive cellular adaptation to new environments.

5.2.2 Experimental challenges and opportunities

Laboratory evolution (directed selection)

Experimental evolution studies have changed our understanding of short-term adaptation in microbial populations. A common strategy employed in these studies is to enrich, over several generations, mutants that are better suited to propagate in a gradually changing environment. In such experiments adaptation to new conditions (improved fitness relative to the ancestral genotype) emerges quickly, within a few to hundreds of generations depending on the type and severity of stress. Perhaps most well-known and celebrated are Richard Lenski's long-term evolution studies in *E. coli*. In an early study, Travisano and Lenski propagated 12 replicate *E. coli* populations for 2,000 generations in a glucose-limited environment [338]. To assay for increased fitness, the authors competed these evolved strains against their ancestor in 11 novel, single-nutrient environments. In cases where the uptake mechanism of the novel nutrient is similar to glucose, the evolved strains exhibited similar levels of increased fitness; in response to nutrients with uptake mechanisms different than glucose, however, the strains behaved more unpredictably, suggesting that each strain achieved adaptation to glucose limitation by an independent mechanism. In follow up to this work, Blount et al. reported the adaptation of *E. coli* to growth in minimal glucose supplemented with citrate [46]. Wild-type *E. coli* cannot utilize citrate as a carbon source under oxic conditions. The adaptive ability of *E. coli* to utilize citrate manifested after 33,127 generations and over twenty years of growth under selective conditions. Adaptation of this novel functionality likely required at least three genetic changes. While the exact location and type of mutations are still unknown, the authors suggest that the adapted *E. coli* strains, which have the ability to metabolize citrate but lack the ability to transport it into the cell, may have activated a cryptic citrate transporter. This study is of particular importance both because it demonstrates the emergence of novel functionality during the course of a laboratory experiment and suggests that mutational events have historical

contingency. Without at least two "potentiating" mutations in the genetic background, citrate metabolism evolved infrequently. Most recently, Barrick et al. compared the rates of genomic evolution and adaptation in *E. coli*. They confirm the long-standing observation that adaptation slows considerably after several thousand generations [210] and demonstrate that the rate of genomic mutation remains relatively constant for as many as 20,000 generations [34]. Surprisingly, most mutations observed in this experiment were beneficial. Taken as a whole, this body of work illuminates important evolutionary mechanisms and raises important questions regarding the reproducibility of adaptation, duration in the periods of rapid evolution followed by stasis, and the role of chance events like mutation and drift in adaptive evolution. Later on we will discuss how complex genetic interactions also play an important role in defining the constraints of adaptive evolution [211].

Studies in *Pseudomonas fluorescens* SBW25 have also elegantly demonstrated the evolution of novel phenotypes under laboratory conditions [37]. Depending on conditions of growth (shaken or static), *Psuedomonas* genotypes retain close resemblance to their ancestor or they diversify into a range of sub-types that can occupy unique environmental niches. Remarkably, when subjected to a fluctuating environment that alternately favors one of two phenotypes, *P. fluorescens* evolves bet-hedging mechanisms that permit stochastic switching between the two phenotypes. Similar to Lenski's finding, Beaumont *et al.* report a limited number of genotypic differences (nine in this case) between the evolved strain and its ancestor. Surprisingly, the final requisite mutation can be attributed to a single non-synonymous mutation in the large subunit of carbamoylphosphate synthetase (*CarB*), a central enzyme in the pyrimidine and arginine biosynthetic pathways. Similar to Lenskis finding, the final mutation required an accumulation of previous, potentiating mutations, suggesting that complex epistatic interactions between genotypes drives the evolution of novel phenotypes. The methodology in this experiment is of particular interest for future experimental evolution studies, as the authors achieve this new phenotype rapidly by imposing bottlenecks on the population structure at each s̈electionëvent.

A consistent conclusion from these laboratory studies is that significant improvements

in fitness are gained through simple mutations in few to single genes. Functional mutations typically reside within the coding region of genes and presumably alter the kinetics or substrate specificity of proteins, although recent studies suggest that simple mutations in some non-coding elements, such as riboswitches, can also confer selective advantage [339]. Such findings, however, are inconsistent with comparative genomic studies, which suggest that regulatory rewiring of cis-regulatory elements is a primary source of early diversification between species. Theoretical insights similarly suggest that modularity, which is the result of regulatory programs, is critical for the robustness and adaptability of living systems [190]. Many of the adaptive features that allow an organism to robustly respond to changes in its environment are encoded at the level of gene regulation – how genes are turned on and off, where and when they are expressed, and what controls their expression (forming the modularity observed in living systems). This raises interesting questions regarding failure of laboratory studies to enrich mutants with improved fitness that results from alterations in GRNs.

One plausible explanation is that the types of selective pressures used in these studies can be generally categorized as simple nutritional stress. If so, selection with repeated patterns of complex, dynamically changing environmental conditions should enrich for regulatory mutants that can reversibly readjust physiology in novel ways. The study that comes closest to this design is the one conducted by Tagkopolous et al [329]. In this study the authors enriched a population of *E. coli* with improved fitness to artificially decoupled changes in oxygen and temperature in less than 100 generations. Although the authors did not characterize the mechanistic underpinnings of this improved fitness phenotype, one can speculate that alterations to the GRN structure is the only plausible mechanism to reverse the naturally evolved relationship among processes independently attuned to handle temperature and oxygen-related physiologies. Regardless, we should recollect that selection is a very powerful tool in evolution. While selection imposed by altering a single factor might facilitate analysis and interpretation of adaptive mechanisms discovered in the lab, it does not accurately mirror processes in natural environments, where multiple variables change

simultaneously. Environmental heterogeneity and complexity are important and ubiquitous factors in the evolution of natural populations. Interactions among genes, mutations, and environmental factors contribute to adaptability, defining the landscape in which organisms evolve. Heterogeneous environments may create rugged fitness landscapes that contain a multitude of local fitness optima, many of which may be explored by a natural population [80]. Simulations suggest that varying environments can actually speed up the rate of evolution [181], especially when new environmental goals are modular sharing features present at early times. Combinatorial and sequential experimental designs, where cells are exposed to varying sequences or combinations of pressures may reveal natural couplings between environmental events that have been learned by cells [24] and in long-term studies may suggest how organisms anticipate and respond to complex environmental changes.

It is also important to keep in mind that regulatory rewiring events, though fast on an evolutionary timescale, may still require over thousands to millions of years, far beyond the scope of a typical laboratory experiment. Approaches that introduce vast amounts of variation into targeted genetic elements of a natural population prior to selection with an appropriate complex perturbation could circumvent this limitation. Such strategies, while utilized in the past [354], have witnessed renewed interest and enthusiasm because it is now possible to generate fully synthetic genes and genomes [131] and comprehensively identify and track mutants in large populations using NextGen sequencing technologies [240].

Model systems to study adaptation

Microorganisms are ideal for studying the molecular basis for physiological adaptation because of the ease with which they can be genetically manipulated and cultured under controlled environments. In addition, they generally have (i) short generation times, (ii) large effective populations, and (iii) small, relatively simple genomes. These properties allow for rapid, high-throughput surveys of genetic fitness landscapes over relatively short timescales. Microorganisms commonly employed for molecular evolution studies include *E. coli*, *B. subtilis*, and *S. cerevisiae*. All of these organisms are well characterized, with completely

sequenced genomes and a wealth of validated functional annotations. Groundbreaking insights into evolution have been made using each of these organisms. Richard Lenski and colleagues demonstrated adaptation to growth on citrate, where the inability of *E. coli* to proliferate on citrate has traditionally been a characteristic hallmark [46]. In *B. subtilis*, several groups have studied adaptation in stress response pathways, including sporulation and competence (Reviewed in [143, 104]). Finally, the yeast community has produced tremendous work linking gene expression to adaptive evolution [114]. Halophillic archaea, like *H. salinarum* sp. NRC-1 , are an especially powerful system in which to study the evolution of cellular stress defense mechanisms as they have evolved in constantly fluctuating and extreme environments, including salinity (10 times that of seawater; 2.5-5.0M), light (¿150mol photons/m-2/s-1), oxygen (¡5 M), temperature (30C50C), nutrients and DNA damaging agents such as UV radiation. Like *E. coli* and *B. subtilis*, *H. salinarum* is relatively simple, readily manipulated, and has a smaller completely sequenced genome (2,400 genes). All of these features make it an appealing system to disentangle the complex adjustments cells undergo in response to variable environmental conditions. In addition, archaea are evolutionarily unique relative to both bacteria and eukarya. Their basal transcriptional and replication machinery, for instance, shares common ancestry with the eukaryotic machinery, whereas their regulatory systems have bacterial character and they retain the capacity for HGT [23, 128, 99]. Studies in *H. salinarum* are revealing mechanisms by which a GRN acquires complexity through duplication of transcription factors. By expansion of two eukaryotic general transcription factor families (six TBPs and seven TFBs, TFIIB orthologs) and subsequent changes to the promoters and coding sequences *H. salinarum* has evolved regulatory programs to modulate the expression of large fractions of genes that allow it to rapidly acclimate to changing environmental conditions [367, 107, 300].

New experimental approaches to study GRN adaptation

We are witnessing unprecedented advances in technologies for probing biological phenomena. Powerful high-resolution and high-throughput experimental and computational tools

are being used to tackle old biological questions and discover new, often-unanticipated avenues for research. These tools have changed ways in which we think about biological systems, generate hypotheses, and execute experiments. When the first complete genome of a microorganism was sequenced fifteen years ago [118], sequencing a several megabase genome in a day was unimaginable. Now, NextGen sequencing technologies continually push the limits of quality, quantity and cost of sequencing runs. As the cost of these technologies decreases, they are becoming more widespread. Together, these tools give investigators unprecedented access to the signatures of evolution; by comparing larger numbers of related and diverse genomes, and as they occur within laboratory experiments. For instance, we can observe within the time frame of a laboratory experiment evolutionary processes that normally occur over thousands to millions of years in natural environments. We can do this by accelerating evolution using automated multiplexed culturing devices to enrich mutants from microbial populations containing large numbers of random or semi-random variation generated using a combination of gene synthesis and mutagenesis technologies. By using NextGen sequencing, whole genome oligonucleotide arrays, metabolic arrays and mass spectrometry, and high-throughput protein quantitation, it will be possible to link phenotypic outcomes to genetic changes and characterize adaptive mechanisms responsible for improved fitness in new environments. Such forward-looking approaches are exemplified by an experiment in which Wang et al identified one mutant with fivefold increased lycopene production within 3 days from a population of 4.3 billion *E. coli* [354]. The key technologies used in this remarkable proof-of-concept study included multiplexed automated genome engineering (MAGE) to introduce combinatorial variation into the *E. coli* EcHW2 genome and clever bottlenecking and selection approaches. The combination of these technologies accelerated the emergence of multistep phenotypic adaptations that may have otherwise been extremely improbable due to transient decreases in fitness or genetic drift [37]. Especially with respect to GRN evolution, the ability to generate vast amounts of genetic diversity rapidly, followed by deep sequencing of mutants with increased fitness will give us a lens into long-term evolutionary processes that would otherwise be inaccessible. Fi-

nally, synthetic approaches that directly manipulate pre-existing cellular architecture can rapidly accelerate evolutionary processes and confirm or repudiate hypothesis of adaptive mechanisms suggested by laboratory evolution or comparative genomics studies. Directly manipulating gene regulatory networks is a powerful approach to disentangle complicated evolutionary scenarios discovered by comparative approaches. Isalan *et al*, for example, studied the effects of regulatory circuit rewiring in *E. coli* by swapping *cis*-regulatory and downstream DNA binding domains of TF genes across the TF hierarchy [165]. They observed that most rewiring events are neutral under standard growth conditions and can even be advantageous under some stress conditions. Although the phenotypes monitored in this experiment may not fully reveal the consequences of GRN rewiring, the authors suggest that higher-order control of GRNs may minimize the impact of transcriptional rewiring and that such cost-free tinkering with gene regulatory circuits may be a common mechanism in the evolution of GRN. It will be interesting to assess how often this is the case in more complex environments, where the effects of misregulation may manifest themselves more dramatically and to validate the molecular phenotypes of GRN rewiring by gene expression profiling.

Evolution of microbial communities

Part of the complexity of natural environments is a product of interspecies dynamics. Members of ecological communities influence each other by altering their natural environment, competing for common resources, cooperating with one another to solve common challenges, complementing each others nutritional needs and preying on one another. Interspecies interactions directly influence natural selection by modifying the selective criteria imposed by the environment. The evolution of a single species is tightly coupled to the co-inhabitants of its environment. Importantly, the evolution of a species in isolation may proceed differently when that same species evolves in the presence of other organisms. Competition between *S. pneumoniae* and *H. infuenzae* for colonization of mucosal surfaces, for example, drives S. pneumoniae to develop opsonophagocytosis-resistance, which is associated with invasive na-

sopharynx diseases. In the absence of *H. influenzae*, this resistance mutation bears a fitness cost; yet, in competition with *H. influenzae*, which actively stimulates neutrophil-mediated opsonophagocytosis, the resistance mutation becomes advantageous [222]. Besides adjusting to varied selective criteria that are the consequence of interspecies dynamics, cells in natural environments have access to vast pools of genetic material with which they can diversify and innovate. The prevalence of horizontal genetic diversity within microbial populations was severely underestimated until recently (Reviewed in [197]). Faced with stressful conditions that are difficult to address with existing molecular parts, an organism may look outward to find genes that increase its likelihood of survival. Many microorganisms maintain conserved pathways through which they become competent to uptake DNA from the environment. In *B. subtilis*, the activation of competence is mediated by noise in expression of the regulator ComK [302, 225]. Given stressful conditions, *B. subtilis* dedicates approximately ten percent of its population to survey the environment for new survival strategies. Such mechanisms can dramatically increase the rate of evolution in asexual populations. Another important property of microbes that has only recently been appreciated is genomic plasticity, which refers to variation in genome architectures across individuals of a natural population. It is becoming increasingly evident that microbial populations are rife with genomic variability, bringing into question the definition of a microbial species itself [314, 4]. Together, genomic plasticity and metabolic flexibility (the ability of organisms to reversibly adjust biochemical capability by regulating gene expression and enzyme activities) increases the adaptive capability of a microbial community.

Insights into mechanisms of adaption and future directions

We have reviewed known mechanisms by which organisms adapt to new environments and made a case for integrating computational and experimental approaches to study this process and discover new mechanisms. It should be noted that while new technologies have increased the resolution and scale at which we can probe evolutionary processes, they do not override the need for well-designed experiments. Carefully designed selection pressures and culturing

technologies will decide the degree to which adaptive mechanisms uncovered in laboratory studies accurately reflect natural processes. The ultimate test of our understanding of how cells adapt will come from surveying variations within related populations that have evolved naturally to deal with varying environments. This is feasible to do as we now have the capability to determine complete metagenome sequences using sequencing technologies. It is important that we iteratively refine our models so findings from the laboratory are brought into close apposition with what we are able to observe in a natural environment. Knowing the mechanisms by which organisms evolve will eventually equip us with better engineering principles and strategies to synthetically alter their capabilities. Such rational reengineering has very important implications for how we can address environmental and health related problems in the future [192].

Intelligence in Microorganisms

An intriguing question emerging from experimental evolution studies and studies of regulatory networks in microbes is whether microorganisms are capable of "intelligence". For centuries, mankind has grappled with the precise nature and defining features of intelligence. Debates have erupted over how to define and measure the extent of intelligence in parts of the biological (and non-biological) world. Alan Turing, for example, famously proposed a test for evaluating the performance of "artificial intelligence": namely, can it be distinguished from the performance of human beings by another human [342]? There have also long been philosophical discussions on what can be considered intelligent. A number of studies have explored whether there are differences in intelligence between human populations [257], whether animals [335] and even plants [340] exhibit intelligent behaviors, whether non-human artificial systems are capable of intelligence [61] and, more recently, whether intelligence spans biological domains including even the simplest of microbes [150, 62, 156, 38].

As an abstract concept, intelligence escapes easy definition. As a linguistic construct, its characteristics have varied substantially across philosophical and cultural contexts. Rather than launch an ontological, epistemological, or semantic inquiry, I ask instead whether there

is scientific utility in assigning intelligence to microbes? Is it possible that mathematical perspectives of complex adaptive systems and recent data-intensive developments in systems biology will help us detect and define microbial intelligence? Would viewing microbes through the lens of intelligence can help us better describe their behavior, harness their intelligence to perform valuable actions and, in the end, possibly extend our understanding of the human brain itself?

The modern biological perspective of intelligence, even at its most fundamental level, tends to associate it with the human brain. In this context, intelligence is a property of the human brain, or a feature that somehow emerges from its activity. Accepting that intelligence may not be exclusively a feature of the human brain, but rather it may be present at least to a degree in all creatures possessing brains or nervous systems, already helps refine the general features of intelligence. However, intelligence may not have to be associated solely with a certain biological organ, such as a brain or a nervous system. Brains and nervous systems may be highly adapted conduits for expressing and integrating multiple intelligent behaviors. Some of these behaviors may be exhibited by other complex adaptive systems present in living organisms that do not have a brain or nervous system. As early as 1995, Hellingwerf suggested that some two-component systems in bacteria comply with the requirements for elements of a neural network [150]. More recently, the so-called biogenic approach of cognition has gained momentum by focusing on the biological origin of cognition and intelligence, abandoning a strict anthropocentric perspective [209, 221].

A recently submitted manuscript provided in Appendix A provides examples of intelligence in the microbial world. It argues that, at least for some specific tasks, microbial intelligence can be compared to human intelligence, i.e., microbial networks can be considered formally "intelligent". Recognizing microbial intelligence may allow us to modify microbial networks or develop new microbial networks capable of intelligent solutions to specific human problems *de novo*. Alternatively, if intelligence (or components thereof) emerges from the dynamics of complex adaptive systems and the human brain is an evolved organ for the encapsulation of intelligent characteristics, then it is possible that there may

be features of intelligence that remain undiscovered.

5.3 Inference, visualization, and dissemination of GRNs

The above section described several directions for experimental studies that accurately reflect evolution in natural environments, or introduce changes into organisms that accelerate the process of evolution itself. To leverage the power of these new biological insights, we must integrate data into intuitive-yet-accurate models within which investigators can interactively explore functional relationships among genes and generate new hypotheses. As system-wide data becomes more available, it is increasingly important to develop computational methods to handle diverse data types, both individually and collectively. Important advances in the area of network inference have been made over the past several decades. From high-resolution kinetic models to abstract Boolean representations, the field of network inference is rapidly growing in sophistication.

5.3.1 Challenges and opportunities: GRN inference

Numerous advances have been made in GRN inference, including what has been described previously in this manuscript. In the past, studying the function or evolution of GRNs presented a formidable challenge, even for model organisms with relatively small genomes. It took considerable effort to develop molecular and computational tools to map all regulatory connections to construct reliable network models. Even in this regard "well-defined" GRNs were sparse and mostly restricted to a select few model organisms including *E. coli* [293, 228], yeast [206] and *B. subtilis* [252, 232]. Despite the obvious limitations that come with poor taxonomic coverage of such model organisms, comparative analysis of their GRNs did identify three major mechanisms driving GRN evolution (see Figure 4.1): (i) alterations in TF-target gene interactions, (ii) gene duplication followed by subfunctionalization or neofunctionalization, and (iii) horizontal gene transfer events between species. Since there were few experimentally derived models of GRNs, investigations relied on *in silico* models whose components include synthetic genes that evolve according to rules encoded in a computer

simulation [351, 247, 313]. Although *in silico* approaches are extremely informative and will continue to yield valuable insights, they grossly underestimate complexity of naturally evolved GRNs [90]. There are now increasingly sophisticated methods for inferring the structure of GRNs from high-throughput systems biology data [48, 126], (Figure 5.1)

It is important to recognize that key to the success of these inference procedures is the quality, quantity and types of data. Well thought out experiment designs are critical to build models that are predictive and mechanistically accurate. These experiments should appreciate the fundamental properties of biological processes in that they are dynamic, probabilistic and conditional. Through application of these approaches there is now an ever-increasing number of experimentally verified GRNs [90]. Such systems level investigation has and will continue to revolutionize how we investigate and understand cellular adaptation to a new environment through the evolution of complex GRNs.

High-throughput methodologies generate a wealth of data. Consolidating diverse data types into coherent models of biological interactions and their functional consequences will be critical for understanding systems-level adaptive processes. This presents a challenge for traditional computational tools, which generally represent biological networks as static, uni-dimensional entities. Evolutionary systems biologists, however, are interested in how the structure of multi-dimensional biological networks changes as a result of adaptation and how these alterations in network properties affect dynamic cellular processes and their associated downstream phenotypes. While systems approaches have been successfully applied to a number of biological problems, including abstract boolean network analysis of cell-cycle attractor states [214], flux-balance models of metabolism (Reviewed in [183]), and causal biclustering methods for transcriptional regulation [50], to elucidate evolutionary mechanisms future models must (1) integrate across diverse biological processes, (2) represent biological information across multiple time scales with varying degrees of resolution, and (3) develop network representations that track changes in network structures over long (evolutionary) timescales. Such advances will be made by harnessing interdisciplinary approaches, leveraging insights and expertise from a number of disciplines, including mathematics, com-

putation, physics, and engineering. The workflow by which systems-methodologies can be applied to experimental evolution studies is depicted in Figure 5.1.

In parallel, advances have been made in statistical machine learning, including the ensemble learning frameworks described in Chapter 2. As the number and diversity of data types increases, there will be opportunities to integrate multiple algorithms that each learn from a single data type into an ensemble. This way multiple aspects of biology can be captured in a single model. The primary challenge will be to build analytical frameworks that accommodate fundamentally different types of predictions and associations (e.g., *cis*-regulatory motifs, co-regulation, protein-protein interaction, kinetic, etc. in a single model). These future models will also have to weigh the level of detail at which they intend to model the biology. Most important, these models should be constructed to address specific biological questions. Finding that appropriate level of abstraction will depend on the problem - there will not be one-size-fits-all solutions.

5.3.2 Challenges and opportunities: GRN visualization and dissemination

Having inferred models from data, a next step is to make that information useful to investigators. This is especially important given the increasing size and complexity of these models. A single EGRIN 2.0model, for example, consists of hundreds of thousands of biclusters and motifs, not to mention accessory functions and additional data. The footprint of each model is on the order of tens of gigabytes of hard drive space and loading the ensemble requires a file-backed memory management system because the models are too large to fit into active memory on current machines (e.g., >8GB). That's not even to mention the computational time and power to construct the model in the first place. To facilitate exploration of the model's predictions by biologists and other researchers in our group, I designed a comprehensive web-framework. All of the model predictions are hosted in a PostgreSQL database that is queried by a web application framework called Django. Django uses a model-view-controller (MVC) framework to dynamically generate content for users requesting a specific page. It consists of (1) an object-relational mapper that manages data models (i.e., Python

classes) and a relational database (the "Model"), (2) a web templating system for generating viewable content from the models (the "View") and (3) a regular-expression-based URL dispatcher (the "Controller"). In addition to the base Django functionalities, we also developed a genome browser, iGB^{web}, to emphasize model dynamics. For every gene in the model, the user can examine the relative influence of each GRE on co-regualtion of the gene in context of each corem to which it belongs (see Figure 2.10). A paper describing iGB^{web} is provided in Appendix B. The EGRIN 2.0 website is available at: http://egrin2.systemsbiology.net

Encoding biological information in a graphical representation has proved valuable for analyzing GRNs and discovering critical topological determinants of information flow in other contexts as well. Models of GRNs can be made accessible through visualization software with user-friendly interfaces, such as Cytoscape [309], BioTapestry [216], and Gephi [36]. By interoperating with other software and databases on the desktop or via the internet using frameworks such as Gaggle [310] and Firegoose [32], scientists can explore the complex networks and disparate data types to formulate hypotheses and drive experimentation. Application of these frameworks is turning increasingly to the web. Several projects like BioJS [140] - a modular JavaScript library for reusable interfaces - have attempted to increase biologist's ability to interact with their data on the web.

5.4 Towards a dynamical interpretation of genetic co-regulation: what do corems mean?

Corems expand the lexicon of genetic regulation. They make us reconsider what we mean by the term "co-regulation". Would, for example, a map of the direct interactions between every TF and every gene be sufficient to understand expression of the genome? From a dogmatic position centered around operons and regulons, it should. If genes are regulated by a common factor, they should be co-expressed. One might have to account for some combinatorial interactions that alter expression (Figure 5.2), but at least in microbes combinatorial interactions are somewhat rare. We would also expect from this perspective that genes with common functions will be physically bound by the same factor. While this is

certainly the case for many genes and functions, this limited perspective misses an entire side to co-regulation that has an enormous impact on cellular fitness. Corems can model these relationships.

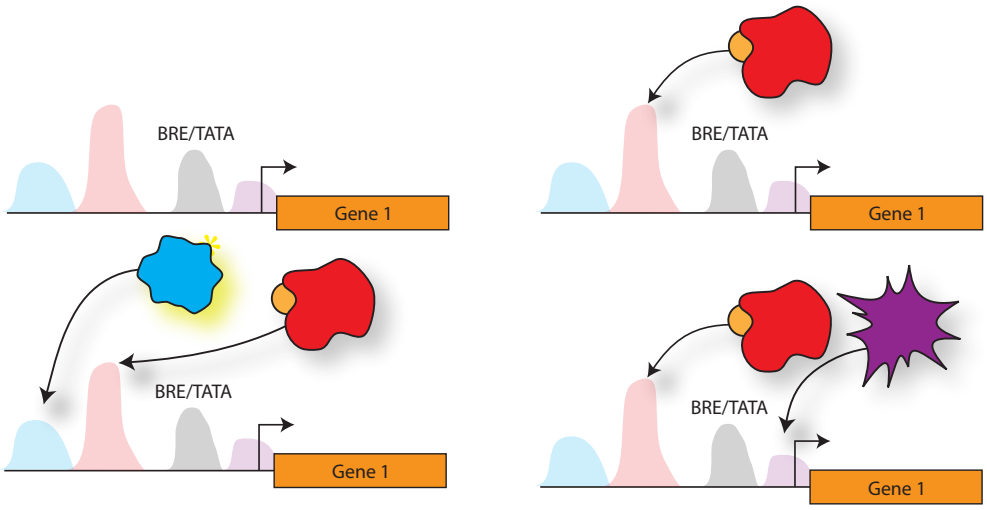


Figure 5.2: Multiple TFs can bind to a gene promoter. Binding of multiple TFs can be synergistic to promote transcription (bottom-left) or cause interference that reduces expression (bottom-right).

What has been missed in these definitions is the role of dynamics in transcriptional regulation - both at the level of physical interactions between TFs and DNA at gene promoters, and the dynamical properties that influence TF activity, like allosteric binding to small molecules. Most studies have emphasized strictly defined physical binding of common TFs to be "co-regulation". I have two concerns with this definition: (1) not all genes bound by a common TF are co-expressed. This is very clear from gene expression data. Many members of regulons are not highly co-expressed. At the moment, this cannot be accounted for by combinatorial influences. This means that it is difficult (if not impossible) to predict whether two genes will be co-expressed knowing that they are bound by a common TF. (2) Genes bound by different TFs can be co-expressed across many environments (more tightly than genes bound by a common TF). This is one of the observations that motivated

formalization of corems. It turns out that oftentimes these genes (even though they have no evidence for common direct transcriptional regulation) also have highly similar impact on fitness (functioning in related pathways). This suggest that there are transcriptional programs beyond direct regulation that contribute to the emergence of genetic expression modules. Two mechanisms for this behavior are proposed in Figure 5.5 and Figure 5.4.

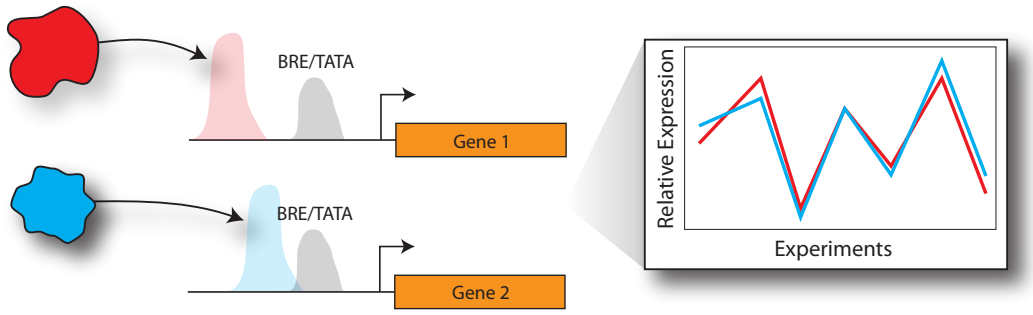


Figure 5.3: Despite being physically bound by different, genes can nevertheless be tightly co-expressed across a broad range of environmental conditions. Oftentimes we have also discovered that genes regulated in this way have similar impacts on fitness, suggesting they play conserved functional roles. Is it possible that there are higher-order mechanisms responsible for coordinating their co-regulation?

While it remains to be demonstrated directly whether these observations have a mechanistic basis, it suggests that coherent expression and coordination of the genome is a system-level property. There are many factors beyond transcription initiation to consider. With a fully detailed model of the cell (all parameters, rates, mechanisms, etc.), prediction of cellular expression state may be possible from GRN topology. Until that time, our results suggest that we take a more phenomenological view of co-regulation - focusing instead on the "emergent" modular states that are defined by (sometimes indirect) co-regulation.

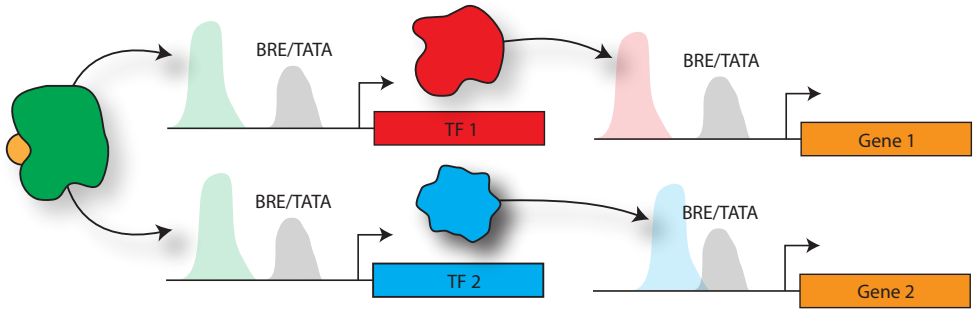


Figure 5.4: One conceivable way two genes with distinct direct inputs could produce similar expression is if the TFs themselves were coordinated by a similar factor (i.e., the TFs are isomorphic). This is likely to occur in biological networks and is identifiable directly from GRN topology.

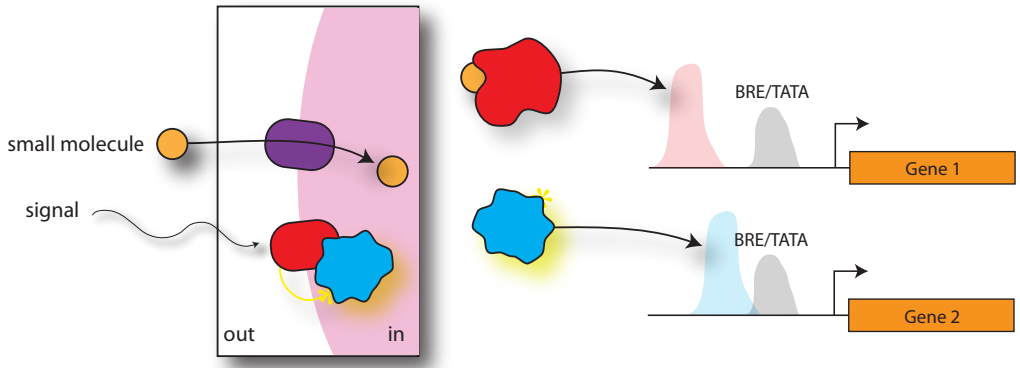


Figure 5.5: A more complicated way two genes with distinct direct inputs could be co-expressed is through allosteric coordination of TFs. TF activity is often regulated post-translationally. In prokaryotes, this is often by binding to small molecules, called allosteric regulators. If these small molecules cause TF activities to be similar (especially if the two small molecules were related by chemical transformations such that they often change together), the genes regulated by those TFs could exhibit similar expression patterns.

5.5 Beyond the GRN

In Chapter 1 I referred to an article by Neidhardt and Savageau entitled, "Regulation Beyond the Operon". It was clear even from the earliest investigations that regulation of the genome was more complicated than co-linear arrangement of genes. I suggest that we also move beyond gene regulation - beyond the GRN. What this thesis work suggests is that genetic regulation involves the entire biochemical milieu of the cell, especially metabolite and protein dynamics. At the moment, we can generate insight by modeling "emergent" features of the system (like corems) - with limited understanding of how these features are generated. Future work needs to explicitly model and integrate these layers. This is becoming possible with technological advances in metabolomics and proteomics. Like sequencing of the genome, generation of a completely accurate GRNs may produce more questions than it would answer.

BIBLIOGRAPHY

[1] August 4 1999. *Parallel Computing Technologies: 5th International Conference, PaCT-99, St. Petersburg, Russia, September 6-10, 1999 Proceedings*. Springer, Berlin ; New York, 1999 edition edition, August 1999.

[2] W N Abouhamad and M D Manson. The dipeptide permease of escherichia coli closely resembles other bacterial transport systems and shows growth-phase-dependent expression. *Molecular microbiology*, 14(5):1077–1092, December 1994.

[3] Murat Acar, Bernardo F. Pando, Frances H. Arnold, Michael B. Elowitz, and Alexander van Oudenaarden. A general mechanism for network-dosage compensation in gene circuits. *Science*, 329(5999):1656–1660, September 2010. WOS:000282098100045.

[4] Mark Achtman and Michael Wagner. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, 6(6):431440, June 2008. WOS:000255953300010.

[5] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761764, August 2010.

[6] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods*, 10(5):413420, May 2013.

[7] H. Aiba, F. Nakasai, S. Mizushima, and T. Mizuno. Evidence for the physiological importance of the phosphotransfer between the two regulatory components, EnvZ and OmpR, in osmoregulation in escherichia coli. *Journal of Biological Chemistry*, 264(24):1409014094, August 1989.

[8] Daniela Albanesi, Georgina Reh, Marcelo E Guerin, Francis Schaeffer, Michel Debarbouille, Alejandro Buschiazzo, Gustavo E Schujman, Diego de Mendoza, and Pedro M Alzari. Structural basis for feed-forward transcriptional regulation of membrane lipid homeostasis in staphylococcus aureus. *PLoS pathogens*, 9(1):e1003108, January 2013.

[9] Lilia Alberghina and Hans V. Westerhoff. *Systems Biology: Definitions and Perspectives*. Springer, October 2007.

[10] A Paul Alivisatos, Miyoung Chun, George M Church, Ralph J Greenspan, Michael L Roukes, and Rafael Yuste. The brain activity map project and the challenge of functional connectomics. *Neuron*, 74(6):970974, June 2012.

[11] Eric J Alm, Katherine H Huang, Morgan N Price, Richard P Koche, Keith Keller, Inna L Dubchak, and Adam P Arkin. The MicrobesOnline web site for comparative genomics. *Genome research*, 15(7):1015–1022, July 2005.

[12] U. Alon, M. G. Surette, N. Barkai, and S. Leibler. Robustness in bacterial chemotaxis. *Nature*, 397(6715):168171, January 1999.

[13] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450461, June 2007. WOS:000246603400012.

[14] L Caetano M Antunes, Rosana B R Ferreira, Michelle M C Buckner, and B Brett Finlay. Quorum sensing in bacterial virulence. *Microbiology (Reading, England)*, 156(Pt 8):22712282, August 2010.

[15] A Arkin, J Ross, and H H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. *Genetics*, 149(4):16331648, August 1998.

[16] Seyed M Assadi, Murat Ycel, and Christos Pantelis. Dopamine modulates neural networks involved in effort-based decision-making. *Neuroscience and biobehavioral reviews*, 33(3):383393, March 2009.

[17] Toshiyuki Nakagaki Atsushi Tero. A method inspired by physarum for solving the steiner problem. *IJUC*, 6:109123, 2010.

[18] Maria Avila, David M Ojcius, and Ozlem Yilmaz. The oral microbiota: living with a permanent guest. *DNA and cell biology*, 28(8):405411, August 2009.

[19] Ana Babic, Ariel B. Lindner, Marin Vulic, Eric J. Stewart, and Miroslav Radman. Direct visualization of horizontal gene transfer. *Science*, 319(5869):1533–1536, March 2008. WOS:000253943800041.

[20] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283291, June 2004. WOS:000222538400004.

[21] T. L. Bailey and M. Gribskov. Methods and statistics for combining motif match scores. *J Comput Biol*, 5(2):211221, 1998.

[22] TL Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, page 2836, 1994.

[23] N. S. Baliga, Y. A. Goo, W. V. Ng, L. Hood, C. J. Daniels, and S. DasSarma. Is gene expression in halobacterium NRC-1 regulated by multiple TBP and TFB transcription factors? *Molecular Microbiology*, 36(5):1184–1185, June 2000. WOS:000087358800018.

[24] Nitin S. Baliga. Systems biology - the scale of prediction. *Science*, 320(5881):1297–1298, June 2008. WOS:000256441100029.

[25] Nitin S. Baliga, Richard Bonneau, Marc T. Facciotti, Min Pan, Gustavo Glusman, Eric W. Deutsch, Paul Shannon, Yulun Chiu, Rueyhung Sting Weng, Rueichi Richie Gan, Pingliang Hung, Shailesh V. Date, Edward Marcotte, Leroy Hood, and Wailap Victor Ng. Genome sequence of haloarcula marismortui: a halophilic archaeon from the dead sea. *Genome Res*, 14(11):22212234, November 2004.

[26] NS Baliga, SP Kennedy, WV Ng, L Hood, and S DasSarma. Genomic and genetic dissection of an archaeal regulon. *Proc Natl Acad Sci USA*, 98(5):25212525, 2001.

[27] NS Baliga, M Pan, YA Goo, EC Yi, DR Goodlett, K Dimitrov, P Shannon, R Aebersold, WV Ng, and L Hood. Coordinate regulation of energy transduction modules in halobacterium sp. analyzed by a global systems approach. *Proc Natl Acad Sci USA*, 99(23):1491314918, 2002.

[28] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3:78, 2007.

[29] Tarun Bansal, Robert C Alaniz, Thomas K Wood, and Arul Jayaraman. The bacterial signal indole increases epithelial-cell tight-junction resistance and attenuates indicators of inflammation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):228233, January 2010.

[30] A. L. Barabasi and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–U15, February 2004. WOS:000188602400012.

[31] I. Barak and A. J. Wilkinson. Where asymmetry in gene expression originates. *Molecular Microbiology*, 57(3):611620, August 2005. WOS:000230303500002.

[32] J Christopher Bare, Paul T Shannon, Amy K Schmid, and Nitin S Baliga. The firegoose: two-way integration of diverse data from different bioinformatics web resources with desktop applications. *BMC bioinformatics*, 8:456, 2007.

[33] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. NCBI GEO: mining tens of millions of expression profilesdatabase and tools update. *Nucleic acids research*, 35(Database issue):D760765, January 2007.

[34] Jeffrey E. Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E. Lenski, and Jihyun F. Kim. Genome evolution and adaptation in a long-term experiment with escherichia coli. *Nature*, 461(7268):1243U74, October 2009. WOS:000271190800040.

[35] Nicholas H. Barton, Derek E. G. Briggs, Jonathan A. Eisen, David B. Goldstein, Nipam H. Patel, and & 2 more. *Evolution*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y, 1st edition edition, June 2007.

[36] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media*, March 2009. Gephi is an open source software for graph and network analysis. It uses a 3D render engine to display large networks in real-time and to speed up the exploration. A flexible and multi-task architecture brings new possibilities to work with complex data sets and produce valuable visual results. We present several key features of Gephi in the context of interactive exploration and interpretation of networks. It provides easy and broad access to network data and allows for spatializing, filtering, navigating, manipulating and clustering. Finally, by presenting dynamic features of Gephi, we highlight key aspects of dynamic network visualization.

[37] Hubertus J. E. Beaumont, Jenna Gallie, Christian Kost, Gayle C. Ferguson, and Paul B. Rainey. Experimental evolution of bet hedging. *Nature*, 462(7269):90–U97, November 2009. WOS:000271419200038.

[38] Eshel Ben Jacob, Israela Becker, Yoash Shapira, and Herbert Levine. Bacterial linguistic communication and social intelligence. *Trends in microbiology*, 12(8):366372, August 2004.

[39] AF Bennett, KM Dao, and RE Lenski. Rapid evolution in response to high-temperature selection. *Nature*, 346(6279):7981, July 1990. WOS:A1990DM39600067.

[40] Albert F. Bennett and Richard E. Lenski. An experimental test of evolutionary trade-offs during temperature adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 104:8649–8654, May 2007. WOS:000246697800013.

[41] H C Berg and P M Tedesco. Transient response to chemotactic stimuli in escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 72(8):32353239, August 1975.

[42] David B Berry, Qiaoning Guan, James Hose, Suraiya Haroon, Marinella Gebbia, Lawrence E Heisler, Corey Nislow, Guri Giaever, and Audrey P Gasch. Multiple means to the same end: the genetic basis of acquired stress resistance in yeast. *PLoS genetics*, 7(11):e1002353, November 2011.

[43] T. K. Blackwell and H. Weintraub. Differences and similarities in DNA-binding preferences of MyoD and e2a protein complexes revealed by binding site selection. *Science*, 250(4984):11041110, November 1990.

[44] Y Blat and N Kleckner. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell*, 98(2):249–259, July 1999.

[45] F R Blattner, G Plunkett, 3rd, C A Bloch, N T Perna, V Burland, M Riley, J Collado-Vides, J D Glasner, C K Rode, G F Mayhew, J Gregor, N W Davis, H A Kirkpatrick, M A Goeden, D J Rose, B Mau, and Y Shao. The complete genome sequence of escherichia coli k-12. *Science (New York, N.Y.)*, 277(5331):14531462, September 1997.

[46] Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. Historical contingency and the evolution of a key innovation in an experimental population of escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23):78997906, June 2008. WOS:000256781800002.

[47] R Bonneau, DJ Reiss, P Shannon, L Hood, NS Baliga, and V Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, 7(5):R36, 2006.

[48] Richard Bonneau. Learning biological networks: from modules to dynamics. *Nature Chemical Biology*, 4(11):658664, November 2008. WOS:000260315000010.

[49] Richard Bonneau, Nitin S. Baliga, Eric W. Deutsch, Paul Shannon, and Leroy Hood. Comprehensive de novo structure prediction in a systems-biology context for the archaea halobacterium sp. NRC-1. *Genome Biol*, 5(8):R52, 2004.

[50] Richard Bonneau, Marc T. Facciotti, David J. Reiss, Amy K. Schmid, Min Pan, Amardeep Kaur, Vesteinn Thorsson, Paul Shannon, Michael H. Johnson, J. Christopher Bare, William Longabaugh, Madhavi Vuthoori, Kenia Whitehead, Aviv Madar, Lena Suzuki, Tetsuya Mori, Dong-Eun Chang, Jocelyne DiRuggiero, Carl H. Johnson, Leroy Hood, and Nitin S. Baliga. A predictive model for transcriptional control of physiology in a free living. *Cell*, 131(7):13541365, December 2007.

[51] F. C. Boogerd, F. J. Bruggeman, R. C. Richardson, A. Stephan, and H. V. Westerhoff. Emergence and its place in nature: A case study of biochemical networks. *Synthese*, 145(1):131164, May 2005.

[52] Fred C. Boogerd, Frank J. Bruggeman, and Robert C. Richardson. Mechanistic explanations and models in molecular systems biology. *Foundations of Science*, 18(4):725744, November 2013.

[53] Fred C Boogerd, Hongwu Ma, Frank J Bruggeman, Wally C van Heeswijk, Rodolfo Garca-Contreras, Douwe Molenaar, Klaas Krab, and Hans V Westerhoff. AmtB-mediated NH3 transport in prokaryotes must be active and as a consequence regulation of transport by GlnK is mandatory to limit futile cycling of NH4(+)/NH3. *FEBS letters*, 585(1):2328, January 2011.

[54] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):23012309, December 2011.

[55] Robert B Bourret and Ann M Stock. Molecular information processing: lessons from bacterial chemotaxis. *The Journal of biological chemistry*, 277(12):96259628, March 2002.

[56] Nathan R Brady, Anne Hamacher-Brady, Hans V Westerhoff, and Roberta A Gottlieb. A wave of reactive oxygen species (ROS)-induced ROS release in a sea of excitable mitochondria. *Antioxidants & redox signaling*, 8(9-10):16511665, October 2006.

[57] Leo Breiman. Bagging predictors. In *Machine Learning*, page 123140, 1996.

[58] Leo Breiman. Random forests. *Machine Learning*, 45(1):532, October 2001.

[59] Aaron N Brooks, David J Reiss, Antoine Allard, WeiJu Wu, Diego M Salvanha, Christopher L Plaisier, Sriram Chandrasekaran, Min Pan, Amardeep Kaur, and Nitin S Baliga. A systemlevel model for the microbial regulatory genome. *Molecular Systems Biology*, 10(7), July 2014.

[60] Aaron N. Brooks, Serdar Turkarslan, Karlyn D. Beer, Fang Yin Lo, and Nitin S. Baliga. Adaptation of cells to new environments. *Wiley interdisciplinary reviews. Systems biology and medicine*, 3(5):544–561, September 2011.

[61] Rodney Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139159, 1991.

[62] F J Bruggeman, W C van Heeswijk, F C Boogerd, and H V Westerhoff. Macromolecular intelligence in microorganisms. *Biological chemistry*, 381(9-10):965972, October 2000.

[63] Frank J Bruggeman, Fred C Boogerd, and Hans V Westerhoff. The multifarious short-term regulation of ammonium assimilation of escherichia coli: dissection using an in silico replica. *The FEBS journal*, 272(8):19651985, April 2005.

[64] Molly K. Burke, Joseph P. Dunham, Parvin Shahrestani, Kevin R. Thornton, Michael R. Rose, and Anthony D. Long. Genome-wide analysis of a long-term evolution experiment with drosophila. *Nature*, 467(7315):587U111, September 2010. WOS:000282273100038.

[65] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 418429, 2000.

[66] P. Bhlmann and B. Yu. Analyzing bagging. *Annals of Statistics*, (30):927–961, 2002.

[67] L. N. Calhoun and Y. M. Kwon. The effect of long-term propionate adaptation on the stress resistance of salmonella enteritidis. *Journal of Applied Microbiology*, 109(4):1294–1300, October 2010. WOS:000281895200019.

[68] Javier Carrera, Santiago F Elena, and Alfonso Jaramillo. Computational design of genomic transcriptional networks with adaptation to varying environments. *Proceedings of the National Academy of Sciences of the United States of America*, 109(38):1527715282, September 2012.

[69] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander, and R. A. Young. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell*, 12(2):323337, February 2001. WOS:000170348000007.

[70] Samuel Chaffron, Hubert Rehrauer, Jakob Pernthaler, and Christian von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959, July 2010. WOS:000279404700009.

[71] Erika Check Hayden. Technology: The $1,000 genome. *Nature*, 507(7492):294295, March 2014.

[72] Byung-Kwan Cho, Stephen Federowicz, Young-Seoub Park, Karsten Zengler, and Bernhard . Palsson. Deciphering the transcriptional regulatory logic of amino acid metabolism. *Nat Chem Biol*, 8(1):6571, January 2012.

[73] Byung-Kwan Cho, Stephen A Federowicz, Mallory Embree, Young-Seoub Park, Donghyuk Kim, and Bernhard Palsson. The PurR regulon in escherichia coli k-12 MG1655. *Nucleic acids research*, 39(15):6456–6464, August 2011.

[74] Kevin B Clark. Arrhenius-kinetics evidence for quantum tunneling in microbial "social" decision rates. *Communicative & Integrative Biology*, 3(6):540544, 2010.

[75] Kevin B. Clark. On classical and quantum error-correction in ciliate mate selection. *Communicative & Integrative Biology*, 3(4):374378, July 2010.

[76] Kevin B. Clark. Origins of learned reciprocity in solitary ciliates searching grouped 'courting' assurances at quantum efficiencies. *Bio Systems*, 99(1):2741, January 2010.

[77] Kevin B. Clark. Social biases determine spatiotemporal sparseness of ciliate mating heuristics. *Communicative & Integrative Biology*, 5(1):311, January 2012.

[78] Kevin B. Clark. Ciliates learn to diagnose and correct classical error syndromes in mating strategies. *Frontiers in Microbiology*, 4:229, 2013.

[79] Jose C Clemente, Luke K Ursell, Laura Wegener Parfrey, and Rob Knight. The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):12581270, March 2012.

[80] Tim F. Cooper and Richard E. Lenski. Experimental evolution with e. coli in diverse resource environments. i. fluctuating environments promote divergence of replicate populations. *Bmc Evolutionary Biology*, 10:11, January 2010. WOS:000275250400001.

[81] W. C. Corning and R. von Burg. Protozoa. In W. C. Corning, J. A. Dyal, and A. O. D. Willows, editors, *Invertebrate Learning*, page 49122. Springer US, January 1973.

[82] J W Costerton, Z Lewandowski, D E Caldwell, D R Korber, and H M Lappin-Scott. Microbial biofilms. *Annual Review of Microbiology*, 49(1):711745, 1995.

[83] Gregory E Crawford, Ingeborg E Holt, James C Mullikin, Denise Tai, Robert Blakesley, Gerard Bouffard, Alice Young, Catherine Masiello, Eric D Green, Tyra G Wolfsberg, Francis S Collins, and National Institutes Of Health Intramural Sequencing Center. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):992–997, January 2004.

[84] John F Cryan and Timothy G Dinan. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nature reviews. Neuroscience*, 13(10):701712, October 2012.

[85] Marie E Csete and John C Doyle. Reverse engineering of biological complexity. *Science (New York, N.Y.)*, 295(5560):16641669, March 2002.

[86] Shiladitya DasSarma, Brian R Berquist, James A Coker, Priya DasSarma, and Jochen A Muller. Post-genomics of the model haloarchaeon halobacterium sp. NRC-1. *Saline Systems*, 2:3, March 2006.

[87] Julian Davies. Origins and evolution of antibiotic resistance. *Microbiologia (Madrid)*, 12(1):916, 1996. BCI:BCI199699083815.

[88] Roger Day and Alex Lisovich. DAVIDQuery: Retrieval from the DAVID bioinformatics data resource into r, 2010.

[89] John Day-Richter, Midori A Harris, Melissa Haendel, Gene Ontology OBO-Edit Working Group, and Suzanna Lewis. OBO-editan ontology editor for biologists. *Bioinformatics (Oxford, England)*, 23(16):21982200, August 2007.

[90] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, October 2010. WOS:000281908700011.

[91] Riet De Smet and Kathleen Marchal. An ensemble biclustering approach for querying gene expression compendia with experimental lists. *Bioinformatics*, 27(14):19481956, July 2011.

[92] JAGM de Visser and D. E. Rozen. Clonal interference and the periodic selection of new beneficial mutations in escherichia coli. *Genetics*, 172(4):2093–2100, April 2006. WOS:000237225800008.

[93] Janos Demeter, Catherine Beauheim, Jeremy Gollub, Tina Hernandez-Boussard, Heng Jin, Donald Maier, John C Matese, Michael Nitzberg, Farrell Wymore, Zachariah K Zachariah, Patrick O Brown, Gavin Sherlock, and Catherine A Ball. The stanford microarray database: implementation of new analysis tools and open source release of software. *Nucleic acids research*, 35(Database issue):D766–770, January 2007.

[94] Glynn Dennis, Jr, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. DAVID: Database for annotation, visualization, and integrated discovery. *Genome biology*, 4(5):P3, 2003.

[95] Gyanendra P Dubey and Sigal Ben-Yehuda. Intercellular nanotubes mediate bacterial communication. *Cell*, 144(4):590600, February 2011.

[96] John E Dueber, Gabriel C Wu, G Reza Malmirchegini, Tae Seok Moon, Christopher J Petzold, Adeeti V Ullal, Kristala L J Prather, and Jay D Keasling. Synthetic protein scaffolds provide modular control over metabolic flux. *Nature biotechnology*, 27(8):753759, August 2009.

[97] Zo Dumas, Adin Ross-Gillespie, and Rolf Kmmerli. Switching between apparently redundant iron-uptake mechanisms benefits bacteria in changeable environments. *Proceedings. Biological sciences / The Royal Society*, 280(1764):20131055, August 2013.

[98] D. J. Earl and M. W. Deem. Evolvability is a selectable trait. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):1153111536, August 2004. WOS:000223276700003.

[99] D. R. Edgell and W. F. Doolittle. Archaea and the origin(s) of DNA replication proteins. *Cell*, 89(7):995998, June 1997. WOS:A1997XG83000002.

[100] J. L. Edwards, E. J. Brown, S. Uk-Nham, J. G. Cannon, M. S. Blake, and M. A. Apicella. A co-operative interaction between neisseria gonorrhoeae and complement receptor 3 mediates infection of primary cervical epithelial cells. *Cellular Microbiology*, 4(9):571584, September 2002. WOS:000178117300002.

[101] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):1486314868, December 1998.

[102] H El-Samad, J P Goff, and M Khammash. Calcium homeostasis and parturient hypocalcemia: an integral feedback perspective. *Journal of theoretical biology*, 214(1):1729, January 2002.

[103] B. Enjalbert, A. Nantel, and M. Whiteway. Stress-induced gene expression in candida albicans: Absence of a general stress response. *Molecular Biology of the Cell*, 14(4):14601467, April 2003. WOS:000182185200018.

[104] J. Errington. Regulation of endospore formation in bacillus subtilis. *Nature Reviews Microbiology*, 1(2):117–126, November 2003. WOS:000220402500013.

[105] Amandine Everard and Patrice D Cani. Diabetes, obesity and gut microbiota. *Best practice & research. Clinical gastroenterology*, 27(1):7383, February 2013.

[106] Marc T. Facciotti, Wyming L. Pang, Fang-yin Lo, Kenia Whitehead, Tie Koide, Ken-ichi Masumura, Min Pan, Amardeep Kaur, David J. Larsen, David J. Reiss, Linh Hoang, Ewa Kalisiak, Trent Northen, Sunia A. Trauger, Gary Siuzdak, and Nitin S. Baliga. Large scale physiological readjustment during growth enables rapid, comprehensive and inexpensive systems analysis. *BMC Syst Biol*, 4:64, 2010.

[107] Marc T. Facciotti, David J. Reiss, Min Pan, Amardeep Kaur, Madhavi Vuthoori, Richard Bonneau, Paul Shannon, Alok Srivastava, Samuel M. Donohoe, Leroy E. Hood, and Nitin S. Baliga. General transcription factor specified global gene regulation in archaea. *Proceedings of the National Academy of Sciences of the United States of America*, 104(11):4630–4635, March 2007. WOS:000244972700069.

[108] Jeremiah J. Faith, Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, January 2007.

[109] W R Farmer and J C Liao. Improving lycopene production in escherichia coli by engineering metabolic control. *Nature biotechnology*, 18(5):533537, May 2000.

[110] Michael J Federle. Autoinducer-2-based chemical communication in bacteria: complexities of interspecies signaling. *Contributions to microbiology*, 16:1832, 2009.

[111] David Fell. *Understanding the Control of Metabolism*. Portland Pr, London; Miami; Brookfield, VT, 1 edition edition, November 1996.

[112] G. Feller and C. Gerday. Psychrophilic enzymes: Hot topics in cold adaptation. *Nature Reviews Microbiology*, 1(3):200208, December 2003. WOS:000220431600013.

[113] Georges Feller. Life at low temperatures: is disorder the driving force? *Extremophiles*, 11(2):211–216, March 2007. WOS:000244336400001.

[114] T. L. Ferea, D. Botstein, P. O. Brown, and R. F. Rosenzweig. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 96(17):9721–9726, August 1999. WOS:000082098500052.

[115] Richard P. Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6-7):467488, June 1982.

[116] S Fields and O Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 1989.

[117] S. E. Finkel and R. Kolter. Evolution of microbial diversity during prolonged starvation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7):40234027, March 1999. WOS:000079507900121.

[118] RD Fleischmann, MD Adams, O. White, RA Clayton, EF Kirkness, AR Kerlavage, CJ Bult, JF Tomb, BA Dougherty, JM Merrick, K. McKenney, G. Sutton, W. Fitzhugh, C. Fields, JD Gocayne, J. Scott, R. Shirley, LI Liu, A. Glodek, JM Kelley, JF Weidman, CA Phillips, T. Spriggs, E. Hedblom, MD Cotton, TR Utterback, MC Hanna, DT Nguyen, DM Saudek, RC Brandon, LD Fine, JL Fritchman, JL Fuhrmann, NSM Geoghagen, CL Gnehm, LA McDonald, KV Small, CM Fraser, HO Smith, and JC Venter. Whole-genome random sequencing and assembly of haemophilus-influenzae rd. *Science*, 269(5223):496512, July 1995. WOS:A1995RL49500017.

[119] Enrique Flores and Antonia Herrero. Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nature reviews. Microbiology*, 8(1):3950, January 2010.

[120] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In Paul Vitnyi, editor, *Computational Learning Theory*, number 904 in Lecture Notes in Computer Science, page 2337. Springer Berlin Heidelberg, January 1995.

[121] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):11891232.

[122] Eileen Fung, Wilson W Wong, Jason K Suen, Thomas Bulter, Sun-gu Lee, and James C Liao. A synthetic gene-metabolic oscillator. *Nature*, 435(7038):118122, May 2005.

[123] Jochen Frster, Iman Famili, Patrick Fu, Bernhard Palsson, and Jens Nielsen. Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome Research*, 13(2):244253, February 2003.

[124] Rodrigo S. Galhardo, P. J. Hastings, and Susan M. Rosenberg. Mutation as a stress response and the regulation of evolvability. *Critical Reviews in Biochemistry and Molecular Biology*, 42(5):399–435, January 2007.

[125] Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muiz-Rascado, Hilda Solano-Lira, Vernica Jimenez-Jacinto, Verena Weiss, Jair S Garca-Sotelo, Alejandra Lpez-Fuentes, Liliana Porrn-Sotelo, Shirley Alquicira-Hernndez, Alejandra Medina-Rivera, Irma Martnez-Flores, Kevin Alquicira-Hernndez, Ruth Martnez-Adame, Csar Bonavides-Martnez, Juan Miranda-Ros, Araceli M Huerta, Alfredo Mendoza-Vargas, Leonardo Collado-Torres, Blanca Taboada, Leticia Vega-Alvarado, Maricela Olvera, Leticia Olvera, Ricardo Grande, Enrique Morett, and Julio Collado-Vides. RegulonDB version 7.0: transcriptional regulation of escherichia coli k-12 integrated within genetic sensory response units (gensor units). *Nucleic acids research*, 39(Database issue):D98105, January 2011.

[126] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, July 2003. WOS:000183914700043.

[127] AP Gasch, PT Spellman, CM Kao, O Carmel-Harel, MB Eisen, G Storz, D Botstein, and PO Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):42414257, 2000.

[128] E. P. Geiduschek and M. Ouhammouch. Archaeal transcription and its regulators. *Molecular Microbiology*, 56(6):1397–1407, June 2005. WOS:000229181800001.

[129] B. Gelber. Investigations of the behavior of paramecium aurelia. i. modification of behavior after training with reinforcement. *Journal of Comparative and Physiological Psychology*, 45(1):5865, February 1952.

[130] D. Gevers, K. Vandepoele, C. Simillion, and Y. Van de Peer. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends in Microbiology*, 12(4):148154, April 2004. WOS:000221051600003.

[131] Daniel G. Gibson, John I. Glass, Carole Lartigue, Vladimir N. Noskov, Ray-Yuan Chuang, Mikkel A. Algire, Gwynedd A. Benders, Michael G. Montague, Li Ma, Monzia M. Moodie, Chuck Merryman, Sanjay Vashee, Radha Krishnakumar, Nacyra Assad-Garcia, Cynthia Andrews-Pfannkoch, Evgeniya A. Denisova, Lei Young,

Zhi-Qing Qi, Thomas H. Segall-Shapiro, Christopher H. Calvey, Prashanth P. Parmar, Clyde A. Hutchison, Hamilton O. Smith, and J. Craig Venter. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987):5256, July 2010. WOS:000279402700028.

[132] Ezequiel Gleichgerrcht, Agustn Ibez, Mara Roca, Teresa Torralva, and Facundo Manes. Decision-making cognition in neurodegenerative diseases. *Nature reviews. Neurology*, 6(11):611623, November 2010.

[133] Christian Godon, Gilles Lagniel, Jaekwon Lee, Jean-Marie Buhler, Sylvie Kieffer, Michel Perrot, Hlian Boucherie, Michel B. Toledano, and Jean Labarre. The h2o2 stimulon in saccharomyces cerevisiae. *Journal of Biological Chemistry*, 273(35):2248022489, August 1998.

[134] Eunhye Goo, Charlotte D Majerczyk, Jae Hyung An, Josephine R Chandler, Young-Su Seo, Hyeonheui Ham, Jae Yun Lim, Hongsup Kim, Bongsoo Lee, Moon Sun Jang, E Peter Greenberg, and Ingyu Hwang. Bacterial quorum sensing, cooperativity, and anticipation of stationary-phase stress. *Proceedings of the National Academy of Sciences of the United States of America*, 109(48):1977519780, November 2012.

[135] SJ Gould and RC Lewtonin. Spandrels of san-marco and the panglossian paradigm - a critique of the adaptationist program. *Proceedings of the Royal Society Series B-Biological Sciences*, 205(1161):581598, 1979. WOS:A1979HN99900010.

[136] Stephen Jay Gould. *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton & Company, New York, September 1990.

[137] S. Gribaldo and C. Brochier-Armanet. The origin and evolution of archaea: a state of the art. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 361(1470):10071022, June 2006. WOS:000238359500010.

[138] John R. Guest, Jeffrey Green, Alistair S. Irvine, and Stephen Spiro. The FNR modulon and FNR-regulated gene expression. In *Regulation of Gene Expression in Escherichia coli*, page 317342. Springer US, January 1996.

[139] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biol*, 8(2):R24, 2007.

[140] John Gmez, Leyla J. Garca, Gustavo A. Salazar, Jose Villaveces, Swanand Gore, Alexander Garca, Maria J. Martn, Guillaume Launay, Rafael Alcntara, Noemi Del Toro Aylln, Marine Dumousseau, Sandra Orchard, Sameer Velankar, Henning

Hermjakob, Chenggong Zong, Peipei Ping, Manuel Corpas, and Rafael C. Jimnez. BioJS: An open source JavaScript framework for biological data visualization. *Bioinformatics*, page btt100, February 2013.

[141] F. He H. V. Westerhoff. Understanding principles of the dynamic biochemical networks of life through systems biology. 2014.

[142] George Hajishengallis. Porphyromonas gingivalis-host interactions: open war or intelligent guerilla tactics? *Microbes and infection / Institut Pasteur*, 11(6-7):637645, June 2009.

[143] L. W. Hamoen, G. Venema, and O. P. Kuipers. Controlling competence in bacillus subtilis: shared use of regulators. *Microbiology-Sgm*, 149:9–17, January 2003. WOS:000180536500002.

[144] J. Handelsman. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669+, December 2004. WOS:000225854100005.

[145] Clinton H Hansen, Robert G Endres, and Ned S Wingreen. Chemotaxis in escherichia coli: a molecular model for robust precise adaptation. *PLoS computational biology*, 4(1):e1, January 2008.

[146] N Hao, M Behar, T C Elston, and H G Dohlman. Systems biology analysis of g protein and MAP kinase signaling in yeast. *Oncogene*, 26(22):32543266, May 2007.

[147] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47C52, December 1999. WOS:000084014100007.

[148] Fei He, Vincent Fromion, and Hans V Westerhoff. (im)perfect robustness and adaptation of metabolic networks subject to metabolic and gene-expression regulation: marrying control engineering with metabolic control analysis. *BMC systems biology*, 7:131, 2013.

[149] Guimei He, Beibei He, Paul A. Racey, and Jie Cui. Positive selection of the bat interferon alpha gene family. *Biochemical Genetics*, 48(9-10):840846, October 2010. WOS:000281670700012.

[150] K J Hellingwerf, P W Postma, J Tommassen, and H V Westerhoff. Signal transduction in bacteria: phospho-neural network(s) in escherichia coli? *FEMS microbiology reviews*, 16(4):309321, July 1995.

[151] U. Hentschel, J. Hopke, M. Horn, A. B. Friedrich, M. Wagner, J. Hacker, and B. S. Moore. Molecular evidence for a uniform microbial community in sponges from different oceans. *Applied and Environmental Microbiology*, 68(9):4431–4440, September 2002. WOS:000177718000036.

[152] Ann M Hermundstad, Kevin S Brown, Danielle S Bassett, and Jean M Carlson. Learning, memory, and the role of neural network architecture. *PLoS computational biology*, 7(6):e1002063, June 2011.

[153] C. F. Higgins, S. C. Hyde, M. M. Mimmack, U. Gileadi, D. R. Gill, and M. P. Gallagher. Binding protein-dependent transport systems. *Journal of Bioenergetics and Biomembranes*, 22(4):571592, August 1990.

[154] David J. Hinkle and David C. Wood. Is tube-escape learning by protozoa associative learning? *Behavioral Neuroscience*, 108(1):9499, 1994.

[155] Peter W. Hochachka and George N. Somero. *Biochemical Adaptation: Mechanism and Process in Physiological Evolution*. Oxford University Press, New York, 1 edition edition, January 2002.

[156] S M Hoffer, H V Westerhoff, K J Hellingwerf, P W Postma, and J Tommassen. Autoamplification of a two-component regulatory system results in "learning" behavior. *Journal of bacteriology*, 183(16):49144917, August 2001.

[157] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):25542558, April 1982.

[158] Taha Hosni, Chiaraluce Moretti, Giulia Devescovi, Zulma Rocio Suarez-Moreno, M' Barek Fatmi, Corrado Guarnaccia, Sandor Pongor, Andrea Onofri, Roberto Buonaurio, and Vittorio Venturi. Sharing of quorum-sensing signals and role of interspecies communities in a bacterial plant disease. *The ISME journal*, 5(12):18571870, December 2011.

[159] Araceli M. Huerta, Heladia Salgado, Denis Thieffry, and Julio Collado-Vides. RegulonDB: A database on transcriptional regulation in escherichia coli. *Nucleic Acids Research*, 26(1):55–59, January 1998.

[160] GW Huisman and R. Kolter. Sensing starvation - a homoserine lactone-dependent signaling pathway in escherichia coli. *Science*, 265(5171):537539, July 1994. WOS:A1994NY21600035.

[161] Vn Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, September 2010.

[162] T Ideker, V Thorsson, A F Siegel, and L E Hood. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of computational biology: a journal of computational molecular cell biology*, 7(6):805817, 2000.

[163] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299, September 1996.

[164] J. Ihmels, S. Bergmann, M. Gerami-Nejad, I. Yanai, M. McClellan, J. Berman, and N. Barkai. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, 309(5736):938–940, August 2005. WOS:000231101400048.

[165] Mark Isalan, Caroline Lemerle, Konstantinos Michalodimitrakis, Carsten Horn, Pedro Beltrao, Emanuele Raineri, Mireia Garriga-Canut, and Luis Serrano. Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, 452(7189):840–U2, April 2008. WOS:000255026000041.

[166] Nobuyoshi Ishii, Kenji Nakahigashi, Tomoya Baba, Martin Robert, Tomoyoshi Soga, Akio Kanai, Takashi Hirasawa, Miki Naba, Kenta Hirai, Aminul Hoque, Pei Yee Ho, Yuji Kakazu, Kaori Sugawara, Saori Igarashi, Satoshi Harada, Takeshi Masuda, Naoyuki Sugiyama, Takashi Togashi, Miki Hasegawa, Yuki Takai, Katsuyuki Yugi, Kazuharu Arakawa, Nayuta Iwata, Yoshihiro Toya, Yoichi Nakayama, Takaaki Nishioka, Kazuyuki Shimizu, Hirotada Mori, and Masaru Tomita. Multiple high-throughput analyses monitor the response of e. coli to perturbations. *Science*, 316(5824):593597, April 2007.

[167] F Jacob and J Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3:318356, June 1961.

[168] R. Jaenicke. Protein stability and molecular adaptation to extreme conditions. *European Journal of Biochemistry*, 202(3):715–728, December 1991. WOS:A1991GX99500003.

[169] Ken F Jarrell and Mark J McBride. The surprisingly diverse ways that prokaryotes move. *Nature reviews. Microbiology*, 6(6):466476, June 2008.

[170] D. D. Jensen. Experiments on learning in paramecia. *Science (New York, N.Y.)*, 125(3240):191192, February 1957.

[171] Carl Hirschie Johnson, Tetsuya Mori, and Yao Xu. A cyanobacterial circadian clockwork. *Current Biology*, 18(17):R816–R825, September 2008. WOS:000259108600026.

[172] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–1502, June 2007.

[173] Anagha Joshi, Riet De Smet, Kathleen Marchal, Yves Van de Peer, and Tom Michoel. Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4):490496, February 2009.

[174] Mario Juhas, Lutz Wiehlmann, Birgit Huber, Doris Jordan, Joerg Lauber, Prabhakar Salunkhe, Anna Silke Limpert, Franz von Gtz, Ivo Steinmetz, Leo Eberl, and Burkhard Tmmler. Global regulation of quorum sensing and virulence by VqsR in pseudomonas aeruginosa. *Microbiology (Reading, England)*, 150(Pt 4):831841, April 2004.

[175] D Kahn and H V Westerhoff. Control theory of regulatory cascades. *Journal of theoretical biology*, 153(2):255285, November 1991.

[176] Dale Kaiser. Signaling in myxobacteria. *Annual review of microbiology*, 58:7598, 2004.

[177] Alex T Kalinka and Pavel Tomancak. linkcomm: an r package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics (Oxford, England)*, 27(14):20112012, July 2011.

[178] Frits Kamp and Hans V. Westerhoff. Molecular machines and energy channelling. In G. Rickey Welch and James S. Clegg, editors, *The Organization of Cell Metabolism*, number 127 in NATO ASI Series, page 357365. Springer US, January 1986.

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

[180] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389401, July 2012.

[181] Nadav Kashtan, Elad Noor, and Uri Alon. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):1371113716, August 2007. WOS:000249064700035.

[182] Rees Kassen. Toward a general theory of adaptive radiation insights from microbial experimental evolution. In C. D. Schlichting and T. A. Mousseau, editors, *Year in Evolutionary Biology 2009*, volume 1168, page 322. 2009. WOS:000268507900001.

[183] K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5):491496, October 2003. WOS:000186448200007.

[184] A. Kaur, M. Pan, M. Meislin, M. T. Facciotti, R. El-Gewely, and N. S. Baliga. A systems view of haloarchaeal strategies to withstand stress from transition metals. *Genome Research*, 16(7):841854, July 2006. WOS:000238712400004.

[185] Amardeep Kaur, Phu T. Van, Courtney R. Busch, Courtney K. Robinson, Min Pan, Wyming Lee Pang, David J. Reiss, Jocelyne DiRuggiero, and Nitin S. Baliga. Coordination of frontline defense mechanisms under severe oxidative stress. *Molecular Systems Biology*, 6:393, July 2010. WOS:000284524200005.

[186] M. Kellis, N. Patterson, B. Birren, B. Berger, and E. S. Lander. Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, 11(2-3):319–355, 2004. WOS:000222588300008.

[187] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003. WOS:000182853100033.

[188] S. P. Kennedy, W. V. Ng, S. L. Salzberg, L. Hood, and S. DasSarma. Understanding the adaptation of halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Research*, 11(10):16411650, October 2001. WOS:000171456000006.

[189] Arkady B. Khodursky, Jonathan A. Bernstein, Brian J. Peter, Virgil Rhodius, Volker F. Wendisch, and Daniel P. Zimmer. Escherichia coli spotted double-strand DNA microarrays. In Michael J. Brownstein and Arkady B. Khodursky, editors, *Functional Genomics*, number 224 in Methods in Molecular Biology, pages 61–78. Humana Press, January 2003.

[190] H. Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, November 2004. WOS:000224832600010.

[191] Dorthe Kixmller, Henrik Strahl, Andy Wende, and Jrg-Christian Greie. Archaeal transcriptional regulation of the prokaryotic KdpFABC complex mediating k(+) uptake in h. salinarum. *Extremophiles*, 15(6):643652, November 2011.

[192] Tie Koide, Wyming Lee Pang, and Nitin S. Baliga. The role of predictive modelling in rationally re-engineering biological systems. *Nature Reviews Microbiology*, 7(4):297–305, April 2009. WOS:000264179900015.

[193] Tie Koide, David J. Reiss, J Christopher Bare, Wyming Lee Pang, Marc T. Facciotti, Amy K. Schmid, Min Pan, Bruz Marzolf, Phu T. Van, Fang-Yin Lo, Abhishek Pratap, Eric W. Deutsch, Amelia Peterson, Dan Martin, and Nitin S. Baliga. Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol*, 5:285, 2009.

[194] Alexey Kolodkin, Fred C Boogerd, Nick Plant, Frank J Bruggeman, Valeri Goncharuk, Jeantine Lunshof, Rafael Moreno-Sanchez, Nilgun Yilmaz, Barbara M Bakker, Jacky L Snoep, Rudi Balling, and Hans V Westerhoff. Emergence of the silicon human and network targeting drugs. *European journal of pharmaceutical sciences: official journal of the European Federation for Pharmaceutical Sciences*, 46(4):190197, July 2012.

[195] Alexey Kolodkin, Evangelos Simeonidis, Rudi Balling, and Hans V Westerhoff. Understanding complexity in neurodegenerative diseases: in silico reconstruction of emergence. *Frontiers in physiology*, 3:291, 2012.

[196] Alexey Kolodkin, Evangelos Simeonidis, and Hans V Westerhoff. Computing life: Add logos to biology and bios to physics. *Progress in biophysics and molecular biology*, 111(2-3):6974, April 2013.

[197] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annual Review of Microbiology*, 55:709–742, 2001. WOS:000171732600027.

[198] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223, November 2002. WOS:000179200900058.

[199] Anders Krogh and Peter Sollich. Statistical mechanics of ensemble learning. *Physical Review E*, 55(1):811825, January 1997.

[200] Natalja Kurbatova, Tomasz Adamusiak, Pavel Kurnosov, Morris A Swertz, and Misha Kapushesky. ontoCAT: an r package for ontology traversal and search. *Bioinformatics (Oxford, England)*, 27(17):2468–2470, September 2011.

[201] Donn Kushner. *Microbial life in extreme environments*. Academic Press, 1978.

[202] Peter Langfelder and Steve Horvath. WGCNA: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9:559, 2008.

[203] Michael T Laub and Mark Goulian. Specificity in two-component signal transduction pathways. *Annual review of genetics*, 41:121–145, 2007.

[204] Federico M. Lauro, Roger A. Chastain, Lesley E. Blankenship, A. Aristides Yayanos, and Douglas H. Bartlett. The unique 16s rRNA genes of piezophiles reflect both phylogeny and adaptation. *Applied and Environmental Microbiology*, 73(3):838–845, February 2007. WOS:000244263800021.

[205] Jintae Lee, Arul Jayaraman, and Thomas K Wood. Indole is an inter-species biofilm signal mediated by SdiA. *BMC microbiology*, 7:42, 2007.

[206] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, October 2002. WOS:000178791200051.

[207] Madeleine Leisner, Kerstin Stingl, Erwin Frey, and Berenike Maier. Stochastic switching to competence. *Current opinion in microbiology*, 11(6):553559, December 2008.

[208] Karen Lemmens, Tijl De Bie, Thomas Dhollander, Sigrid C. De Keersmaecker, Inge M. Thijs, Geert Schoofs, Ami De Weerdt, Bart De Moor, Jos Vanderleyden, Julio Collado-Vides, Kristof Engelen, and Kathleen Marchal. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in escherichia coli. *Genome Biol*, 10(3):R27, 2009.

[209] J W Lengeler. Metabolic networks: a signal-oriented approach to cellular models. *Biological chemistry*, 381(9-10):911920, October 2000.

[210] R E Lenski. Quantifying fitness and gene stability in microorganisms. *Biotechnology (Reading, Mass.)*, 15:173–192, 1991.

[211] RE Lenski and M. Travisano. Dynamics of adaptation and diversification - a 10,000-generation experiment with bacterial-populations. *Proceedings of the National Academy of Sciences of the United States of America*, 91(15):6808–6814, July 1994. WOS:A1994NY34800014.

[212] Martin J. Lercher and Csaba Pal. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Molecular Biology and Evolution*, 25(3):559–567, March 2008. WOS:000253491100010.

[213] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4936–4942, April 2005. WOS:000228195800006.

[214] F. T. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):47814786, April 2004. WOS:000220761200013.

[215] Wen-Hsiung Li. *Molecular Evolution*. Sinauer Associates, Sunderland, Mass, January 1997.

[216] W. J. R. Longabaugh, E. H. Davidson, and H. Bolouri. Computational representation of developmental genetic regulatory networks. *Developmental Biology*, 283(1):116, July 2005. WOS:000230418400001.

[217] Richard Christiaan Looijen. *Holism and Reductionism in Biology and Ecology: The Mutual Dependence of Higher and Lower Level Research Programmes*. Rijksuniversiteit Groningen, 1998.

[218] Luis Lopez-Maury, Samuel Marguerat, and Juerg Baehler. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8):583593, August 2008. WOS:000257758400009.

[219] Irma Lozada-Chavez, Vladimir Espinosa Angarica, Julio Collado-Vides, and Bruno Contreras-Moreira. The role of DNA-binding specificity in the evolution of bacterial regulatory networks. *Journal of Molecular Biology*, 379(3):627–643, June 2008. WOS:000256586500021.

[220] N. M. Luscombe and J. M. Thornton. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology*, 320(5):991–1009, July 2002. WOS:000177459300007.

[221] Pamela Lyon. The biogenic approach to cognition. *Cognitive processing*, 7(1):1129, March 2006.

[222] Elena S. Lysenko, Rebeccah S. Lijek, Sam P. Brown, and Jeffrey N. Weiser. Within-host competition drives selection for the capsule virulence determinant of streptococcus pneumoniae. *Current Biology*, 20(13):12221226, July 2010. WOS:000280024300033.

[223] Sylvie Ltoff, Philippe Delepelaire, and Ccile Wandersman. The housekeeping dipeptide permease is the escherichia coli heme transporter and functions with two optional peptide binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(34):12891–12896, August 2006.

[224] Qun Ma, Alicia Fonseca, Wenqi Liu, Andrew T Fields, Meaghan L Pimsler, Aline F Spindola, Aaron M Tarone, Tawni L Crippen, Jeffery K Tomberlin, and Thomas K Wood. Proteus mirabilis interkingdom swarming signals attract blow flies. *The ISME journal*, 6(7):13561366, July 2012.

[225] Hedia Maamar, Arjun Raj, and David Dubnau. Noise in gene expression determines cell fate in bacillus subtilis. *Science*, 317(5837):526–529, July 2007. WOS:000248339800049.

[226] Werner K. Maas and A. J. Clark. Studies on the mechanism of repression of arginine biosynthesis in escherichia coli: II. dominance of repressibility in diploids. *Journal of Molecular Biology*, 8(3):365370, March 1964.

[227] Sandra Macfarlane, Bahram Bahrami, and George T Macfarlane. Mucosal biofilm communities in the human intestinal tract. *Advances in applied microbiology*, 75:111143, 2011.

[228] M Madan Babu and Sarah A Teichmann. Evolution of transcription factors and the gene regulatory network in escherichia coli. *Nucleic acids research*, 31(4):12341244, February 2003.

[229] Martin Mahner and Mario Bunge. Function and functionalism: A synthetic perspective. *Philosophy of Science*, 68(1):7594, 2001.

[230] Shaun Mahony and Panayiotis V Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research*, 35(Web Server issue):W253258, July 2007.

[231] Kira S. Makarova, Nick V. Grishin, and Eugene V. Koonin. The HicAB cassette, a putative novel, RNA-targeting toxin-antitoxin system in archaea and bacteria. *Bioinformatics*, 22(21):2581–2584, November 2006. WOS:000241629600001.

[232] Y. Makita, M. Nakao, N. Ogasawara, and K. Nakai. DBTBS: database of transcriptional regulation in bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Research*, 32:D75–D77, January 2004. WOS:000188079000014.

[233] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):1198011985, October 2003. WOS:000186024300013.

[234] YI Manin. The computable and the non-computable. (vychislimoe i nevychislimoe. January.

[235] Nils N Mank, Bork A Berghoff, and Gabriele Klug. A mixed incoherent feed-forward loop contributes to the regulation of bacterial photosynthesis genes. *RNA biology*, 10(3):347352, March 2013.

[236] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[237] Daniel Marbach, James C Costello, Robert Kffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, DREAM5 Consortium, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, August 2012.

[238] Stephen Maren, K Luan Phan, and Israel Liberzon. The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nature reviews. Neuroscience*, 14(6):417428, June 2013.

[239] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7, 2006.

[240] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. T. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang,

Y. Wang, M. P. Weiner, P. G. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376380, September 2005. WOS:000231849100045.

[241] Bruz Marzolf, Eric W Deutsch, Patrick Moss, David Campbell, Michael H Johnson, and Timothy Galitski. SBEAMS-microarray: database software supporting genomic expression analyses for systems biology. *BMC bioinformatics*, 7:286, 2006.

[242] H. H. McAdams, B. Srinivasan, and A. P. Arkin. The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics*, 5(3):169178, March 2004. WOS:000189334500012.

[243] Kathleen E McGinness, Tania A Baker, and Robert T Sauer. Engineering controllable protein degradation. *Molecular cell*, 22(5):701707, June 2006.

[244] Simon McGregor, Vera Vasas, Phil Husbands, and Chrisantha Fernando. Evolution of associative learning in chemical networks. *PLoS computational biology*, 8(11):e1002739, 2012.

[245] Jerome T Mettetal, Dale Muzzey, Carlos Gmez-Uribe, and Alexander van Oudenaarden. The frequency dependence of osmo-adaptation in saccharomyces cerevisiae. *Science (New York, N.Y.)*, 319(5862):482484, January 2008.

[246] Tom Michoel, Riet De Smet, Anagha Joshi, Yves Van de Peer, and Kathleen Marchal. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst Biol*, 3:49, 2009.

[247] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, March 2004. WOS:000220000100049.

[248] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824827, October 2002. WOS:000178791200058.

[249] Amir Mitchell, Gal H. Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan, and Yitzhak Pilpel. Adaptive prediction of environmental changes by microorganisms. *Nature*, 460(7252):220U80, July 2009. WOS:000267761000033.

[250] Binny M Mony, Paula MacGregor, Alasdair Ivens, Federico Rojas, Andrew Cowton, Julie Young, David Horn, and Keith Matthews. Genome-wide dissection of the quorum sensing signalling pathway in trypanosoma brucei. *Nature*, 505(7485):681685, January 2014.

[251] Fantine Mordelet and Jean-Philippe Vert. SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76i82, August 2008.

[252] Samadhi Moreno-Campuzano, Sarath Chandra Janga, and Ernesto Perez-Rueda. Identification and analysis of DNA-binding transcription factors in bacillus subtilis and other firmicutes - a genomic approach. *Bmc Genomics*, 7:147, June 2006. WOS:000239344500001.

[253] Dale Muzzey, Carlos A Gmez-Uribe, Jerome T Mettetal, and Alexander van Oudenaarden. A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell*, 138(1):160171, July 2009.

[254] T Nakagaki, H Yamada, and T Ueda. Interaction between cell shape and contraction pattern in the physarum plasmodium. *Biophysical chemistry*, 84(3):195204, May 2000.

[255] Kazu Nakazawa, Michael C Quirk, Raymond A Chitwood, Masahiko Watanabe, Mark F Yeckel, Linus D Sun, Akira Kato, Candice A Carr, Daniel Johnston, Matthew A Wilson, and Susumu Tonegawa. Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science (New York, N.Y.)*, 297(5579):211218, July 2002.

[256] Frederick C. Neidhardt. *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, Washington, D.C, 2 edition edition, May 1996.

[257] Ulric Neisser, Gwyneth Boodoo, Thomas J. Bouchard Jr., A. Wade, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, Robert Perloff, Robert J. Sternberg, and Susana Urbina. Intelligence: Knowns and unknowns. *American Psychologist*, 51(2):77101, 1996.

[258] K. E. Nelson, R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, L. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399(6734):323329, May 1999. WOS:000080547800050.

[259] Robert J Nichols, Saunak Sen, Yoe Jin Choo, Pedro Beltrao, Matylda Zietek, Rachna Chaba, Sueyoung Lee, Krystyna M Kazmierczak, Karis J Lee, Angela Wong, Michael Shales, Susan Lovett, Malcolm E Winkler, Nevan J Krogan, Athanasios Typas, and Carol A Gross. Phenotypic landscape of a bacterial cell. *Cell*, 144(1):143–156, January 2011.

[260] Kenneth W Nickerson, Audrey L Atkin, and Jacob M Hornby. Quorum sensing in dimorphic fungi: farnesol and beyond. *Applied and environmental microbiology*, 72(6):38053813, June 2006.

[261] Philippe Noirot and Marie-Franoise Noirot-Gros. Protein interaction networks in bacteria. *Current opinion in microbiology*, 7(5):505512, October 2004.

[262] Pavel S Novichkov, Thomas S Brettin, Elena S Novichkova, Paramvir S Dehal, Adam P Arkin, Inna Dubchak, and Dmitry A Rodionov. RegPrecise web services interface: programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics. *Nucleic acids research*, 40(Web Server issue):W604–608, July 2012.

[263] Pavel S Novichkov, Dmitry A Rodionov, Elena D Stavrovskaya, Elena S Novichkova, Alexey E Kazakov, Mikhail S Gelfand, Adam P Arkin, Andrey A Mironov, and Inna Dubchak. RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic acids research*, 38(Web Server issue):W299–307, July 2010.

[264] You-Kwan Oh, Bernhard O Palsson, Sung M Park, Christophe H Schilling, and Radhakrishnan Mahadevan. Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *The Journal of biological chemistry*, 282(39):2879128799, September 2007.

[265] A V Oleskin. [biosocial phenomena in unicellular organisms (exemplified by data concerning prokaryota)]. *Zhurnal obshche biologii*, 70(3):225238, June 2009.

[266] Maureen A O'Malley and John Dupr. Towards a philosophy of microbiology. *Studies in history and philosophy of biological and biomedical sciences*, 38(4):775779, December 2007.

[267] Aharon Oren. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems*, 4:2, April 2008.

[268] Aharon Oren. Industrial and environmental applications of halophilic microorganisms. *Environmental technology*, 31(8-9):825–834, August 2010.

[269] H. A. Orr. Theories of adaptation: what they do and don't say. *Genetica*, 123(1-2):313, February 2005. WOS:000227550900002.

[270] Yan Ouyang, Carol R. Andersson, Takao Kondo, Susan S. Golden, and Carl Hirschie Johnson. Resonating circadian clocks enhance fitness in cyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):86608664, July 1998. BCI:BCI199800360930.

[271] Maureen A. OMalley. Philosophy and the microbe: a balancing act. *Biology & Philosophy*, 28(2):153159, March 2013.

[272] Maureen A. OMalley and John Dupr. Size doesnt matter: towards a more inclusive philosophy of biology. *Biology & Philosophy*, 22(2):155191, March 2007.

[273] D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7):3807–3812, March 1999. WOS:000079507900084.

[274] H Parkinson, M Kapushesky, M Shojatalab, N Abeygunawardena, R Coulson, A Farne, E Holloway, N Kolesnykov, P Lilja, M Lukk, R Mani, T Rayner, A Sharma, E William, U Sarkans, and A Brazma. ArrayExpressa public database of microarray experiments and gene expression profiles. *Nucleic acids research*, 35(Database issue):D747750, January 2007.

[275] P R Patnaik. Are microbes intelligent beings?: An assessment of cybernetic modeling. *Biotechnology advances*, 18(4):267288, July 2000.

[276] Ivan Petrovich Pavlov. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, 1927.

[277] J. Piette, H. Nyunoya, C. J. Lusty, R. Cunin, G. Weyens, M. Crabeel, D. Charlier, N. Glansdorff, and A. Pirard. DNA sequence of the carA gene and the control region of carAB: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoyl-phosphate synthetase in escherichia coli k-12. *Proc Natl Acad Sci U S A*, 81(13):41344138, July 1984.

[278] Morgan N. Price, Adam M. Deutschbauer, Jeffrey M. Skerker, Kelly M. Wetmore, Troy Ruths, Jordan S. Mar, Jennifer V. Kuehl, Wenjun Shao, and Adam P. Arkin. Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol Syst Biol*, 9:660, 2013.

[279] Morgan N. Price, Katherine H. Huang, Eric J. Alm, and Adam P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, 33(3):880892, 2005.

[280] Priscilla E M Purnick and Ron Weiss. The second wave of synthetic biology: from modules to systems. *Nature reviews. Molecular cell biology*, 10(6):410422, June 2009.

[281] Mark J Quinton-Tulloch, Frank J Bruggeman, Jacky L Snoep, and Hans V Westerhoff. Trade-off of dynamic fragility but not of robustness in metabolic pathways in silico. *The FEBS journal*, 280(1):160173, January 2013.

[282] R. J. Ram, N. C. VerBerkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, M. Shah, R. L. Hettich, and J. F. Banfield. Community proteomics of a natural microbial biofilm. *Science*, 308(5730):1915–1920, June 2005. WOS:000230120000042.

[283] Lennart Randau. RNA processing in the minimal organism nanoarchaeum equitans. *Genome Biology*, 13(7):R63, July 2012.

[284] David Reiss, Nitin Baliga, and Richard Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7(1):280, 2006.

[285] David J. Reiss, Marc T. Facciotti, and Nitin S. Baliga. Model-based deconvolution of genome-wide DNA binding. *Bioinformatics*, 24(3):396403, February 2008.

[286] D A Relman and S Falkow. The meaning and impact of the human genome sequence for microbiology. *Trends in microbiology*, 9(5):206208, May 2001.

[287] Vanessa K Ridaura, Jeremiah J Faith, Federico E Rey, Jiye Cheng, Alexis E Duncan, Andrew L Kau, Nicholas W Griffin, Vincent Lombard, Bernard Henrissat, James R Bain, Michael J Muehlbauer, Olga Ilkayeva, Clay F Semenkovich, Katsuhiko Funai, David K Hayashi, Barbara J Lyle, Margaret C Martini, Luke K Ursell, Jose C Clemente, William Van Treuren, William A Walters, Rob Knight, Christopher B Newgard, Andrew C Heath, and Jeffrey I Gordon. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science (New York, N.Y.)*, 341(6150):1241214, September 2013.

[288] Forest Rohwer and Rebecca Vega Thurber. Viruses manipulate the marine environment. *Nature*, 459(7244):207–212, May 2009. WOS:000266036100031.

[289] Edmund T. Rolls and Alessandro Treves. *Neural Networks and Brain Function*. Oxford University Press, Oxford ; New York, 1 edition edition, January 1998.

[290] S. M. Rosenberg. Evolving responsively: Adaptive mutation. *Nature Reviews Genetics*, 2(7):504515, July 2001. WOS:000169681600011.

[291] Simon Rumpel, Joseph LeDoux, Anthony Zador, and Roberto Malinow. Postsynaptic receptor trafficking underlying a form of associative learning. *Science (New York, N.Y.)*, 308(5718):8388, April 2005.

[292] S. L. Rutherford. From genotype to phenotype: buffering mechanisms and the storage of genetic information. *Bioessays*, 22(12):10951105, December 2000. WOS:000165552500007.

[293] H Salgado, S Gama-Castro, M Peralta-Gil, E Diaz-Peredo, F Sanchez-Solano, A Santos-Zavaleta, I Martinez-Flores, V Jimenez-Jacinto, C Bonavides-Martinez, J Segura-Salazar, A Martinez-Antonio, and J Collado-Vides. Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue):D394–7, 2006.

[294] H Salgado, G Moreno-Hagelsieb, T F Smith, and J Collado-Vides. Operons in escherichia coli: genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):66526657, June 2000.

[295] Gabino Sanchez-Perez, Alex Mira, Gabor Nyiro, Lejla Pasic, and Francisco Rodriguez-Valera. Adapting to environmental changes using specialized paralogs. *Trends in Genetics*, 24(4):154–158, April 2008. WOS:000255346300002.

[296] Terence D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2(6):459473, 1989.

[297] Bettina E Schirrmeister, Alexandre Antonelli, and Homayoun C Bagheri. The origin of multicellularity in cyanobacteria. *BMC evolutionary biology*, 11:45, 2011.

[298] Amy K Schmid, Min Pan, Kriti Sharma, and Nitin S Baliga. Two transcription factors are necessary for iron homeostasis in a salt-dwelling archaeon. *Nucleic acids research*, 39(7):2519–2533, April 2011.

[299] Amy K Schmid, David J Reiss, Amardeep Kaur, Min Pan, Nichole King, Phu T Van, Laura Hohmann, Daniel B Martin, and Nitin S Baliga. The anatomy of microbial cell state transitions in response to oxygen. *Genome research*, 17(10):13991413, October 2007.

[300] Amy K. Schmid, David J. Reiss, Min Pan, Tie Koide, and Nitin S. Baliga. A single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability. *Molecular Systems Biology*, 5:282, 2009.

[301] S. J. Schrag, V. Perrot, and B. R. Levin. Adaptation to the fitness costs of antibiotic resistance in escherichia coli. *Proceedings of the Royal Society B-Biological Sciences*, 264(1386):1287–1291, September 1997. WOS:A1997XY66000005.

[302] Daniel Schultz, Peter G. Wolynes, Eshel Ben Jacob, and Jose N. Onuchic. Deciding fate in adverse times: Sporulation and competence in bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):2102721034, December 2009. WOS:000272795300006.

[303] Wolfgang Schumann. The bacillus subtilis heat shock stimulon. *Cell Stress & Chaperones*, 8(3):207217, July 2003.

[304] Benno Schwikowski, Peter Uetz, and Stanley Fields. A network of proteinprotein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, December 2000.

[305] E Segal, M Shapira, A Regev, D Pe'er, D Botstein, D Koller, and N Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, 2003.

[306] E Segal, R Yelensky, and D Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(Suppl 1):273–282, 2003. Evaluation Studies.

[307] Giovanni Seni, John Elder, Robert Grossman, and & 0 more. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan and Claypool Publishers, San Rafael, Calif., February 2010.

[308] M Angeles Serrano, Marin Bogu, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci U S A*, 106(16):64836488, April 2009.

[309] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003. 22959694 1088-9051 Journal Article.

[310] Paul T Shannon, David J Reiss, Richard Bonneau, and Nitin S Baliga. The gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC bioinformatics*, 7:176, 2006.

[311] Dilara I Sharif, John Gallon, Chris J Smith, and Ed Dudley. Quorum sensing in cyanobacteria: N-octanoyl-homoserine lactone release and response, by the epilithic colonial cyanobacterium gloeothece PCC6909. *The ISME journal*, 2(12):11711182, December 2008.

[312] Emi Shudo, Patsy Haccou, and Yoh Iwasa. Optimal choice between feedforward and feedback control in gene expression to cope with unpredictable danger. *Journal of theoretical biology*, 223(2):149160, July 2003.

[313] Mark L. Siegal, Daniel E. L. Promislow, and Aviv Bergman. Functional and evolutionary inference in gene networks: does topology matter? *Genetica*, 129(1):83–103, January 2007. WOS:000242816200008.

[314] JM Smith, NH Smith, M. Orourke, and BG Spratt. How clonal are bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 90(10):4384–4388, May 1993. WOS:A1993LC72000014.

[315] L M Smith, J Z Sanders, R J Kaiser, P Hughes, C Dodd, C R Connell, C Heiner, S B Kent, and L E Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, June 1986.

[316] Stephen L Smith and Jon Timmis. An immune network inspired evolutionary algorithm for the diagnosis of parkinson's disease. *Bio Systems*, 94(1-2):3446, November 2008.

[317] Jacky L Snoep, Coen C van der Weijden, Heidi W Andersen, Hans V Westerhoff, and Peter Ruhdal Jensen. DNA supercoiling in escherichia coli is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase i and DNA gyrase. *European journal of biochemistry / FEBS*, 269(6):16621669, March 2002.

[318] H Soest. Dressuryersuche mit ciliaten und rhabdocoelen turbellarien. *Zeitschrift fur Vergleichende Physiologic*, 24:720748, 1937.

[319] J. Soppa, A. Baumann, M. Brenneis, M. Dambeck, O. Hering, and C. Lange. Genomics and functional genomics with haloarchaea. *Archives of Microbiology*, 190(3):197–215, September 2008. WOS:000258527100002.

[320] Matan Sorek, Nathalie Q Balaban, and Yonatan Loewenstein. Stochasticity, bistability and the wisdom of crowds: a model for associative learning in genetic regulatory networks. *PLoS computational biology*, 9(8):e1003179, 2013.

[321] JC Spain and PA Vanveld. Adaptation of natural microbial communities to degradation of xenobiotic compounds - effects of concentration, exposure time, inoculum, and chemical-structure. *Applied and Environmental Microbiology*, 45(2):428435, 1983. WOS:A1983QB28500012.

[322] Franois St-Pierre and Drew Endy. Determination of cell fate selection during phage lambda infection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):2070520710, December 2008.

[323] Achim Stephan. Varieties of emergence. 1999.

[324] Achim Stephan. The dual role of emergence in the philosophy of mind and in cognitive science. *Synthese*, 151(3):485498, August 2006.

[325] C K Stover, X Q Pham, A L Erwin, S D Mizoguchi, P Warrener, M J Hickey, F S Brinkman, W O Hufnagle, D J Kowalik, M Lagrou, R L Garber, L Goltry, E Tolentino, S Westbrock-Wadman, Y Yuan, L L Brody, S N Coulter, K R Folger, A Kas, K Larbig, R Lim, K Smith, D Spencer, G K Wong, Z Wu, I T Paulsen, J Reizer, M H Saier, R E Hancock, S Lory, and M V Olson. Complete genome sequence of pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406(6799):959964, August 2000.

[326] W. R. Streit and R. A. Schmitz. Metagenomics - the key to the uncultured microbes. *Current Opinion in Microbiology*, 7(5):492498, October 2004. WOS:000224575700009.

[327] Jesse Stricker, Scott Cookson, Matthew R Bennett, William H Mather, Lev S Tsimring, and Jeff Hasty. A fast, robust and tunable synthetic gene oscillator. *Nature*, 456(7221):516519, November 2008.

[328] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561–568, January 2011.

[329] Ilias Tagkopoulos, Yir-Chung Liu, and Saeed Tavazoie. Predictive behavior within microbial genetic networks. *Science*, 320(5881):13131317, June 2008. WOS:000256441100037.

[330] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533538, July 2010.

[331] E. L. Tatum and Joshua Lederberg. Gene recombination in the bacterium escherichia coli. *Journal of Bacteriology*, 53(6):673–684, June 1947.

[332] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):3336, January 2000. WOS:000084896300009.

[333] S. A. Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nature Genetics*, 36(5):492–496, May 2004. WOS:000221183000022.

[334] B H ter Kuile and H V Westerhoff. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS letters*, 500(3):169171, July 2001.

[335] EL Thorndike. Animal intelligence: An experimental study of the associate processes in animals. *American Psychologist*, 53(10):11251127, 1998.

[336] Brian Tjaden, Rini Mukherjee Saxena, Sergey Stolyar, David R. Haynor, Eugene Kolker, and Carsten Rosenow. Transcriptome analysis of escherichia coli using highdensity oligonucleotide probe arrays. *Nucleic Acids Research*, 30(17):37323738, September 2002.

[337] I Torrecilla, F Legans, I Bonilla, and F Fernndez-Pias. A calcium signal is involved in heterocyst differentiation in the cyanobacterium anabaena sp. PCC7120. *Microbiology (Reading, England)*, 150(Pt 11):37313739, November 2004.

[338] M. Travisano and R. E. Lenski. Long-term experimental evolution in escherichia coli .4. targets of selection and the specificity of adaptation. *Genetics*, 143(1):1526, May 1996. WOS:A1996UH28300003.

[339] Pier-Luc Tremblay, Zarath M Summers, Richard H Glaven, Kelly P Nevin, Karsten Zengler, Christian L Barrett, Yu Qiu, Bernhard O Palsson, and Derek R Lovley. A c-type cytochrome and a transcriptional regulator responsible for enhanced extracellular electron transfer in geobacter sulfurreducens revealed by adaptive evolution. *Environmental microbiology*, 13(1):13–23, January 2011.

[340] Anthony Trewavas. Mindless mastery. *Nature*, 415(6874):841, February 2002.

[341] Brian B. Tuch, David J. Galgoczy, Aaron D. Hernday, Hao Li, and Alexander D. Johnson. The evolution of combinatorial gene regulation in fungi. *Plos Biology*, 6(2):352364, February 2008. WOS:000254928400022.

[342] Alan M Turing. Computing machinery and intelligence. *Mind*, page 433460, 1950.

[343] Serdar Turkarslan, David J Reiss, Goodwin Gibbins, Wan Lin Su, Min Pan, J Christopher Bare, Christopher L Plaisier, and Nitin S Baliga. Niche adaptation by expansion and reprogramming of general transcription factors. *Molecular systems biology*, 7:554, 2011.

[344] Stijn van Dongen and Cei Abreu-Goodger. Using MCL to extract clusters from networks. *Methods in molecular biology (Clifton, N.J.)*, 804:281295, 2012.

[345] Wally C van Heeswijk, Hans V Westerhoff, and Fred C Boogerd. Nitrogen assimilation in escherichia coli: putting molecular data into a systems perspective. *Microbiology and molecular biology reviews: MMBR*, 77(4):628695, December 2013.

[346] J. van Helden, B. Andr, and J. Collado-Vides. A web site for the computational analysis of yeast regulatory sequences. *Yeast*, 16(2):177187, January 2000.

[347] Jan-Willem Veening, Wiep Klaas Smits, and Oscar P Kuipers. Bistability, epigenetics, and bet-hedging in bacteria. *Annual review of microbiology*, 62:193210, 2008.

[348] Jan-Willem Veening, Wiep Klaas Smits, and Oscar P. Kuipers. Bistability, epigenetics, and bet-hedging in bacteria. In *Annual Review of Microbiology*, volume 62, page 193210. 2008. WOS:000259968000012.

[349] Gregory J Velicer. Social strife in the microbial world. *Trends in microbiology*, 11(7):330337, July 2003.

[350] C. Vieille, K. L. Epting, R. M. Kelly, and J. G. Zeikus. Bivalent cations and amino-acid composition contribute to the thermostability of bacillus licheniformis xylose isomerase. *European Journal of Biochemistry*, 268(23):62916301, December 2001. WOS:000172540800034.

[351] A. Wagner. Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. *Biological Cybernetics*, 74(6):557–567, June 1996. WOS:A1996UQ95400009.

[352] Andreas Wagner. *Robustness and Evolvability in Living Systems:*. Princeton University Press, Princeton, N.J.; Woodstock, 1 edition edition, July 2007.

[353] G. P. Wagner and L. Altenberg. Perspective: Complex adaptations and the evolution of evolvability. *Evolution*, 50(3):967–976, June 1996. WOS:A1996UY55500001.

[354] Harris H. Wang, Farren J. Isaacs, Peter A. Carr, Zachary Z. Sun, George Xu, Craig R. Forest, and George M. Church. Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, 460(7257):894–U133, August 2009. WOS:000268938300039.

[355] Ilan Wapinski, Jenna Pfiffner, Courtney French, Amanda Socha, Dawn Anne Thompson, and Aviv Regev. Gene duplication and the evolution of ribosomal protein gene regulation in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 107(12):5505–5510, March 2010. WOS:000275898300044.

[356] Christopher M Waters and Bonnie L Bassler. Quorum sensing: cell-to-cell communication in bacteria. *Annual review of cell and developmental biology*, 21:319346, 2005.

[357] J D WATSON and F H CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[358] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440442, June 1998.

[359] Grzegorz Wegrzyn and Alicja Wegrzyn. Genetic switches during bacteriophage lambda development. *Progress in nucleic acid research and molecular biology*, 79:148, 2005.

[360] Jevin D. West, Theodore C. Bergstrom, and Carl T. Bergstrom. The eigenfactor MetricsTM: A network approach to assessing scholarly journals. *College & Research Libraries*, 71(3):236244, May 2010.

[361] Hans V. Westerhoff. *Thermodynamics and Control of Biological Free-Energy Transduction*. Elsevier Science Ltd, Amsterdam ; New York : New York, NY, USA, July 1987.

[362] Hans V Westerhoff. Signalling control strength. *Journal of theoretical biology*, 252(3):555567, June 2008.

[363] Hans V Westerhoff, Miguel A Aon, Karel van Dam, Sonia Cortassa, Daniel Kahn, and Marielle van Workum. Dynamical and hierarchical coupling. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1018(2):142146, 1990.

[364] Hans V Westerhoff, Alexey Kolodkin, Riaan Conradie, Stephen J Wilkinson, Frank J Bruggeman, Klaas Krab, Jan H van Schuppen, Hanna Hardin, Barbara M Bakker, Martijn J Mon, Katja N Rybakova, Marco Eijken, Hans J P van Leeuwen, and Jacky L Snoep. Systems biology towards life in silico: mathematics of the control of living cells. *Journal of mathematical biology*, 58(1-2):734, January 2009.

[365] W. Brian Whitaker, Michelle A. Parent, Lynn M. Naughton, Gary P. Richards, Seth L. Blumerman, and E. Fidelma Boyd. Modulation of responses of vibrio parahaemolyticus o3:k6 to pH and temperature stresses by growth at different salt concentrations. *Applied and Environmental Microbiology*, 76(14):4720–4729, July 2010. WOS:000279611500016.

[366] Kenia Whitehead, Adrienne Kish, Min Pan, Amardeep Kaur, David J Reiss, Nichole King, Laura Hohmann, Jocelyne DiRuggiero, and Nitin S Baliga. An integrated systems approach for understanding cellular responses to gamma radiation. *Molecular systems biology*, 2:47, 2006.

[367] Kenia Whitehead, Min Pan, Ken-ichi Masumura, Richard Bonneau, and Nitin S. Baliga. Diurnally entrained anticipatory behavior in archaea. *Plos One*, 4(5):e5485, May 2009. WOS:000265933800015.

[368] M. A. Woelfle, O. Y. Yan, K. Phanvijhitsiri, and C. H. Johnson. The adaptive value of circadian clocks: An experimental assessment in cyanobacteria. *Current Biology*, 14(16):1481–1486, August 2004. WOS:000223586900026.

[369] C. R. Woese, G. J. Olsen, M. Ibba, and D. Soll. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, 64(1):202+, March 2000. WOS:000085790100010.

[370] Denise M. Wolf, Lisa Fontaine-Bodin, Ilka Bischofs, Gavin Price, Jay Keasling, and Adam P. Arkin. Memory in microbes: Quantifying history-dependent behavior in a bacterium. *Plos One*, 3(2):e1700, February 2008. WOS:000260586500044.

[371] Arno Wouters. Viability explanation. *Biology and Philosophy*, 10(4):435457, 1995.

[372] Arno Wouters. Four notions of biological function. *Studies in History and Philosophy of Science Part C*, 34(4):633668, 2003.

[373] Arno G. Wouters. Design explanation: determining the constraints on what can be alive. *Erkenntnis*, 67(1):6580, July 2007.

[374] Arno G. Wouters. Biologys functional perspective: Roles, advantages and organization. In Kostas Kampourakis, editor, *The Philosophy of Biology*, number 1 in History, Philosophy and Theory of the Life Sciences, page 455486. Springer Netherlands, January 2013.

[375] Arno Gerhard Wouters. *Explanation without a cause*. Zeno, The Leiden-Utrecht Research Institute of Philosophy, 1999.

[376] Yangle Wu, Xiaomeng Zhang, Jianglei Yu, and Qi Ouyang. Identification of a topological characteristic responsible for the biological robustness of regulatory networks. *Plos Computational Biology*, 5(7):e1000442, July 2009. WOS:000269220100008.

[377] Elisabeth J. Wurtmann, Alexander V. Ratushny, Min Pan, Karlyn D. Beer, John D. Aitchison, and Nitin S. Baliga. An evolutionarily conserved RNase-based mechanism for repression of transcriptional positive autoregulation. *Molecular Microbiology*, 92(2):369382, April 2014.

[378] Joao B. Xavier. Social interaction in synthetic and natural microbial communities. *Molecular Systems Biology*, 7(1), January 2011.

[379] T M Yi, Y Huang, M I Simon, and J Doyle. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9):46494653, April 2000.

[380] Sung Ho Yoon, David J. Reiss, J Christopher Bare, Dan Tenenbaum, Min Pan, Joseph Slagel, Robert L. Moritz, Sujung Lim, Murray Hackett, Angeli Lal Menon, Michael W W. Adams, Adam Barnebey, Steven M. Yannone, John A. Leigh, and Nitin S. Baliga. Parallel evolution of transcriptome architecture during genome reorganization. *Genome Res*, 21(11):18921904, November 2011.

[381] Sung Ho Yoon, Serdar Turkarslan, David J. Reiss, Min Pan, June A. Burn, Kyle C. Costa, Thomas J. Lie, Joseph Slagel, Robert L. Moritz, Murray Hackett, John A. Leigh, and Nitin S. Baliga. A systems level predictive model for global gene regulation of methanogenesis in a hydrogenotrophic methanogen. *Genome Res*, 23(11):18391851, November 2013.

[382] Haiyuan Yu and Mark Gerstein. Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(40):1472414731, October 2006. WOS:000241069300017.

[383] Alexander Zaborin, Kathleen Romanowski, Svetlana Gerdes, Christopher Holbrook, Francois Lepine, Jason Long, Valeriy Poroyko, Stephen P. Diggle, Andreas Wilke, Karima Righetti, Irina Morozova, Trissa Babrowski, Donald C. Liu, Olga Zaborina, and John C. Alverdy. Red death in caenorhabditis elegans caused by pseudomonas aeruginosa PAO1. *Proceedings of the National Academy of Sciences of the United States of America*, 106(15):63276332, April 2009.

[384] Chunmei Zhai, Ping Zhang, Fei Shen, Changxin Zhou, and Changhong Liu. Does microcystis aeruginosa have quorum sensing? *FEMS Microbiology Letters*, 336(1):3844, November 2012.

[385] Guishan Zhang, Fan Zhang, Gang Ding, Jie Li, Xiaopeng Guo, Jinxing Zhu, Liguang Zhou, Shichun Cai, Xiaoli Liu, Yuanming Luo, Guifeng Zhang, Wenyuan Shi, and Xiuzhu Dong. Acyl homoserine lactone-based quorum sensing in a methanogenic archaeon. *The ISME journal*, 6(7):13361344, July 2012.

[386] Jindan Zhou and Kenneth E. Rudd. EcoGene 3.0. *Nucleic Acids Res*, 41(Database issue):D613D624, January 2013.

[387] Ilana Zilber-Rosenberg and Eugene Rosenberg. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS microbiology reviews*, 32(5):723735, August 2008.

[388] E. R. Zinser and R. Kolter. Escherichia coli evolution during stationary phase. *Research in Microbiology*, 155(5):328336, June 2004. WOS:000222736200005.

[389] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005.

Appendix A

MACROMOLECULAR NETWORKS AND INTELLIGENCE IN MICROORGANISMS

Appendix A has been published:

Westerhoff H*, Brooks AN*, Simeonidis E*, Garca-Contreras R*, Boogerd F, He F, Jackson VJ, Goncharuk V, Kolodkin A. (2014) Macromolecular networks and intelligence in microorganisms. *Front.Microbiol.* 5:379.

* Indicates equal contribution

A.1 Abstract

Living organisms persist by virtue of complex interactions among many components organized into dynamic, environment-responsive networks that span multiple scales and dimensions. Biological networks constitute a type of Information and Communication Technology (ICT): they receive information from the outside and inside of cells, integrate and interpret this information, and then activate a response. Biological networks enable molecules within cells, and even cells themselves, to communicate with each other and their environment. We have become accustomed to associating brain activity particularly activity of the human brain with a phenomenon we call "intelligence". Yet, four billion years of evolution could have selected networks with topologies and dynamics that confer traits analogous to this intelligence, even though they were outside the intercellular networks of the brain. Here, we explore how macromolecular networks in microbes confer intelligent characteristics, such as memory, anticipation, adaptation and reflection and we review current understanding of how network organization reflects the type of intelligence required for the environments in

which they were selected. We propose that, if we were to leave terms such as "human" and "brain" out of the definition of "intelligence", all forms of life from microbes to humans exhibit some or all characteristics consistent with "intelligence". We then review advances in genome-wide data production and analysis, especially in microbes, that provide a lens into microbial intelligence and propose how the insights derived from quantitatively characterizing biomolecular networks may enable synthetic biologists to create intelligent molecular networks for biotechnology, possibly generating new forms of intelligence, first *in silico* and then in vivo.

A.2 Introduction

For centuries, mankind has grappled with the precise nature and defining features of intelligence. Debates have erupted over how to define and measure the extent of intelligence in parts of the biological (and non-biological) world. Alan Turing, for example, famously proposed a test for evaluating the performance of "artificial intelligence": namely, can it be distinguished from the performance of human beings by another human [342]? There have also long been philosophical discussions on what can be considered "intelligent". A number of studies have explored whether there are differences in intelligence between human populations [257], whether animals [335] and even plants [340] exhibit intelligent behaviors, whether non-human artificial systems are capable of intelligence [61] and, more recently, whether intelligence spans biological domains including even the simplest of microbes [150, 62, 156, 38]. For the purposes of this discussion, however, and in the interest of brevity, we limit ourselves to systems of biological nature.

As an abstract concept, "intelligence" escapes easy definition. As a linguistic construct, its characteristics have varied substantially across philosophical and cultural contexts. Here, we do not attempt a definition of intelligence; rather, we discuss how some features (like decision making) commonly associated with a brain can also be found in the microbial world. Rather than launch an ontological, epistemological, or semantic inquiry, we instead focus on the scientific utility of assigning intelligence to microbes. We review how the mathemat-

ical perspectives of complex adaptive systems and recent data-intensive developments in systems biology offer insight and help structure this problem. Finally, we consider whether viewing microbes through the lens of "intelligence" can help us better describe their behavior, harness their intelligence to perform valuable actions and, in the end, possibly extend our understanding of the systems biology underlying the functions of the human brain.

A.2.1 What is "intelligence"?

The modern biological perspective on "intelligence", even at its most fundamental level, tends to associate it with the human brain. In this context, "intelligence" is a property of the human brain, or a feature that somehow emerges from its activity. Accepting that intelligence may not be exclusively a feature of the human brain, but rather it may be present at least to a degree in all creatures possessing brains or nervous systems, already helps refine the general features of intelligence. However, intelligence may not have to be associated solely with a certain biological organ, such as a brain or a nervous system. Brains and nervous systems may be highly adapted conduits for expressing and integrating multiple intelligent behaviors. Some of these behaviors may be exhibited by other complex adaptive systems present in living organisms that do not have a brain or nervous system. As early as 1995, Hellingwerf suggested that some two-component systems in bacteria comply with the requirements for elements of a neural network [150]. More recently, the so-called biogenic approach of cognition has gained momentum by focusing on the biological origin of cognition and intelligence, abandoning a strict anthropocentric perspective [209, 221]. This is the central paradigm around which we base our analysis.

A.2.2 How does intelligence emerge?

A small molecule at room temperature cannot be intelligent; it cannot store information about its past with implications for its behavior in some future. Large macromolecules, such as proteins and polynucleotides, may store information as, for example, Gibbs free energy in metastable states, where interactions between their structural components can differ

depending on the way they were folded some time ago. The primary difference between small and large molecules with respect to information storage is that small molecules have a sufficiently small number of structural microstates (i.e. conformations) such that all of these states are visited by the molecule on time scales relevant for biochemistry (10 milliseconds), i.e. they are ergodic [361]. However, large molecules may not visit all of their microstates, even on equivalent or greater time scales. In principle, phosphorylation, dephosphorylation and other chemical modifications may increase the possible number of microstates [178]. High energy nucleic acid and protein complex states called chromatin, for example, may take hours, if not days, to relax after refolding.

Information storage within an object requires that the object be away from its equilibrium state for a sufficient period of time. This can be achieved transiently by bringing the object into a high free energy state, with the relaxation back to the equilibrium state being slow. Or, it may be achieved permanently by making this process permanent (at the cost of Gibbs free energy), such as in the terminal phosphoryl bond in ATP. More generally, in open systems, Gibbs free energy harvested from the environment can be used to maintain the non-equilibrium state. Such free-energy transductions require nonlinear interactions of multiple components: they require complexity [361] and so does intelligence.

Vis--vis memory, intelligence is an emergent property of a complex system; a feature that is not reducible to the parts of the system in isolation. Intelligence emerges when a systems components interact. For example, the intelligence (or intelligent-like behavior) we observe inside a single cell emerges from interactions among thousands of non-intelligent macromolecules. Similarly, the intelligent behavior of a microbial society is not simply the sum of the behavior of intelligent cells; rather, it is a property that emerges from the interactions amongst many of them. In the human brain, intelligence emerges from interactions of nearly 90 billion neurons.

While, in practice, it is not trivial (or yet possible) to specify the interactions leading to intelligence, a promising start would be to catalog all of the interacting components (molecules, microorganisms, neurons), thereby defining the topology of the interactions as

a network. Experimentally, this would correspond to performing ChIP-on-chip, yeast two-hybrid experiments or antibody pull-down experiments. However, as we will show, this does not suffice to establish a basis for intelligence. It is not the mere existence of a network that begets intelligent behavior a rock can be full of networked structures in the form of bonds among its component molecules and ions, yet it is not intelligent. Rather, it is the dynamics of the interactions in a system that generate the system-level property we call intelligence. Somehow, nonlinearities in the interactions and their indirect and incomplete, yet nonzero, reciprocities are important.

Although we have discovered many of the components of living systems, e.g. neurons and their connectivity in the brain [10, 6] and macromolecules and their interactions in the cell, we still have no clear view on how they collectively contribute to intelligence. One reason for this failure is that the complete picture may be too complex to be perceived fully by our human brains. With computer simulation, however, it should be possible to reconstruct the emergence of these properties. Even then, it is debatable whether our brain, biased by its very human nature, will be able to identify and appreciate all forms of intelligence, especially those that are dissimilar to our own. Identifying unfamiliar forms of intelligence is the transcendental challenge of this paper one that would have enormous implications for synthetic biology and engineering. We start by describing features of microbial systems that are analogous to familiar forms of human intelligence.

A.3 Systems biology of intelligence: reconstructing the emergence of intelligence from component properties of the system

Systems biology can be defined as a science that aims to understand how biological function that is absent from macromolecules in isolation emerges when these macromolecules exist as components of a system [9, 364]. The concepts of System, Function and Emergence are central in this context.

The notion of function plays an important role in (systems) biology. Yet, often this concept is ill-defined. Because the word “function” has strong teleological connotations,

many biologists hasten to clarify that they invoke neither purpose nor intention when they use the notion of function. The subtle reasoning that accompanies these notions, however, is often overlooked [375, 217], not in the least because the term "function" is used in various ways. Here, we adopt the perspective of Wouters, who distinguished four principal kinds of biological function [375]. In short, he argues that the term "function" is used to refer to: (i) function as activity; (ii) function as role; (iii) function as advantage; and (iv) function as selected effect [372, 374]. Mahner and Bunge arrived independently to a similar set of functions [229]. Considering the cognitive functions that are discussed in this study (decision making, robust adaptation, association, anticipation, self-awareness and problem solving), the first three definitions are the most useful. The fourth definition is used in evolutionary biology and it features in historical evolutionary explanations. Defining "function" is important to understand the explanations of biological systems we craft. We need, for instance, to distinguish mechanistic explanations and design explanations. Mechanistic explanations categorize a system into a number of functional components; they describe how these components are arranged, how their activities are organized in time, and relate these features to some phenotype [52]. Mechanistic models are mathematical models related to the activities of cellular reaction networks involving transport, metabolism, signal transduction or gene expression. However, mechanisms only suffice to explain how the features are brought about (how they work). Understanding why certain mechanisms exist (rather than other, alternative organizations) requires design explanations [371, 373]. These explanations typically contrast observed organizations with conceivable alternatives in an attempt to identify invariances (or "laws") that can account for our observations. Delineating the difference between these two types of explanations relates to how we attribute function to systems (e.g. "function as an activity " versus "function as an advantage"). A human brain comprised by neurons, a microbial community comprised by different species and individual organisms or an individual cell comprised by molecules are all semi-open systems. They all selectively interact with their environments by way of mass and energy exchange, where the decrease of free energy in the environment is coupled to the increase of the order of

the biosystem itself (decreasing its own entropy), or with the maintenance of the biosystem against the activity of the many processes that tend to dissipate it [361]. Systems of artificial intelligence are semi-open as well. They all need an external energy source to maintain their existence. In other words, there is always a flow of mass and energy through the system, and then a certain function emerges. The function in which we are interested here is "intelligence". Intelligence consists of many features that allow a system to adapt to its environment. Together with other functions of the system, intelligence emerges from interactions among system components. As an emergent property, it satisfies three theses, as expounded by Stephan: (i) physical monism; (ii) synchronic determinism; and (iii) systemic (organizational) property [323]. The thesis of physical monism restricts the nature of the systems elements and states, so that the system consists of only physical entities and interactions, denying any supernatural influences this is how we describe our system ab initio: we neglect all supernatural influences *de juro*. The thesis of synchronic determinism restricts the way systemic properties and the systems microstructure are related to each other and states that there can be no difference in systemic properties without changes in the structure of the system or in the properties of the components: features of intelligence are underlined exactly by the changes in the system (firing between neurons, chemical reactions between molecules, electrical current between components of a computer); in other words, differences in systemic properties should be measurable at least in principle and, with the advent of genomics and the other -omics, also in practice. It is noteworthy that this thesis also implies that the inverse statement is invalid: a change in a systems microstructure or properties does not necessarily yield a change in its behavior or properties. The thesis of being a systemic (organizational) property means that a property is not exhibited by elements in isolation; interactions must keep the elements out of their non-informative equilibrium state.

If emergence is weak, it simply satisfies just the three theses stated above. According to Stephan [323, 324], strong emergence would satisfy one additional criterion irreducibility. In general, there are three conditions for irreducibility, but it has been argued that for

biochemical networks only one condition is relevant [51]: if the properties of parts (say A, B, and C) in their relationship (RABC) within the system as a whole (together constituting an explanation of the systemic property at hand) do not follow from the properties of parts (A,B,C) or simpler subsystems (AB,BC,AC) in isolation, it is a strongly emergent property. It should be noted that in this definition of strong emergence, the deduction base does not include systemic knowledge, such as the state of the system. Cognitive-like capabilities of a single microbial cell might then be irreducible in the sense that these properties cannot be deduced from the full knowledge of the behavior of the parts of the system in isolation or in configurations simpler than the one prevailing within the whole system. In fact, all features of microbial intelligence described in this study are expected to be irreducible in this sense, and therefore strongly emergent.

It is worthwhile to compare our notion of strong emergence with that from philosophy of mind. In philosophy of mind, mental properties like human intelligence are considered strongly emergent; contrary to our contention here, however, the underlying reason for this limitation is that the property does not follow from the behavior of the parts and their interactions within the system. By contrast, we assert that microbial intelligence, or in principle any systemic property, can be mechanistically explained if the properties and behaviors of the parts and their relationship within the system are fully known, i.e., when full knowledge of the state of the system is available. For this reason, any microbial property can, in principle, be mechanistically explained and, thus, can also be reconstructed in mathematical models of the underlying mechanism provided that knowledge of the system is fully available. Properties that are declared strongly emergent because of a limited deduction base are still calculable if the behavior of all relevant components and their mutual interactions within the system are available [51].

The limited deduction base of strong emergence provides the opportunity to rank emergent systemic properties according to the strength of emergence, which can be clarified as follows: in principle, every single component of the system, albeit indirectly, interacts with all other components. Let us consider an example of two abstract proteins A and B binding

to each other inside the cell. The binding reaction between proteins A and B might depend on the presence of other proteins. For example, transporters and structural proteins forming intracellular compartments keep proteins A and B together or separate. Other proteins (e.g. chaperones) might modulate the interaction directly by chemical modification of the interacting proteins. Binding between proteins A and B can also depend on environmental parameters, like intracellular pH. However, the pH is the result of proteins that regulate the uptake and pumping out of ions and different buffering molecules. In turn, ion transport processes are coupled to ATP hydrolysis and thus are dependent on the Gibbs free energy flux through the cell. Thus, the interaction between two components in the cell depends to a variable extent on the state of the whole system. In other words, system component properties are state dependent. The greater their state dependency is, the greater the degree of irreducibility of the system (non-deducibility), implying stronger emergence [194, 195].

The ability of a system to "choose the best option to solve a question and to anticipate the future" and, thus, to be intelligent might be state-dependent to a very high extent. Nevertheless, the intelligent response can be reconstructed in a computer model if we have complete knowledge of all interactions between system components. Similarly to other forms of emergent behavior, intelligent behavior is somehow predetermined by the system itself and by applied stimuli. Theoretically, with very precise mathematical description of all system components, all interactions among them and boundary conditions, the emergent intelligent behavior reconstructed in a computer model should become an accurate description of the behavior of a real system. Thus, a computer model can potentially predict (prevision) the manifestation of intelligence. But, in practice, we are not able to get the complete information necessary about a real system because, first of all, a real system is semi-open and some tiny event somewhere on the planet may have a" butterfly effect" on interactions among system components and break all model predictions. Secondly, even in a complete and very precise model, there might exist some bifurcation points where certain sets of the parameters might allow multiple solutions. So, in this sense, the intelligent response is not 100% predetermined, but not because of the "free will" of the system.

A description of how components interact with and affect each other can be represented as a network: metabolic networks, signal transduction networks, gene expression networks, anatomical networks, microbial ecological networks, etc. One can generate and model these networks using various approaches. For example, one can determine the kinetic rules of how network components interact and express the rates of these interactions in terms of mathematical relationships, e.g. differential equations. Then, one can integrate all equations and solve them for the whole system. As a result, one may be able to simulate the dynamic behavior of the network and, thus, reconstruct its emergent properties in silico. For example, the response of the nuclear receptor network to the cortisol signal has been modelled in a kinetic ODE-based model [196]. The intelligent properties of the physiological network emerged in the computer model; for instance, the modelled system was able to learn from previous stress and anticipate the next cortisol pulse.

The example above shows how intelligent behavior can emerge from just one feedback and one feedforward loop. In reality, the network can be much more complicated and contain many such loops. Biologically inspired "intelligence" models and algorithms have been extensively developed in the fields of artificial intelligence and optimization with many real-world applications, such as artificial immune systems [316], evolutionary algorithms, artificial neural networks [289] and the Kirdin kinetic machine [1]. For instance, feedback and feedforward loops, based on the architecture of neurons (including synapses and dendrites), are crucial for understanding the functional connectivity in the brain that is usually modelled by the artificial neural networks [289]. Neural networks are mainly classified into two groups, i.e. (i) the feedforward neural networks (FFNNs) where data is propagated from input to output using combinatorial machines, e.g. radial basis function (RBF), multilayer perceptron (MLP), self-organizing map (SOM); and (ii) the recurrent neural networks (RNNs). Several important feedforward loop motifs have been identified in both neuronal connectivity networks and transcriptional gene regulation networks [248], despite these networks operating on different spatial and temporal scales. This similarity in motifs may reflect a fundamental similarity in the evolved designs of both types of networks: to reject

transient input fluctuations/noises and activate output only if the input is persistent, a so-called persistence detector [13]. In addition, a multi-input feedforward structure is identified in the neuronal network of the nematode C. elegans, which serves as a so-called coincidence detector: the output is activated only if stimuli from two or more different inputs occur within a certain period of time [181, 13]. Another biological example appears in the retina, where a hierarchical feedforward cortical architecture is used for the pre-processing of visual information [296]. Although successful in practical applications, pure FFNNs are expected to be rare in the human neural system. On the other hand, recurrent neural networks have immediate biological application (i.e. self-organizing dynamic systems) and can describe complex nonlinear dynamics, including both feedforward and feedback structures. Nevertheless, very few real applications have been studied based on RNNs. Until recently, RNNs have been employed to study short-term memory and brain-like memory. This is because RNNs allow the output of a neuron to influence its input, either directly or indirectly, via its effect on other neurons. This allows the network to reflect the input presented to it, but also its own internal activity at any given time. In intracellular macromolecular network organization, a distinction has been made between dictatorial and democratic hierarchies, where only in the latter case the metabolite concentrations close to the systems output are able to influence gene expression [363, 317]. The two types of hierarchy may affect FFNNs and RNNs, respectively.

Learning and memory are two important, counterposed features of "intelligence". The former assimilates new information, requiring flexibility in the network to produce complex dynamics; the latter retains old information, requiring stability in the network with sufficient storing capacity. Tradeoffs between the two can be modeled and observed using neural networks. A recent study [152], for example, investigated the relationship between the neural network architecture (e.g. parallel and layered networks) and performance mediated through feedforward neural networks. Another study indicated that classical feedforward networks with gradient descent learning algorithms are not sufficient to describe complex memory and learning dynamics, because real brain dynamics (e.g. memory) are more

complex than fixed point attractors, i.e., characterized by cyclic and chaotic regimes. Hence, classical feedforward networks with gradient descent learning algorithms may not converge when complex nonlinear dynamics (e.g. bifurcation) exist. In this case, RNNs may be a more appropriate choice for describing memory-like structures. In addition, feedback structures can increase network stability and exhibit the paradoxical property of near-perfect adaptation, where many properties of the system remain constant even when the system is subject to an environmental challenge or strong change in other network properties [148].

These examples provide a high-level overview of how to reconstruct and understand the emergence of intelligence using information about component relationships, even when intelligence is strongly emergent. In the next section, we refine our understanding of intelligence in microbes by detailing examples of microbes exhibiting specific characteristics of intelligence.

A.4 Manifestations of intelligence in the microbial world

A.4.1 Decision-making

Decision-making in humans is a vital process undertaken on a daily basis. It is a complex process that involves the coordinated activity of an extended neural network, including several different areas of the brain. Making a decision requires the execution of several subtasks, such as outcome appraisal, cost-benefit analysis, and error perception, before finally selecting and implementing the optimal action. These processes can also be influenced by several factors such as personal preference, reward evaluation, reinforcement learning and social cooperation [16, 132]. In the microbial world, decisions are made by monitoring the current state of the system, by processing this information and by taking action with the ability to take into account several factors such as recent history, the likely future conditions and the cost and benefit of making a particular decision. At the population level, microbes are also capable of hedging their bets, by having individuals of an isogenic population in different states even when experiencing the same environmental conditions, and they are also able to make collective decisions that cause the entire population to respond in a particular

way. Microbes are able to make decisions based on different criteria of information and also to perform the decision-making using different mechanisms, utilizing different types of molecular networks.

It can be argued that even simple biological systems like viruses are capable of decision-making when interacting with their host under certain conditions. A well-studied example is the bacteriophage lambda lysis/lysogeny decision upon infection of *E. coli*. The decision is regulated at the genetic level by a bistable switch, formed by mutual repression [359]. The decision is made based on the conditions of the host cell and the number of phages present. However, stochastic effects are also thought to play a role, either through stochasticity in the expression and regulation of the lambda switch system [15] or through differences between host cell environments prior to infection [322]. The fact that microbes experience stochasticity, due in part to low molecule numbers and the probabilistic nature of molecular interactions, adds layers of complexity to the decision-making process, for example the need to discriminate between signal and noise. With relatively recent technological advances, experimental measurements of stochasticity are more readily obtained and it has been found to affect some decision-making systems. This should be of no surprise, as stochasticity is at the basis of all time dependent processes high molecule numbers and linearity being the forces that remove stochasticity from observation [361].

One of the earliest known systems where a microbe makes decisions is that of ammonia transport and assimilation in *E. coli* [345]. The ammonium transporter (AmtB), the ammonium assimilating enzymes glutamate dehydrogenase (GDH) and glutamine synthetase (GS), and the helper enzyme glutamate synthase (GOGAT) are the main players in ammonium transport and assimilation at low environmental ammonium availability. A decision needs to be made between high-cost, high-accumulation transport by AmtB, low-cost, low-affinity assimilation by GDH, and high-cost, high-affinity assimilation by GS/GOGAT. In making this decision, *E. coli* factors several tradeoffs: (i) maintaining intracellular ammonium at levels sufficient for growth; keeping in check energy costs (ii) of transport and (iii) of assimilation; (iv) minimizing a futile cycle generated by ammonium-ammonia move-

ment across the membrane; and (v) preventing or minimizing the wastage of ATP by the simultaneous action of biosynthetic GS and degradative GDH. This delicate decision is made in *E. coli* through the action of a complex hierarchical regulatory network, simultaneously involving gene expression, signal transduction, metabolic regulation and transport [175, 63, 53, 345].

Many prokaryotic cells are able to move through liquids or over moist surfaces by using a variety of motility mechanisms (swimming, swarming, gliding, twitching, floating) and mostly use complex sensory devices to control their movements [169]. The decision of microbes to move towards nutrient sources or away from toxic compounds is another observation that appears "intelligent". The most studied system is that of chemotaxis in *E. coli*, with common features in other prokaryotes and eukaryotes. In order to make this decision, the cell monitors the environment by means of multiple receptors in the cell membrane. The information of the ligand binding to the receptor, and the processing of this information inside the cell, is achieved by means of a signaling pathway involving methylation and phosphorylation, as opposed to the genetic switch seen in the lysis/lysogeny decision [55]. The level of phosphorylated CheY, the downstream protein of the signaling pathway, determines which movements the cell undertakes: when phosphorylated CheY is bound to the flagellar motor (i.e. when an attractor ligand is present) it rotates counter-clockwise, resulting in a straight swimming movement; in the absence of phosphorylated CheY the unbound flagellar motor rotates clockwise, resulting in a tumbling motion. Using this mechanism, organisms make a biased-random walk, with the length of the periods of straight swimming dependent on the signal, resulting in movement towards or away from different stimuli.

Pseudomonas aeruginosa has been shown to make its decisions about which of its two siderophore-dependent iron acquisition systems to use when faced with iron limitation based on the cost-to-benefit ratios of the two options [97]. The two mechanisms have different costs and benefits to the cell: one mechanism, using the pyoverdine siderophore, has a high iron scavenging efficiency (since pyoverdine has a high iron affinity, Ka = 1024 - 1032 M-1), but comes at a high cost, requiring the expression of at least 14 genes, hence utilizing

high amounts of nucleotides, amino acids, ATP, and other cellular resources. The other mechanism, using the siderophore pyochelin, has a lower cost to the cell because of a reduced biosynthetic pathway consisting only of 7 genes, hence requiring the utilization of few cellular resources, but has a much reduced efficiency of iron-acquisition (since its affinity to iron is relatively low, Ka = 105 - 106 M-1). Here, information processing and decision making is achieved by the finely tuned parameters of the two systems feedback loops that enable them to exhibit different sensitivities. The parameters of the feedback loop for the high-cost, high-efficiency system limit its use to extreme iron limitation conditions and the parameters of the feedback loop for the low-cost, low-efficiency system enable it to be utilized in more moderate iron limitation, thereby optimizing the cost-benefit ratio.

The decision of Bacillus subtilis to become transformation-competent (i.e. able to take up DNA) is made at an individual level; yet, the mechanism by which it occurs results in a reproducible portion of the population making the decision to become competent. The decision making regulatory system is a bistable switch that operates near a critical threshold that, once passed, leads to a committed decision to become competent[225, 207]. Due to this system operating close to the threshold, stochastic fluctuations in the levels of one protein, ComK, are able to push the cell over the threshold to begin the transition to competence [225]. As this is based on stochasticity, it will only occur in a portion of the cells in a population. Since this results in different phenotypes from an isogenic population of cells in the same environment, it is considered to be a bet-hedging strategy [348]. Although each individual may not be in the optimal state for the given conditions, the population as a whole gains an advantage by becoming more adaptable.

Through the above examples of decision-making in microbes, it can be seen that there are several common features that are analogous to processes involved in human decision-making. Although the network components may vary (gene-expression regulation, signaling pathways, metabolism, transport), the networks involved and the parameters controlling their interactions allow the microbes to monitor their environment, process the information and react, effectively making a decision in an “intelligent” manner by taking into account

such factors as the cost-benefit ratio and population survival strategies. We note, however, that decision-making in microbes is not limited to the examples contained here. More importantly, the mechanisms for generating decision-making behaviors are not confined to the particular mechanisms described. Recent work aimed at constructing genome-wide protein interaction networks, for example, has revealed many additional molecules and interconnections that play important roles in these processes [261].

A.4.2 Robust adaptation

An important feature of "intelligence" in microbes is the robust adaptation to changes in environments. Such robust adaptions include homeostasis, as well as adaptive tracking of nutrient sources [275] and evasion of harmful compounds (e.g., bacterial chemotaxis, mentioned previously). Almost all adaptation mechanisms involve feedback or feedforward regulation structures (or motifs). These can be relevant for signaling, gene regulatory and metabolic networks, where homeostasis can be introduced via fine-tuning of rate constants in feedback and feedforward motifs. Relatively long-term adaptations often involve changes in genetic expression, such as gene mutations, transcription/translation activities or rewiring of gene regulatory networks for a review see [60]. Examples include adaptation to salt conditions, temperature or asymmetric cell division. Short-term adaptation, on the other hand, typically involves regulation mediated by (i) protein-protein interactions and covalent modifications (e.g. phosphorylations) in signal transduction pathways; or (ii) allosteric or more direct substrate-product effects in metabolic networks. Of all the adaptive regulations, robust perfect adaptation is of particular interest. It describes an organisms response to an external perturbation by returning state variables to their original values before perturbation. For example, perfect adaptation has been reported in bacterial (e.g. *E. coli*) chemotaxis [41, 12, 379, 145], osmotic-stress adaptations [253] and MAP-kinase regulation [146, 245]. Such perfect adaption behaviors are thought to be introduced through a time integral on the controlled variable in the network, which corresponds to a specific control system structure, i.e. an integral feedback control [85]. A recent in silico study [224] identi-

fied an alternative topology that can also ensure perfect adaptation through an incoherent feedforward structure, where a positive regulation cancels out the effect of a simultaneous negative regulation, hence the overall output is insensitive to the input signal. Because it has been difficult to experimentally discriminate between perfect and strong adaptation and because at least some of the proposed mechanisms for perfect adaptation require biochemically unrealistic features (including zero order degradation of proteins [148]), the evidence for truly perfect adaptation needs to be revisited. In many cases, adaptation may be less perfect, with robustness being strong, but limited. Here, it would help if robustness were quantified [281]. In non-robust "proportional" [148] regulations, the appearance of a specific signal or environmental condition can be a direct indicator/predictor of a particular response. The feedforward regulatory mechanism, then, is introduced to respond directly to the signal rather than to the disturbance. Feedforward regulatory structures were observed in gene regulatory networks in the regulation of membrane lipid homeostasis[233, 8], in bacterial photosynthesis genes for optimal free-energy supply [235], and in the heat shock response in *E. coli* [312].

Different regulation mechanisms in living cells often occur at multiple levels simultaneously with a hierarchical structure [362]. For example, in a microbial metabolic network, the regulation of a reaction rate can be achieved by the modulation of (i) enzyme activity through a substrate or product effect, or through an allosteric effect, i.e. metabolic regulation; (ii) enzyme covalent modification via signal transduction pathway; or (iii) enzyme concentration via gene expression, gene-expression regulation. Such multi-level regulation corresponds to different control loops in a control system, which may ensure the robustness versus perturbations at various frequencies, as employed in engineering system design. Let us consider an unbranched metabolic pathway, with the first enzyme inhibited by the end-product via both allosteric/metabolic and gene-expression regulation. If the flux demand on the end-product module increases rapidly, the concentration of the end-product decreases rapidly. Often, as a result of the allosteric effect of the end-product directly on the first enzyme, the activity of that first enzyme increases quickly too. This metabolic control of

enzyme activity is a fast actuator of the system. However, if there is a further increase in the flux demand, the first enzyme may lose its regulatory capacity since its activity may be approaching its maximum capacity (kcat). At this stage, the system has a second adaptation through gene expression that is slow but leads to an increase in the concentration of the first enzyme, which then decreases the direct stimulation of the catalytic activity of the first enzyme. The regulation of the first enzyme is then bi-functional in dynamic terms [85]: the metabolic regulation rapidly buffers against high frequency perturbations, but possibly with small amplitude or capability, while the gene-expression regulation is slow to adapt, but may be able to accommodate very large constant perturbations [334].

When interpreting metabolic and gene-expression regulation separately as specific "control system structures", the former was recently identified as more of a "proportional control" action [379, 102] with limited range and the latter as more of an "integral control" action with potentially a wider range, but acting more slowly [148]. Such control engineering interpretations can also be linked with classical Metabolic Control Analysis (MCA) [111] and Hierarchical Control Analysis (HCA) [175]. The relatively fast metabolic regulation is related to the direct "elasticities" of MCA, while the slow gene-expression regulation corresponds to the indirect "elasticities" of HCA [148].

A.4.3 Association and anticipation

Associative learning allows one to model how two or more features in the world co-vary and respond accordingly. This type of learning provides context, in the sense that it specifies how several features in the environment, or within cells, change together. It implies that the learner has a mechanism to encode mutual information. In humans and animals, this type of learning has been associated with experimental settings where, for example, a subject is conditioned (often through an auditory or visual cue) to activate unconditioned responses (like salivation) after presenting the subject with a conditioned stimulus (e.g. a bell) simultaneous to the unconditioned stimulus (e.g. dinner) that usually elicits the unconditioned response. After a period of learning the association, the unconditioned response (salivation)

can be achieved in the absence of the unconditioned stimulus (simply ringing the bell). Conditioned behaviors like this have been well studied in humans and other animals since the pioneering work of Ivan Pavlov [276]. Recently, the molecular mechanisms responsible for encoding these behaviors in neurons have been defined [238]. In general, these mechanisms rely on the plasticity of neurons to reinforce electrochemical couplings, such as changing the localization and abundance of glutamate and NMDA receptors at synapses [255, 291]. The development of recurrent artificial neural networks, for example Hopfield networks [157], has provided a computational model for studying the processes of associative memory.

Associative learning allows learners to structure dependencies that exist in the world. Pavlovs dog, for example, salivates because of the linkage the dog has learned between bell and dinner; even though the association is entirely manufactured in this case. Outside of contrived laboratory conditioning, associative learning occurs when environmental variables are physically coupled, or somehow co-vary non-randomly. For example, the increase in the level of light (photons) at sunrise, signals associated changes in the environment, such as increase in temperature, change in O2 availability, etc. Organisms leverage these physical associations to better adjust their physiology in specific environments [50], to employ easily-measured proxies as indications for other phenomena (like the bell for Pavlovs dog) and, in some cases, even use the cues themselves to prepare or "anticipate" subsequent alterations to the environment. Investigators have asked recently whether organisms like microbes, which do not have nervous systems, can also exhibit associative learning and anticipation. Several experimental studies and modeling efforts have suggested that, indeed, microbes can learn associations, both as communities and individually. Studies, furthermore, suggest that gene regulatory networks can encode associative learning. One of the most comprehensive examples of this phenomenon comes from a study of the bacterium *E. coli* [329]. As a microbe that lives both in the soil and the guts of mammals, *E. coli* has to adjust its physiology to environments that vary with respect to important biological parameters, such as temperature and oxygen availability. Since many of these environmental parameters do not change randomly, but rather in coupled ways (e.g., increase of temperature in the oral

cavity and corresponding decrease in oxygen availability in the gut), *E. coli* is able to take advantage of this predictable physical association to direct its physiology accordingly. In this study, the authors demonstrated that transcriptional responses in elevated temperatures are highly similar to those observed in oxygen perturbation experiments, even though the second stimulus is absent (much in the same way that Pavlovs dog can be stimulated to salivate simply by ringing a bell). More impressively, they showed that *E. coli* can "re-learn" these associations. Relative to ancestral *E. coli*, evolved strains grow better in environments where temperature and oxygen are decoupled (in this case inverted). This study demonstrates (1) that microbes have both the capacity for associative learning, and (2) that the learned associations are plastic. A similar study in yeast suggested that previous lifestyle plays an important role in adaptation to severe stress, re-emphasizing the existence of associative learning in microbes [42].

It is important to note, however, that time-scale for this "learning" is on the order of evolutionary processes and most likely involves genetic changes. This has an analogy in the development of "fixed" hard-wired neuronal connections in a brain or cultural learning in human society. In the example, it took many generations for *E. coli* to learn about the altered association between oxygen and temperature and, presumably, much longer for the natural situation to be canalized. Critical questions for future studies will include whether gene regulatory networks encode associations that are capable of being learned within the lifetime of an individual bacterium; a case in point was made for ammonia assimilation in *E. coli* [150, 62]. A recent modeling study suggested that gene regulatory networks composed of bistable elements with stochastic dynamics can exhibit associative learning, although the number of learnable associations may scale as the square root of the number of bistable elements [320]. Similar results have been obtained in the context of chemical networks [244] and other transcriptional networks [68]. Additional experiments, however, are required to evaluate whether the dynamics of cellular networks with multiple stable states are sufficient to encode and retrieve contextual associations. Hellingwerf showed learning behavior should be possible in realistic mono-stable *E. coli* networks [150].

Among microbial populations, associative learning seems to be commonplace. Mechanisms and examples of associative learning in microbial communities have been discussed extensively elsewhere [38, 378]. Typically, associative learning in microbial populations involves some sort of social communication (such as quorum sensing, discussed in section A.4.5). This type of networked communication is highly plastic and eminently reminiscent of neuronal activities. Other examples of association and anticipation in the microbial world are exhibited by pathogenic bacteria such as *Pseudomonas aeruginosa*, which is an important human, animal, and plant opportunistic pathogen and, perhaps, the bacterial species that has most genes devoted to regulatory purposes [325]. In the context of human digestive tract infections, this bacterium senses several compounds released by the host tissues, such as interferon, opioids, and metabolites like adenosine, which are all released into the intestinal tissues and lumen during surgical injury, ischemia and inflammation. In addition, it senses the extracellular levels of phosphorus, which decrease severely when the patients condition deteriorates. Hence, when the bacterium senses high concentrations of host-released compounds together with a decrease in phosphate levels, it anticipates the vulnerability of the patient and turns on several virulence determinants that frequently lead to lethal sepsis [383]. Recently, it was demonstrated that in the genus *Burkholderia*, quorum sensing allows the activation of cellular enzymes required for production and secretion of oxalic acid, which serves to counteract ammonia-mediated alkaline toxicity during the stationary phase, hence anticipating a stress situation and triggering a preventive strategy that helps cells better adapt to the oncoming harsh environmental conditions [134].

The capacity for associative learning among microbes may be one of the reasons why we are able to reverse engineer them. Since microbes do not respond to stimuli independently, but rather their internal networks direct common responses to diverse but related environmental signals, regulatory networks in microbes can be reconstructed by simply measuring their response across a broad range of conditions. Gene regulatory networks, for example, can be inferred in three simple steps: (i) perturb cells across a broad range of relevant conditions; (ii) measure their transcriptional response in each environment; and (iii) cluster

similar gene expression patterns observed reproducibly across environments. Mining for genetic similarities among genes sharing a particular expression pattern, such as common *cis*-regulatory elements in their promoter regions, in turn helps link these transcriptional modules to some of the molecular mechanisms responsible for regulating them. In practice, such approaches allow the construction of gene regulatory networks directly from transcriptome measurements [284]. It should be recognized, however, that the networks thus reconstructed are incomplete, as they forego the signal transduction and metabolic networks that are part of the actual regulation [334].

A.4.4 Associative learning in protozoa

Early investigation of intelligent traits in microbes, such as associative learning and memory, occurred in ciliated protozoa. While early studies concluded that ciliates are capable of associative learning, several experimental design flaws have led to skepticism about these conclusions. In 1937, for example, Soest et al. reported that the ciliate *Stentor* contracts if exposed to light after conditioning with simultaneous luminous stimuli and electrical shock [318]. The authors concluded that *Stentor* exhibited classical condition response; their study, however, lacked important controls, such as training *Stentor* with the administration of shocks alone [81]. A similar study suggested that paramecia perform instrumental conditioning [129]. The author observed that paramecia attached preferentially to a bare wire that had been baited previously with bacteria compared to a wire that had not been baited. It was demonstrated later, however, that the behavior likely resulted from increased bacterial concentration near the wire rather than as a consequence of associative learning [170]. Even the paradigmatic example of learned escape from the bottom end of narrow capillary tubes into a larger volume of media by *Stentor* and *Paramecium* has been refuted. Subsequent to the initial report of this behavior, it was noticed that the strategy simply entailed decreased upward swimming. In fact, the same behavior was observed when the task was reversed, demonstrating that this behavior is unlikely to be the result of associative learning [154]. It should be noted, however, that these examples may have been insufficient

to meaningfully test associative learning, since they did not reflect abilities required by protozoa in their natural environments.

Contemporary research has focused instead on ecologically salient intelligent behaviors, such as mate selection, foraging and hunting [75, 74, 77, 78]. This new wave of research has renewed interest in ciliate intelligence. More significantly, it has reinforced the claim that ciliate protozoa indeed have remarkable learning abilities, including complex cooperation and competition behaviors usually attributed to higher organisms. The observations imply an ability to learn and adjust mating strategies using Hebbian-like associative learning behavioral heuristics.

The ciliate *Spirostomum ambiguum*, for example, learns to advertise mating fitness to suitors and rivals during the preconjugal courtship. Fitter suitors - "conspicuous consumers" - advertise their status by avoiding exchange of preconjugal touches, despite the metabolic cost of swimming away. Less fit individuals - "prudent savers" - on the other hand, wait for favorable opportunities for partner conjugation, conserving energy and exhibiting lower avoidance frequencies. Interestingly, both "conspicuous consumers" and "prudent savers" learn to switch between the two strategies, apparently tuning their behavioral heuristics and switching frequencies to optimize mate selection [76].

Less fit individuals are even capable of "cheating" in this system. These individuals take advantage of a fit suitors "conspicuous consumer" behavior. A less fit individual positioned between a fit suitor and potential mate may, for example, corrupt the "conspicuous consumers"contraction-reversal movements (e.g., flip the signal from avoidance to conjugation). The "cheater" can physically interact with these signals, since they are spread as vibrations through viscous media. As a result, the "cheater" can conjugate with a mating partner that has been "aroused" by the fit suitor's actions. The signal would be easy to take advantage of if it were scripted in a binary encoding (e.g., 0 - no contraction, 1 - contraction); however, suitors appear to encode a low probability of contracting and reversing simultaneously, in addition to simple contraction and reversal behaviors. This would make ciliate mating signals resemble a quantum bit flip channel used in quantum computing [**?**, **?**].

Encoding mating communication with a contraction-reversal qubit would make it far more robust to "cheating" behaviors of competitors.

Evolution of error-correction systems that counteract degradation of mating signals is quite remarkable. These mechanisms must account for nonrandom color noise created by mixing of vibrations emitted by mating rivals and suitors, as well as random ecological white noise [75]. It would seem that ciliates have developed coding schemes to diagnose, decrease, and counteract mating-signal errors due to noisy information processing [78].

These findings suggest that quantum computing concepts may be required to understand emergence of intelligent communication in microbes. Without the concept of qubits, for example, we would have been unable to describe the complex encoding of ciliate mate selection behaviors. Quantum computing was first proposed in the 1980s [234, 115], so one has to wonder how the expansion of our knowledge horizons may influence our understanding of intelligence in all forms of life in the future.

A.4.5 Quorum sensing and self-awareness in microbial populations and communities

Quorum sensing is a widespread type of bacterial cell-cell communication between individuals of the same or different species [356, 205, 158]. The accepted paradigm for this kind of communication is that individual cells steadily produce and release several kinds of small diffusible molecules (signals), called auto-inducers. In parallel, each cell has the ability to sense the presence of those molecules, by means of receptors/transcriptional modulator proteins that bind the auto-inducers and, once complexed with them, trigger a global transcriptional response that leads to crucial changes in the expression of several phenotypes and behaviors. An important property of quorum sensing communication is that the response is only achieved after one specific signal (i.e., cell number) threshold is exceeded. The response is mediated by a positive feedback loop of auto-inducer production, since genes for the enzymes that biosynthesize the signals are under their own control. There is a plethora of behaviors and phenotypes controlled by quorum sensing systems, including light production by several species of the Vibrio genus, competence (i.e. the ability to uptake

and incorporate foreign DNA), biofilm formation, synthesis of secondary metabolites and the production of virulence factors.

Self-awareness can be described as the ability to recognize oneself as an individual separate from the environment and other individuals. Quorum sensing provides the entire bacterial network with the ability to recognize and adjust itself collectively once a specific population threshold is exceeded. This is specific for all individuals of a certain organism and even strain. Quorum sensing, therefore, can be viewed as a kind of self-awareness among isogenic bacterial populations.

Signaling related to specific environmental cues is interwoven with quorum sensing signaling; for example in *Pseudomonas aeruginosa*, the iron availability signal network and the quorum sensing system communicate and influence each other [174]. In addition, bacteria can sense quorum sensing signals of other species [110] and act in accordance with the population sizes of competing or mutualistic species, including cells of eukaryotic or pluricellular organisms, such as their hosts [29, 158, 224]. Thus, microbial networks have the ability to distinguish themselves from similar networks in other species. Most of the bacterial cell-cell communication described to date exclusively involves the release of autoinducers to the extracellular medium and the sensing of those molecules by other cells; phenomena that depend on the diffusion of signals and therefore lack directionality. Since, in a well-mixed environment such as a stirred liquid culture of planktonic cells, one cell can sense the autoinducer produced by any other cell, communication among network components should be uniform. This is in contrast to communication among molecule types in signal transduction networks and among cells in neuronal networks. In the latter cases, each member interacts directly with a limited set of other network components, creating clusters and functional domains that, together, form a structured network with non-trivial topological features and a higher-than-random complexity. The situation changes in more realistic environments, such as in bacterial biofilms, which are known to be the preferred lifestyle of several bacterial species [82]. Those biofilms can be composed of a single bacterium species, but more often are complex ecologies of single-cell organisms that may include hundreds of different

species of algae, bacteria, protozoa, fungi and viruses. They collectively generate and embed themselves in an extracellular polymeric matrix that provides structure and protection. In such environments, cell-cell communication could be more specifically performed among clusters of cells organized in different spatial and functional biofilm domains. Recently, the discovery of bacterial communication networks of multiple cells of *Bacillus subtilis* that are directly connected to others by bacterial nanotubes was reported [95]. These structures are able to mediate the exchange of non-conjugative plasmids, metabolites and even enzymes, and can be formed in an interspecies manner between *B. subtilis* and *Staphylococcus aureus* or even the phylogenetically more distant *E. coli*. The authors speculated that these kinds of networks may represent a major form of bacterial communication in nature. If so, they may constitute complex and intricate structured bacterial communication networks with high potential to exhibit intelligent behavior.

Some features of self-awareness can be manifested already at a lower level of social organization of microorganisms. Thus, bacteria of the same species are capable of assembling together and isolating themselves from other species. This advanced social organization would be reflected in cooperation; for example in swarming motility (coordinated translocation of many bacterial cells), in collective repairing of holes in biofilm, in collective capture and digestion of food, etc. Microorganisms can cooperate for collective aggression through the coordinated production of antibiotics. There are even "bacteria-altruists", who sacrifice themselves to become food for their brethren [265]. However, at the opposite extreme, there also exist "microbe-cheaters", which can disrupt cooperative systems by acquiring a disproportionate share of group-generated resources while making relatively small contributions [349].

Gram negative bacterial pathogens, such as *P. aeruginosa*, *E. coli* enteropathogenic strains and several Vibrio species, and Gram positive pathogens, such as *Staphylococcus aureus*, use QS to coordinate expression of several virulence determinants [14]. Beyond prokaryotes, QS is also used by eukaryotic pathogens, like the fungi *Candida albicans* [260], and even more complex microbes, such as parasitic protozoa like *Trypanosoma brucei* [250].

Although QS systems have been studied mostly in microbial pathogens, it has been discovered recently that several harmless free-living bacteria, such as cyanobacteria [311, 384] and methanogenic Archaea [385], also possess QS communication systems. Unlike pathogenic organisms, however, these microbes appear to use QS to achieve robust adaptation to environmental change. This is accomplished by redirecting metabolic fluxes at high cellular densities to optimize energy and resource consumption. In this sense, QS allows communities of related microbes to anticipate and prepare for nutrient scarcity [311, 385]. QS may even play a key role in establishing biofilms and initiating cellular blooms of cyanobacteria [384].

In free living bacteria, QS contributes to cell differentiation and establishment of multicellular populations. A classic example of QS-mediated cell differentiation in bacteria is starvation-induced reproductive fruiting body development in *myxobacteria*. In *Myxococcus xanthus*, for example, soluble quorum-sensing A-signal assesses starvation and mediates the initial stages of cell aggregation [176]. Finally, filamentous cyanobacteria exhibit one of the most complex cell differentiation processes observed in bacteria. These microbes can differentiate into four different cell types, including: (i) multicellular filaments that branch in multiple dimensions (trichomes); (ii) specialized nitrogen fixing cells called heterocysts; (iii) spore-like cells called akinetes; and (iv) hormogonia, which are small motile filaments that are important for dispersal [119, 297]. So far, calcium cell signaling has been implicated in development of heterocysts [337]. Given that QS was recently discovered in these organisms [311], it will be interesting to see what, if any, role QS plays in these differentiation pathways. Multicellularity, even in microbial populations, is an adaptation that allows cells to perform complex tasks and exhibit intelligent behaviors, like coordinating community-wide responses to environmental change. QS clearly plays a role in establishing multicellularity in microbes, but may also be the chemical language for communication of that intelligence.

The complexity of bacterial biofilms is equally striking. These rich ecosystems provide an environment for microbes to demonstrate their individual and collective intelligences. The human oral cavity, for example, contains hundreds of different bacterial, viral and fungal

species. These species establish complex relationships, including both competitive and cooperative behaviors. We call the biofilm formed by these microbes the "dental plaque". While many plaque species are commensal, some may become pathogenic in response to environmental triggers. A sudden shift in biofilm composition or dynamics may lead to dental caries and several other periodontal diseases [18]. Among the dental plaque residents, *Porphyromonas gingivalis* is of particular concern. This species is a predominant contributor to human periodontitis. It employs several intricate mechanisms to subvert the innate immune system of the host. In fact, these evasive strategies are so clever that they have been compared to military tactics used in "guerilla" wars [142]. Complex microbial communities are located in the gut of mammals as well. These highly dynamic, species-rich communities help modulate the host's immune system. They are implicated in several human diseases, including chronic inflammatory diseases, such as Crohn's disease [227, 79], as well as obesity [287] and diabetes [105]. Some evidence even suggests that microbes may alter human brain function and behavior [84]. The ability of the microbiome to influence human intelligence has earned it the title, "the forgotten organ" [286]. Together, these results suggest that symbiotic microbiota may have played an important role in the evolution of plants and animals, leading some to contend that the unit of selection in evolution may be the holobiont, i.e., "the animal or plant with all of its associated microorganisms" [387].

Finally, it is worthwhile to note that philosophers of biology are beginning to appreciate the remarkable microbial capacities for cooperation and communication [272, 266, 271].

A.4.6 Problem solving

An essential feature of any intelligent system is that, in addition to storing information and incorporating new knowledge from experiences, it must have the ability to use that knowledge to solve new problems. Generally, the more complex a problem a system can solve, the more intelligent it is considered. In this regard, some microorganism networks show problem solving abilities that can even match or surpass those shown by human beings: the slime mold *Physarum polycephalum* in its plasmodium configuration a large multi-nuclear

amoeba-like cell consisting of a dendritic network of pseudopodia has the ability to connect two different food sources located at different points using the minimum-length pathway in a labyrinth, which optimizes its foraging efficiency [254]. The mold is able to create solutions with comparable efficiency, fault tolerance and cost to those of human infrastructure networks, such as the Tokyo rail system, but, unlike humans, the mold achieves optimal solutions solely by a process of selective reinforcement of the preferred routes and the simultaneous removal of redundant connections, without any centralized control or explicit global information. This striking mold ability was captured in a mathematical model, which the authors claim can provide a starting point to improve routing protocols and topology control for self-organized networks used for human transport and communication systems [17]. This is a perfect example of applied microbial intelligence with the potential to improve human engineering.

A.5 Learning from intelligence in the microbial world

Given the examples of the previous section, it is likely that, at least for some specific tasks, microbial "intelligence" can be compared to human intelligence, and microbial networks could be considered formally as "intelligent". Recognizing microbial intelligence can allow us to potentially modify microbial networks or to develop new microbial networks capable of intelligent solutions to specific human problems *de novo*. If intelligence (or components thereof) emerges from the dynamics of complex adaptive systems and the human brain is an evolved organ for the encapsulation of intelligent characteristics, it is possible that there are features of intelligence that remain undiscovered.

A.5.1 A deeper understanding of the microbial world

One important and exciting domain of synthetic biology is the manipulation and design of microbial metabolism for chemical production in the energy, biomedicine and food industry [280]. Such design relies on effective control and adaptation of metabolism (e.g. pathway flux) in response to intracellular or environmental perturbations. In an engineered

genetic-metabolic circuit, there are many parameters that can be used for design purposes. Promoter characteristics, such as tightness, strength or regulatory sites, can be engineered in the transcriptional control, and the engineering of ribosome binding sites or RNA degradation can be used to control the expression levels of proteins. Well-known examples are the genetic control of lycopene production in *E. coli* [109] and the design of gene-metabolic oscillators [122, 327]. Designing scaffold proteins in the protein-protein interaction domain has been studied for the control of metabolic flux [96]. Recent studies [148, 141] showed that although gene-expression regulation can increase the robustness of an intermediate metabolite concentration, it rarely makes the metabolic pathway infinitely robust. For perfect adaptation to occur, the protein degradation reactions should be zero-order in the concentration of the protein or the living cell should enter stationary phase after a period of growth. The former scenario is rarely observed biologically; nevertheless, in some situations, protein degradation rates can be controlled by adding or removing a degradation tag to the gene sequence [243]. In this way, a relatively small degradation rate may be obtained in an engineered gene-metabolic network, and near-perfect adaptation behavior can be achieved with a quasi-integral control structure.

A.5.2 Microbial vs. human intelligence

Our paper collects various examples of the intelligent features discovered in the microbial world. Microbial intelligence emerges from the dynamic interactions among macromolecules. Intelligence is a strong form of emergence; its reconstruction requires information of state-dependent component properties. The more state-dependent information we need, the stronger the emergence is. The degree of state-dependency of the component property is determined by the presence of other components in the system affecting this property, on the flux of matter through the system and on the history of the system [194, 195, 196]. In this context, we can scale and compare the strength of emergence of intelligence for different complex adaptive systems, e.g. for microorganisms, animals or humans.

In bacteria, there are many potential intracellular interactions that can affect the state-

dependent property of a certain molecule. For example, the ability of a single transcription factor to bind a response element might depend on the presence of other transcription factors and their ligands, on components involved in intracellular trafficking of these ligands, on molecules providing ATP-convertible free energy for this trafficking and for receptor synthesis and even on molecules maintaining pH, viscosity, macromolecular crowding, etc. Thus, the emergence of intelligence that is raised due to interactions in an intracellular microbial network can be very strong indeed. On the other hand, the number of neurons affecting the firing of a single neuron in the human brain is tremendously high; and this is before we consider the intracellular interactions occurring in each and every neuronal cell, all of which contribute to the strength of the emergence of intelligence in our brains. Are these intelligences even comparable? We intuitively feel that the intelligence in microorganisms and in humans is different. The physiological adaptive behavior of microorganisms is not stable and disappears when the environment does not support this behavior. Programs of adaptive behavior are imprinted on the population genome. When adaptation is lost, new training is required to regain this adaptation. Microorganisms exhibit some features of elastic behavior, but they do not have the conditional reflexes of higher animals. In an evolutionary context, in animals the elementary reflection of the environment is replaced by perceptive reflection and animals gain different forms of individually-adapted behavioral changes co-tuned to the changes in the environment. Animal activity toward objects develops depending on the objects animals have already dealt with. This correlates with anatomical changes; the cerebral cortex emerges in addition to basal ganglia that cause a crucial shift in animal behavior. Basal ganglia enable signal reception and turn on inherited behavioral programs. The cerebral cortex, in its turn, enables analysis and integration of external signals, reflection on external objects and situations, building up of new connections and, ultimately, development of the behavior that is based, not on the inherited programs, but rather on the animals perception of external reality. With the development of the cerebral cortex, new forms of individual behavior based on objective reflection of the environment are formed.

Further development of the cerebral cortex takes place in humans. Aside from both inherited programs and individually gained experience, humans develop a third form of behavior: the ability to transfer collective experience from one human being to another. The transfer of collective experience includes the knowledge gained at school, at work, in life, etc. Animals are born with the inherited programs and enrich these programs through individual experience. Humans might be born with the poorest instinctive inborn programs, but can develop their mental processes, not only through personal experience, but also through learning from collective experience. Human individuals are able to communicate with each other and even, through the media of oral tradition and written history, with their predecessors. Nevertheless, in the context of scaling the degree of the strength of emergence, the complexity of the human brain does not change immensely compared to the brain of an animal. Rather, the new behavior emerges from the changes in the design, and not from a tremendous increase of interacting components.

Intelligence is a strongly emergent property in both microorganisms and animals, including humans. Still, there is a difference in the way these intelligences are manifested. Thus, humans study microorganisms and debate about microbial intelligence, and bacteria, while supremely adapted and aware of their environments, are probably not even aware of us and our endeavors.

A.5.3 The way forward

Most aspects of human intelligence are also exhibited by microorganisms at least to some degree, except those that depend on reading, writing and listening. The examples we presented regarding quorum sensing and problem solving were from multicellular networks. The question remains whether networks at any single, more molecular level, such as intracellular signaling, also exhibit most aspects of intelligence. It has been proposed that intracellular quorum sensing occurs during mitochondrial apoptosis [56]. The hierarchy of regulatory networks involved in ammonia assimilation is a candidate for rich intelligent behavior. The molecular information is now so complete [345] that it may well be possible to

develop the existing replica models [63] into a full representation. These may then be used to determine the extent to which our present molecular network understanding suffices to demonstrate that these networks should be expected to exhibit almost all types of intelligent behavior [150, 62]. This could then also help with experimental design driving subsequent experimental testing. Similarly, such mathematical representations may also be used to search for new aspects of intelligence that we, as humans, do not recognize as such, for example adjustable robustness, random creativity facilitated by deterministic chaos in the networks, productive noise thereby, and read-only memory. Many of these aspects may be useful for synthetic biology; a synthetic biology that will give rise to much more sustainable, productive systems.

Appendix B

IGBWEB: AN INTERACTIVE GENOME BROWSER FOR THE WEB

Appendix B has been submitted:

Salvanha, DM, Brooks AN, Reiss DJ, Vêncio, RZN, Baliga NS. iGBweb: an interactive genome browser for the web. Submitted.

B.1 Abstract

Summary/Motivation: iGBweb is a web tool that enables interactive visualization of systems-scale data in the context of the genome. iGBweb implements a suite of features for large-scale data exploration within an easy-to-embed, browser-independent web application. The application is open-source, highly-customizable and easily-integrated into existing HTML5 web applications. Multiple data types can be imported, rendered, paired, and synchronously animated by iGBweb, including: heat maps, line charts and a novel positional quantitative-string track. iGBweb is the first genome browser to allow direct visualization and animation of dynamic biological processes.

Availability and Implementation: The iGBweb front end is implemented using HTML5 resources (e.g. AJAX, D3.js, SVG, CSS). A Java and MySQL back-end web-service is also available for data retrieval if necessary. Source code (LGPL license), documentation, and examples are available at: http://igbweb.systemsbiology.net.

B.2 Introduction

Biological systems can now be studied across multiple scales by integrating and modeling diverse kinds of molecular measurements. The high-throughput technologies used in these studies generate very large amounts of complex data. Effective representation, integration, and interactive analyses are critical to extract biological insight from these data. We describe a tool that allows end-users to interact with integrated visualizations of dynamic biological processes in the context of the genome.

iGBweb is an interactive web application module that integrates biological data using multiple genomic tracks. The user interface consists of a minimal Javascript library that provides a straightforward web environment for developers and end-users alike. Data can be uploaded to iGBweb using several strategies, enabling rapid development. Unlike other genome browsers, iGBweb can animate paired data from experiments with a dynamic context (e.g., changes in transcript expression levels and transcription factor binding across a time-course). iGBweb is also readily embedded in pre-existing web resources, encouraging developers to extend their own web approaches using our tool.

B.3 Implementation

iGBweb is a client-side application with an optional server-side component that leverages modern web technologies. The iGBweb front-end is cross-platform, browser independent, and readily combined with other web technologies. The front-end user interface collects data from a back-end data repository, rendering the data as track(s) in the browser. The front-end module was implemented using the D3.js library (Data-Driven Documents; [54]). The UI was implemented as independent modules to provide scalability and modularity for web integration (including genome, focus, and context viewers, as well as view controllers). Since the modules are semi-independent, they can be added or re-implemented as necessary. New track types can be developed easily as well.

iGBweb requires a server-side implementation for data retrieval. Developers can use their own back-end (as simple as a structured JSON file) or our pre-defined back-end tem-

plate. This optional server-side component uses Java (JAX-RS) to expose RESTFul retrieval services that consume a MySQL database. Alternatively, developers can connect to their own database using Ajax calls.

B.4 Available Features

B.4.1 Features

iGBweb provides a suite of features to represent diverse biological data types along a genomic axis. The iGBweb software represents genome-anchored data as two-dimensional graphics, including heat maps, bar charts, data points and segments. In addition to encoding a suite of diverse graphical representations, iGBweb also includes an intuitive API that encourages development of additional features to enhance end-user productivity. All of the graphical properties (including track style attributes such as color, symbol, and scale) are customizable and can be programmatically changed, if necessary. Each of these design features was implemented in iGBweb to make it easy for developers to implement a user-friendly genome browser that can be customized to meet individual project requirements.

Beyond bringing standard genome browser visualizations to the web, iGBweb includes two features that are not currently supported by other genome browsers. First, iGBweb can represent scaled, positional quantitative strings. It can map any ASCII character to genomic coordinates. This feature facilitates representation of data abstractions, such as position-specific scoring matrices that are often used to represent DNA or protein sequence motifs and dot-bracket notation to encode 2D RNA structure (Use case 1). Second, iGBweb can animate data transitions. This feature is particularly useful to facilitate analysis of disparate data sets (e.g. ChIP-chip, microarray or time-series), which can be combined, paired, and rendered synchronously, allowing the end-user to visualize connections and correlations in their data (Use case 2). The software includes data transition modes for four track-types, including: heat maps, quantitative positional, quantitative segment and quantitative strings. Both of these novel features aim to modernize the genome browser, making it a core tool for hypothesis generation in systems biology.

B.4.2 Use cases

To illustrate the flexibility, simplicity, and utility of the iGBweb we describe three use cases. The examples are presented in order of increasing complexity, demonstrating the full range of capabilities offered by iGBweb. The first example integrates quantitative gene expression data with character representations of RNA secondary structure predictions; the second pairs ChIP-chip and gene expression data to explore the effects of a transcription factor knockout; finally, the third demonstrates how easy it is to extend pre-existing web applications using iGBweb. Each example reveals how iGBweb enables intuitive visualizations to be developed quickly and easily. Online tutorials are available to guide the developer through configuration of iGBweb for examples 1 and 2, including all steps required to implement iGBweb, from reading structured data into the genome browser to configuring tracks to display the data.

Use case 1.

One of the novel features of iGBweb is its ability to render genomic data as scaled and/or dynamically modulated ASCII characters. For the first example, we visualized data from a study by Randau [283] that revealed an abundance of processed small RNAs (sRNAs) in the hyperthermophilic archaeon Nanoarchaeum equitans. RNA secondary structures were displayed by iGBweb in standard dot-bracket notation. We visualized the authors RNA-seq data together with RNA structure predictions from all 39 families in the Rfam database, and structures for tRNAMet and tRNAVal from the paper. In this case, simultaneous visualization of both data types not only revealed the abundance of sRNAs noted by Randau, but also highlighted potentially interesting correlations between RNA secondary structure and expression levels. This complete example can be fully developed in less than one hour.

Use case 2.

The second example illustrates how iGBweb enables intuitive, dynamic visualization for paired measurements. This example includes data integration (as in Use case 1.), as well

as animation to visualize paired transitions across two kinds of data. We paired ChIP-chip binding data for the E. coli transcription factor PurR in two conditions (+/- supplemental adenine) with gene expression measurements for both wild type and PurR mutants in these conditions (GEO GSE26591; [73]). The data are visualized in iGBweb as two genome tracks: a heatmap for gene expression; and quantitative segments for ChIP-chip data, including vertical bars to indicate TF binding sites. Integrated, visual representation allows users to detect important patterns in their data; for example, at the *carA* promoter (a gene involved in pyrimidine biosynthesis) the visualization clearly reveals that PurR binding in the presence of adenine results in repression of *carA*. This type of visualization can easily be extended to other types of experiments with multiple data types and/or dynamics, including time course studies.

Use case 3.

As a last example, we demonstrate how easy it is to plug the iGBweb into preexisting web apps. We implemented iGBweb to facilitate exploration of a computational model that predicted the condition-dependent magnitude of influence of gene regulatory elements (GREs) at each nucleotide residue across the genome (position-based numeric values; EGRIN 2.0 [Brooks AN, DJ Reiss, et al. Accepted MSB]). These predictions were contained in a PostgreSQL database with an associated Django web app. iGBweb was embedded into this preexisting web framework using Javascript to replace static representations of the model predictions. Given a gene name, iGBweb queries the EGRIN 2.0 database to retrieve information about the gene, including model predictions about the gene promoter region. Using this information, iGBweb produces a line plot with a dynamic representation of changes in predicted GRE-associated influences across that region. The implementation allows the user to cycle through condition-dependent predictions and select which GREs are plotted. In addition, the user can activate the novel iGBweb positional quantitative-string feature to represent the genomic sequence within the line plot. All of these features were easily customizable on the front-end. Implementation of iGBweb in this large web resource took

less than one week (including extensions to the default iGBweb features).

B.4.3 Documentation

The iGBweb source code is freely available at http://igbweb.systemsbiology.net. The site includes tutorials for the Use cases, API documentations and examples source codes that can be used as a template.

B.4.4 Future Directions

As genomic technologies evolve, the ability to adapt data visualization platforms for new data types will become increasingly important. We propose a public track repository that will allow iGBweb developers to keep up with the most recent trends in visualization technologies using iGBweb. The resource will enable developers to contribute track implementations, tutorials, as well as data. This, in turn, will enable a community of developers to create user-friendly, flexible, plug-and-play genome browsers that will advance biological understanding.

VITA

Aaron Brooks was enrolled in the Molecular and Cellular Biology Program at the University of Washington where he was a member of Nitin Baliga's lab at the Institute for Systems Biology.

Aaron studies gene regulatory networks in prokaryotes. Blending hypothesis-driven experimentation, large-scale data analysis, and theoretical insights from complexity science, his research aims to reveal how regulatory networks function and evolve.

Aaron holds degrees in Biochemistry (BS) and Political Science (BA) from the University of New Mexico, where he worked with David Bear in Cell Biology and Physiology and Terran Lane in Computer Science. Aaron participated in the Santa Fe Institute's Complex Systems Summer School. His research has received numerous accolades, including recognition from the U.S. Department of Energy, National Science Foundation, and the Goldwater Scholarship Program.

Aaron enjoys being outside.