Structural and parametric identification of bacterial regulatory networks

Diana Stefan

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques et informatique**

Arrêté ministérial : 7 Aout 2006

Présentée par

Diana STEFAN

Thèse dirigée par **Hidde DE JONG**
et codirigée par **Hans GEISELMANN** et **Eugenio CINQUEMANI**

préparée au sein **de l'équipe IBIS, INRIA Grenoble - Rhône - Alpes**
et de **l'École doctorale Mathématiques, Science et Technologies de l'Information, Informatique (MSTII)**

Identification structurelle et paramétrique des réseaux de régulation bactériens

Thèse soutenue publiquement le **30 juin 2014**,
devant le jury composé de :

Dr. Giancarlo FERRARI TRECATE
Professeur associé, Université de Pavia , Rapporteur
Dr. Hubert CHARLES
Professeur, INSA Lyon, Rapporteur
Dr. William NASSER
Directeur de recherche, CNRS, INSA Lyon, Examinateur
Dr. Hidde DE JONG
Directeur de recherche, INRIA Grenoble-Rhône-Alpes, Directeur de thèse
Dr. Johannes GEISELMANN
Professeur, Université Joseph Fourier de Grenoble, Co-Directeur de thèse
Dr. Eugenio CINQUEMANI
Chargé de recherche, INRIA Grenoble-Rhône-Alpes, Co-Directeur de thèse

ACKNOWLEDGEMENTS

First and foremost, I want to present my sincere gratitude to my advisor Hidde DE JONG. His full support and significant contribution during my entire PhD, his availability at any time and his kindness and patience in teaching me great science and guiding the steps of my early carrier allowed me to successfully obtain my PhD degree. As well, thanks to Hidde, I am now emotionally attached to what is commonly only a work place, the INRIA laboratory.

My deepest acknowledgements and sympathy go to my second advisor Eugenio CINQUEMANI for his guidance, inspiring ideas and professional investment to make my PhD experience productive and stimulating and whose help had been essential for the completion of my work. I am honoured to have been his first PhD student and to learn from such a resourceful and enthusiastic teacher.

I would also like to thank my third advisor Johannes GEISELMANN for the opportunity he has given to me to learn and practice microbiology, allowing my work to become also my passion. I deeply appreciate the immense experience and pertinent advices he has brought to my thesis, but also his bright personality.

I would like to thank Delphine Ropers for all scientific discussions and practical advices during my PhD.

Many thanks go to Corinne PINEL, Stéphane PINHAL, Julien DEMOL, Stéphan LACOUR and Omaya DUDIN for their availability and for helping me with (not so obvious) biological experiments.

I am also grateful to Françoise de CONNINCK, Valentin ZULKOWER, Edith GRAC, Nils GIORDANO, the students of the IBIS team, the MISTIS team and the NANO-D team in INRIA for the scientific exchanges and for the pleasant moments spent at work and outside work.

Finally, I express my very profound gratitude to my fiancé, Bruno Roberto FRANCISCATTO, for providing me with unfailing support and constant encouragements throughout the years of my PhD. I also thank my heart friends, Aline, Flora, Vincent, Bertrand, Vitor and Alina for their backing and care during both good and difficult moments. My parents receive my warmest thanks as they have always been my enlightenment. Most important, they love me unconditionally. And they introduced me to science matters.

This work was supported by the Rhône-Alpes region (cluster ISLE, PhD grant to DS), the Investissements d'Avenir Bio-informatique programme under project RESET (ANR-11-BINF-0005), the INRIA/INSERM project ColAge, and the Agence Nationale de la Recherche under project GeMCo (ANR-2010-BLAN-0201-02).

RÉSUMÉ en français

L'interprétation des grandes quantités de données d'expression génique récemment générées par des techniques expérimentales à haut-débit exige des outils mathématiques et informatiques fiables pour l'inférence des interactions régulatrices. Nous nous intéressons à l'inférence des interactions régulatrices et l'amélioration des résultats de l'inférence en identifiant les informations précises fournies par les données expérimentales.

Nous avons développé une approche expérimentale et computationnelle intégrée pour l'inférence de modèles quantitatifs de promoteurs bactériens à partir des données d'expression génique temporelles mesurée par l'intermédiaire de gènes rapporteurs fluorescents. Nous montrons comment les effets physiologiques globaux et les concentrations de protéines peuvent être estimés à partir des données de fluorescence et intégrés dans des méthodes d'inférence, à la fois structurelle et paramétrique, des fonctions de régulation génique. Nous avons validé notre approche sur un module central dans le réseau de régulation contrôlant la motilité et le système de chimiotactisme chez *Escherichia coli*.

L'approche proposée est orthogonale aux méthodes déjà existantes pour l'inférence des réseaux de régulation à partir de données temporelles d'expression génique et peut être intégré avec plusieurs autres méthodes proposées dans la littérature.

MOTS CLÉS

Inférence des réseaux de régulation géniques, algorithmes d'identification structurelle et paramétrique, biologie des systèmes, modélisation des réseaux de régulation bactériens, estimation de paramètres

TITLE in english

STRUCTURAL AND PARAMETRIC IDENTIFICATION OF BACTERIAL REGULATORY NETWORKS

ABSTRACT in English

The interpretation of the large amounts of gene expression data yielded recently by high-throughput experimental techniques requires more reliable mathematical and computational tools for the inference of regulatory interactions. We focus on the inference of regulatory interactions and improving the results of inference by pinpointing the precise information provided by the experimental data.

We developed an integrated experimental and computational approach for the inference of quantitative models of bacterial promoters from time-series gene expression data measured by means of fluorescent reporter genes. We show how global physiological effects and protein concentrations can be estimated from fluorescence data and integrated into methods for the inference of both structural and parametric gene regulation functions. We validated our approach on a central module in the regulatory network controlling motility and the chemotaxis system in *Escherichia coli*.

The proposed approach is orthogonal to existing methods for regulatory networks inference from time-series gene expression data and can be combined with several other methods proposed in the literature.

KEYWORDS

Inference of genetic regulatory networks, algorithms for structural and parametric identification, systems biology, modelling of bacterial regulatory networks, parameter estimation

LONG ABSTRACT in English

High-throughput technologies yield large amounts of data about the steady-state levels and the dynamical changes of gene expression in bacteria. An important challenge for the biological interpretation of these data consists in deducing the topology of the underlying regulatory network as well as quantitative gene regulation functions from such data. A large number of inference methods have been proposed in the literature and have been successful in a variety of applications, although several problems remain.

We focus here on improving two aspects of the inference methods. First, transcriptome data reflect the abundance of mRNA, whereas the components that regulate are most often the proteins coded by the mRNAs. Although the concentrations of mRNA and protein correlate reasonably during steady-state growth, this correlation becomes much more tenuous in time-series data acquired during growth transitions in bacteria because of the very different half-lives of proteins and mRNA. Second, the dynamics of gene expression is not only controlled by transcription factors and other specific regulators, but also by global physiological effects that modify the activity of all genes. For example, the concentrations of (free) RNA polymerase and the concentration of ribosomes vary strongly with growth rate. We therefore have to take into account such effects when trying to reconstruct a regulatory network from gene expression data.

We propose here a combined experimental and computational approach to address these two fundamental problems in the inference of quantitative models of the activity of bacterial promoters from time-series gene expression data. We focus on the case where the dynamics of gene expression is measured *in vivo* and in real time by means of fluorescent reporter genes.

Our network reconstruction approach accounts for the differences between mRNA and protein half-lives and takes into account global physiological effects. When the half-lives of the proteins are available, the measurement models used for deriving the activities of genes from fluorescence data are integrated to yield estimates of protein concentrations. The global physiological state of the cell is estimated from the activity of a phage promoter, whose expression is not controlled by any transcription factor and depends only on the activity of the transcriptional and translational machinery. We apply the approach to a central module in the regulatory network controlling motility and the chemotaxis system in *Escherichia coli*. This module comprises the *fliA*, *flgM*

and *tar* genes. FliA is a sigma factor that directs RNA polymerase to operons coding for components of the flagellar assembly. The effect of FliA is counteracted by the anti-sigma factor FlgM, itself transcribed by FliA. The third component of the network, *tar*, codes for the aspartate chemoreceptor protein Tar and is directly transcribed by the FliA-containing RNA polymerase holoenzyme. The FliA-FlgM module is particularly well-suited for studying the inference problems considered here, since the network has been well-studied and protein half-lives play an important role in its functioning.

We stimulated the FliA-FlgM module in a variety of wild-type and mutant strains and different growth media. The measured transcriptional response of the genes was used to systematically test the information required for the reliable inference of the regulatory interactions and quantitative predictive models of gene regulation.

Our results show that for the reliable reconstruction of transcriptional regulatory networks in bacteria it is necessary to include global effects into the network model and explicitly deduce protein concentrations from the observed expression profiles. Our approach should be generally applicable to a large variety of network inference problems and we discuss limitations and possible extensions of the method.

RÉSUMÉ SUBSTANTIEL en français

Les technologies expérimentales à haut débit produisent de grandes quantités de données sur les niveaux d'expression des gènes dans les bactéries à l'état d'équilibre ou lors des transitions de croissance. Un défi important dans l'interprétation biologique de ces données consiste à en déduire la topologie du réseau de régulation ainsi que les fonctions de régulation quantitatives des gènes. Un grand nombre de méthodes d'inférence a été proposé dans la littrature. Ces méthodes ont été utilisées avec succès dans une variété d'applications, bien que plusieurs problèmes persistent.

Nous nous intéressons ici à l'amélioration de deux aspects des méthodes d'inférence. Premièrement, les données transcriptomiques reflètent l'abondance de l'ARNm, tandis que, le plus souvent, les composants régulateurs sont les protéines codées par les ARNm. Bien que les concentrations de l'ARNm et de protéines soient raisonnablement corrélées à l'état stationnaire, cette corrélation devient beaucoup moins évidente dans les données temporelles acquises lors des transitions de croissance à cause des demi-vies très différentes des protéines et des ARNm. Deuxièmement, la dynamique de l'expression génique n'est pas uniquement contrôlée par des facteurs de transcription et d'autres régulateurs spécifiques, mais aussi par des effets physiologiques globaux qui modifient l'activité de tous les gènes. Par exemple, les concentrations de l'ARN polymérase (libre) et les concentrations des ribosomes (libres) varient fortement avec le taux de croissance. Nous devons donc tenir compte de ces effets lors de la reconstruction d'un réseau de régulation à partir de données d'expression génique.

Nous proposons ici une approche expérimentale et computationnelle combinée pour répondre à ces deux problèmes fondamentaux dans l'inférence de modèles quantitatifs de promoteurs bactériens à partir des données temporelles d'expression génique. Nous nous intéressons au cas où la dynamique de l'expression génique est mesurée *in vivo* et en temps réel par l'intermédiaire de gènes rapporteurs fluorescents. Notre approche d'inférence de réseaux de régulation tient compte des différences de demi-vie entre l'ARNm et les protéines et prend en compte les effets physiologiques globaux. Lorsque les demi-vies des protéines sont connues, les modèles expérimentaux utilisés pour dériver les activités des gènes à partir de données de fluorescence sont intégrés pour estimer les concentrations des protéines. L'état physiologique global de la cellule est estimé à partir de l'activité d'un promoteur de phage, dont l'expression n'est contrôlée par

aucun des facteurs de transcription et ne dépend que de l'activité de la machinérie dexpression génique.

Nous appliquons l'approche à un module central dans le réseau de régulation contrôlant la motilité et le système de chimiotactisme chez *Escherichia coli*. Ce module est compose des genes *fliA*, *flgM* et *tar*. FliA est un facteur sigma qui dirige l'ARN polymérase vers les opérons codant pour des composants de l'assemblage des flagelles. L'effet de FliA est contrecarré par le facteur anti-sigma FlgM, lui-même transcrit également par FliA. Le troisième composant du réseau, *tar*, code pour la protéine récepteur chimiotactique de l'aspartate, Tar, et est directement transcrit par FliA associé à l' holoenzyme ARN polymérase. Le module FliA-FlgM est particulièrement bien adapté pour l'étude des problèmes d'inférence considérés ici, puisque le réseau a été bien étudié et les demi-vies des protéines jouent un rôle important dans son fonctionnement.

Nous avons stimulé le module FliA-FlgM dans une variété de souches de type sauvage et mutantes et dans des milieux de croissance différents. La réponse transcriptionnelle des gènes mesurée a été utilisée pour tester systématiquement les informations requises pour l'inférence fiable des interactions régulatrices et des modèles prédictifs quantitatifs de la régulation des gènes.

Nos résultats montrent que, pour la reconstruction fiable de réseaux de régulation transcriptionelle chez les bactéries, il est nécessaire d'inclure les effets globaux dans le modèle de réseau et d'en déduire de manière explicite les concentrations des protéines à partir des profils d'expression observés, car la demi-vie de l'ARNm et des protéines sont très différentes. Notre approche reste généralement applicable à une grande variété de problèmes d'inférence de réseaux et nous discutons les limites et les extensions possibles de la méthode.

Contents

1. Introduction

1.1 Context

1.1.1 Bacteria in their environment

Bacteria are single-cell organisms that, under favorable conditions such as nutrient abundance, can grow and divide rapidly. The number of bacteria on earth is estimated to be $4 - 6 \cdot 10^{30}$ (Whitman et al., 1998) and their total amount of carbon or biomass equals the estimated total carbon in plants. Typically few micrometers in dimension, these numbers are indicative of their pervasiveness in our world. Their contribution is everywhere enormous, in soils, water and air, recycling all types of matter. Bacteria not only inhabit human organisms, we even rely on them to metabolize and absorb essential nutrients, to fight pathogens, and to train and improve our immune function (Peterson et al., 2009; Turnbaugh et al., 2007).

Bacteria can adapt to many different environments, for example they survive extreme stress conditions associated with high concentrations of toxic metals or radioactive environments (Daly et al., 2004; Keller and Zengler, 2004; Schaechter et al., 2006). Recent studies of the human microbiome have begun to characterize the bacterial communities living in our gastrointestinal tract (Peterson et al., 2009). These communities thrive in the extreme conditions they have to deal with.

One of the microbes of the human gut flora, which can be grown easily and inexpensively in a laboratory setting, is *Escherichia coli*. The bacterium produces menaquinones (Conly et al., 1994; Suvarna et al., 1998), known as K2 vitamin, which plays an important and complex role in hemostasis. However, some serotypes have developed ways to thwart the immune system and can cause serious food poisoning in their hosts (Brown et al., 2008; Tenaillon et al., 2010). *Escherichia coli* has served for over 60 years as a model organisms for microbiology and biotechnology research (Keseler et al., 2013; Salgado et al., 2013). Recently, this bacterium has been engineered to create "living

materials", assembling rows of gold nanowires. The resulting network conducts electricity and could be worth exploring for use in energy applications, such as batteries and self-healing materials (Chen et al., 2014).

1.1.2 Responses of bacteria to changes in their environment

Bacteria have developed many original solutions to respond to often rapid changes in their environment. Bacterial stress responses enable cells to survive adverse and fluctuating conditions in their environment, such as the depletion of nutrients, changes in pH and temperature, high population density (Storz and Hennge-Aronis, 2000). When nutrients become exhausted, cell membranes become thicker for protection and cell division is dramatically reduced or turned off to prevent energy expenditure.

Another example of a bacterial stress response is motility (Berg, 2004). This is thought to be one of the most impressive evolutionary aspects of bacterial behavior, as it allows bacteria to populate regions rich in nutrients and to avoid repellents (Berg, 2004). Many bacterial species swim by rotating external filamentous organelle, known as flagella. The flagellum has three main components (as described by Berg (2004)), a basal body integrated in the cell wall (''the motor''), a short joint structure (''the hook'') and a long filament (''the propeller''). In *E. coli*, each flagellum is driven at its base by a reversible flagellar motor that is powered by a chemical gradient across the membrane and propels bacteria in a particular direction. When flagellar rotation is counterclockwise, the flagella push the bacterium forward, allowing a reasonably smooth "run". When the rotation is clockwise, the flagella pull in opposing directions and the bacteria "tumble" (Berg (2004) and Figure 1.1). Sensing chemicals is achieved through a complicated protein chemotaxis system (Porter et al., 2011) that controls the flagellar motor, allowing bacteria to migrate towards environments that are optimal for growth. If the movement is directed towards an attractant, the running period is prolonged and bacteria swim towards the attractant. On the contrary, if a repellent is encountered, the bacteria tumble, which prevents them from swimming towards the repellent.

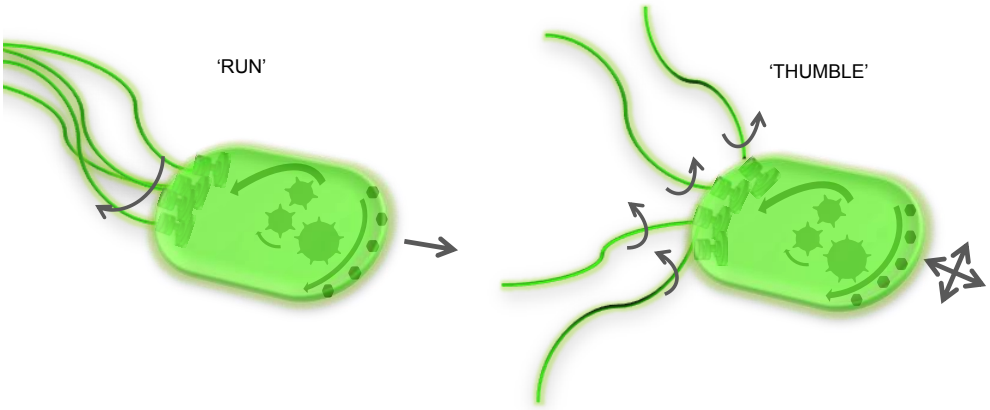

Figure 1.1: Bacterial motility. Most of bacterial species can move by rotating their flagella when sensing chemical gradients. The complex system coordinating movement and chemical sensing is the chemotaxis system. Cells switch continuously between two modes of movement: a "run" mode in which the bacterium swims forward, by rotating their flagella counterclockwise and a "tumble" mode in which the bacterium randomly changes direction, by rotating their flagella clockwise. The rotary motor located at the base of each flagellum controls the direction of rotation (Berg, 2004). The complex chemotaxis system, triggered by signals emitted by the transmembrane receptor proteins, dictates the direction of rotation of the flagella.

1.1.3 Responses controlled by complex regulatory networks

How does a bacterial cell initiate and coordinate its adaptive responses when changes in their environment occur? This is achieved by highly sophisticated, complex regulatory networks.

For example, the synthesis and function of the flagellar and motility system is based on the coordinated expression of more than 50 genes (Chilcott and Hughes, 2000). The expression of these genes responds to environmental stimuli and, in addition, to signals that are coupled to the morphological development of the flagella. The flagellar genes are organized in a transcriptional hierarchy of three operon classes (Chevance and Hughes, 2008; Kutsukake et al., 1990; Macnab, 1996a) as shown in Figure 1.2. The class 1 operon, *flhDC*, encodes the proteins FlhC and FlhD, which form a heteromultimeric complex initiating the transcription of the entire flagellar cascade through the class 2 operons. These operons encode the structural proteins required for flagellar

hook assembly as well as the main regulator of the class 3 operons, the sigma factor FliA (σ^{28}). When bound to core RNA polymerase, FliA directs the transcription of the class 3 operons, the lower level of the regulatory hierarchy. However, FliA activity is inhibited by the anti-sigma factor FlgM, which binds to FliA and thus prevents its association with RNA polymerase, delaying class 3 genes expression and completion of the flagellar assembly. FlgM is transcribed from both a class 2 and class 3 promoter and can be excreted once the intermediary hook basal body structure is constructed.

Cellular processes are very complex, but it seems that such processes can often be broken down into a limited number of recurring patterns of connectivity. The transcriptional cascade in Figure 1.2 can be decomposed into elementary network motifs, such as the SUM input FeedForwardLoop (Alon, 2007). That is, the master regulator FlhDC activates a second regulator, FliA, and both activate, in an additive fashion, the operons that produce the flagella motor. This specific network motif prolongs flagella expression following deactivation of the master regulator, highlighting the regulatory hierarchy and timing of the control of the flagella assembly (Kalir and Alon, 2004; Kalir et al., 2005).

1.1.4 Experimental measurements

Recent advances in molecular biology and in biophysics have led to new technologies for measuring cellular processes at the molecular level and in real time, including DNA microarrays and RNA sequencing, gene reporter systems, quantitative RT-PCR and mass-spectrometry based measurement of proteins. This allows the stress responses of bacteria to be monitored and may give insights into the functioning of the regulatory networks controlling these responses.

Measurements of the transcriptome of a bacterial cell by means of DNA microarrays or RNA sequencing produce quantitative information on the state of the entire transcriptional program of an organism at any time during an experiment (Dharmadi and Gonzalez, 2004). These approaches provide a relative quantification of mRNA abundance and cannot be obtained *in vivo*, in the sense that the actual measurements are carried out on molecules extracted from the cells. Techniques such as quantitative RT-PCR also measure relative mRNA concentrations, though usually not on a genome-wide scale (Saunders and Lee (2013), see White et al. (2011) for an exception). Fluorescent

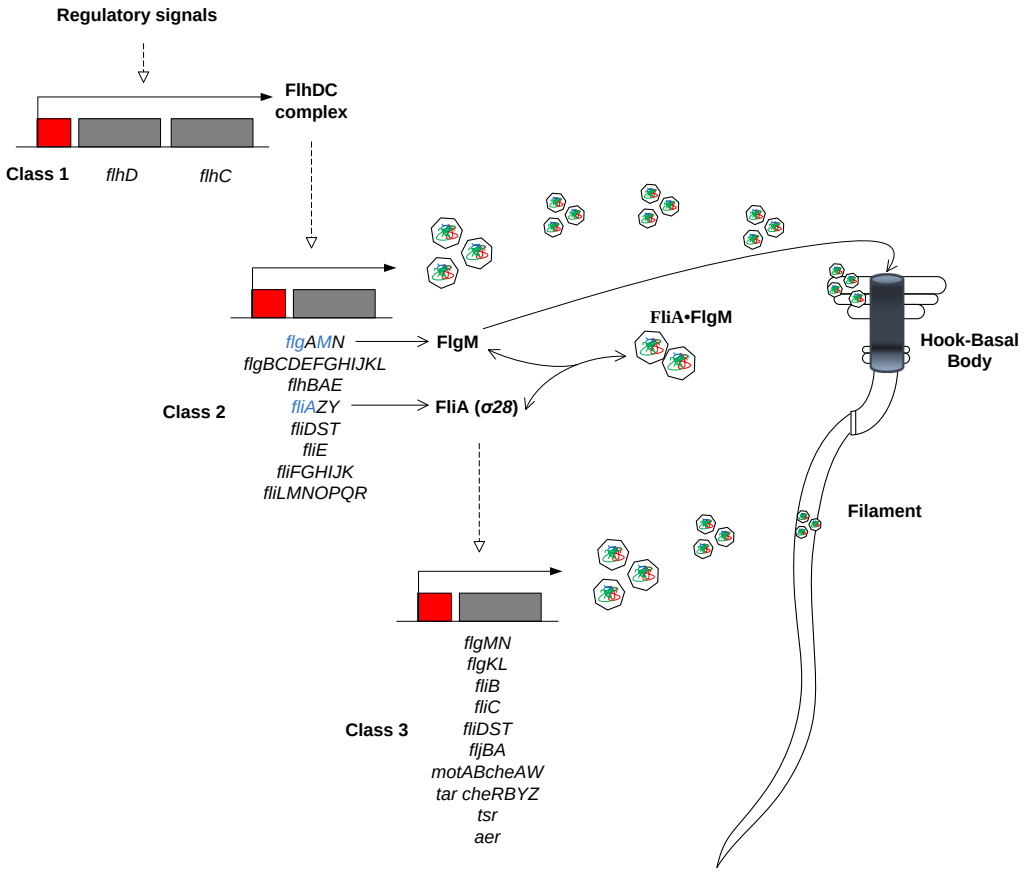

Figure 1.2: The hierarchical transcriptional network controlling flagellar assembly and the synthesis of the chemotaxis system (Karlinsey et al., 2000a). Class 1 genes encode the proteins FlhD and FlhC, transcriptional activators that are required for expression of all the other flagellar genes. Class 2 genes encode structural components of the Hook-Basal Body (HBB) complex, as well as regulatory proteins, including the sigma factor σ^{28} and the anti-σ^{28} factor FlgM. σ^{28} is required for the expression of all class 3 gene promoters, including those for flagellin and those related to chemotaxis and motility. FlgM binds to and inhibits σ^{28} until completion of the HBB. FlgM is then secreted from the cell, and σ^{28}-dependent (class 3) gene expression is initiated.

reporter genes allow measuring the activities of the promoter regulating the transcription of genes *in vivo* and in real time. Although quantitative proteomics has much advanced recently (Picotti and Aebersold, 2012), *in vivo* and real time measurements of proteins on a genome-wide scale are not yet possible.

Reporter gene technology, in particular, allows the direct observation/quantification of

the activity of genes in the cell. For motility gene data obtained by means of fluorescent reporter systems this reflects the expression of the different classes of genes in the flagellar cascade (Kalir and Alon, 2004). We can observe the precise temporal expression profiles of the most important genes involved in flagellar assembly. However, discovering the structure and functioning of the regulatory network is not directly possible from the raw experimental data.

1.1.5 Network inference

Computational tools are needed to uncover regulatory interactions from the large amounts of experimental observations, as well as to construct dynamic models of the functioning of bacteria. This problem is commonly defined as reverse engineering (Tegner et al., 2003), network reconstruction (MacCarthy et al., 2005) or network inference (Faith and Gardner, 2005; Gardner et al., 2003).

Different modeling formalisms have been proposed for inferring the topology of gene regulatory networks from gene expression data (de Jong, 2002; Hecker et al., 2009; Villaverde et al., 2013). These models represent regulatory networks in different ways, for example, as (oriented) graphs (Bayesian networks), discrete dynamic systems (Boolean networks) and continuous dynamic systems (differential equations). Specific inference algorithms for each of these formalisms have been developed, reconstructing the activity of one gene as a function of the activity of other genes. Although this has resulted in powerful methods, a number of recurrent problems remain (Marbach et al., 2010; Prill et al., 2010).

Examples of such issues are the dimensionality problem, or the problem of discriminating direct from indirect regulations. Dimensionality issues arise, e.g. when the number of variables (genes) is much larger than the number of experimentally observed quantities. This is typical for *in vivo* time course gene expression measurements for quantitative dynamic model inference and results in a multitude of regulatory structures consistent with the data. Moreover, measurements are typically noisy and do not capture the entire dynamics of the gene expression. Using additional constraints, biological background knowledge and clever simplifications can help analyze the prohibitively large space of possible solutions in a time-efficient manner (de Smet and Marchal, 2010). Second, a regulator, *e.g.*, a transcription factor may directly control

the expression of a target gene by fixing to its promoter region, but can also affect it indirectly, by regulating the expression of an intermediary gene whose product in turn regulates the expression of the target gene. Such direct and indirect regulatory influences are not easily distinguishable from gene expression data (Basso et al., 2005; Rice et al., 2005). This is thought to be due to the fact that the inference algorithms rely on specific assumptions about the underlying network topology (e.g. cyclic or acyclic network structures, including feedback loops or cascades). However, network motif analysis makes it possible to quantitatively assess how the difficulty of distinguishing direct and indirect connections affects inference methods (Marbach et al., 2010). These different problems may lead to systematic errors in predicting regulatory interactions and, as a result, compromise the performance of network inference algorithms (Marbach et al., 2010).
Several other fundamental problems of this sort exist in model inference. In my PhD work I have focused on two other recurrent problems in network inference that will be introduced in the next section.

1.2 Problem Statement: two problems in network inference

Usually, network inference algorithms rely on transcriptome data generated for instance, through DNA microarray analysis or RNA sequencing. These relative RNA concentrations characterize the transcriptome state of the cell as well as the activity of the promoters initiating the transcription of the genes. The problem of inferring regulatory interactions is that in general the active regulator is not mRNA but protein. At steady-state, mRNA and protein concentrations are relatively well correlated (Lu et al., 2007; Taniguchi et al., 2010). However, this is not expected to occur when the two vary over time. mRNA and protein have different half-lives and their concentrations evolve on different time-scales. For instance, in bacteria, mRNA half-lives are on the order of a few minutes (Bernstein et al., 2002), whereas most of the proteins are stable (Larrabee et al., 1980; Mosteller et al., 1980) and the degradation rate is dominated by growth dilution, taking place on the time-scale of a cell cycle. The effect of rapid responses in gene expression, taking place within a single generation, may thus persist

over several generations (Maier et al., 2009; Taniguchi et al., 2010). As a consequence of this temporal decorrelation of mRNA and protein concentrations, inference of regulatory networks relying exclusively on time-series transcriptome data may potentially lead to spurious results.

Another important problem derives from the fact that the dynamics of gene expression is not only controlled by transcription factors, small regulatory RNAs, and other specific regulators, but also by global physiological effects influencing the rates of transcription and translation of all genes (Berthoumieux et al., 2013b; Gerosa et al., 2013; Keren et al., 2013; Klumpp and Hwa, 2008). Most gene expression studies have been based on the assumption that cells produce similar levels of total RNA per cell, without including standardized controls that would reveal global transcriptional amplification or repression. For instance, cells can globally up-regulate their gene expression program, producing two to three times more total RNA and generating larger cells. In the conventional approach to expression analysis, normalized amounts of RNA would be introduced into the assay, thus masking changes in the activity of gene expression machinery. Ignoring such changes, for example in experiments with important variations of the growth rate, can lead to artefacts in inferring regulatory interactions (Lovén et al., 2012; Regenberg et al., 2006). Unfortunately, global physiological parameters characterizing changes in the activity of the gene expression machinery, such as the concentrations of (active) RNA polymerase and ribosome, are difficult to quantify in a direct way.

The aim of my PhD work is (I) to propose a combined experimental and computational approach to address the above two fundamental problems in the inference of quantitative models of regulatory interactions from time-series data and (II) the application of these methods to real data of gene expression from the regulatory network controlling motility in *E. coli*. This network has been well-studied and is therefore particularly suitable as a test-case. From a methodological point of view, several issues will be addressed during the project, such as the choice of appropriate modeling formalisms, the study and integration of time-series data in dynamic models, the design of informative experiments, the application of effective algorithms for model identification on real data, and the interpretation of the identification results to learn about the structure and functioning of the network. In order to generate a set of rich observations

of the system, I will experimentally probe the motility network in a variety of wild-type and mutant genetic backgrounds and in different growth media by means of fluorescent gene reporter measurements.

1.3 Principal research questions and approaches

1.3.1 Fluorescent reporter gene measurements

Our first research question is how to measure bacterial gene expression *in vivo* and in real time in order to quantify time-varying gene expression changes in the motility network. We have chosen to use fluorescent reporter gene techniques (de Jong et al., 2010; Giepmans et al., 2006; Southward and Surette, 2002), since they allow gene expression to be monitored with high precision and temporal resolution. Fluorescent reporter genes consist of transcriptional fusions of a gene encoding a fluorescent protein, e.g., GFP or mCherry, to the promoters of the target genes, on (low-copy) plasmids or on the chromosome. Chromosomal reporters avoid a number of potential artifacts, such as a change in plasmid copy number across different growth phases, but they are more difficult to construct and the intensity of the fluorescence signal may be close to the background fluorescence, especially when GFP is used. In this study, we have chosen to use plasmidic reporters of the motility genes, available in a reporter library (Zaslaver et al., 2006). The copy number of these plasmids was shown previously to be stable in media supporting different growth rates (Berthoumieux et al., 2013b; Zaslaver et al., 2006).

1.3.2 Reconstruction of protein concentrations and global regulatory effects

The experimental protocol allows monitoring the transcriptional response of a biological network. How can we reconstruct quantities of interest for our purpose from these data, notably protein concentrations and global physiological effects?
In the case of fluorescent reporter gene system, primary absorbance and fluorescence signals can be transformed into what are commonly called in the literature promoter activities (Zaslaver et al., 2006), using kinetic measurement models of gene expression (de Jong et al., 2010; Huang et al., 2008; Ronen et al., 2002; Wang et al., 2008). More

precisely, reporter gene data allow one to deduce protein synthesis rates and under certain conditions to consider them proportional to mRNA concentrations and promoter activities. These quantities reflect the transcriptional activity of the gene, but they can also be used to reconstruct the concentration of the protein, using information on the protein half-life (de Jong et al., 2010).

As explained in Section 1.2, large-scale differences in gene expression over time or across conditions may also reflect global changes in cellular physiology. The approach described in Berthoumieux et al. (2013b) allows the global state of the cell to be monitored in real-time and *in vivo* during the growth transition. I will therefore use a GFP reporter driven by a constitutive promoter, not regulated by any transcription factor, to assay the time-varying physiological state of the cell. For example, a plasmid expressing a GFP reporter for a phage promoter (Oppenheim et al., 2005), not regulated by any protein in the host cell, can provide this type of information. The variations in the activity of the constitutive promoter reflect changes in the overall physiological state of the cell, including the RNA polymerase and ribosome concentrations, as well as pool sizes of amino acids and nucleotides.

1.3.3 Transcriptional response of the FliA-FlgM module

We apply the experimental approach to a central module in the regulatory network controlling the synthesis of flagella and the chemotaxis sensing system in *Escherichia coli* (Chevance and Hughes, 2008; Kalir et al., 2001; Macnab, 1996a). This module comprises the FliA and FlgM transcription factors and their targets. FliA or σ^{28} is a sigma factor that directs RNA polymerase to operons coding for the flagellar filament and the chemotaxis sensing system controlling the flagellar motor. The effect of FliA is counteracted by the anti-sigma factor FlgM. As typical examples of FliA-dependent genes we study *flgM*, the gene encoding FlgM, and *tar*. The latter gene encodes the aspartate chemoreceptor protein Tar, which activates the flagellar motor component (Berry and Armitage, 2008; Macnab, 1996b). The FliA-FlgM module forms a checkpoint in the temporally-organized expression cascade. It is particularly well-suited for investigating the inference problems considered here, since the interactions in this network have been well-studied and protein stability has been found to play an important role in its functioning.

We will experimentally excite the FliA-FlgM module in a variety of wild-type and mutant conditions, in different growth media, and measure the transcriptional response of the genes. Promoter activities and protein concentrations will be computed for the *fliA*, *flgM* and *tar* genes, as well as the activity of the constitutive phage promoter pRM, to account for global regulatory effects.

1.3.4 Inference of quantitative models of the FliA-FlgM module

How can the data on the transcriptional response of the FliA-FlgM module be used to systematically test the information required for the reliable inference of the regulatory interactions (structure) and quantitatively predictive models (parameters) of gene regulation?

For the structural inference problem I will use a previously described inference method (Porreca et al., 2010a). I will notably test if the use of *fliA* and *flgM* promoter activities, instead of their protein concentrations, allows us to retrieve the expected pattern of regulatory interactions. Furthermore, I will assess to which extent the results can be improved when considering protein concentrations instead of promoter activities. In addition, analysis will consider the presence or absence of a model factor for the potential effects of global physiology. In order to compute protein concentrations we will use measured or estimated half-lives, while global physiological effects will be measured by means of a constitutive promoter.

In order to quantify to which extent the addition of the latter information improves the identification of quantitative models of promoter activity, I will construct kinetic models of the regulation of FliA-dependent genes. Using heuristic global optimization methods, such as genetic algorithms, I will then estimate parameter values from the data and asses the quantitative fit in the different conditions.

1.4 Contributions

The interpretation of the large amounts of data yielded recently by high-throughput experimental techniques requires more reliable mathematical and computational tools for the inference of regulatory interactions.

In this work, we have made explicit the relation between experimental data and physiological quantities by means of mathematical models of gene expression, calling into question two assumptions that are commonly made in the inference of regulatory interactions and quantitative gene regulation functions from time-series data. The first assumption is that transcriptome data alone are sufficient to capture the time-varying state of gene expression. Often, the regulators of gene expression are proteins, while mRNA and protein concentrations are not correlated in dynamic experiments. As a consequence, currently it is not possible to fully exploit the information contained in time-series transcriptome data (Marbach et al., 2010). A second implicit assumption in the analysis of transcriptome data is that gene regulation can be reduced to the action of transcription factors and other specific regulators. This ignores the fact that the activity of the transcriptional and translational machinery, as well as other global physiological effects, may drastically change over the course of an experiment, a fact that has been well-documented for microorganisms (Dennis et al., 2004; Scott and Hwa, 2011). This may lead to erroneous interpretations and the inference of spurious regulatory interactions (Lovén et al., 2012).

The main contribution of this thesis is an integrated experimental and computational approach for addressing the above two problems, in the context of time-series measurements of gene expression by means of fluorescent reporter genes. We notably show how global physiological effects and protein concentrations can be estimated from fluorescence data and integrated into methods for the inference of structural and parametric gene regulation functions. This work relies on solid results obtained previously. The reconstruction of protein concentrations from real-time promoter activities by means of kinetic models as well as the quantification of global physiological effects by means of reporter genes have been proposed before (Berthoumieux et al., 2013b; de Jong et al., 2010; Gerosa et al., 2013; Keren et al., 2013). For instance, Gerosa et al. (2013) have developed quantitative models to dissect global and specific regulation of *E. coli* genes involved in arginine biosynthesis (Gerosa et al., 2013).

To our knowledge, the work presented here is the first systematic study of how the integration of information on both global physiological effects and protein concentrations can improve the inference of regulatory interactions and the identification of gene regulation functions from time-series data. It is important to emphasize that the proposed approach is orthogonal to existing methods for the inference of regulatory networks from time-series gene expression data and can be combined with any of the large variety of methods proposed in the literature.

We have validated our approach by analyzing a central module of the motility network in *E. coli*. The FliA-FlgM module has been very well-studied and has characteristics that make it particularly suitable for our purpose, such as short half-lives due to export of certain proteins from the cell and proteolysis. The secretion and degradation rates change across conditions, depending on the strength of induction of the flagella synthesis network. As a consequence, the FliA and FlgM concentrations are expected to vary during the course of an experiment and across the experimental conditions. This yields a rich and challenging data set for testing how accounting for the distinction between cellular responses on the level of mRNA and protein influences the results of the inference process.

Furthermore, we use the reporter gene data not only for deducing the regulatory structure but also for quantifying the regulation function of two FliA-dependent motility genes, not known to be regulated by any other transcription factors. When progressively solving the problems mentioned above, by integrating information about the activity of the gene expression machinery and computing estimates of protein concentrations from promoter activities, both the structure and the dynamics of the regulation of the *tar* and *flgM* promoters could be identified successfully. We emphasize that, when using available measurements of FliA and FlgM half-lives, this was achieved without increasing the number of parameters in the models and is therefore not simply a consequence of increasing the degrees of freedom. The results confirmed the important roles played by global physiological effects and the active regulation of FliA and FlgM half-lives in shaping the dynamics of FliA-dependent promoters.

We believe that the approach proposed in this work has broad practical applicability for exploiting and analyzing transcriptome data and improving network inference in a variety of organisms.

1.5 Thesis overview

The manuscript of this thesis will be organized as follows:

- Chapter 2 (State of the art) will review existing inference methods for dynamical models. We will also describe the experimental techniques that can be currently used to monitor gene expression data in microbial cells. Finally, data analysis methods that allow the transformation of reporter gene data into biologically relevant quantities will be presented.

- Chapter 3 (Results) will present the results obtained during this thesis. We will investigate how reconstructing protein concentrations from promoter activities and monitoring the global physiological state of the cell may improve the structural and parametric inference of gene regulatory networks. The approach will be exemplified by means of a central module of the motility network in *E. coli* . A paper containing the work presented in this chapter is in preparation.

- Chapter 4 will summarize the conclusions drawn from the current work and will present perspectives and future improvements for the inference of quantitative models of regulatory networks in biology.

2. State of the Art

Inference of gene regulatory networks generally deals with the problem of reconstructing interactions among genes from experimental data. Different types of data exist, and correspondingly, different interaction models and different methods for their inference are most adapted to the data considered. Depending on the context, the problem goes under alternative names, notably reverse-engineering and network identification.

In this chapter we present the various methods employed in the literature for inferring gene regulatory networks. In Section 1 we introduce common experimental techniques that provide high-throughput measurements of regulatory molecules such as mRNAs or proteins. Then, we present the inference algorithms that have been developed to exploit these experimental data, their strengths and weaknesses. In the last section (Section 3), we address the data processing methods and the development of measurement models for reporter gene techniques.

2.1 Experimental techniques for measuring gene expression

Measurements of the transcriptome and proteome of bacterial cells by means of DNA microarrays, RNA sequencing, and other high-throughput or quantitative technologies have created huge amounts of data on the state of the transcriptional program in different growth conditions and genetic backgrounds, over the time course of an experiment. Large efforts have been made to develop such experimental methods. This section offers a short overview of the experimental technologies as well as examples of their application.

2.1.1 High-throughput transcriptomics

Transcriptomics allows the monitoring of the genome-wide transcriptional response of the cell to an environmental stimulus or a genetic modification.

The most common technique used for producing data for the inference of gene networks are microarrays. There are several types of microarrays, but DNA microarrays are by far the most widely used.

The principle of DNA microarrays is based on the complementarity of messenger RNA (mRNA) and the DNA strand from which it is transcribed. Therefore mRNA will bind to single -stranded DNA molecules with the same sequence as the originally transcribed gene.

To determine gene expression patterns in a cell, the messenger RNA molecules in a sample are extracted. Each mRNA molecule is then reverse transcribed by a reverse transcriptase (RT) thus generating a complementary DNA (cDNA). In addition, the cDNA is labeled in the process, typically using a fluorescent nucleotide. Next, the labeled cDNAs obtained from cellular target mRNA are added to the microarray where they hybridize with their complementary single-stranded DNA probe fixed on the microarray substrate (Dharmadi and Gonzalez, 2004). The fluorescence intensity measured at a particular spot in the array reveals the amount of the gene transcript which was present in the sample. Many choices of DNA microarray platforms are available, such as cDNA microarrays or oligonucleotide microarrays (Lockhart et al., 1996; Marshall, 2004; Schena et al., 1995). Starting with the dot-blots (Southern et al., 1999), DNA microarrays have evolved to filter arrays (nylon membrane support), and to a glass slides format, which has the advantage of high-probe density (Dharmadi and Gonzalez, 2004).

Informative high-throughput datasets have been obtained by means of DNA microarrays. Many studies on bacterial genetics make use of these datasets, especially for characterizing bacterial responses to environmental changes, trying to explore the highly complex regulatory networks and transcriptional regulation or genetic and metabolic engineering (see review of Dharmadi and Gonzalez (2004)). For instance, global transcriptional profiling of *E. coli* in acetate cultures was investigated using DNA microarrays on glass slides (Oh et al., 2002). Other examples of the use of DNA microarrays datasets for *E. coli* can be found in the literature, such as the identification of regulatory networks (Faith et al., 2007) or metabolic engineering (Park et al., 2005).

A limitation of DNA microarray technology is that it only measures the relative concentration of mRNA. Moreover, while microarrays have contributed to our understanding

of transcription regulation (van Vliet, 2010), they have a limited precision of measurements due to variability from the array fabrication process, systematic errors in the hybridization process and heterogeneity of experimental design procedures (Dharmadi and Gonzalez, 2004). This also increases the difficulty of combining and interpreting different available microarray datasets (Bloom et al., 2009).
While DNA microarrays thus have limitations in their applicability, sequencing technologies recently became available for the detection and quantification of transcripts in microorganisms. In addition to measuring mRNA levels, RNA-seq technology has also been used for identifying small regulatory RNAs (Waters and Storz, 2009). For example, Perkins et al. (2009) in a study on *Salmonella enterica* serovar Typhi, used RNA-seq information to identify novel noncoding RNA sequences and new members of regulons. A limitation of current sequencing technologies is that the mRNA has to be extracted from a bacterial sample and measurements cannot be made *in vivo* or used to investigate single-cell dynamics. If one wants to evaluate a dynamic response, sampling time will become critical for the identification of changes in gene expression.

2.1.2 Quantitative proteomics

DNA microarrays are used to estimate mRNA levels, a molecule that rapidly changes its concentration in response to regulatory signals. However, the biologically active regulator is usually not mRNA, but protein (Cox and Mann, 2011). Although in steady-state condition proteins and mRNA concentrations are moderately correlated (Lu et al., 2007; Taniguchi et al., 2010), if one wants to evaluate changes in gene expression dynamically over time, this is not expected to be the case. Mass spectrometry (MS)-based proteomics has emerged as an universal method for the measurement of proteins (Bensimon et al., 2012; Schmidt et al., 2009; Wepf et al., 2009). Protein samples are extracted from cells and digested into peptides. The resulting peptide mixture is separated, typically by high-performance liquid chromatography (HPLC) and converted to gas phase ions by using the electrospray (Fenn et al., 1989) or matrix-assisted laser ionization (Hillenkamp et al., 1991) methods. Next, the mass spectrometer scans the entire mass range every few seconds. The data analysis software then isolates the selected peptides in the mass spectra, fragments them and measures the mass spectra of each of the fragments with a high resolution. Peptide-based proteomics does

not directly identify proteins, but reconstructs them from the obtained mass spectra of the fragments (Nesvizhskii and Aebersold, 2005). Using this methodology, the first complete proteome quantification for yeast has recently been published (de Godoy et al., 2008). One of the main perspectives is to combine proteomics and data provided by other high-throughput technologies in order to create comprehensive datasets (Cox and Mann, 2011). Ishii et al. (2007) integrated proteome with transcriptome and metabolome data for the investigation of responses of *E. coli* cells to environmental and genetic perturbations.

Although proteomics can determine the absolute amount of each of the proteins in a sample or their relative change between several conditions, some major limitations remain. Efforts have to be made to increase its throughput compared to other large-scale technologies, by reducing measuring time on improved instruments (Picotti et al., 2010; Reiter et al., 2011). Furthermore, special attention needs to be given to the development of simplified sample preparation protocols and computational analysis software. As for high-throughput techniques, protein extraction from bacterial samples and the long processing times do not allow direct *in vivo* and real-time measurements (Picotti and Aebersold, 2012).

2.1.3 RT-qPCR

Real-time polymerase chain reaction (RT-PCR) is a recent technology developed for molecular biology and medicine, based on DNA labeling dyes to quantify changes in RNA expression levels. Quantitative RT-PCR has become one of the most popular methods for the analysis of gene expression and verification of microarray results. While high-throughput microarray analysis allows large-scale analysis of gene expression profiles, the reverse transcription (RT) followed by the polymerase chain reaction (PCR) are often used to validate their findings (Bustin and Nolan, 2004; Bustin et al., 2013). In addition, relative quantification or absolute quantification of gene expression compared to standards that are run in parallel can be performed.

RT-qPCR protocols consist of several steps. First, RNA is isolated from sample cells, then mRNA is reverse-transcribed to cDNA. Second, the synthesized cDNA is amplified using specific PCR primers for the gene of interest. The PCR reaction also contains a fluorophore that specifically binds to double-stranded DNA, which allows

real-time monitoring of the amplification reaction. The fluorophores that are currently available for quantitative PCR can be broadly classified as non-specific and sequence-specific DNA-associating dyes (Mackay, 2007; Saunders and Lee, 2013). The most commonly used reporter is the SYBR Green dye, a non-specific probe. Once it is bound to the double-stranded DNA, it emits a fluorescent signal increasing with the accumulation of PCR products. The other specific reporters such as TaqMan rely on Fluorescence Resonance Energy Transfer (FRET) from the dye molecule to the quencher, both attached to a specific oligoprobe (Holden and Wang, 2008). When the oligoprobe hybridizes with its template DNA, the fluorophores (dye molecule and quencher) are released and fluorescence emissions can be detected.
In conclusion, RT PCR has been mostly used for microarray data validation, but also for the absolute and relative quantification of the number of plasmid copies in Lee et al. (2006). However, the method uses cell lysate and does not allow *in vivo* measurements. Moreover, the technology is not easily parallelized, therefore it is only used for detailed study and validation of results. In addition, preparation of mRNA involves additional steps, may lead to the loss of some initial mRNA, and it is more difficult to assess the quality of the final product (Saunders and Lee, 2013).

2.1.4 Reporter genes

Current reporter gene technologies, based on Green Fluorescent Proteins (GFPs) (Southward and Surette, 2002) and other fluorescent and luminescent reporter proteins, provide an excellent means to measure gene expression *in vivo* and in real time, in contrast to the other techniques presented in this chapter. The underlying principle of the technology is the fusion of the promoter region and possibly (part of) the coding region of a gene of interest and a reporter gene. The expression of the reporter gene generates a visible fluorescence or luminescence signal that is proportional to the actual number of fluorescent molecules (Rosenfeld et al., 2006), it is easy to detect and reflects the expression of a gene of interest when its promoter is activated. Reporter gene constructions have enabled the real-time tracking of gene expression dynamics in single cells, which provides valuable information about the functioning of cells (Golding et al., 2005; Longo and Hasty, 2006). Blue, yellow, cyan, and red fluorescent proteins have been engineered, allowing the expression of several genes to be studied in parallel, in the

same cell. For example, three fluorescent reporter genes (encoding cyan, yellow and red fluorescent protein) were used to investigate how noise is transmitted in a gene network (Pedraza and van Oudenaarden, 2005).

Reporter constructs can be located on plasmids or on the genome, depending on the

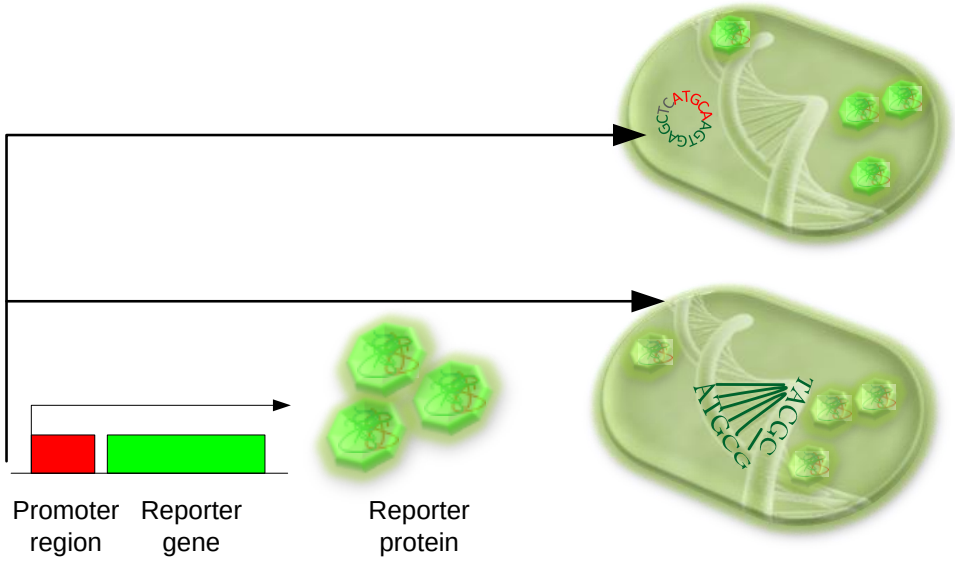

Figure 2.1: Reporter gene systems. The promoter region of a gene of interest is fused with a reporter gene. When the promoter is activated, the reporter gene is transcribed and produces the reporter protein as a direct quantification of the strength of the promoter. The reporter system can be introduced in cells on a plasmid or directly integrated into the chromosome.

problem studied (de Jong et al., 2010). Plasmidic reporters have the advantage of being easy to construct and generating a strong signal, compared to chromosomal constructions, but their copy number may change with the experimental conditions (Lin-Chao and Bremer, 1986), thus introducing biases in the interpretation of the data. By measuring the plasmid copy number in a cell with qRT-PCR technology, one can correct this bias (Berthoumieux et al., 2013b). In addition, protein concentrations can be reconstructed from fluorescent reporter gene measurements by means of existing data processing methods (de Jong et al., 2010), when information about proteins half-lives is available.

Many studies use fluorescent reporter gene data to quantify gene expression in *E. coli* .

For example, (Ronen et al., 2002) used fluorescent reporter genes to investigate its SOS DNA repair system. Kalir et al. (2001, 2005) analyzed the dynamics of the gene network regulating flagella motif. To create a tool for accurate, high-resolution analysis of transcription networks, Zaslaver et al. (2006) constructed a library of transcriptional fusions of *gfp* to the intergenic regions containing promoters, in *E. coli* K12 strain, on a low-copy plasmid. Dynamic measurements have been obtained using this library on a genomic scale, in a diauxic shift experiment (Zaslaver et al., 2006).

2.2 Inference of gene regulatory networks

Existing methods for gene network inference can be classified in different ways, depending on the criterion used for the classification. Several review papers explain the principles of the different inference methods. In Faith and Gardner (2005), inference methods are divided into 'physical interaction' approaches, that aim at identifying interactions among transcription factors and their target genes, and 'influence interaction' approaches, that try to relate the expression of a gene to the expression of other genes. Stelling (2004) and Doyle III and Stelling (2006) classify inference methods based on the type of models they pertain, distinguishing between interaction-based, constraint-based and mechanism-based models. A relevant classification of modelling and estimation approaches is provided in de Jong (2002). Discussions of modelling methods and details on the computational aspects can be also found in Beer and Tavazoie (2004), Ambesi and Bernardo (2006) and Markowetz and Spang (2007). A mixed classification of algorithms based on methodological approach, modelling context and performance on different inference problems results from the DREAM challenge (Dialogue for Reverse Engineering Assessments and Methods), a large and cooperative effort toward the assessment of inference performance (Greenfield et al., 2010; Marbach et al., 2010, 2012) in the form of a competitive game among participant methods.
In what follows we will review gene network inference algorithms associated with different interaction modelling formalisms, mostly following the recent review by Bansal et al. (2007). In every case we will discuss the main approaches proposed in the literature along with their advantages and limitations.

2.2.1 Inference of gene clusters

A first approach to gene network inference is to try group genes that may be functionally related. Genes whose expression appears to be altered in a coordinated manner in response to one or more experimental perturbations are grouped together and considered to be related to the cell function that has been probed experimentally. Typically in this context, gene expression screening is performed by way of microarray experiments, and functional gene grouping is operated by way of statistical methods called clustering.

Clustering relies on the concept of similarity among expression patterns or profiles (Eisen et al., 1998). As a similarity metrics, a correlation coefficient, most commonly the Pearson coefficient, is used:

$$r_{ij} = \frac{\sum_{k=1}^{M} (x_i(k)x_j(k))}{\sqrt{\sum_{k=1}^{M} (x_i^2(k)) \sum_{k=1}^{M} (x_j^2(k))}} \tag{2.1}$$

where x_i and x_j are gene expression measurements taken in M different conditions and r_{ij} is the pairwise correlation coefficient computed between gene i and gene j. All pairwise correlation coefficients r_{ij} for all possible gene pairs ij are computed for a set of n profiles. If the expression patterns of two genes are perfectly correlated (i.e., they are identical up to shifting and scaling) then $r_{ij} = 1$; in the opposite situation, when the variables are linearly independent, $r_{ij} = 0$. Correspondingly, gene pairs with large enough correlation coefficients are deemed to be functionally related (e.g. above a suitable threshold between 0 and 1). Based on this principle, suitable methods for grouping genes into similarity clusters have been developed (Amato et al., 2006; Eisen et al., 1998). Gene clusters obtained in this way can be seen as fully connected subgraphs of the graph with all genes as the graph nodes. Links among nodes within a cluster are thus functional relationships and, although no direct mutual regulation among clustered genes is implied by this grouping, clustered genes are often presumed to undergo some form of mutual interaction. A common subsequent analysis step is to annotate each cluster with a functional category representative of that cluster and use this categorization for further inference (Guthke et al., 2005).

A common issue in gene clustering is the choice of the policy for associating genes to one cluster or another. In particular, the number of clusters is a priori unknown and can be chosen automatically based on data or manually, depending on the clustering algorithm used (Amato et al., 2006; Eisen et al., 1998).
Similarity measures alternative to (2.1) have also been proposed, and the choice has important effects on the clustering results. However, validation of results still remains an opened issue(Allison et al., 2006; Handl et al., 2005). Most current clustering algorithms do not provide estimates of the significance of the results returned. The validation of clustering results is therefore often based on a manual and subjective exploration process, such as visual inspection and prior biological knowledge to select what is considered the most "appropriate" result. Recently, clustering has been applied to metabolic pathways analysis (Milone et al., 2014), incorporating prior knowledge into the cluster formation itself and show important improvements in the convergence and performance.
Most clustering methods developed for gene networks (D'haeseleer et al., 2000; Stuart et al., 2003) are derived from hierarchical clustering (Eisen et al., 1998). The clustering algorithm developed by Eisen et al. (1998) has been applied to a variety of systems such as in Spellman et al. (1998) to identify cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* from microarray data, and in Bansal et al. (2007) to recover network structure from steady-state and time-series data by both *in silico* and experimental data analysis. Although the accuracy in identifying correct interactions was fairly low, based on a large dataset (*S. cerevisiae* steady-state dataset, see Bansal et al. (2007)) some known interactions were indeed recovered. Newer, sophisticated methods such as distance correlation have theoretical advantages over Pearson coefficient (Székely and Rizzo, 2009) and have been tested on protein networks (Roy and Post, 2012).
Recently, methods for clustering have been widely coupled to more complex inference algorithms to reduce the dimensionality of the search space before network inference (cMonkey, Reiss et al. (2006)).

2.2.2 Inference of interaction graphs

Clustering approaches based on the Pearson coefficient may be effective for linear correlations, but their performance decreases for nonlinear systems (Villaverde and Banga,

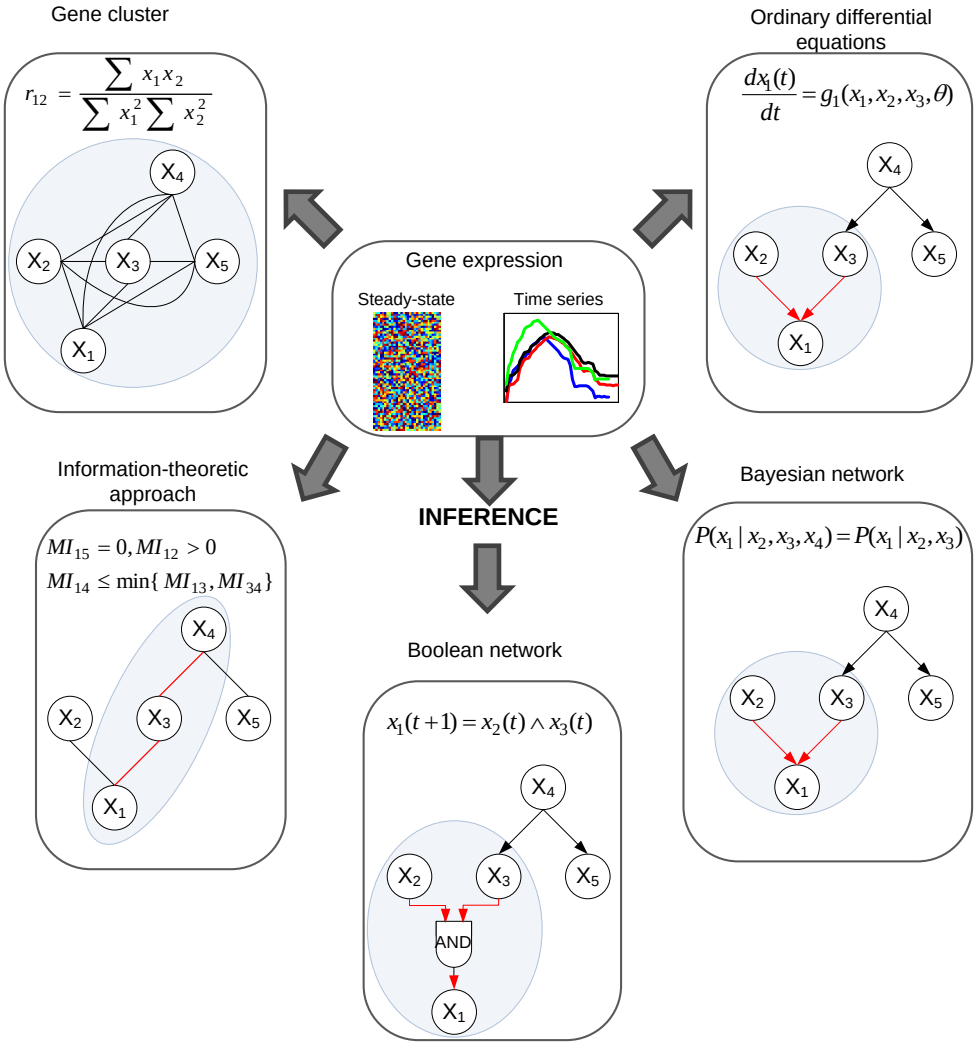

Figure 2.2: Approaches for inferring gene regulatory networks. In gene clusters, all dependencies between genes are typically assessed by Pearson correlations. In information-theoretic networks, MI is 0 for statistically independent variables and DPI is used to select direct regulatory interactions. In Boolean networks, the state of a gene is computed as a simple Boolean rule from the activities of other genes. Bayesian networks employ probability distributions to determine regulatory effects between genes. In the case of ODE models, the activity of one gene is computed as a function (g) of the level of its regulators. (Inspired by (Bansal et al., 2007))

2014). In addition, being the focus on functional relationships, no distinction is usually made between direct and indirect interactions. A related problem is that of inferring graphs of interactions, where nodes are genes but edges represent actual regulatory

effects between gene pairs. To discriminate between direct and indirect interactions, rather than looking at Pearson correlation (Eq. 2.1), information-theoretic concepts such as Mutual Information (MI, Shannon (1948)) can be employed instead. Based on this, in general agreement with literature terminology, we will discuss inference of interaction graphs in terms of methods based on an information-theoretic approach. The underlying principle of information-theoretic approaches is the concept of entropy, describing the uncertainty of a single random variable:

$$H_i = -\sum_{x_i \in \chi} p(x_i) log(p(x_i)) \tag{2.2}$$

which captures the a priori variability of the expression of the gene, and more generally, the joint entropy

$$H_{ij} = -\sum_{x_i \in \chi} \sum_{x_j \in \chi} p(x_i, x_j) log(p(x_i, x_j)) \tag{2.3}$$

which quantifies the variability of the random variables involved. Mutual information, MI_{ij}, between gene i and gene j is computed as:

$$MI_{ij} = H_i + H_j - H_{ij} \tag{2.4}$$

It is a measure of dependencies between variables, the higher the value, the stronger the mutual dependency. If two variables are statistically independent, their joint entropy is $H_{ij} = H_i + H_j$ (i.e. $p(x_i x_j) = p(x_i)p(x_j)$) and the mutual information is zero. Conversely, full dependency corresponds to $H_{ij} = 0$. In practice, these quantities can be computed e.g. from microarray gene expression profile data (Butte and Kohane, 2000), see further below. Similar to the correlation approach, choosing a MI threshold provides a mechanism to identify potential regulatory interactions, the smaller the threshold, the higher the hit rate at the price of more probable false hits.

Different from the correlation approach, use of the so-called data processing inequality allows one to prune many indirect interactions among genes. The data processing inequality states that if genes (i, k) interact indirectly through j, and no alternative path exists between genes (i, k), then $MI_{ik} \leq \min(MI_{ij}, MI_{jk})$. Thus, when the data processing inequality is verified, indirect interactions can be excluded from the set of estimated interactions in favor of direct interactions. A well known example of inference method based on the DPI is ARACNE (Basso et al., 2005; Margolin et al., 2006). In ARACNE, MI_{ij} is estimated from data for all pairs of observed genes i and j by

the use of a Gaussian kernel (Steuer et al., 2002). Based on Montecarlo simulations, a threshold MI_0 corresponding to a suitable p-value for testing the independence of Gaussian variables is determined. All gene pairs with mutual information below this threshold are considered independent. Putative interactions are then checked for false positives, i.e. dependence relationships that are not direct are sought based on the DPI. This allows elimination of indirect interactions but may also eliminate direct interactions (Margolin et al., 2006), depending on the data and the tuning of certain algorithm parameters. Automatic choice of MI and DPI threshold parameters suggested along with ARACNE gives a good sensitivity-accuracy compromise (Bansal et al., 2007).
In order to further improve the accuracy vs. sensitivity tradeoff, other approaches have been investigated, leading e.g. to the design of minimum redundancy (Meyer et al., 2007), entropy reduction (Samoilov, 1997; Villaverde et al., 2013) and continuous three-way mutual information (Luo et al., 2008) methods.
Since M is symmetric ($M_{ij} = M_{ji}$) and mutual information does not provide any information about directionality of regulation, the interaction network is still reconstructed in the form of an undirected graph G, where edges represent statistical dependencies observed in the data, but do not carry information about causality of regulation. In order to establish causal relationships from the inferred associations between interacting nodes, i.e. discriminate between regulatory and target genes, additional information is necessary. Toward this aim, one example is Context Likelihood of Relatedness (CLR) algorithm (Faith et al., 2007) where a distinction is made between the roles of transcription factors and target genes. This algorithm includes an adaptive background correction to reduce false corrections and indirect influences. CLR calculates the statistical likelihood of each MI value and then compares the MI of a transcription factor/gene pair to the background distribution of all possible transcription factor/gene pairs that include either the transcription factor or its target. Retaining only regulations whose associated MI is significantly higher than the background values, apparent interactions stemming from e.g. activity of one transcription factor weakly varying with the expression of several genes, or expression of one gene weakly varying with the activity of several transcription factors, are filtered out of the reconstructed network.
In practice, the ability to estimate the MI between genes from experimental data, and hence the practicality of the methods discussed above, depends on the quality of the data, in particular, on the relevance of certain statistical independence assumptions.

Estimation of MI can be performed from multiple independent steady-state datasets as well as from dynamic data, as long as the sampling time is long enough to consider subsequent measurements statistically independent (Bansal et al., 2007). Computation of MI indices then involves the probability mass function $p(x)$, which is typically unknown in reverse-engineering problems and needs to be reconstructed from the data. This is often done by histogramming, i.e., partitioning the data into equally sized bins and then counting the frequency of appearence of the data in every bin (Steuer et al., 2002). This step is important and may affect the gain in performance from Pearson correlation-based methods, quantifying only linear dependencies, and mutual information-based methods (Steuer et al., 2002).

In real-world applications, ARACNE has been shown to perform well on large and medium size steady-state datasets (human Bcells and *S. cerevisiae*, see Bansal et al. (2007)), yielding results similar to clustering algorithms. On small time-varying datasets, ARACNE had a poor performance, as expected because of its requirement of statistically independent time-points (Bansal et al., 2007). Therefore, a new version of ARACNE better suited for time-course data has been recently developed (Zoppoli et al., 2010).

Using a compendium of microarray expression profiles in *E. coli*, CLR was used in (Faith et al., 2007) not only to reconfirm known regulations, but also to discover several novel interactions, some of which were validated experimentally. Still, direction of regulation is well defined only for interactions discovered between one transcription factor and one target gene, while it is undefined for interactions found between transcription factors. In combination with other methods, CLR was reported to be an effective inference algorithm for the DREAM4 100-gene *in silico* network inference challenge (Greenfield et al., 2010).

Thus, despite the variously successful applications and the development of methods such as CLR, the main challenge of the approaches developed in this framework remains that of establishing the causality of interactions. In the next section, a framework that allows for a natural treatment of the direction of regulations is discussed.

2.2.3 Inference of Boolean networks

Boolean networks are discrete network models (or logical networks). Boolean networks were first described by Kauffman (1969) and were later on recognized as a natural framework for gene regulation modelling (Bornholdt, 2008; Kauffman et al., 2003; Thomas, 1973). The state of a gene (nodes in the network) can be approximated or described by a Boolean variable $x_i \in \{0, 1\}$. The gene can be "inactive"(0) or "active"(1) (de Jong, 2002). Boolean networks are usually represented as directed graphs, where the edges (interactions between nodes) are represented by activation/inhibition Boolean functions. Through these functions, the state of a gene is determined on the basis of the states of its parent genes by applying basic Boolean operations (AND, OR, NOT). Boolean networks can capture the dynamics of a regulatory system on a discrete time grid as follows:

$$x_i(t+1) = g_i(x_1(t), ...x_k(t)) \tag{2.5}$$

where $x_i(t)$, the state of the gene i at time t changes to a state, $x_i(t+1)$ at time $t+1$ following Boolean function g_i. Each gene can be active or inactive, so the state space consists of 2^N states. The boolean function uses states for k nodes (regulators), so the number of possible functions is 2^{2^k} (de Jong, 2002). Although rapidly increasing with the number of genes in the network, the state space and the number of interaction functions is finite. Therefore, Boolean networks evolve towards a steady state or a cycle of states, called *attractor* (Klipp et al., 2009). Reverse engineering a Boolean network means finding a Boolean function for each gene in the network such that the observed discrete data are explained by the logics of the model. REVEAL (REVerse Engineering ALgorithm, Liang et al. (1998)) is one of the various algorithms used for the inference of Boolean networks. REVEAL uses mutual information to identify a reduced set of inputs that describe the activity of an output gene and then determines the Boolean interaction functions from the data. The algorithm was shown to perform well on a network of 50 genes, each gene having up to 3 regulators, when using for the analysis only the reduced state transitions collection (100 out of 10^{15}).

At a closer look, Boolean networks cannot fully describe real gene expression profiles by their binary abstraction, therefore Boolean networks are inherently limited. As well, there are important behaviors that cannot be modeled using Boolean framework, such as amplification or addition of signals. Probabilistic Boolean networks (Akutsu

et al., 2000; Shmulevitch et al., 2002) have been developed as an alternative to classic deterministic Boolean networks to deal with data uncertainty, allowing more than one possible state transition Boolean function to describe gene interaction from noisy expression data. Nevertheless, Boolean networks provide a good qualitative interpretation of gene regulation and can be used to simulate gene regulatory functions (Hecker et al., 2009).

2.2.4 Inference of Bayesian networks

Bayesian Networks (BN) are probabilistic graph-based methods characterizing the expression of every gene i in a regulatory network by a random variable X_i. The interactions among these are represented as joint probability distributions $P(X_1, ..., X_N)$ and encoded in the structure of a directed acyclic graph G, whose nodes are the random variables X_i. The joint probability density is expressed as a product of conditional probabilities by applying Bayes' theorem: $P(A, B) = P(B \parallel A) * P(A) = P(A \parallel B) * P(B)$. This allows one to write

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(X_i = x_i \parallel X_{j_1} = x_{j_1}, ..., X_{j_p} = x_{j_p}). \tag{2.6}$$

The p genes (with p generally depending on i) that appear in the conditional probability for gene i represent the actual regulators of gene i, i.e. the directed edges of the associated graph G (also called the parents of node i). A crucial assumption is that the model obeys a Markov property stating that each variable X_i is conditionally independent of its non-descendants given its parents. If this causality condition does not hold, similar to information-theoretic approaches, the edges of the BN graph no longer represent direct causal interactions but only statistical dependencies (Bansal et al., 2007).

Inferring a Bayesian network model means finding the directed acyclic graph G that best explains a gene expression dataset, in the sense of maximizing a probabilistic scoring function related to Eq. (2.6) over the candidate network topologies (Bansal et al., 2007; N. Friedman, 2004; Villaverde and Banga, 2014). Existing methods typically explore the multidimensional space of possible graphs (i.e., regulatory networks) exhaustively (Faith et al., 2007; Marbach et al., 2012).

In reverse engineering applications, neither prior nor posterior probabilities are known (Villaverde and Banga, 2014). Markov Chain Monte Carlo (MCMC) techniques (Gelfand and Smith, 1990; Geman and Geman, 1984) are often used to evaluate Eq. (2.6) and other relevant probabilities. Still, the search for the best BN model is an NP-hard problem (Chickering, 1996). Hence, heuristic solutions are often implemented. In the context of the DREAM project, MCMC BN inference was found to be computationally very costly, making this approach better fitted for smaller networks (Marbach et al., 2012).

In addition, the BN learning problem is usually underdetermined. Several strategies exist to deal with this underdetermination, such as model averaging, bootstrapping or if available, adding *a priori* knowledge to select the most likely model structure (Bansal et al., 2007; Villaverde and Banga, 2014). BN inference is well suited to integrate heterogeneous datasets (Klipp et al., 2005), which makes BN modelling appealing when compared to the network inference methods reviewed above.

However, the main limitation of Bayesian network models is that they cannot contain cycles and thus they cannot represent cellular network feedback loops (Bansal et al., 2007; Doyle and Stelling, 2006). To overcome this limitation, Dynamic Bayesian networks were developed along with inference methods using time-series data (Yu et al., 2004).

A well-known BN inference algorithm is Banjo (Yu et al., 2004). This algorithm addresses both static and dynamic BNs and hence can infer gene networks from steady-state and dynamic data, based on heuristic approaches for the exploration of the possible network topologies. The algorithm was tested on simulated and real microarray data (Bansal et al., 2007). Banjo showed poor performance when applied on steady-state data from a limited number of experiments (*E. coli* steady-state dataset, see (Bansal et al., 2007)) when compared to interaction graph inference via ARACNE. In presence of more data from a larger number of steady-state experiments (*S. cerevisiae*, see (Bansal et al., 2007)), inference accuracy of Banjo proved to be considerably better (*S. cerevisiae*, see (Bansal et al., 2007)) although the number of correct inferred interactions remained limited. Unfortuntately, applicability of Banjo is limited for increasing size of the dataset (number of experiments) due to computational complexity (HumanBcells and *S. cerevisiae*, see (Bansal et al., 2007)). In the case of dynamic data analysis, relative to the number of network genes, analysis based on both synthetic

and experimental data (*E. coli* steady-state, see (Bansal et al., 2007)) showed that Banjo (Yu et al., 2004) requires a large number of experiments for the estimation of the necessary probability distributions.

2.2.5 Inference of Ordinary Differential Equation models

We have seen so far inference of models describing qualitative relationships among genes, whether in the form of statistical relationships or in the form of regulatory logics. What all these frameworks are missing is quantitative information concerning the strength with which genes regulate each other, and how this reflects into quantitative time-course dynamics. In this section we look at Ordinary Differential Equation (ODE)-type modelling, which will also constitute the modelling framework of choice in the following chapters.

In the context of gene networks, ODE models describe the dynamics underlying time-course gene expression in terms of a time-varying network state. The state generally represents concentrations of gene products (proteins, and sometimes corresponding mRNAs) and transcription factors that control the regulation of the genes of interest. Thus, this approach provides a more detailed and complex representation of the functioning of the biological network (Villaverde et al., 2013) and is better suited for the analysis of and inference from time-course quantitative (population-average) gene expression data.

In their simplest form, ODE models of gene regulation have the representation

$$\frac{dx_i(t)}{dt} = g_i(x_1, ..., x_N, \theta_i, u(t)) - \lambda_i x_i(t), \qquad i = 1, \ldots, N \tag{2.7}$$

where $x_i(t)$ is the time-varying concentration of the product of gene i, λ_i is a degradation rate, g_i is the synthesis rate function, θ_i is a vector of characteristic parameters and u represents a network input or perturbation (chemical treatments, genetic modifications, etc.) for a network with N genes (Bansal et al., 2007).

From a gene interaction viewpoint, these models encode the regulatory effects of every gene on gene i into the analytical form of g_i, thus enabling a link with the previously discussed model classes. While the regulatory structure can in principle be captured by suitably defined parameters θ_i, it is generally more appropriate to keep the concepts of structure and continuous-type parameters distinct. The initial state of the system, $x_i(t_0)$ with $i = 1, \ldots, N$, often is a steady-state solution of Eq. (2.7) (for an

appropriate constant pre-experimental input) corresponding to the system being at a dynamical equilibrium at the beginning of the experiment (de Jong, 2002). Inference of ODE models most often refers to the problem of estimating parameters θ_i from time-course or multiple steady-state experimental measurements of concentrations x_i from one or more system perturbations u, for a given structure of functions g_i. However, much work has also been dedicated to reconstructing or selecting the structure of the g_i, whence notably the set of regulatory interactions among the network genes, from within suitable model families (Porreca et al., 2010a).

Unlike other approaches discussed in this chapter, in accordance with its deterministic nature, ODE modelling usually does not rely on a statistical characterization of gene regulatory interactions.

Model inference requires first of all the choice of an appropriate mathematical representation of g_i, whether a model family or a fixed model structure, usually from the class of low-order polynomials or a combination of Hill functions (see Aracena (2008); de Jong (2002); Porreca et al. (2010a); Szederkényi et al. (2011); Yang et al. (2007)). Second, it requires the estimation of θ_i for each gene i. If the functional form of the model is not fixed, structure and parameter estimation are intertwined problems, and different methods try to isolate the two with different expedients, see Section [subsection on sign pattern] for a specific example.

Given a model structure, the identification of nonlinear ODE parameters from gene expression data is a challenging problem, since the ODE system often does not have an explicit solution (de Jong, 2002). In general, without suitable constraints, there are multiple solutions, i.e. the ODE system is not uniquely identifiable from data. A common distinction is made between structural identifiability and practical identifiability. Despite lack of agreement in the literature, the first generally refers to impossibility to distinguish different parameter values for the given model and observed outputs, no matter the abundance of the data, while the second refers to inherent limitations in estimation accuracy due to the quality of the data (Berthoumieux et al., 2013a). Indeed, time-course gene expression data are usually sparse and associated with large noise (Raue et al., 2009), which makes estimation of structurally identifiable parameters quite uncertain.

Parameter estimation is generally expressed as an optimization problem, where the objective functions minimized quantify the distance between the observations and the

model-predicted values. Thus, model inference is performed using optimization techniques (Bansal et al., 2007). Convex optimization, when applicable, provides a unique minimum and scales well with the dimension of the problem (number of unknown parameters, etc.) (Boyd and Vandenberghe, 2004). However, identification of nonlinear ODE parameters often results in non-convex problems, revealing a series of difficulties if standard local optimization methods are used, such as converging to local solutions or bad scalability for large systems. Global optimization methods have been developed to seek globally optimal solutions (see Floudas and Gounaris (2009) for a review; Banga (2008); Banga et al. (2005); Vilas et al. (2012)). However, their computational complexity increases rapidly with the problem size (Miró et al., 2012).

Models inferred from data carry information of the interactions among genes and hence provide implicitly a signed graph of regulatory interactions. Different from previously discussed methods for network inference, however, ODE models (Eq. 2.7) with parameters inferred from data can then be used to predict quantitatively the time response of the network to different internal perturbations (e.g. gene knock-out or over-expression) and external stimuli (environmental changes).

Among the best known algorithms for ODE network model identification are Network Identification by multiple Regression (NIR), Microarray Network Identification (MNI) and Time-Series Network Identification (TSNI) Bansal et al. (2006); Cantone et al. (2009); di Bernardo et al. (2005); Gardner et al. (2003). NIR and MNI analyse steady-state mRNA measurements, whereas TSNI uses time-series datasets.

The network is modeled as a system of linear ODEs (de Jong, 2002) expressing the synthesis rate of every transcript as a linear function of the concentration x_i of all other cell transcripts and a network perturbation u. For one experiment with given perturbation u, at measurement time t_k, with $k = 1, \ldots, M$, the system satisfies

$$\dot{x_i}(t_k) = \sum_{j=1}^{N} a_{ij} x_j(t_k) + b_i u(t_k), \tag{2.8}$$

where, for N genes, $i = 1, ..., N$. The coefficients a_{ij} and b_i (collectively captured by θ_i in Equation 2.7) quantify the effect of gene j and perturbation u, respectively, on gene i. If u is constant and the system has reached steady-state, then $\dot{x_i}(t_k) = 0$ and, for all

i, Equation 2.8 simplifies to

$$\sum_{j=1}^{N} a_{ij} x_j(t_k) = -b_i u(t_k) \tag{2.9}$$

From steady-state gene expression data ($M = 1$ and measurements of x_i for different constant values of u), the NIR algorithm (Gardner et al., 2003) computes the edges a_{ij} by solving the multiple linear regression (2.9) assuming a priori information about genes that have been directly perturbed (i.e. terms $b_i u$ known). The user can choose the maximum number of regulators per gene (i.e., the number of edges to a node). After solving linear regression (2.9), the algorithm returns the estimated matrix of interaction strenghts a_{ij}. If the noise in the data is small, this method does not require large datasets. It has been tested in (Bansal et al., 2006) showing good performance compared to ARACNE or Banjo even when a reduced number of experiments are available.

Similarly, MNI algorithm (di Bernardo et al., 2005) is based on Equation 2.9 and uses steady-state data. However, each microarray experiment can result from any kind of perturbation and knowledge about $b_i u$ is not necessary. First, MNI computes the a_{ij} from the gene expression data D and determines a model of the regulatory interactions between genes. Then a test dataset $\{x_1^d, ..., x_N^d\}$ representing the perturbed expression of the genes is used to compute $b_i u$ from Equation 2.9, with u a simple constant. The network model initially identified in the algorithm (trained on the dataset D) is used as a filter to predict a better model from the test perturbation data. A $b_i \neq 0$ quantifies that gene i is directly affected by the perturbation. The algorithm returns a ranked list of genes, where most likely targets of the perturbation have high values of b_i.

The TSNI algorithm (Bansal et al., 2006) relies on dynamical time-series data and identifies both the network structure a_{ij} and the targets of perturbation, that is the b_i, by solving in this case the linear regression corresponding to a discrete-time version of Equation 2.8. The algorithm assumes that a single perturbation experiment is performed and M time points following the perturbation are measured, in contrast to M different steady-state conditions considered for NIR and MNI (it can, however, be easily generalized to multiple dynamical experiments). The algorithm is capable of correctly inferring the structure (a_{ij}) and the targets of perturbation (b_i) of small gene networks. In larger networks instead, its performance in recovering the structure is not

very good, probably because the network is not fully observable and one perturbation experiment does not yield sufficient information for the inference (SOS *E. coli* network, see Bansal et al. (2006, 2007)).

Many other ODE-based algorithms have been proposed in the literature. Inferelator (Bonneau et al., 2006) is an inference algorithm for deriving genome-wide transcriptional regulatory interactions, and has been applied to predict a large portion of the regulatory network of the archaeon Halobacterium NRC-1. The algorithm infers regulatory interactions for genes and gene clusters from mRNA or protein expression levels and uses standard regression and l^1 shrinkage techniques to select models for the expression of a gene or cluster of genes as a function of the levels of their regulators. In (Bonneau et al., 2006), many novel gene interactions were predicted, and in several cases the inferred regulatory interactions were validated by experimental tests. The Inferelator was also able to predict mRNA levels of 80% of the genes in the genome over new experimental conditions in *Halobacterium salinarium* (Bonneau et al., 2007). Greenfield et al. (2010) demonstrate complementarity between this method and the mutual information CLR (Section 2.2.2) algorithm. Based on application to *in silico* time-series data of the DREAM4 competition, their combined use significantly improves the ability of selecting valid regulatory interactions compared to both methods alone. Moreover, the duo is able to accurately predict the response of the system to new conditions (new double knock-out perturbations).

Inference of Boolean-like models and the sign pattern analysis method

In the spirit of mixed ODE-boolean modeling, an original method of network inference has been proposed in Porreca et al. (2010a), explicitly based on the idea of transcription rate functions encoding the logics of regulation of the target genes. The resulting ODE models are referred to as "Boolean-like" models. The proposed inference algorithm tackles the identification of both structure and parameters of kinetic models of gene regulatory networks from time-course gene expression data. A modeling framework is considered where the dynamic equations are described in terms of a class of gene activation rules known as unate functions (Aracena, 2008; Comet et al., 2013). These functions reflect interactions where each gene is exclusively either an activator or an inhibitor for the expression of any given target gene (though a regulator may

be an activator and a repressor of distinct target genes). According to Grefenstette et al. (2006), unate functions provide biologically realistic dynamics for gene networks. The majority of the known gene activation rules (Kauffman et al., 2004; Nikolajewa et al.) are modeled by nested canalizing functions (Jarrah et al., 2007), a class of unate functions. Unate function modelling and analysis of biochemical networks has been discussed also in (Murrugarra and Laubenbacher, 2011; Raeymaekers, 2002).
In Porreca et al. (2010a), the properties of these functions are exploited in order to develop a two-step identification algorithm. Boolean-like ODE models are used to describe the evolution of the product of gene i (x_i):

$$\dot{x}_i = g_i(x) - \gamma_i(x) \tag{2.10}$$

where $x = (x_1, ..., x_N)$, N is the number of genes in the network, while $g_i(x) \geq 0$ and $\gamma_i(x) \geq 0$ are the synthesis and the degradation rates of the product of gene i.
The nonlinear model for the synthesis rate is

$$g_i(x) = k_{0,i} + k_{1,i} b_i(x) \tag{2.11}$$

where $k_{0,i} \in \mathbb{R}_+$ and $k_{1,i} \in \mathbb{R}_+$ are constants and $b_i(x) : \mathbb{R}_+^n [0, 1]$ quantifies the regulatory effects of the gene products in the network on the expression of gene i by algebraic combinations of Hill activation or repression functions (Keller, 1995; Yang et al., 2007), namely

$$\sigma^+(x_j) = \frac{{x^d}_j}{{x^d}_j + \theta^d}, \qquad \sigma^-(x_j) = 1 - \sigma^+(x_j). \tag{2.12}$$

Due to the assumption of unate structure, every function b_i, and hence every corresponding g_i, is monotonically increasing or decreasing in every state variable x that is an effective regulator of target i, while it is independent of x_j if the product of gene j does not regulate expression of gene i. These monotonicity properties can be captured by a sign pattern, i.e. an N-tuple $p = (p_1, ..., p_N) \in \{-1, 0, 1\}^N$ (depending on i) where, for $j = 1, ..., N$ p_j is -1 if gene j acts as an inhibitor for gene i expression, it is 1 if gene j acts as an activator for gene i expression, and 0, if gene j has no effect on the expression of gene i.
The problem tackled by the two-stage algorithm is the reconstruction of every g_i. To this purpose, and unlike other algorithms in this section, here the authors assume measurements of time-varying protein concentrations and promoter activities. As measurements of x and corresponding $g_i(x)$ are considered to be available, the decay rate

$\gamma_i(x)$ can be ignored in the reconstruction of g_i. In practice, when only concentration measurements or synthesis rate measurements are available, the dataset required by the algorithm can be completed from the available data provided knowledge of γ_i, see Porreca et al. (2010b).

In the first step, the algorithm isolates families of consistent model structures, i.e. families of so-called consistent sign patterns, by testing hypothetical sign patterns and rejecting those corresponding to monotonicity properties of g_i that are falsified by the experimental data. This reduces considerably the number of plausible interactions. In addition, the family of consistent patterns can be arranged in a hierarchical fashion, and is fully characterized by a small set of minimal possible topologies of the network. In the second step, quantitative identification of the networks returned by the first step is performed. By solving a nonlinear regression problem (estimation of θ, d, $k_{0,i}$ and $k_{1,i}$ and selection of the best specific model structure among those with a given sign pattern), models of minimal complexity explaining the data with sufficient accuracy are returned.

The method has been tested on an *in silico* network and on real data from synthetic network (IRMA, see Cantone et al. (2009)) and compared to TSNI. The signed directed graphs inferred from IRMA by TSNI were less accurate than those of Porreca et al. (2010a), where an analysis of sensitivity to noise of network reconstruction performance was also provided based on *in silico* data.

Performance of existing inference algorithms, data requirements and perspective for novel developments are largely discussed by DREAM reports and in Bansal et al. (2007); Hecker et al. (2009); Villaverde et al. (2013). Reliable inference from gene expression data still remains an open subject. One of the main conclusions is that, indeed, the performance of current network-inference methods is strongly dependent on the properties of the network that is being inferred and cannot be analyzed in isolation of the data that made it necessary. Employing modeling formalisms and algorithms that train on different features of the data and merging results seems to be a good way to improve the performance of inference. Another important point is that models and inference methods should be interpretable in terms of biological relevance of the results. For instance, while steady-state methods can assume protein and mRNA concentrations to be correlated, this can lead to spurious inference results from time-series datasets. However, the access to dynamic measurements for these chemical species is

not always easy and computational methods need to be developed in order to account for these issues. Next section addresses this point in a way that will be used for the results of this thesis.

2.3 Reporter gene data analysis

2.3.1 Measurement models

The reconstruction of biologically-relevant quantities from reporter gene data requires measurement models making explicit the relations between the concepts and the assumptions under which these relations hold (de Jong et al., 2010). Measurement models can describe the expression of the gene of interest in two steps (de Jong et al., 2010):

$$\frac{d}{dt}m(t) = g(t) - (\mu(t) + \gamma_m)\, m(t), \qquad m(0) = m_0, \tag{2.13}$$

$$\frac{d}{dt}p(t) = \kappa_p\, m(t) - (\mu(t) + \gamma_p)\, p(t), \qquad p(0) = p_0, \tag{2.14}$$

where $m(t), p(t)$ are the mRNA and protein concentrations, respectively, $\mu(t)$ is the time-varying growth rate, κ_p is the protein synthesis rate constant, and γ_m, γ_p are the degradation constants of mRNA and protein, respectively. A similar measurement model can be written for the reporter protein:

$$\frac{d}{dt}n(t) = g(t) - (\mu(t) + \gamma_n)\, n(t), \qquad n(0) = n_0, \tag{2.15}$$

$$\frac{d}{dt}r(t) = \kappa_r\, n(t) - (\mu(t) + \gamma_r)\, r(t), \qquad r(0) = r_0, \tag{2.16}$$

with analogous meanings for the variables and parameters. By construction of the transcriptional fusions, the mRNA synthesis rates or promoter activities of the gene of interest and the reporter gene are equal. This promoter activity is denoted by $g(t)$.
Two common assumptions make it possible to simplify the above models. First of all, typical mRNA half-lives in bacteria are on the order of a few minutes (Bernstein et al., 2002), whereas typical cell doubling times range from tens of minutes to hours (Larrabee et al., 1980; Mosteller et al., 1980). This motivates $\gamma_m, \gamma_n \gg \mu(t)$. Second, the mRNA concentrations evolve on a much faster time-scale than the protein concentrations, so that the former can be assumed to be in quasi-steady state: $dm(t)/dt = dn(t)/dt = 0$.

As a consequence, $m(t) = g(t)/\gamma_m$ and $n(t) = g(t)/\gamma_n$, and the models of Eqs. 2.13-2.16 simplify to the following reduced models:

$$\frac{d}{dt}p(t) = \hat{k}_p\, g(t) - (\mu(t) + \gamma_p)\, p(t), \qquad p(0) = p_0, \tag{2.17}$$

$$\frac{d}{dt}r(t) = \hat{k}_r\, g(t) - (\mu(t) + \gamma_r)\, r(t), \qquad r(0) = r_0, \tag{2.18}$$

with $\hat{k}_p = \kappa_p/\gamma_m$ and $\hat{k}_r = \kappa_r/\gamma_n$. We define the synthesis rate of the reporter protein

$$f(t) = \hat{k}_r\, g(t). \tag{2.19}$$

This quantity is proportional to the synthesis rate of the protein of interest, with proportionality constant $\alpha = (\kappa_r/\kappa_p)(\gamma_m/\gamma_n)$, *i.e.*,

$$f(t) = \alpha\, \hat{k}_p\, g(t). \tag{2.20}$$

Therefore, if $\kappa_p = \kappa_r$ (true for translational fusions) and $\gamma_m = \gamma_n$, then $f(t)$ also equals the synthesis rate of the protein of interest. As explained in de Jong et al. (2010), $f(t)$ can be directly computed from the absorbance and fluorescence signals. The quantity is usually called promoter activity in the literature or more generally the activity of the gene, motivated by the fact that it is proportional to $g(t)$. Promoter activity is also proportional to the mRNA concentration of the gene of interest. This simply follows from the fact that $f(t)$ is proportional to $\hat{k}_p\, g(t)$ and the latter expression equals $k_p\, m(t)$ by Eq. 2.14.

2.3.2 Constitutive promoters

One of the limitations of the above measurement model is that it assumes that k_p, k_r (the protein and reporter protein synthesis rate constants) are constants and do not depend on the time-varying activity of the ribosomes. The model also does not distinguish between the contributions of specific transcription regulators and the activity of RNA polymerase to the promoter activity $g(t)$. In order to address these limitations, the measurement models can be easily generalized (Berthoumieux et al., 2013b) by positing

$$g(t) = k_m\, g_{global}(t)\, g_{specific}(t), \tag{2.21}$$

and by replacing k_p by $k_p(t)$, and k_r by $k_r(t)$. Analogously to Eq. 2.19, the generalized expression for the synthesis rate of the reporter protein becomes:

$$f(t) = \left(k_m \, \hat{k}_r(t) \, g_{global}(t)\right) \, g_{specific}(t), \tag{2.22}$$

which is decomposed in a part due to the activity of the gene expression machinery ($k_m \, \hat{k}_p(t) \, g_{global}(t)$) and a part due to specific effects of transcription regulators ($g_{specific}(t)$). By the same reasoning as for the classic measurement models, this expression remains proportional to the synthesis rate of the protein of interest (with proportionality constant $(\kappa_r/\kappa_p)(\gamma_m/\gamma_n)$).
If we consider a reporter gene with a constitutive promoter that has the same ribosome-binding site as the reporter of the gene of interest, following Eq. 2.22, we have:

$$g_{const}(t) = k_m^{const} \, g_{global}(t), \tag{2.23}$$

and, correspondingly,

$$f_{const}(t) = k_m^{const} \, \hat{k}_r(t) \, g_{global}(t), \tag{2.24}$$

Therefore, when measuring both $f(t)$ (by means of the reporter of the gene of interest) and $f_{const}(t)$ (by means of the reporter of a constitutively expressed gene), global physiological effects due to the activity of the gene expression machinery and specific effects due to transcription factors and other regulators can be separated.

2.3.3 Data processing

As described in Section Experimental data of this chapter, *in vivo* and real-time gene expression profiles can be obtained by means of fluorescent reporter gene systems monitored in an automated, thermostated reader. The absorbance or the optical density measured at 600 nm quantifies the biomass, while the fluorescence signal emitted at 520 nm, when excited at 485 nm, is proportional to the number of GFP molecules. The absorbance is expressed in dimensionless units, whereas fluorescence intensities have specific relative fluorescence units (RFU). In this section we describe how, by means of the measurement models previously described, we derive promoter activities and protein concentrations from the absorbance and fluorescence data (Berthoumieux et al., 2013b; de Jong et al., 2010).

2.3.4 Reconstruction of promoter activities

The corrected absorbance and fluorescence data are used to compute promoter activities (synthesis rates) and protein concentrations, following the measurement models in de Jong et al. (2010). From Eqs. 2.18-2.19 it follows that

$$f(t) = \frac{d}{dt}r(t) + (\mu(t) + \gamma_r)\, r(t). \tag{2.25}$$

The growth rate $\mu(t)$ can be estimated from the absorbance, that is,

$$\mu(t) = \frac{d}{dt}A(t)\frac{1}{A(t)} = \frac{d\ln A(t)}{dt}. \tag{2.26}$$

The time-varying GFP concentration in the bacterial population, $r(t)$, can also be estimated from the absorbance and fluorescence, making the usual assumptions that the fluorescence is proportional to the number of GFP molecules and the absorbance proportional to the biomass:

$$r(t) \sim \frac{I(t)}{A(t)}. \tag{2.27}$$

We arbitrarily set the proportionality constant in Eq. 2.16 to 1, thus expressing the reporter protein concentration in units RFU (and the synthesis rate in units RFU min^{-1}). Substituting the expressions for $r(t)$ and $\mu(t)$ into Eq. 2.25 and after some basic computations (de Jong et al., 2010) we obtain:

$$f(t) = \frac{\frac{d}{dt}I(t)}{A(t)} + \gamma_r\frac{I(t)}{A(t)}. \tag{2.28}$$

The definition is equivalent to other definitions in the literature (Ronen et al., 2002) when $\mu(t) >> \gamma_r$. The expression is evaluated using estimates of $A(t)$, $I(t)$, and $dI(t)/dt$ obtained by means of cubic smoothing splines (de Jong et al., 2010).

2.3.5 Reconstruction of protein concentrations

In order to reconstruct the concentration of a protein of interest, the same measurement models are used, in particular Eq. 2.17. The term $\hat{k}_p\, g(t)$ was seen to be proportional to $f(t)$, following Eq. 2.20. We arbitrarily set the proportionality constant in Eq. 2.20 to 1, and we rewrite:

$$\frac{d}{dt}p(t) = f(t) - (\mu(t) + \gamma_p)\, p(t), \quad p(0) = p_0, \tag{2.29}$$

With the definition of the initial protein concentration and additional information on the half-life of the protein (degradation constant γ_p) $p(t)$ can be computed by numerical solution of the above ODE.

3. Inference of quantitative models of bacterial promoters from time-series gene expression data

This chapter will present the results of the PhD thesis. We will show how the transcriptional response of the genes in the FliA-FlgM module and global regulatory effects have been measured by means of fluorescent reporter genes, in a variety of wild-type and mutant conditions, in different growth media. We will present the mathematical models developed to describe FliA-dependent gene expression. Furthermore, we will illustrate how these data were used to systematically test the information required for the reliable inference of the regulatory interactions and quantitatively predictive models of gene regulation. In a first step, we tested if the use of FliA and FlgM promoter activities, instead of their protein concentrations, allows the expected pattern of regulatory interactions to be inferred, and a quantitative model of the activity of FliA-dependent genes to be identified from the data. In a second step, we introduced global regulatory effects, measured by means of a reporter gene driven by a constitutive promoter. In a third step we estimated the concentrations of FliA and FlgM from the observed promoter activities and physiologically plausible half-lives of the proteins. The results had been further refined in a fourth step, by taking into account that FliA and FlgM half-lives may vary across conditions, in the range of physiologically valid values.

We also describe in detail the experimental methods used either to produce or validate the biological data on the central module controlling motility in *E. coli*, along with the experimental conditions, the strains and the inference and modeling frameworks.

3.1 Results

3.1.1 Monitoring the transcriptional response of the FliA-FlgM module

The more than 60 genes responsible for motility in bacteria are structured in a transcriptional hierarchy of three operon classes, which has been mapped in detail for *Escherichia coli* and *Salmonella enterica* (Chevance and Hughes, 2008; Kalir et al., 2001; Kutsukake et al., 1990; Macnab, 1996a). The single class 1 operon *flhDC* encodes the proteins FlhD and FlhC, that form a heteromultimeric complex activating σ^{70}-dependent transcription of the class 2 operons. The latter encode the proteins making up the flagellar motor structure as well as a major regulator of the class 3 operons, the sigma factor FliA (σ^{28}). When bound to core RNA polymerase, FliA directs the transcription of a total of **5** class 3 operons (Keseler et al., 2013), which code for the proteins forming the filament structure of the flagellum and the chemotaxis sensing system. The aspartate chemoreceptor Tar is an example of such a class 3 protein. The action of FliA is counteracted by the anti-sigma factor FlgM, which binds to FliA and thus prevents its association with RNA polymerase. FlgM is encoded by the gene *flgM*, which is transcribed from both a class 2 promoter and a class 3 promoter. FlgM can be excreted from the cell through the center of the basal-body structure of the flagellum (Figure 3.1).

The transcriptional hierarchy produces a temporally-arranged order of events during the assembly of the flagella and the chemotactic sensing system (Chevance and Hughes, 2008; Kalir et al., 2001; Kutsukake et al., 1990; Macnab, 1996a). On the highest level of the hierarchy, the transcription of the flagellar master regulator responds to a variety of intracellular signals (Girgis et al., 2007; Pesavento et al., 2008). For instance, the expression of the *flhDC* operon is repressed when the bacteria are grown on minimal medium with glucose (Adler and Templeton, 1967). When glucose is depleted from the environment, however, the signalling molecule cyclic AMP (cAMP) accumulates in the cell, which induces *flhDC* transcription through the intermediary of the cAMP receptor protein Crp (Zhao et al., 2007). In the presence of FlhDC, the class 2 operons, and thus the genes encoding the hook basal-body (HBB) structure as well as FliA and FlgM, are actively transcribed. FlgM sequesters FliA and prevents it from transcribing

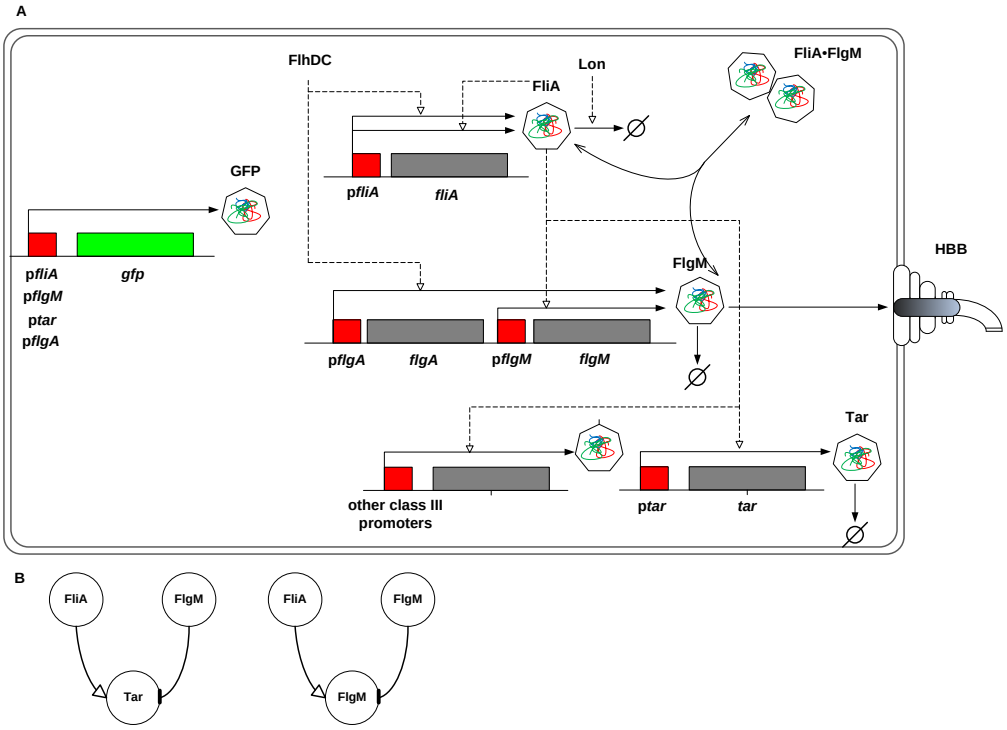

Figure 3.1: FliA-FlgM module. *A:* The regulatory circuit composed of the flagellar-specific transcription factor FliA, a sigma factor also known as σ^{28}, and the anti-sigma factor FlgM forms a check-point in the transcriptional hierarchy of the motility genes in *E. coli*. While *fliA* is transcribed from a single class 2 promoter (p*fliA*), *flgM* is transcribed from both a class 2 and a class 3 promoter (p*flgA* and p*flgM*, respectively). FliA binds to RNA polymerase core enzyme and directs transcription from a total of 5 class 3 promoters (Keseler et al., 2013), including p*tar* and p*flgM*. When bound to FlgM, FliA cannot activate transcription. When the hook basal-body (HBB) structure is in place, however, FlgM is exported from the cell, thus releasing FliA from the inactive complex. FliA is subject to proteolysis by Lon, but FlgM-binding protects FliA from degradation. The *fliA* promoter is auto-regulated by FliA and by a number of other regulators, most importantly the motility master regulator FlhDC. The expression of FlhDC itself is under the control of a variety of regulatory factors, including RpoS, CpxR, CsgD and Crp ∘ cAMP. The activity of the genes is measured by fusion of their promoters to a *gfp* reporter gene on a low-copy plasmid. Genes are shown in grey or green and their promoter regions in red. Regulatory interactions are represented by dashed lines, association and dissociation of FliA and FlgM as well as degradation and export by solid lines. The figure does not explicitly show that *fliA*, *flgM*, and *tar* are included in larger transcriptional units, the *fliAZY*, *flgAMN*, *flgMN* and *tar-tap-cheRBYZ* operons (Keseler et al., 2013). *B:* Pattern of known regulatory interactions for the class 3 genes *tar* and *flgM*.

the class 3 operons (Chilcott and Hughes, 2000). When the HBB structures have been completed, however, FlgM is secreted from the cell, releasing FliA and relieving the repression of the class 3 operons. The FliA-FlgM interactions thus form a check-point between the expression of the class 2 and class 3 operons, ensuring that the filament proteins are produced only when the basal body and the hook, to which the flagellar filaments are attached, are in place.

In order to investigate the regulation of the genes involved in this check-point, we measured the time-varying transcription of *fliA*, *flgM*, and *tar* (as an example of a class 3 gene) in *E. coli*. This was accomplished by means of fluorescent reporter systems, consisting of transcriptional fusions of a *gfp* reporter gene with the promoters of the target genes, carried on a low-copy plasmid. The strains transformed with the reporter plasmids were grown in 96-well microplates, following a previously-established protocol (Section 3.2.1). After an overnight preculture, the bacteria were diluted into fresh medium in the microplate and the absorbance of the cultures and the emitted fluorescence were monitored at 37° C in a thermostated microplate reader for 7 to 16 h, until growth arrest occurred. These kinetic experiments were carried out in different growth media (minimal M9 medium with glucose, LB medium) and in different genetic backgrounds (wild-type and deletion mutants of the global transcription regulators RpoS, CsgD, and CpxR). The timing and the strength of the induction of the hierarchy of motility genes varies among conditions, leading to a different time-varying excitation of the FliA-FlgM module.

While *fliA* and *tar* have a single promoter, this is not the case for *flgM*, which is transcribed from both a class 2 and a class 3 promoter, as discussed above. The fluorescence signal from the class 2 promoter, however, was found to be almost indistinguishable from background levels in all conditions (Figure A.2 in Appendix A), consistent with the observation that most FlgM in the cell derives from the FliA-dependent promoter (Chevance and Hughes, 2008; Gillen and Hughes, 1993). In the analysis that follows, we therefore neglected *flgM* transcription from the class 2 promoter.

As illustrated in Figure 3.2, the primary absorbance and fluorescence signals can be transformed into promoter activities using kinetic models of gene expression (Sec-

tion 3.2.2). More precisely, the reporter gene data allow one to deduce protein synthesis rates (de Jong et al., 2010; Ronen et al., 2002). Under certain conditions, as explained in detail in Chapter 2, the latter are proportional to mRNA concentrations and promoter activities and thus reflect the transcriptional activity of the gene. Following established terminology, we will refer to the measured protein synthesis rates as promoter activities, or more generally, activities of genes.

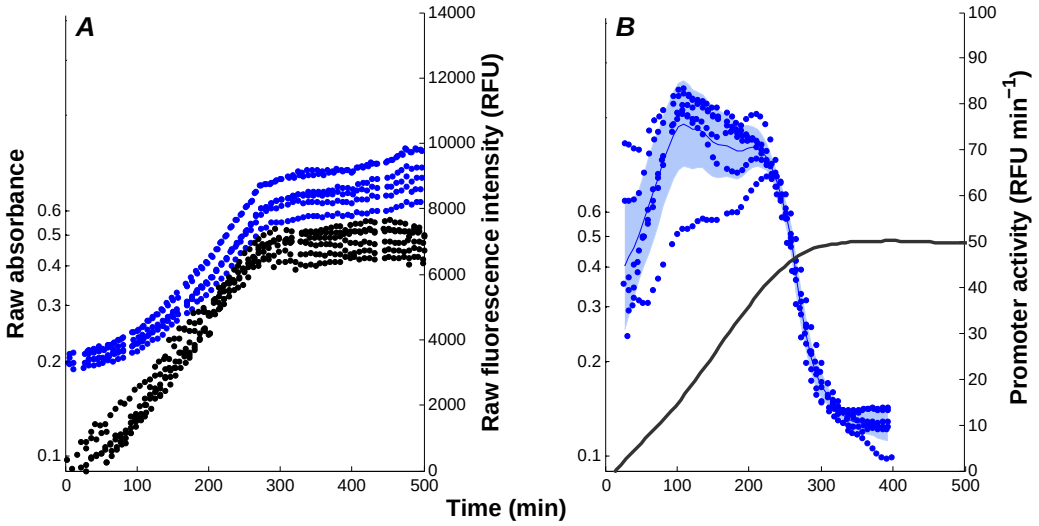

Figure 3.2: Primary data and promoter activities. *A:* Absorbance (•, black) and fluorescence (•, blue) data, corrected for background intensities, obtained with the Δ*cpxR* strain transformed with the p*tar-gfp* reporter plasmid, grown in M9 with glucose. *B:* Activity of the *tar* promoter (•, blue), computed from the primary data as described in Section 3.2.2 and in Chapter 2. The solid line corresponds to the mean of 6 replicate absorbance measurements and the shaded region to the mean of the promoter activities ± twice the standard error of the mean.

In each of the experimental conditions, we have acquired 5 to 8 replicate measurements, which allows for an estimation of the uncertainty in the derived promoter activities. Figure 3.3 shows the results for the five conditions considered here: (i) Δ*rpoS* strain grown in M9 (Δ*rpoS*-M9), (ii) Δ*cpxR* strain grown in M9 (Δ*cpxR*-M9), (iii) Δ*csgD* strain grown in M9 (Δ*csgD*-M9),(iv) Δ*csgD* strain grown in LB (Δ*csgD*-LB), and (v) wild-type strain grown in LB (WT-LB). As expected (Adler and Templeton,

1967), the fluorescence signals in the wild-type strain grown in glucose were mostly not distinguishable from the background fluorescence and therefore this condition was not further considered. In one condition (WT-LB), the activities measured by means of reporter genes were validated using RT-qPCR (Section 3.2.6).

The measured activity profiles in Figure 3.3 show some common features, such as a transient activity peak of the genes during exponential growth, followed by stabilization at a low level after growth arrest. The induction of the individual promoters has a distinct temporal order, corresponding to the level of the promoters in the transcriptional hierarchy (Kalir and Alon, 2004): *fliA*, *flgM*, *tar*. There are also clearly visible differences between the profiles across the conditions though. In M9 medium with glucose the motility genes in the mutant strains are transcribed right from the start, whereas in LB induction occurs only after a number of generations, consistent with previous reports (Adler and Templeton, 1967; Kalir et al., 2001). Moreover, the strength of induction and the duration of the activity peak varies from one condition to the other. For instance, the maximal activity of *tar* varies 10-fold between the WT-LB and Δ*csgD*-LB conditions.

3.1.2 Identification of gene regulation functions from promoter activities

The circuit in Figure 3.1 has been well-studied over several decades and its regulatory structure is well-known (Keseler et al., 2013). This therefore provides an excellent test case for investigating what kinds of information are needed for the reliable inference of regulatory interactions and quantitative regulation functions from gene expression data. In a first step, we tested if we could account for measured time-varying promoter activities while ignoring the distinction between mRNA and protein concentrations as well as the activity of the gene expression machinery and other global physiological effects, as is usually the case.

We expect FliA to be an activator and FlgM an inhibitor of target genes like *tar* and *flgM*. In order to check if this regulatory pattern is consistent with the reporter

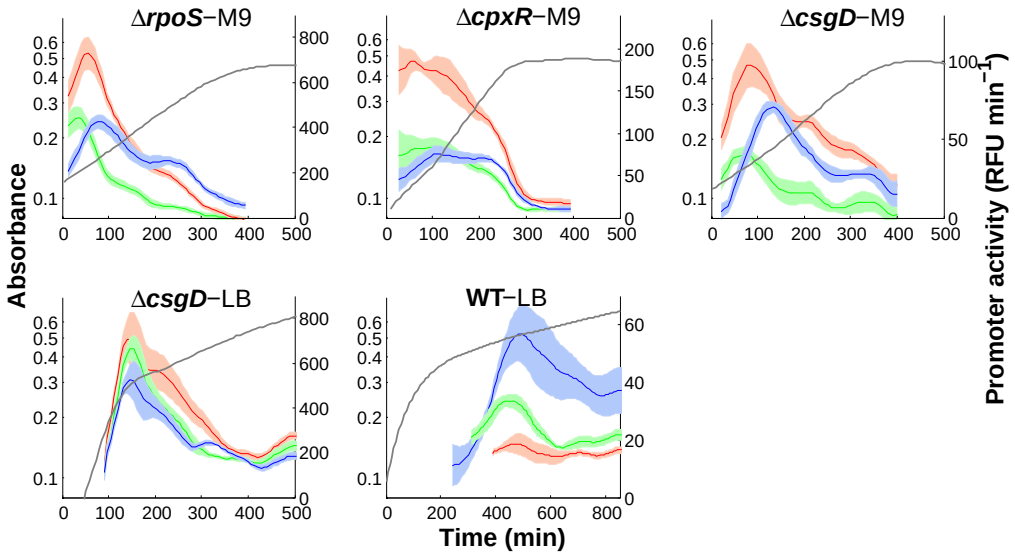

Figure 3.3: Promoter activities of genes in the FliA-FlgM module. The promoter activities of *fliA* (green), *flgM* (red), and *tar* (blue) measured in all experimental conditions considered in this study: Δ*rpoS* strain grown in M9 (Δ*rpoS*-M9), Δ*cpxR* strain grown in M9 (Δ*cpxR*-M9), Δ*csgD* strain grown in M9 (Δ*csgD*-M9), Δ*csgD* strain grown in LB (Δ*csgD*-LB), and wild-type strain grown in LB (WT-LB). The promoter activities have been derived from the primary data as illustrated in Figure 3.2.

gene data, we used minimal sign pattern analysis (Porreca et al., 2010a). This approach exploits time-series data to invalidate patterns of regulatory interactions, based on the assumption that the activity of a gene is a monotonic function of its regulators. The remaining patterns of regulatory interactions are subsumed by so-called minimal patterns. These patterns are minimal in the sense that removing any of the regulators results in an inconsistency with the data, while adding other regulators preserves consistency (see Section 3.2.3 for details on the method).

We applied minimal sign pattern analysis to the reporter gene data in Figure 3.3. In particular, we tested whether the expected regulatory pattern is conserved when replacing the concentrations of FliA and FlgM by the measured promoter activities. In order to check the robustness of the minimal patterns thus obtained, we verified that no regulatory patterns were dismissed because of a single measurement in the time-

series. We found that, both for the *tar* and the *flgM* promoter, the expected regulation by FliA and FlgM is not consistent with the data (Figure 3.4). Intuitively, this can be explained by the fact that, over some interval of time in the condition $\Delta rpoS$, a decrease of the promoter activity of *fliA* and an increase of the promoter activity of *flgM* coincide with an increase of the activity of the target genes.

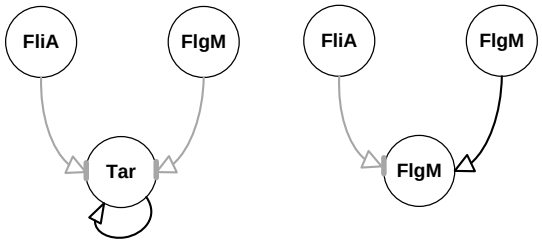

Figure 3.4: Minimal sign patterns for the regulation of *tar* and *flgM* when replacing protein concentrations by promoter activities. For both the regulation of *tar* gene and *flgM* gene, the expected sign pattern (*fliA*, *flgM*, *tar*) = (1,-1,0) is found to be inconsistent with the data. The invalidation of the expected sign pattern is due to the fact that, in $\Delta rpoS$, a decrease of the promoter activity of *fliA* and an increase of the promoter activity of *flgM* corresponds to an increase of the promoter activity of *tar* (and *flgM*). The minimal sign pattern identified for *tar* gene is (0,0,1), meaning that *tar* is necessary for its own regulation. The minimal sign pattern identified for *flgM* gene is (0,1,0), meaning that *flgM* is necessary for its own regulation. Black arcs represent the minimal consistent sign patterns. Every consistent pattern can be obtained from one of these minimal sign patterns by turning some gray arcs into black arcs with either a line (inhibition) or an arrow (activation) end.

Despite this structural problem, we also tested to which extent it is possible to quantitatively predict the activities of *tar* and *flgM* from the activities of their regulators. To this end, we developed a mechanistic model of the regulation of these promoters by FliA and FlgM. The model takes into account the titration of FliA by FlgM and the activation of transcription by (free) FliA. We made a quasi-equilibrium assumption for FliA-FlgM association and dissociation, justified by the fast time-scale on which these reactions occur in comparison with transcription and translation processes (Buchler and Louis, 2008; Bundschuh et al., 2003). Moreover, we chose a Hill function to describe promoter activation and included a basal synthesis rate. The resulting model

is:

$$f(t) = k_0 + k_1 \frac{p_{A,free}(t)^n}{\theta^n + p_{A,free}(t)^n}, \tag{3.1}$$

$$p_{A,free}(t) = \frac{1}{2}\left(-(K + p_M(t) - p_A(t)) + \sqrt{(K + p_M(t) - p_A(t))^2 + 4\,K\,p_A(t)}\right), \tag{3.2}$$

where $f(t)$ is the time-varying promoter activity, $p_{A,free}(t)$ is the concentration of free FliA, θ is a threshold constant for promoter activation, k_0 and k_1 are the basal and maximal synthesis rates, respectively, and n is a Hill constant. The concentration of free FliA is computed from the concentrations $p_A(t)$ and $p_M(t)$ of total FliA and FlgM, respectively, and the FliA-FlgM dissociation constant K. All variables and parameters are non-negative and $n \geq 1$. The concentration variables, as well as θ and K, have the units RFU, while the promoter activity and the rate constants have units RFU min^{-1}. The derivation of the model is described in detail in Section 3.2.4. Notice that the model is in agreement with the expected pattern of regulatory interactions (Figure 3.1*B*).

How well does this model fit the data when the total concentrations of FliA and FlgM in Eq. 3.2, p_A and p_M, are replaced by the measured activities of *fliA* and *flgM*, respectively? We estimated the values of the kinetic parameters $c = (k_0, k_1, n, \theta, K)$ in the regulation model from the data obtained in all five conditions, using a hybrid genetic algorithm that was shown to give good results for nonlinear models in systems biology (Rodriguez-Fernandez et al., 2006). The algorithm minimizes the mean-square error between the observed promoter activities and the predictions of the model of Eqs. 3.1-3.2, while taking into account differences in absolute promoter activity across conditions as well as the time-varying size of confidence intervals (Section 3.2.2).

The predictions of the identified regulation function for *tar* are shown in Figure 3.5. While the fit with the experimental data is quite good for the Δ*csgD*-LB condition and acceptable for the WT-LB condition, the model is not able to account for the peak in *tar* activity in the M9 conditions. The model either predicts no peak or a peak occurring more than an hour before it is observed. When analyzing the estimated parameter values, we observe that the cooperativity parameter n equals 1 and that the

value of the threshold θ is similar with values of the *fliA* activity over all conditions. This means that the the regulation function of the *tar* promoter is essentially a linear transformation of *fliA* activity.

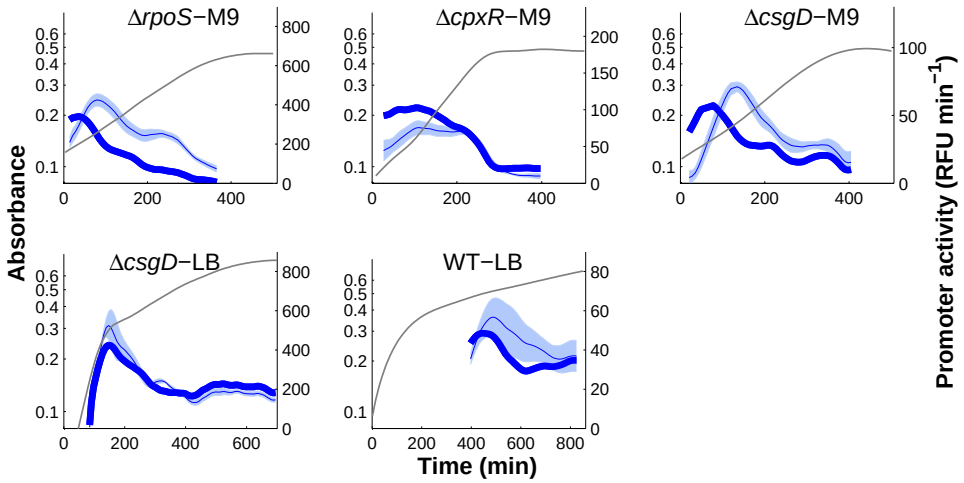

Figure 3.5: Fits of regulation function of *tar* to reporter gene data when replacing protein concentrations by promoter activities. The regulation function of Equations 3.1-3.2 was fit using the promoter activities for *tar*, *fliA*, and *flgM* shown in Figure 3.3, where the latter two replace the concentrations of FliA and FlgM, respectively. The parameters were estimated using a multistart global optimization algorithm (see Section 3.2.5 for details). The best fit (thick solid blue line) returns the value $Q = 33.6$ for the objective function, for the parameter vector $(k_0, k_1, n, \theta, K) = (7.6, 853, 1, 662, 14615)$. The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

We repeated the above analysis for the *flgM* promoter, setting the parameter that is not promoter-specific, the FliA-FlgM dissociation constant K, to the value estimated from the *tar* data. Since the fluorescence signal emitted by the strain carrying the p*flgM*-gfp reporter plasmid is very close to the background levels, and thus unreliable, we eliminate the condition WT-LB. The results are shown in Figure 3.6 and are qualitatively similar to results obtained for *tar*.

In conclusion, replacing protein concentrations by promoter activities in the FliA-FlgM module is insufficient for obtaining reliable models of the promoter activities, either structurally or quantitatively.

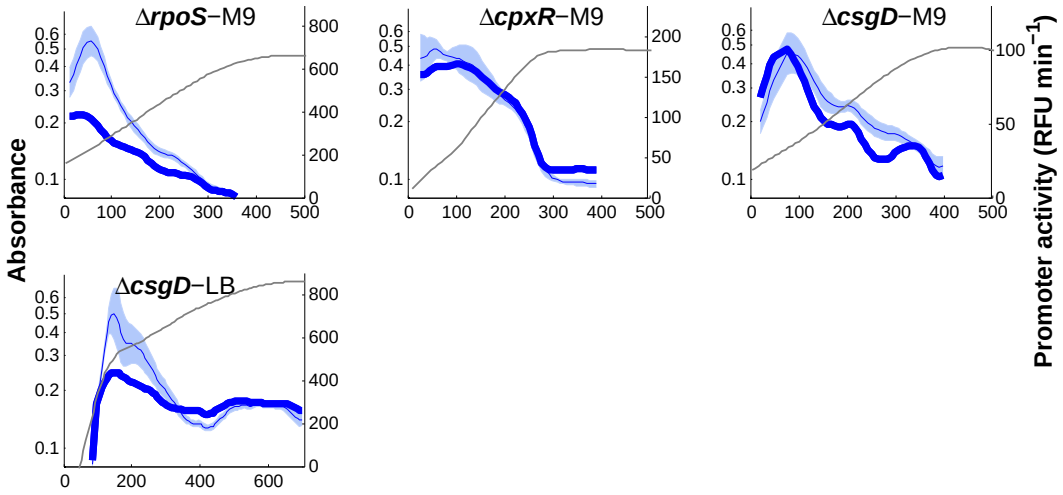

Figure 3.6: Fits of regulation function of *flgM* to reporter gene data when replacing protein concentrations by promoter activities. The regulation function of Equations 3.1-3.2 was fit using the promoter activities for *fliA*, and *flgM* shown in Figure 3.3, where the latter two replace the concentrations of FliA and FlgM, respectively. The parameters were estimated using a hybrid genetic algorithm (see Section 3.2.5 for details). The best fit (thick solid blue line) returns the value $Q = 23$ for the objective function, for the parameter vector $(k_0, k_1, n, \theta, K) = (9, 582, 1, 221, 7307)$. The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

3.1.3 Identification of gene regulation functions from promoter activities including global physiological effects

A possible explanation for the difficulty to identify quantitative regulation functions from information on promoter activities only may be that, in addition to transcription regulators and other specific regulators, the activity of the gene expression machinery also affects gene expression (Bremer and Dennis, 1996; Klumpp et al., 2009; Maloe, 1979). Contrary to FliA and FlgM, which affect specific genes, all motility genes are affected by the activity of the gene expression machinery and other global physiological effects. Figure 3.7 shows the network structure of the FliA-FlgM module when such global physiological effects are taken into account.

The activity of the gene expression machinery includes the abundance and activity of RNA polymerase and ribosome, as well as pools of metabolic precursors, and is

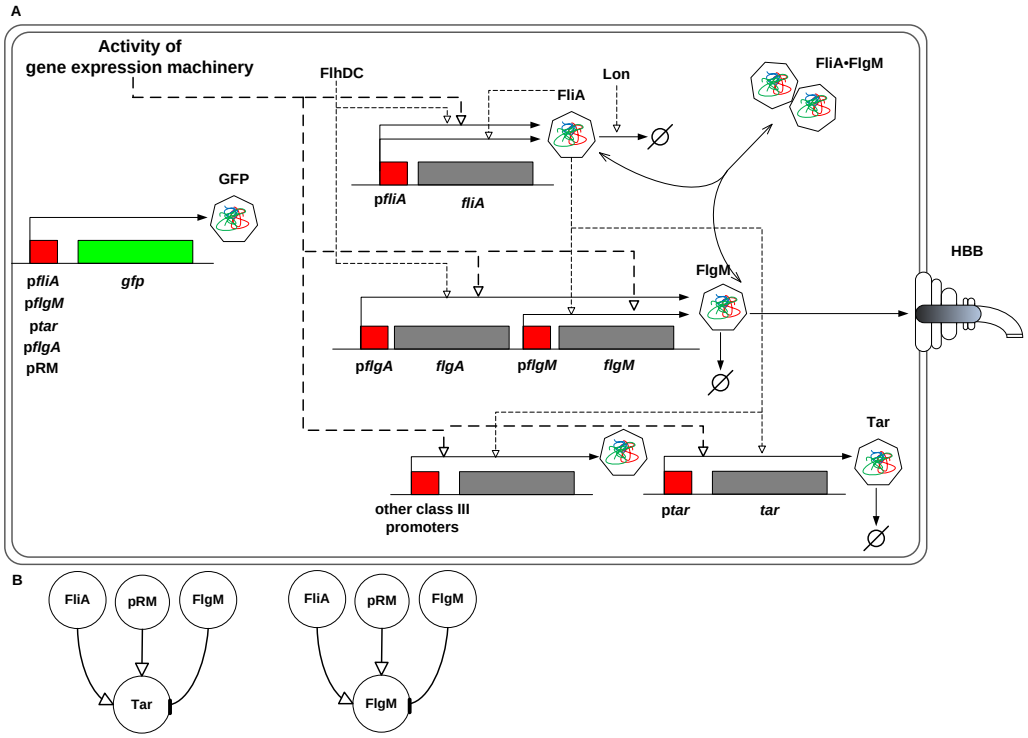

Figure 3.7: FliA-FlgM module extended with activity of the gene expression machinery. *A:* The network is the same as in Figure 3.1, but the regulation of the motility genes by global physiological effects, in particular the activity of the gene expression machinery, has been included. These regulatory interactions are shown by bold, dashed lines. *B:* Pattern of regulatory interactions for the class 3 genes *tar* and *flgM*.

therefore difficult to quantify in a direct way. This has motivated the use of the growth rate and the activity of constitutive genes, whose expression is in principle not controlled by any specific regulators, as indirect read-outs of the global physiological state of the cell (Berthoumieux et al., 2013b; Gerosa et al., 2013; Klumpp et al., 2009). In this study, following (Berthoumieux et al., 2013b), we use the activity of the phage λ promoter pRM, which is constitutive in non-infected *E. coli* cells, as a quantitative measure of the activity of the gene expression machinery, and the global physiological state more generally. In Figure 3.8 the time-varying activity of the constitutively-expressed reporter gene is shown, together with the activity of *tar*.

Does the inclusion of global physiological effects enable the identification of quantitatively predictive gene regulation functions? In order to answer this question, we again applied minimal sign pattern analysis to the reporter gene data, this time includ-

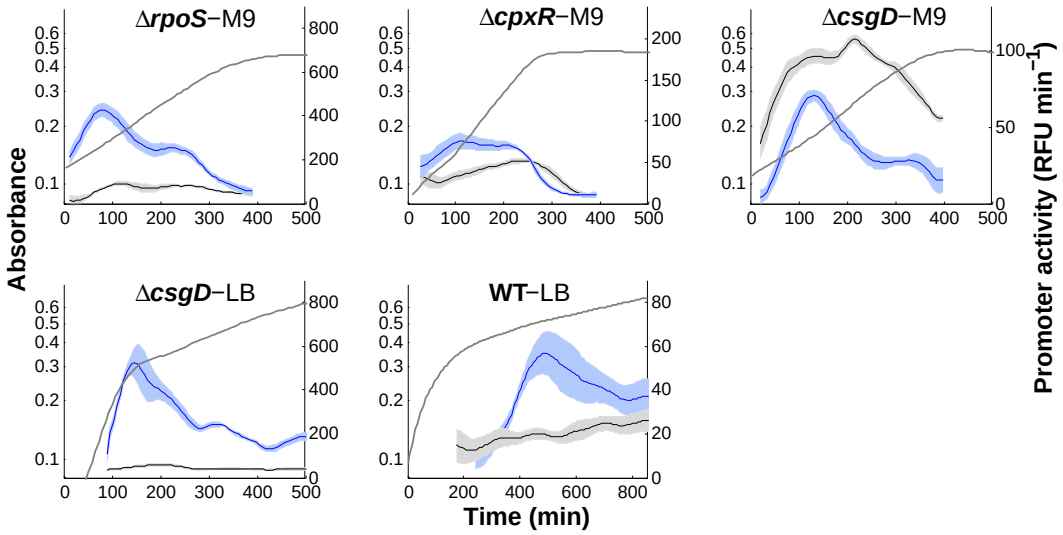

Figure 3.8: Activities of constitutive phage promoter. The activities of the phage λ promoter pRM (black) and the activity of *tar* (blue) measured in all experimental conditions considered in this study. The *tar* promoter activities are the same as shown in Figure 3.3.

ing the activity of the constitutive phage promoter as a proxy for the activity of the gene expression machinery. As in the previous section, the FliA and FlgM concentrations were replaced by the activities of their genes. Whereas the expected pattern of regulatory interactions (activation of the promoter by the gene expression machinery and FliA, inhibition by FlgM) was consistent with the data for *tar*, the analysis again ruled out this pattern for *flgM* (Figure 3.9). This means that, even when including global physiological effects in the analysis, the regulatory structure cannot generally be recovered.

Ignoring the fact that the correct structure could not be recovered for *flgM* regulation, we also checked if the proposed extension improves the capability of the regulation function for FliA-controlled promoters to quantitatively account for the time-varying data. To this end, we multiplied Eq. 3.1 with $f_{const}(t)$, the measured activity of a constitutive promoter:

$$f(t) = f_{const}(t) \left[k_0 + k_1 \frac{p_{A,free}(t)^n}{\theta^n + p_{A,free}(t)^n} \right], \tag{3.3}$$

The fits shown in Figure 3.10, obtained with the parameter estimation approach

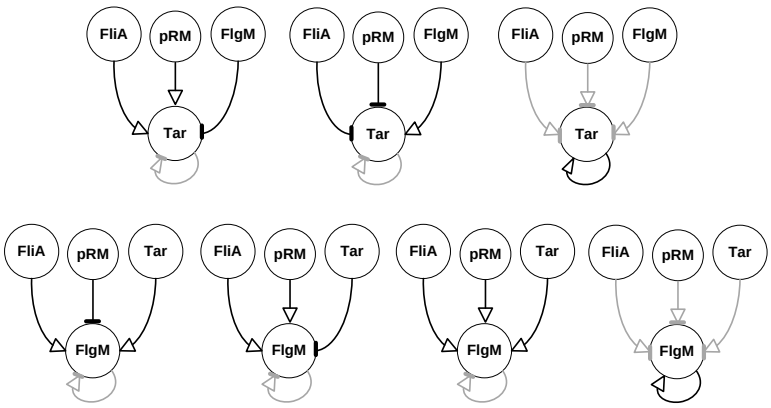

Figure 3.9: Minimal sign patterns for the regulation of *tar* and *flgM* when replacing protein concentrations by promoter activities and including global physiological effects. For the regulation of *tar* gene, the expected sign pattern (*fliA*, *flgM*, *tar*, pRM) = (1,-1, 0, 1) is found to be consistent with the data, e.g. the promoter is activated by the gene expression machinery and FliA and repressed by FlgM. For the regulation of *flgM* gene, the expected sign pattern (1,-1, 0, 1) is found to be inconsistent with the data. Similarly to results presented in Figure 3.4, black arcs represent the minimal consistent sign patterns. Every consistent pattern can be obtained from one of these minimal sign patterns by turning some gray arcs into black arcs with either a line (inhibition) or an arrow (activation) end.

outlined in the previous section, are better than those obtained with a model accounting for the effects of FliA and FlgM only, especially for the $\Delta rpoS$-M9 and $\Delta cpxR$-M9 conditions. The better fit is also reflected in a lower value of the fitting error ($Q = 30$ vs $Q = 33.6$). Notice that the extended model has the same parameters as the model without global physiological effects in Eqs. 3.1-3.2, so that the improvement is not simply due to an increase in the degree of freedom of the model. The parameter estimates are basically the same as for the previous model, though the values obtained for θ are larger than the maximum *fliA* activity (Figure 3.10). Essentially similar results are obtained for *flgM* (Figure 3.11).

In conclusion, although taking into account the activity of the gene expression machinery somewhat improves the results, models obtained are still incorrect from the structural point of view and quantitative predictions of FliA-dependent regulation

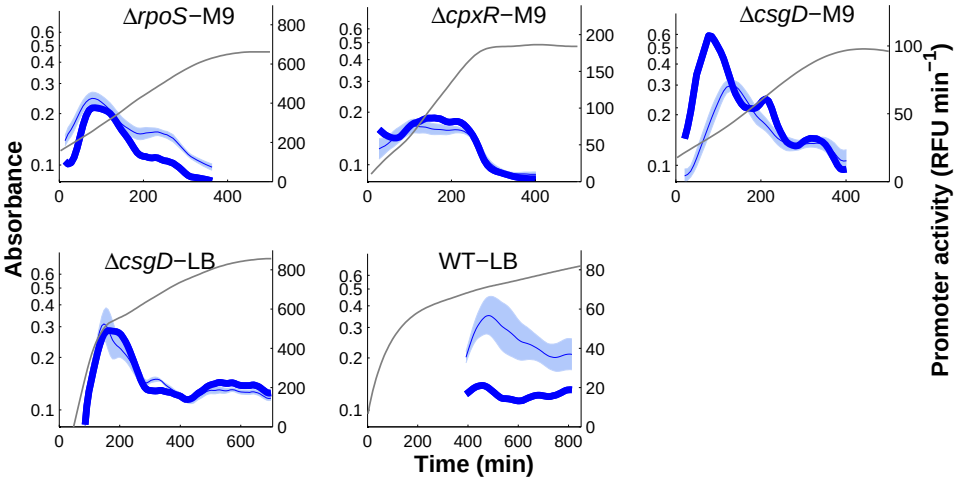

Figure 3.10: Fits of regulation function of *tar* to reporter gene data when replacing protein concentrations by promoter activities and including global physiological effects. The regulation function of Equations 3.2-3.3 was fit using the promoter activities for *tar*, *fliA*, and *flgM* shown in Figure 3.3, where the latter two replace the concentrations of FliA and FlgM, respectively. Moreover, global physiological effects are quantified by the activity of the constitutively expressed pRM promoter (Figure 3.8). The parameters were estimated using a hybrid genetic algorithm (see Section 3.2.5 for details). The best fit (thick solid blue line) returns the value $Q = 30$ for the objective function, for the parameter vector $(k_0, k_1, n, \theta, K) = (0.1, 16, 1.04, 662, 14615)$. The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

functions are still unsatisfactory. As explained in Chapter 1, replacing protein concentrations by promoter activities may not be appropriate, due to the fact that the half-lives of proteins are usually much longer than the half-lives of mRNA, causing the temporal decorrelation of protein concentrations and promoter activities. We therefore investigated how information on protein concentrations can be integrated into the inference process.

3.1.4 Identification of gene regulation functions from estimates of protein concentrations

It is straightforward to provide an estimate of the GFP concentration from the fluorescence and absorbance data, as explained in Section 3.2.2. The results are shown in

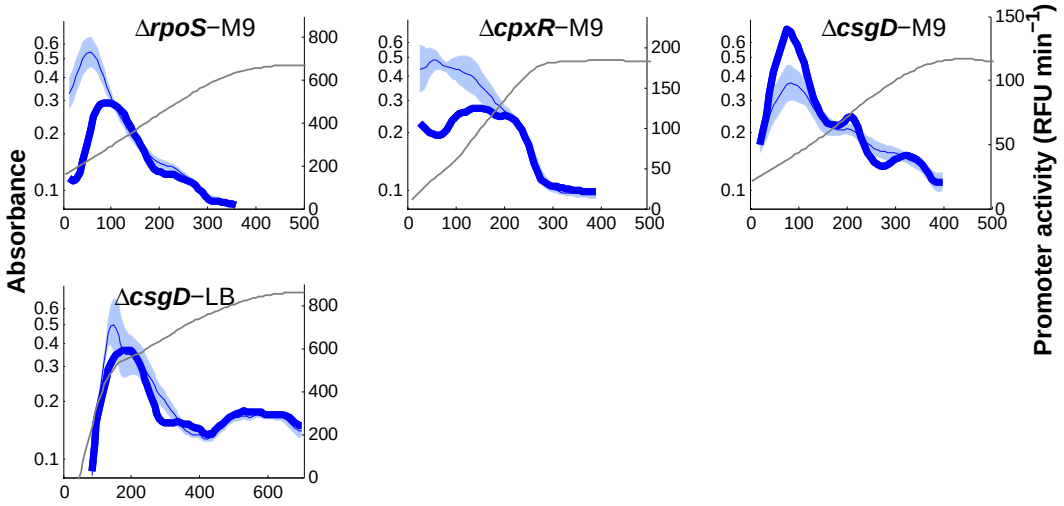

Figure 3.11: Fits of regulation function of *flgM* to reporter gene data when replacing protein concentrations by promoter activities and including global physiological effects. The regulation function of Equations 3.2 - 3.3 was fit using the promoter activities for *fliA*, and *flgM* shown inFigure 3.3, where the latter two replace the concentrations of FliA and FlgM, respectively. Moreover, global physiological effects are quantified by the activity of the constitutively expressed pRM promoter (Figure 3.8). The parameters were estimated using a hybrid genetic algorithm (see Section 3.2.5 for details). The best fit (thick solid blue line) returns the value $Q = 20$ for the objective function, for the parameter vector $(k_0, k_1, n, \theta, K) = (18, 14, 1.2, 817, 7307)$. The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

Figure 3.12 and Figure 3.13. As can be seen, the transcriptional pulse in exponential phase (Figure 3.3), leading to a transient accumulation of mRNA, is seen to be followed by the prolonged presence of stable protein, indicating that the promoter activity may indeed not be a good proxy for the protein concentration. Unfortunately, reporter concentrations are not always representative of the concentrations of proteins of interest, that is, proteins naturally expressed from a promoter. Post-transcriptional regulation and coding bias may cause divergent synthesis rates. The main bias, however, comes from the fact that the two proteins generally have different half-lives and thus different degradation rates (de Jong et al., 2010).

Available data in the literature indicate that the half-lives of FliA and FlgM are much shorter than the 19 h of the GFP reporter. The measured half-lives of FliA and

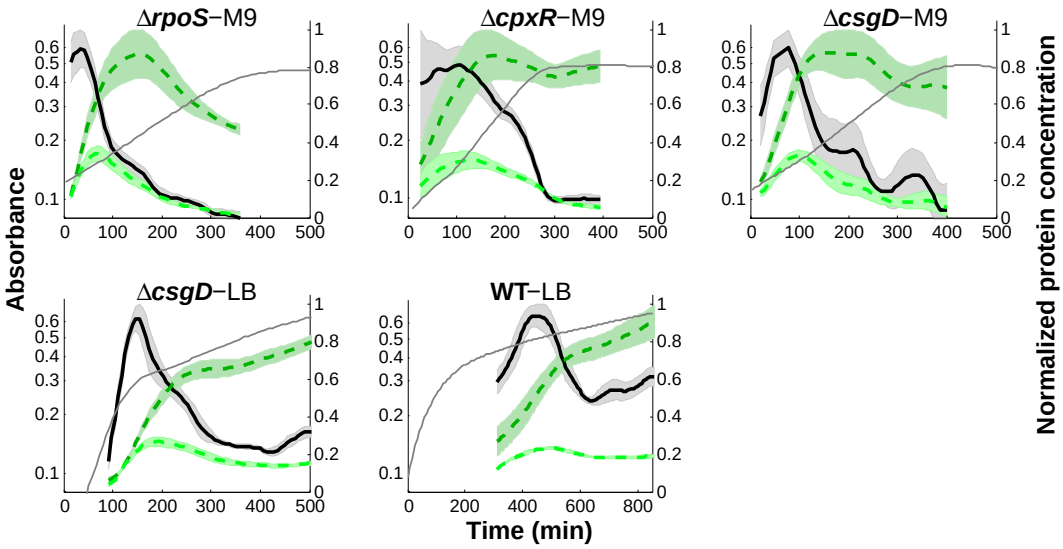

Figure 3.12: Estimates of FliA concentrations from reporter gene data. Concentrations of FliA (dashed line) computed from the *fliA* promoter activity (solid line) in all experimental conditions considered in this study. The *fliA* activities are the same as shown in Figure 3.3. The dark green line represents the concentration of the reporter protein, while the light green line represents the reconstructed concentration for the measured half-live of 30 min. Promoter activity has been normalized with respect to the maximum of the upper limit of its confidence interval in each condition. All protein concentrations have been normalized with respect to the maximum of the upper limit of the confidence interval of the reporter concentration in each condition. The shaded region corresponds to the mean of the promoter activities ± twice the standard error of the mean. Similar estimates of FlgM concentrations can be found in Figure 3.13.

FlgM in *Salmonella enterica* wild-type strains growing in LB were found to be 30 min and 18 min, respectively (Aldridge et al., 2006). These half-lives are much shorter than those commonly found for proteins in *E. coli*. This can be explained by the fact that, in addition to being physically degraded, FlgM is secreted from the cell. Moreover, FliA is subject to active degradation by Lon (Figure 3.1).

How can we exploit this information to reconstruct the protein concentration from the promoter activity? As shown in (de Jong et al., 2010) and Section 3.2.2, if the half-live of the protein of interest is known, then an estimate of its concentration can be reconstructed from the observed promoter activity using a simple kinetic model integrating the effects of protein synthesis and degradation as well as growth dilution of the protein. Figure 3.12 shows the result that is obtained for the FliA concentration, using

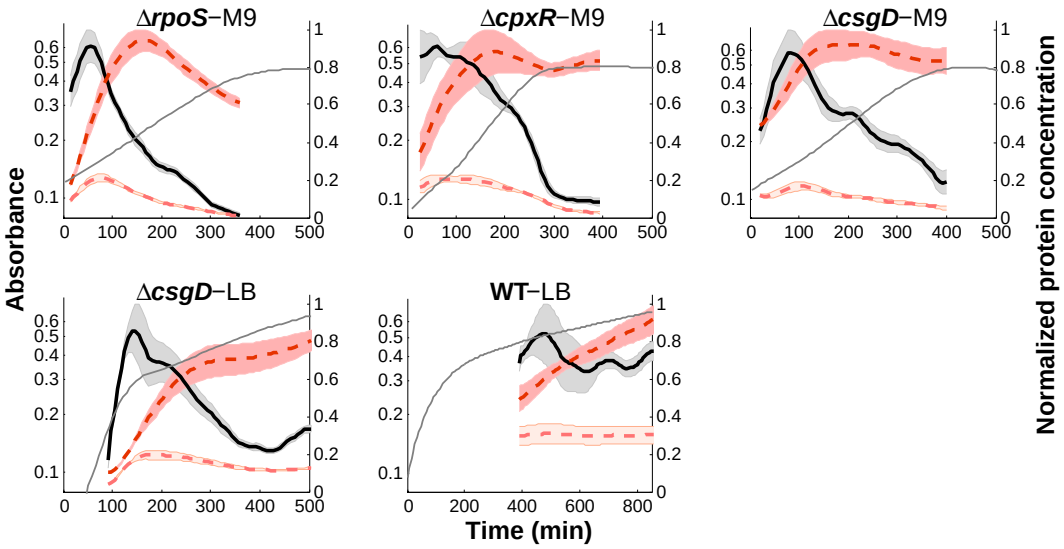

Figure 3.13: Estimates of FlgM concentrations from reporter gene data. Concentrations of FlgM (dashed line) computed from the *flgM* promoter activity (solid line) in all experimental conditions considered in this study. The *flgM* activities are the same as shown in Figure 3.3. The dark red line represents the concentration of the reporter protein, while the light red line represents the reconstructed concentration for the measured half-live of 18 min. Promoter activity has been normalized with respect to the maximum of the upper limit of its confidence interval in each condition. All protein concentrations have been normalized with respect to the maximum of the upper limit of the confidence interval of the reporter concentration in each condition. The shaded region corresponds to the mean of the promoter activities ± twice the standard error of the mean.

the above-mentioned half-life. Although the difference with the promoter activities is less striking than for the GFP concentrations, the computation of the concentration via integration of the corresponding activity smoothens out the rapid variations observed in Figure 3.3 and changes the time-varying profile of the regulators.

A tacit assumption in the computation of protein concentrations from promoter activities is that the half-lives of the proteins are constant over the duration of the experiment. This may not be true in our case, since the apparent half-lives of FliA and FlgM are regulated and depend on the presence of completed HBB structures. Data from the literature indicate that the first FlgM molecules appear in the extracellular medium shortly after the induction of *fliA* (Barembruch and Hengge, 2007; Karlinsey et al., 2000b). Once the cell population stops growing, the rate of assembling new flagella and thus the secretion of FlgM come to a halt as well. Since our kinetic experiments

have focused on the exponential growth phase, and the analysis is limited to the time frame in which *fliA* and *flgM* are expressed, the half-lives of FliA and FlgM have been assumed constant. Does the estimation of time-varying protein concentrations from the promoter activities, by means of a kinetic model and physiologically realistic half-lives, improve the inference of regulatory interactions and gene regulation functions? We performed the same tests as in previous cases, by checking if the minimal sign pattern structures remain consistent with the data when using only reconstructed protein concentrations of FliA and FlgM as regulators of *tar* and *flgM* and if the quantitative fit improves. For both FliA-dependent genes, we find that the model is structurally consistent with the data (Figure B.1 and Figure B.2 in Appendix B). However, the quantitative model of Eqs. 3.1-3.2 identified from the data is not particularly good (Figure B.3 and Figure B.4 in Appendix B). We then verified that a model using the reconstructed FliA and FlgM concentrations as regulators of *tar* and *flgM*, in addition to the activity of the gene expression machinery, is structurally compatible with the data. Minimal sign pattern analysis did not rule out the expected pattern of regulatory interactions, for both FliA-dependent target genes (Figure 3.14 and Figure 3.15). Second, we identified the gene regulation model of Eqs. 3.2-3.3 from the data, with the estimated FliA and FlgM concentrations for p_A and p_M, respectively. As shown in Figure 3.16, the model better captures the quantitative trend in the data, except for the $\Delta csgD$ -LB condition ($Q = 24.7$). Allowing the half-lives to vary around the measured values, which were obtained for a different species in growth conditions that are similar but not identical to ours, results in a very good fit in all conditions ($Q = 24.1$, Figure 3.16). Therefore, even approximately correct half-live values may allow the results of the inference process to be improved.

Interestingly, the estimated parameters show that the regulation function has a slightly different role than when activities are used as placeholders for protein concentrations. Since $n = 2.4$, the regulated term $k_1\, A(t)^n/(\theta^n + A(t)^n)$ in Eq. 3.3 has a (mildly) sigmoid form. The threshold value θ takes a value such that in experiments with strong induction of the flagellar cascade, and thus a strong peak in *fliA* activity ($\Delta rpoS$-M9 and $\Delta csgD$-LB), the regulated term covers the entire range of values from 0 to k_1 (Figure D.1-*C* in Appendix D). That is, contrary to the fits studied in previous

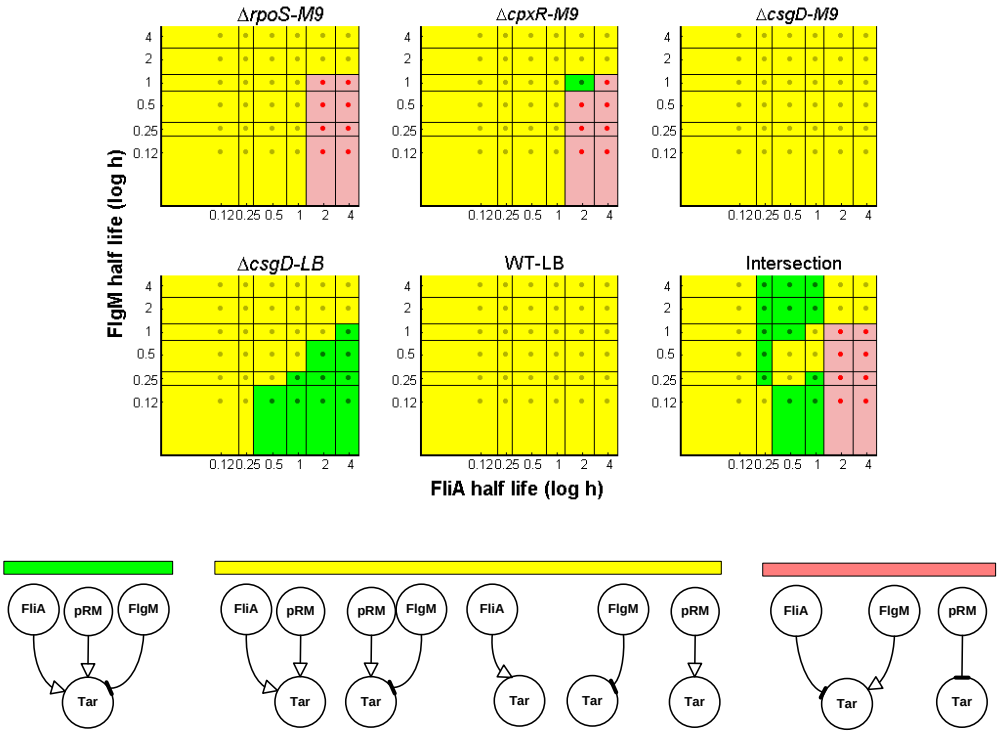

Figure 3.14: Minimal patterns of regulatory interactions for *tar* over a range of physiologically realistic half-lives. The minimal regulatory patterns for the gene *tar* in the motility network of Figure 3.7 as a function of the half-lives of FliA and FlgM. The plots correspond to the five experimental conditions considered (Δ*rpoS*-M9, Δ*cpxR*-M9, Δ*csgD*-M9, Δ*csgD*-LB, and WT-LB) as well as the pooling of the data sets from all five conditions. The dot in the center of each region in the plots corresponds to a tested combination of half-lives of FliA and FlgM, and thus to specific protein concentration profiles computed from the kinetic model of gene expression (Section 3.2.4). The minimal regulatory patterns were obtained by applying the minimal sign pattern algorithm (Porreca et al., 2010a). The color codes represent the different categories of minimal signal patterns inferred. A region is colored green if the expected regulatory patterns is among the minimal sign patterns returned by the algorithm, and yellow if it is compatible with the returned sign patterns. A region is colored red if none of the returned sign patterns is consistent with the data. Two examples of inconsistent sign patterns are shown. The values of the half-lives are represented in `log`.

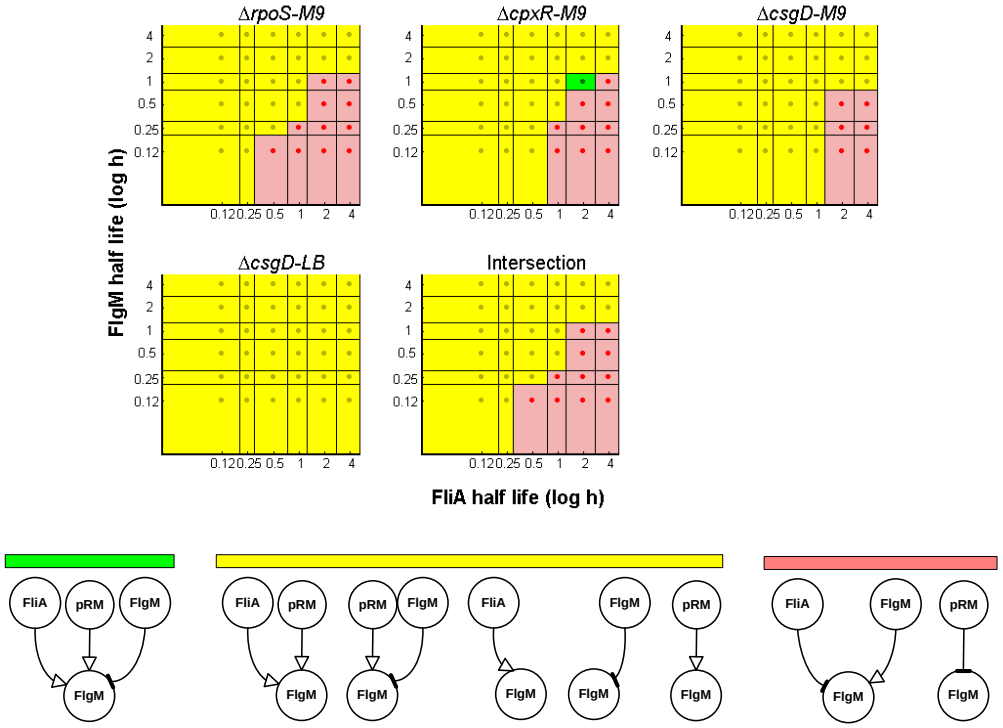

Figure 3.15: Minimal patterns of regulatory interactions for *flgM* over a range of physiologically realistic half-lives. The minimal regulatory patterns for the gene *flgM* in the motility network of Figure 3.7 as a function of the half-lives of FliA and FlgM. Similarly to Figure 3.14, the plots correspond to the four of the experimental conditions considered (Δ*rpoS*-M9, Δ*cpxR*-M9, Δ*csgD*-M9, Δ*csgD*-LB) as well as the pooling of the data sets from all five conditions. The condition WT-LB was not used in the analysis of the regulation of the *flgM* promoter. The dot in the center of each region in the plots corresponds to a tested combination of half-lives of FliA and FlgM, and thus to specific protein concentration profiles computed from the kinetic model of gene expression (Section 3.2.4). The minimal regulatory patterns were obtained by applying the minimal sign pattern algorithm (Porreca et al., 2010a). The color codes represent the different categories of minimal signal patterns inferred. A region is colored green if the expected regulatory patterns is among the minimal sign patterns returned by the algorithm, and yellow if it is compatible with the returned sign patterns. A region is colored red if none of the returned sign patterns is consistent with the data. Two examples of inconsistent sign patterns are shown. The values of the half-lives are represented in `log`.

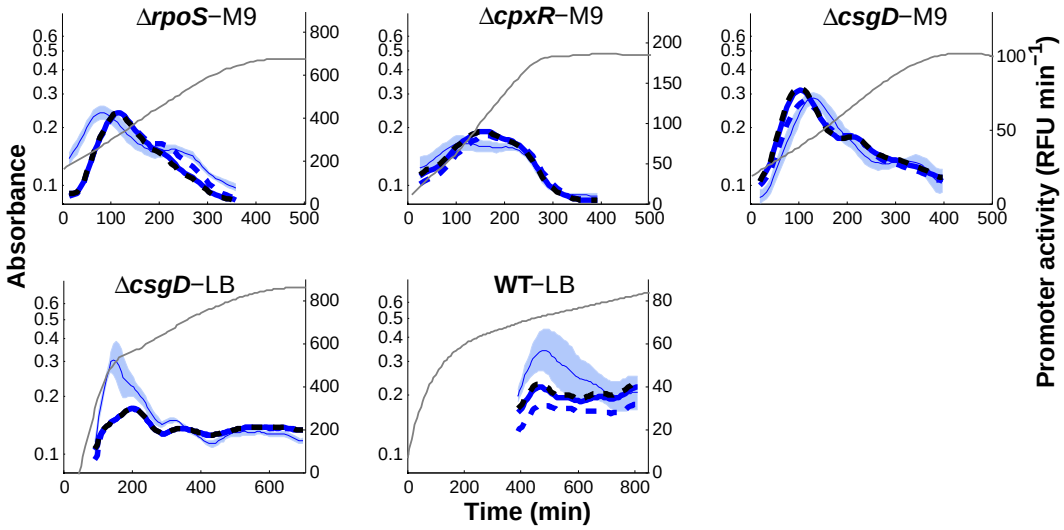

Figure 3.16: Fits of regulation function of *tar* to reporter gene data when reconstructing protein concentrations from the reporter gene data and including global physiological effects. The regulation function of Eqs. 3.2-3.3 was fit to the data using the promoter activity for *tar* (Figure 3.3), concentrations of FliA and FlgM reconstructed from the activities of their promoters for physiologically realistic half-lives (Figure 3.12 and Figure 3.13), and the activity of the constitutively expressed pRM promoter quantifying global physiological effects (Figure 3.8). The parameters were estimated using a hybrid genetic algorithm (see Section 3.2.5 for details). Three fits are shown, namely the best fit for measured half-lives of FliA and FlgM of 30 min and 18 min, respectively (solid line, $Q = 24.7$, $(k_0, k_1, n, \theta, K) = (0.3, 4.6, 2.4, 3030, 223750)$) and two other fits for comparable half-lives (dashed lines, $Q = 24.1$, $(k_0, k_1, n, \theta, K) = (0.2, 4.7, 2.2, 4535, 222800)$ and $Q = 25$, $(k_0, k_1, n, \theta, K) = (0.3, 4.6, 2.4, 2800, 162000)$). The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

sections, at high concentrations of FliA the *tar* promoter is maximally expressed, as expected.

The above analysis ignores a particularity of the FliA-FlgM module, namely that the half-lives vary across growth conditions. Generally speaking, in environmental conditions favoring a larger number of flagella, and thus completed HBB structures, the secretion rate of FlgM is higher and therefore the apparent half-life shorter. In mutant strains without HBB structures, and thus no protein secretion, the FlgM half-life is 3 h (Karlinsey et al., 2000a), while in some conditions half-lives up to 7 min were measured

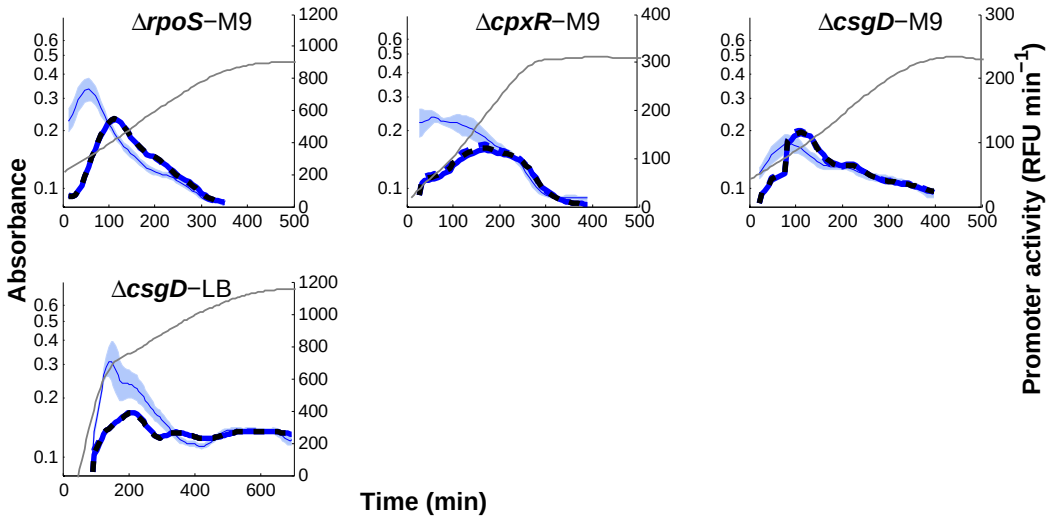

Figure 3.17: Fits of regulation function of *flgM* to reporter gene data when reconstructing protein concentrations from the reporter gene data and including global physiological effects. The regulation function of Eqs. 3.2-3.3 was fit to the data using the promoter activity for *flgM* (Figure 3.3), concentrations of FliA and FlgM reconstructed from the activities of their promoters for physiologically realistic half-lives (Figure 3.12 and Figure 3.13), and the activity of the constitutively expressed pRM promoter quantifying global physiological effects (Figure 3.8). The parameters were estimated using a hybrid genetic algorithm (see Section 3.2.5 for details). Three fits are shown, namely the best fit for measured half-lives of FliA and FlgM of 30 min and 18 min, respectively (solid line, $Q = 27.7$, $(k_0, k_1, n, \theta, K) = (0.3, 6.1, 2, 3170, 223750)$) and two other fits for comparable half-lives (dashed lines, $Q = 26.3$, $(k_0, k_1, n, \theta, K) = (0.4, 5.9, 2.3, 2358, 279500)$ and $Q = 27.1$, $(k_0, k_1, n, \theta, K) = (0.3, 6, 2.1, 2760, 162000)$). The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

(Karlinsey et al. (1998), see Aldridge et al. (2006) for intermediate values). The half-life of FliA, the flagellar sigma factor, is also variable. FliA is subject to active degradation by the Lon protease, but stabilized when bound to FlgM (Figure 3.1). This makes its apparent half-life dependent on the concentration of its anti-sigma factor (Barembruch and Hengge, 2007). The measured half-life of FliA in mutant strains without HBB structures, and thus with maximal protection of by FlgM, is 2 h (Barembruch and Hengge, 2007). In wild-type strains exponentially growing in LB medium, this value may decrease down to 30 min (Aldridge et al., 2006). In summary, the half-lives of

both FliA and FlgM are not identical across all growth conditions considered. While we can give upper and lower bounds on the half-lives, we do not know the exact value in most conditions.

This specific property of the FliA-FlgM module suggests a final extension of the analysis to improve the inference results. We allowed the FliA and FlgM half-lives to vary between physiologically possible bounds in each of the conditions and estimated not only the parameters of the regulation functions, but also the half-lives. In order to reduce the computational complexity of this procedure, we discretized the space of possible half-lives, selecting 27 values each for FliA and FlgM, and we precomputed the protein concentration profiles for each half-life in each of the experimental conditions. The resulting time-course patterns were used for the same analysis as above.

Figure 3.14 shows the results for the structural inference of *tar* regulators. As can be seen, almost all combinations of half-lives are compatible with activation of *tar* by FliA and the gene expression machinery and with inhibition by FlgM. This means that the structure of interactions is robust over a range of half-lives, a desirable property for network inference. Figure 3.18 illustrates that the quantitative regulation function of *tar* activity obtained is more precise than in all other previously considered situations ($Q = 20.9$), while the parameter values are similar to those obtained in the previous sections. Although we substantially relaxed the possible half-live values of FliA and FlgM, it is remarkable that the optimal values are close to the reported values for LB medium (Figure 3.18). This emphasizes the importance of active degradation of FliA and secretion of FlgM for the dynamics of the motility network. Moreover, while the proportion of FliA released by FlgM varies across conditions, most FliA is predicted to be free over the duration of the experiment (Figures C.1, C.2, C.3, C.4 in Appendix C). This is also intuitively expected, as FlgM is actively exported in the exponential growth phase considered. The best fit finds a cooperativity parameter equal to 1.8 (or 1.09 in the case of the second fit considered). *A priori*, positive cooperativity is not expected to occur, since FliA does not form a dimer and has only a single binding site in the promoter region. Buchler and Louis (2008) have shown, however, that the titration of a transcription factor may indeed lead to positive cooperativity (Buchler and Louis, 2008). The conditions they indicate in their analysis ($p_M > K$ and $p_M \approx p_A$) are satisfied here. Like for the fit with the measured half-lives in Figure 3.16, the activity of *tar* varies from 0 to its maximal possible value (Figure D.1-*F* in Appendix D). The

half-lives obtained in the case of the best fit vary around the measured values of half-lives for FliA and FlgM. We obtain a more stable half life for FliA (50 min) and FlgM (27 min) in the Δ*rpoS* -M9 condition. The values of half-live in all conditions can be seen in the legend of Figure 3.18.

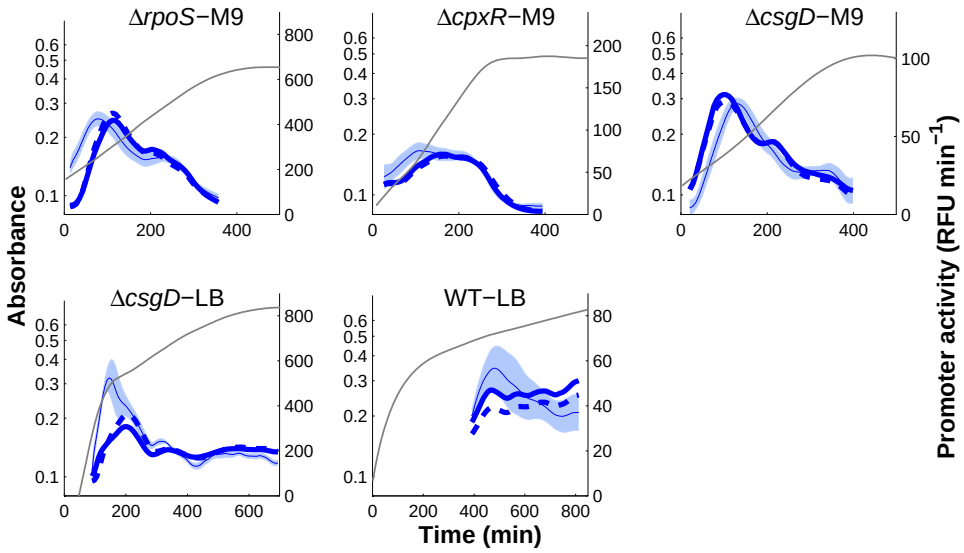

Figure 3.18: Fits of regulation function of *tar* to reporter gene data when reconstructing protein concentrations from the reporter gene data for physiologically realistic half-lives and including global physiological effects. As in Figure 3.16, but the half-lives have now also been estimated from the data, within a physiologically plausible range. Two example fits are shown, namely the best fit for estimated half-lives of FliA and FlgM (solid line, $Q = 20.9$, $(k_0, k_1, n, \theta, K) = (0.2, 5.1, 1.8, 3145, 17204)$) and another example of a high-ranking fit (dashed line, $Q = 21.09$, $(k_0, k_1, n, \theta, K) = (0.12, 8.5, 1.09, 24566, 88350)$). In the case of the best fit, the half-lives of FliA are equal to $(50, 24, 24, 35, 45)$ min in the (Δ*rpoS*, Δ*cpxR*, Δ*csgD*-M9, Δ*csgD*-LB, WT-LB) conditions, respectively, while the half-lives of FlgM are equal to $(27, 18, 24, 18, 18)$ min. In the case of the second fit, the half-lives of FliA are equal to $(60, 30, 24, 60, 30)$ min and the half-lives of FlgM are equal to $(9, 11, 24, 45, 7)$ min in the above experimental conditions, respectively. The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

The improvement in the fit is less evident in the case of *flgM* promoter (results are reported in Figure 3.19). One possible explanation is that the fit (in Figure 3.11) obtained using the promoter activities of *fliA* and *flgM* was already quite good, due

to the fact that the activity of *flgM* is among the regulators. However, the parameter values obtained when using protein concentrations of FliA and FlgM are similar to those obtained for *tar* promoter analysis (Figure D.2-*F* in Appendix D). However, the K parameter value is found to be approximately 10 times the maximum of FlgM concentration. This may be a consequence of the fact that the best fit is achieved for combination of half-lives and K values identified for *tar* regulation.

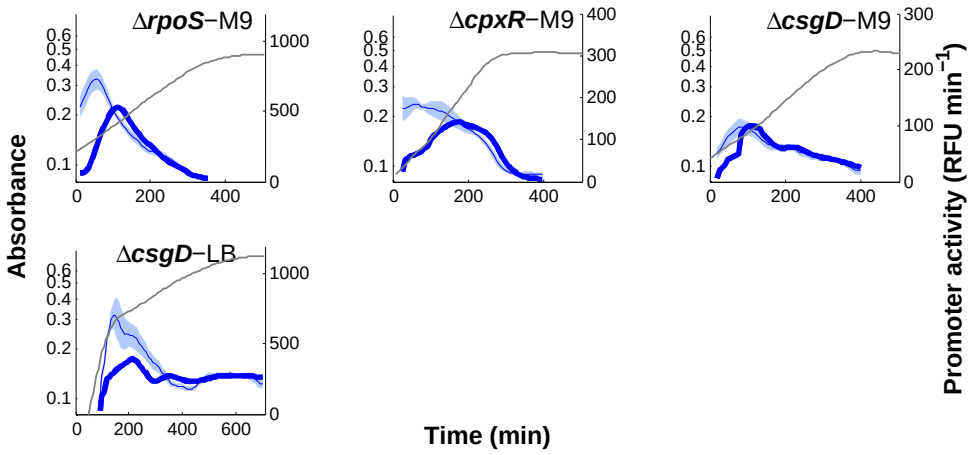

Figure 3.19: Fits of regulation function of *flgM* to reporter gene data when reconstructing protein concentrations from the reporter gene data for physiologically realistic half-lives and including global physiological effects. Similarly to Figure 3.18 the half-lives have been estimated from the data, within a physiologically plausible range. Two examples of fits are shown, namely the best fit for estimated half-lives of FliA and FlgM (solid line, $Q = 25$, $(k_0, k_1, n, \theta, K) = (0.45, 5.9, 2.4, 2930, 222850)$) and another example of a high-ranking fit (dashed line, $Q = 25.4$, $(k_0, k_1, n, \theta, K) = (0.4, 6.1, 2.3, 3030, 279550)$). In the case of the best fit obtained, the half-lives of FliA are equal to $(24\ min, 35\ min, 27\ min, 50\ min)$ in the ($\Delta rpoS$, $\Delta cpxR$, $\Delta csgD$-M9, $\Delta csgD$-LB) conditions, respectively. The half-lives of FlgM are equal to $(27\ min, 11\ min, 11\ min, 9\ min)$ in the ($\Delta rpoS$, $\Delta cpxR$, $\Delta csgD$-M9, $\Delta csgD$-LB) conditions, respectively. In the case of the second fit, the half-lives of FliA are equal to $(24\ min, 35\ min, 27\ min, 27\ h)$ and the half-lives of FlgM are equal to $(18\ min, 13\ min, 20\ min, 27\ min)$ in the above precised experimental conditions, respectively. The mean of the promoter activity of *tar* (thin solid blue line) and confidence intervals (shaded blue regions) are also shown in the figure.

The reconstruction of protein concentrations from transcription data results in much

better inference results for the FliA-FlgM module. The computation of the protein concentrations requires a simple kinetic model, accounting for protein synthesis and degradation, as well as estimates of the protein half-lives. While this increases the complexity of the data analysis procedures, it reflects the actual dynamics of gene expression and is thus critical for exploiting time-series measurements. Moreover, the availability of information on protein half-lives may not be constraining in practice, since even rough half-live estimates from the literature were seen to preserve the expected interaction pattern and provide a significant improvement of the ability of the models to quantitatively describe the time-varying promoter activity.

3.1.5 Determination of conditions in which protein half-lives and global physiological effects are important

The importance of accounting for global physiological effects and protein half-lives was demonstrated above for the regulation of the expression of *tar*. The same analysis was repeated for the regulation of the *flgM* promoter. We found that, for this promoter, the improvement in the fit to the experimental data obtained by including global physiological effects and protein kinetics is much less pronounced than for *tar*. One possible explanation is that the *flgM* activity profile happens to be already well explained using the promoter activities of *fliA* and *flgM* as proxies for the corresponding protein concentrations (Figure 3.6), thus leaving little space for improvement. In addition, from a mathematical viewpoint, we notice that using the promoter activity of *flgM* for the fitting of the same quantity may render the regression problem degenerate. Still, these results raise a more general question: When is it important to take into account protein half-lives and global physiological effects?

To answer this question we performed an *in-silico* analysis where the regulation model of Eqs. 3.2-3.3 is simulated for different protein half-lives and varying strength of the global physiological contribution, using the p*fliA*, p*flgM*, and pRM activity profiles reported in Figures 3.3 and 3.8. Identification is then attempted from the simulated data with models ignoring protein half-lives and global physiology. This enables us to quantify the relevance of the analysis in the previous sections for a variety of realistic scenarios, starting from experimentally measured activities of bacterial promoters.

To evaluate the importance of protein half-lives, we simulated FliA and FlgM concentration profiles for half-lives ranging between 7 minutes and 16 hours. The other

relevant parameters in the model (k_0, k_1, n, θ, K) were fixed in agreement with the best fit obtained for the reference half-lives of 30 min for FliA and 18 min for FlgM, shown in Figure 3.16. More precisely, the relative position of the parameter values within the interval of physiologically plausible values, which may depend on the FliA and FlgM concentrations, as explained in Section 3.2, was conserved across conditions. Activity profiles of *tar* were then generated in accordance with Eqs. 3.2-3.3 based on the experimentally measured pRM activities. We then attempted to identify from these simulated data a gene regulation model accounting for the global physiological effects, but using promoter activities in place of FliA and FlgM concentrations. The results are reported in Figure 3.20.

As can be seen, the quality of the fit decreases with longer half-lives of FliA, but is rather insensitive to the half-life of FlgM. The strong dependency on the half-life of FliA shows that, in general, accounting for slow protein kinetics is important, but that promoter activities can be safely used in place of protein concentrations for very fast-degrading proteins. This is intuitively explained by the fact that fast-degrading protein concentration profiles reproduce promoter activity profiles quite closely, while this is not true in case of slow degradation (Figure 3.12). The relative insensitivity to FlgM half-lives can be explained by the fact that, in the time window considered in our experimental set-up, a good fit requires most FliA to be free (Appendix C). Longer half-lives, and therefore higher concentrations of FlgM, favor lower free FliA concentrations, but this tendency is compensated in the parameter optimization process by higher values for the equilibrium constant K. The actually measured reference half-lives of 18 min for FlgM and 30 min for FliA are located in the upper left corner of Figure 3.20*A*, where fitting residuals are comparably small. Therefore, for networks involving regulators with longer half-lives than the exceptionally short half-lives observed for FliA and FlgM, it will be even more critical to account for protein kinetics.

To evaluate the importance of global physiological effects, starting from the experimentally measured pRM activity profiles, we simulated global physiological effects of different strength. In particular, we rescaled the variations of $f_{const}(t)$ around its temporal mean across all conditions, $\bar{f}_{const}$, by a factor α ranging from 0 (no variability, no regulatory effect) to 1 (measured variability, moderate regulatory effect) and 1.25 (increased variability, strong regulatory effect). That is, synthetic activity profiles of FliA-dependent promoters were generated in accordance with the model

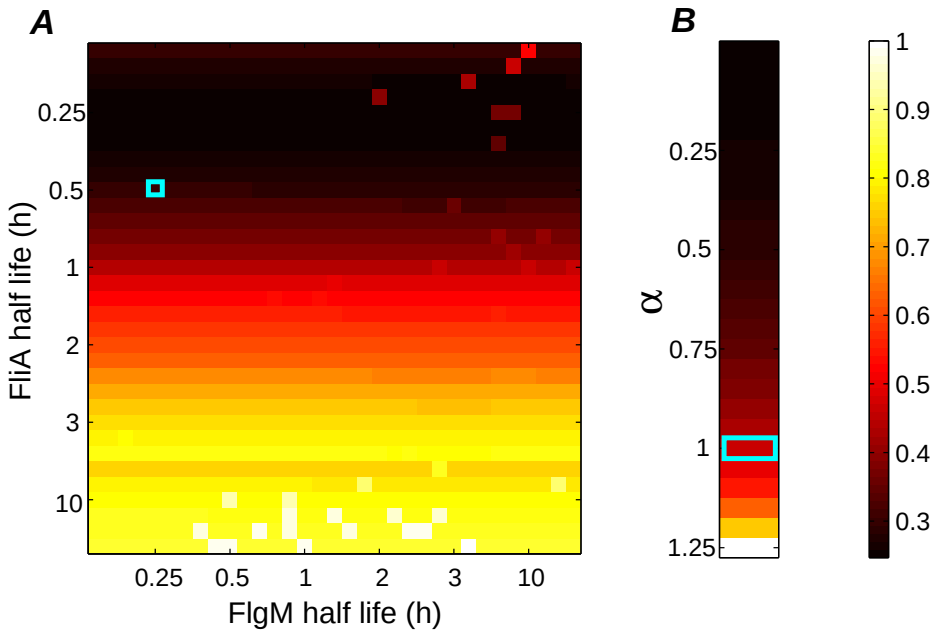

Figure 3.20: Heatmap of the fitting residuals, given by the value of the objective function Q, for simulated data generated for different protein half-lives and for different strengths of global physiological effects. *A:* For all different combinations of 33 half-lives of FlgM (horizontal axis) and FliA (vertical axis), the residual of the fit by a model ignoring protein kinetics is represented by the color code reported in the right bar. The combination corresponding to the measured half-lives in rich LB medium is marked with a light blue square (18 min for FlgM, 30 min for FliA). *B:* For 26 different values of the strength parameter α, defined in Eq. 3.4, the residual of the fit by a model ignoring global physiological effects is represented by the color code. The value corresponding to the real data is marked with a light blue rectangle ($\alpha = 1$).

$$f(t) = \left[\alpha \cdot \left(f_{const}(t) - \bar{f}_{const}\right) + \bar{f}_{const}\right] \cdot \left[k_0 + k_1 \frac{p_{A,free}(t)^n}{\theta^n + p_{A,free}(t)^n}\right], \tag{3.4}$$

with $p_{A,free}(t)$ computed from the FliA and FlgM concentration profiles according to Eq. 3.2. The upper bound of 1.25 for α was chosen so as to avoid negative values of the promoter activity $f(t)$.

Identification results using FliA and FlgM concentrations computed for the reference half-lives of 30 min and 18 min, respectively, but ignoring global physiological effects

are reported in Figure 3.20*B*. It is clear that the misfit of the *tar* promoter activity data increases with the strength of the ignored physiological effects. In particular, with the experimentally observed pRM activity ($\alpha = 1$), the discrepancy between the data and the best model fit is quite significant. This is in agreement with the results of previous sections, where accounting for global physiological effects turned out to be an important step toward improving the quality of the inference results. Neglecting small variations of global physiological state ($\alpha \ll 1$) is safer, but ignoring highly varying global physiological effects ($\alpha > 1$) may have even more severe repercussions on the inference results than those observed in the previous section.

In summary, the simulation study shows that, as expected, the importance of accounting for protein kinetics and global physiological effects depends on the strength of these effects, although the structure of the system itself may also play a role, as illustrated by the different dependency of the quality of the fit on FliA and FlgM concentrations (Figure 3.20*A*). As a general rule, ignoring significant fluctuations of the global physiology or large differences between mRNA and protein half-lives is very likely to result in modelling bias and hence poor inference results. Interestingly, in the previous sections a substantial improvement of the fit of a quantitative regulation function to *tar* activity was already obtained when taking into account concentrations of short-lived proteins and moderately-variable global physiological effects. In the light of the analysis of this section, the contribution of our approach becomes even more fundamental in other systems, bearing in mind that the vast majority of bacterial proteins are much more stable than FliA and FlgM, which are actively degraded and exported from the cell (Figure 3.1).

In the next chapter, we propose further guidelines for experimental design to facilitate the implementation of the approach developed here.

3.2 Methods and materials

3.2.1 Strains and growth conditions

The *E. coli* strains we used in this study are all derived from the wild-type strain BW25113. In particular, we used the Δ*rpoS* , Δ*cpxR* and Δ*csgD* deletion strains of BW25113 taken from the Keio collection (Baba et al., 2006). The mutants were reconstructed in our laboratory (Dudin et al., 2013) in order to eliminate the kanamycin resistance gene present in the original deletion strains (Table 3.1).

Strain	Characteristics	Reference or source
WT	*E. coli* BW25113	Baba et al. (2006)
WTpRM	*E. coli* BW25113 pRM-gfp::intS	This study
Δ*rpoS*	*E. coli* BW25113 Δ*rpoS*	Dudin et al. (2013)
Δ*cpxR*	*E. coli* BW25113 Δ*cpxR*	Dudin et al. (2013)
Δ*csgD*	*E. coli* BW25113 Δ*csgD*	Dudin et al. (2013)

Table 3.1: Strains used in this study.

The wild-type and mutant strains were transformed with low-copy plasmids bearing a *gfp* reporter gene (Table 3.2). The reporter plasmids for the genes *tar*, *fliA*, *flgM*, and *flgA* were selected from the plasmid library developed at the Weizmann Institute (Zaslaver et al., 2006). These low-copy pUA66*gfp* plasmids carry the kanamycin resistance gene and have the origin of replication of the pSC101 plasmid. The promoter regions of the genes of interest control the transcription of the gene encoding the stable GFP-mut2 reporter. The same vector was used to construct a reporter for the constitutive promoter pRM of the phage lambda, by cloning the pRM promoter region contained on the pZE1RM*gfp* plasmid used in Berthoumieux et al. (2013b) into the pUA66*gfp* plasmid backbone. Table E.1 in Appendix E lists the primer sequences used for the construction of this pUA66pRM-*gfp* plasmid using the Gibson Assembly method (Gibson, 2011). The above-mentioned plasmids were transformed into the wild-type and deletion strains of Table 3.1. We verified that the plasmids do not modify the growth of the transformed strains. All strains and plasmids were verified by PCR.

The pRM promoter fused with the *gfp* reporter gene was also inserted into the chromosome of the BW25113 wild-type strain as reference for the qRT-PCR assays. The

Plasmid	Characteristics	Reference or source
pUA66*gfp*	Kan^r, pSC101*ori*, *gfpmut2*	Zaslaver et al. (2006)
pUA66*fliA-gfp*	Kan^r, pSC101*ori*, *fliA* − *gfpmut2*	Zaslaver et al. (2006)
pUA66*flgM-gfp*	Kan^r, pSC101*ori*, *flgM* − *gfpmut2*	Zaslaver et al. (2006)
pUA66*flgA-gfp*	Kan^r, pSC101*ori*, *flgA* − *gfpmut2*	Zaslaver et al. (2006)
pUA66*tar-gfp*	Kan^r, pSC101*ori*, *tar* − *gfpmut2*	Zaslaver et al. (2006)
pUA66pRM-*gfp*	Kan^r, pSC101*ori*, *pRM* − *gfpmut2*	This study

Table 3.2: Plasmids used in this study.

WTpRM strain was constructed by using a linear DNA recombination protocole of Sharan et al. (2009). The pRM promoter region along with the gene encoding the GFPmut3 reporter were introduced into the *intS* loci on the chromosome of the BW25113 WT strain, by means of the λ Red system. pRM-gfpmut3 was recovered from the pZE1RMgfp plasmid used in Berthoumieux et al. (2013b).The recombineering protocols use the bacteriophage λ Red system that includes the phage recombination genes *gam*, *bet* and *exo*. The protein coded by *gam*, Gam, prevents *E. coli* nuclease from degrading linear DNA fragments (Karu et al., 1975; Murphy, 1991) thus allowing preservation of transformed linear DNA *in vivo*. The *bet* gene product, Beta, is a ssDNA binding protein that promotes annealing of two complementary DNA molecules (Karakousis et al., 1998), and the *exo* gene product, Exo, has a 5 to 3 dsDNA exonuclease activity (Cassuto et al., 1971). Working together these latter two proteins insert linear DNA at the desired target, creating genetic recombinants.

For all experiments, the strains were recovered from glycerol stock and grown overnight (16 h) at 37° C in LB rich medium or M9 minimal medium (Miller, 1972) supplemented with 0.3% glucose and mineral trace elements. For the preculture of strains containing plasmids, kanamycin (50 μg/ml) was added. The overnight cultures were diluted (10- to 100-fold) into a 96-well microplate, so as to obtain an adjusted initial OD_{600} of 0.2. The wells of the microplate contain 150 μl of the above medium, to which was added 1.2% of the buffering agent HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) for maintaining a constant external pH. The wells were covered with 60 μl of mineral oil to prevent evaporation. The microplate cultures were then grown for about 16 h at 37° C, with agitation at regular intervals, in a microplate reader (Fusion Alpha, Perkin-Elmer).

3.2.2 Experimental monitoring of gene expression in real time and data analysis

The expression of the fluorescent reporter genes in different genetic backgrounds and different growth media was monitored *in vivo* and in real time. About 150 readings each of absorbance and fluorescence were obtained during a typical experiment using the Perkin-Elmer microplate reader. The absorbance measured at 600 nm quantifies the biomass, while the fluorescence signal emitted at 520 nm, when excited at 485 nm, is proportional to the number of GFP molecules. In order to compute promoter activities and protein concentrations from these data, data analysis procedures were designed and implemented in MATLAB, completing earlier work (Berthoumieux et al., 2013b; de Jong et al., 2010; Ronen et al., 2002). These analysis procedures take into account for the specific half-life of the fluorescent reporter protein and implement procedures for subtracting the autofluorescence background.

3.2.2.1 Background subtraction

We first corrected the absorbance for the background absorbance of the growth medium. The corrected absorbance signal $A(t)$ is computed as

$$A(t) = A_u(t) - A_b(t), \tag{3.5}$$

where $A_u(t)$ is the primary absorbance signal and $A_b(t)$ the absorbance of the growth medium (M9 or LB, depending on the experiment).
The fluorescence signal was corrected for autofluorescence generated by wild-type bacteria carrying the pUA66*gfp* plasmid without any promoter driving the expression of *gfp* or no plasmid at all (in practice these two measures of the autofluorescence gave the same result). The autofluorescence depends on the (time-varying) population size. Since the culture generating the fluorescence signal of interest and the culture generating the autofluorescence signal may not be exactly synchronized, direct subtraction of the autofluorescence background is not always possible. We used a calibration procedure, such that the corrected signal $I(t)$ is defined by

$$I(t) = I_u(t) - s(A(t)), \tag{3.6}$$

where $I_u(t)$ is the primary fluorescence level and s a calibration function, mapping absorbance levels to autofluorescence levels. The calibration function is obtained by

fitting a cubic smoothing spline to the autofluorescence generated by bacteria carrying the promoterless pUA66*gfp* plasmid or no plasmid at all as a function of the absorbance. Splines have the advantage that they can be evaluated for any absorbance within the observed range and easily extrapolated beyond this range. Figure 3.21 gives an example of background correction of absorbance and fluorescence data, in the case of the *tar* reporter in the $\Delta cpxR$ mutant strain.

3.2.2.2 Computation of promoter activity and protein concentrations

Following the measurement model in (de Jong et al., 2010), we describe the expression of the gene of interest and of the reporter protein as follows (Chapter 2):

$$\frac{d}{dt}p(t) = f(t) - (\mu(t) + \gamma_p)\, p(t), \quad p(0) = p_0, \tag{3.7}$$

$$\frac{d}{dt}r(t) = f(t) - (\mu(t) + \gamma_r)\, r(t), \quad r(0) = r_0, \tag{3.8}$$

where $p(t)$ and $r(t)$ are the concentrations of the protein of interest and of the reporter protein, respectively, $\mu(t)$ is the time-varying growth rate, and γ_p, γ_r $[min^{-1}]$ are the degradation constants of the protein of interest and the reporter protein, respectively. Notice that in the case of FlgM, protein degradation includes both physical degradation of the protein and secretion through the cell membrane. The reporter protein concentration $r(t)$ and the promoter activity $f(t)$ can be computed by means of the formulas:

$$r(t) = \frac{I(t)}{A(t)}, \tag{3.9}$$

$$f(t) = \frac{d}{dt}r(t) + (\gamma_r + \mu(t))\, r(t) = \frac{\frac{d}{dt}I(t)}{A(t)} + \gamma_r \frac{I(t)}{A(t)}, \tag{3.10}$$

The reporter concentration is expressed in units RFU and the promoter activity in units RFU min^{-1}, as is usual for this kind of measurements (see (Berthoumieux et al., 2013b) and Chapter 2). The growth rate is easily estimated from the time-varying absorbance, using the standard relation $\mu(t) = d \ln A(t)/dt$.

We used cubic smoothing splines (`csaps` function in MATLAB) to fit the fluorescence and absorbance data and obtain estimates of $A(t)$, $I(t)$, $dA(t)/dt$, and $dI(t)/dt$. The half-life of the GFPmut2 reporter used in this study is 18 h ($\gamma_r = 0.0006 \pm 0.0001$)). The maturation time of GFPmut2 is short enough (4 min, Zaslaver et al. (2006)) to be

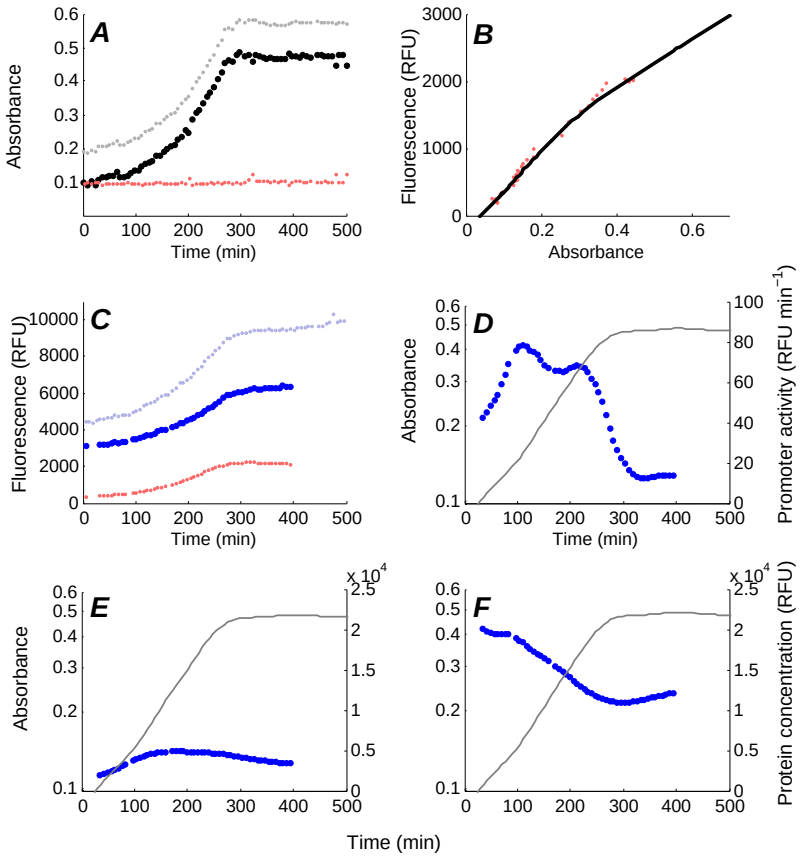

Figure 3.21: Illustration of data analysis procedures. Absorbance and fluorescence data acquired with the Δ*cpxR* mutant strain carying a pUA66*tar-gfp* plasmid, grown in M9 with glucose. *A:* Primary (uncorrected) absorbance (•, grey), background absorbance (•, red), and corrected absorbance (•, black). *B:* Calibration curve obtained by measuring the autofluorescence of the wild-type strain without plasmid. Primary fluorescence data are plotted against (corrected) absorbance data and the curve is obtained by fitting a smoothing spline. *C:* Primary fluorescence data (•, grey), and the corrected fluorescence (•, blue) obtained after subtracting the fluorescence of the background (•, red) as in Eq. 3.6. *D:* Promoter activity of *tar* (•, blue) computed from the corrected absorbance (-, grey) and corrected fluorescence by means of Eq. 3.10. *E:* Protein concentration of *tar* (•, blue) computed for a half-life of 2 h from the corrected absorbance (-, grey) and corrected fluorescence measurements using Eq. 3.7. *F:* Protein concentration of *tar* (•, blue) computed for a half-life of 18 h from the corrected absorbance (-, grey) and corrected fluorescence measurements using Eq. 3.7.

ignored.

In order to reconstruct the concentration of a protein of interest, we again use Eq. 3.7. The promoter activity, $f(t)$, is proportional to the synthesis rate, as explained in detail in Chapter 2. When the degradation constant is known, we can compute the protein concentration by numerical integration, starting from the initial concentration p_0. This initial concentration is obtained from the reporter gene data, by realizing that the bacterial cells at the beginning of the experiment are rediluted cells from a preculture grown in the same medium. In particular, assuming that gene expression in the preculture is at steady-state, it follows from Eq. 3.7 and Eq. 3.8 that

$$p(0) = p(T) = \frac{\mu(T) + \gamma_r}{\mu(T) + \gamma_p} r(T), \tag{3.11}$$

where $\mu(T)$ is the growth rate of the preculture at the time of redilution (at the time T), $p(T)$ and $r(T)$ are the corresponding concentrations of the protein of interest and reporter protein, respectively. Usually, the bacteria in the preculture are in stationary phase, so $\mu(T) = 0$. Eq. 3.7 was solved by numerical integration using the `quad` function in Matlab.

In the case of the motility network there are two complications that slightly modify this general scheme. First, the half-lives of FliA and FlgM are variable over the time-course of the experiment. During exponential growth, when the motility genes are expressed, FliA and FlgM have short half-lives, due to proteolysis and secretion, respectively. During stationary phase, at the end of the preculture, this is no longer the case and FliA and FlgM have larger half-lives. As a consequence, when computing the initial protein concentrations from the reporter concentrations at time T, we need to take protein degradation constants γ_p' corresponding to these larger half-lives. Second, in some experimental conditions, notably in rich medium like LB, the activity of the *fliA*, *flgM*, and *tar* promoter is negligible in the first few hours of the experiment (Kalir et al., 2001). As a consequence, the fluorescence intensity in the corresponding reporter strains is indistinguishable from the background fluorescence. We assume the promoter activity of the genes to be 0 in this case and back-extrapolate the observed promoter activities at earlier times towards 0. Figure 3.22 illustrates the effects of variable half-lives and extrapolation of promoter activities on the computation of FliA and FlgM concentrations in a WT strain. Moreover, in various experimental conditions (rich medium) the activity of the promoters can only be observed when the fluorescence

intensity overreaches the value of background fluorescence. When fluorescence intensity of FliA and FlgM was not observable before actual expression of genes we have assumed it to be 0 and we have interpolated the promoter activity values towards 0 for this part of the experiment. Figure 3.22 shows an example of effects of variable half-lives and interpolation of promoter activities on the computation of protein concentrations of FlgM and FliA in a WT strain.

For each of the derived quantities $r(t)$, $f(t)$, and $p(t)$, confidence intervals (defined as ± 2 standard errors of the mean) were computed from 6-7 experimental replicates.

3.2.3 Computation of minimal consistent sign patterns of regulatory interactions

In this section we adopt a notation, where vector $x = (x_1, \ldots, x_n)$ indicates concentration of regulators (e.g., FliA and FlgM), but may comprise in addition regulatory effects such a global cell physiology, depending on the context. What follows applies identically to all target genes of interest, *i.e. tar* and *flgM*. Let $f(x)$ be the promoter activity of one gene of interest.
We use the approach introduced in ((Porreca et al., 2010a)), which exploits time-series data of protein concentrations and promoter activities (protein synthesis rates) to infer patterns of regulatory interactions. The method relies on two assumptions:

1. $f(x)$ is monotonic in every x_j, with $j = 1, \ldots, n$;
2. A set of measurements $D = \{(\bar{x}^k, \bar{f}^k) : \ k = 1, \ldots, m\}$ of the concentration vectors x and the corresponding target promoter activities f are available, along with confidence intervals $\bar{f}^k \pm \epsilon_k$ and $\bar{x}^k \pm e_k$.

Assumption 1 of the method reflects the hypothesis that a regulator (*e.g.*, a transcription factor, but also the gene expression machinery) cannot operate both as a repressor and as an activator of a specific target gene (see Porreca et al. (2010a) and references therein), while it is allowed to operate as a repressor for one gene and as an activator for another gene. This corresponds to assuming that the activity of a gene is a monotone nondecreasing function of activators and a monotone nonincreasing function of repressors. Any such regulatory pattern can be encoded in terms of a sign pattern,

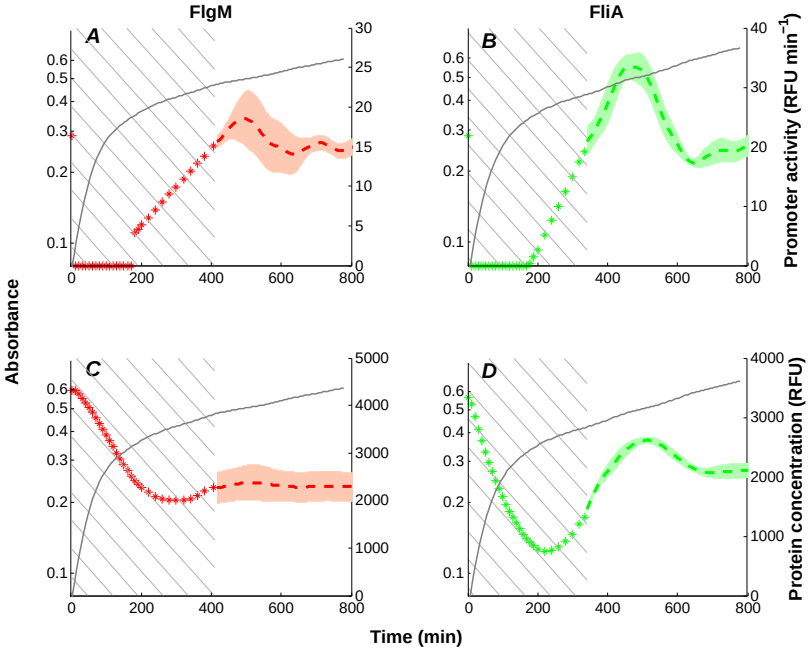

Figure 3.22: Effect of variable half-lives and promoter activity extrapolation on initial protein concentrations in a WT strain. *A*: The observed promoter activtity of *flgM* (dashed line, red) and its etrapolation (*, red). *B*: The observed promoter activtity of *fliA* (dashed line, green) and its etrapolation (*, green). *C*: The effect on the computation of the protein concentration of FlgM when taking into account the extrapolation of its promoter activity in *A* and the initial condition. Protein concentration was computed for an initial half-life of 3 h and a short half life of 18 min. *D*: The effect on the computation of the protein concentration of FliA when taking into account the extrapolation of its promoter activity in *B* and the initial condition. Protein concentration was computed for an initial half-life of 2 h and a short half life of 30 min. The hatched regions correspond to the regions where the activity of promoters was extrapolated, i.e. $[0, 410\ min]$ for *flgM* and $[0, 340\ min]$ for *fliA*. The promoter activities have been derived from the primary data as illustrated in Figure 3.2. The shaded regions correspond to the mean of the promoter activities and protein concentrations, respectively, $\pm$ twice the standard error of the mean. The absorbance is drawn in solid, grey lines.

i.e., a vector containing one entry per regulator, taking value $+1$ for activators, -1 for repressors, and 0 for factors that do not affect the expression of the gene under consideration. We may thus define the sign pattern $\pi = (\pi_1, \ldots \pi_n)$ of f by posing $\pi_j = 1$ if f is increasing in x_j, $\pi_j = -1$ if f is decreasing in x_j, and $\pi_j = 0$ if f is independent of x_j, $j = 1, \ldots, n$. The sign pattern encodes the directed, signed graph of the regulation of the gene under consideration by all possible regulators in the net-

work (compare Figure 3.23 B-E). As for assumption 2, data may come from several gene reporter experimental scenarios (different strains and media) and is provided in the required form by the processing of the previous Section 3.2.2, where $(\bar{x}^k, \bar{f}^k)$ is the measurement average at time t_k, while e_k and ϵ_k are fixed to twice the standard error of the mean $(\bar{x}^k, \bar{f}^k)$. Also observe that dependence of confidence intervals on index k is explicitly taken into account.

The rationale of the procedure for eliminating hypotheses from the set of all candidate sign patterns is the following (Porreca et al., 2010a). Given any two concentration vectors x' and x'', the implication

$$\pi_j(x''_j - x'_j) \geq 0,\ j = 1, \ldots, n \Rightarrow f(x'') \geq f(x')$$

is satisfied by the very definition of the sign pattern π of f. Therefore, for a hypothetical sign pattern $\bar{\pi}$ and perfect measurements ($\epsilon_k = e_k = 0$ for all k), any two data points $(\bar{x}', \bar{f}')$ and $(\bar{x}'', \bar{f}'')$ that falsify the implication allow one to conclude that $\bar{\pi}$ is not the sign pattern of f. In particular, if $\bar{f}'' < \bar{f}'$, the sign pattern $\bar{\pi}$ defined by $\bar{\pi}_j = 1$ if $\bar{x}''_j > \bar{x}'_j$, $\bar{\pi}_j = -1$ if $\bar{x}''_j < \bar{x}'_j$, and $\bar{\pi}_j = 0$ otherwise, is inconsistent with the data. In addition, any subpattern of $\bar{\pi}$, i.e. a pattern $\tilde{\pi}$ whose nonzero entries are equal to the corresponding entries of $\bar{\pi}$ (denoted with $\tilde{\pi} \sqsubseteq \bar{\pi}$), is also inconsistent with the data, since the implication above is still violated.

For instance, in the network module considered in this paper, the assumption that both FlgM and FliA activate *tar* can be rejected if two measurement times are found such that, for increasing concentrations of FlgM and FliA, the promoter activity of *tar* is decreased. The algorithm makes the above verifications in a computationally efficient way and returns, for every target gene, a set of minimal sign patterns. The minimal sign patterns are regulatory patterns consistent with the data, having the properties that removal of any interaction results in an inconsistent pattern, whereas addition of a regulator (activator or repressor) preserves the consistency. This test is easily robustified to account for measurement uncertainties, see Figure 3.23 for a graphical example on a network resembling *tar* regulation.

For any data point $(\bar{x}, \bar{f}) \in D$, let $(\hat{x}, \hat{f})$ and $(\check{x}, \check{f})$ indicate the confidence bounds $\hat{f} = \bar{f} + \epsilon$ and $\check{f} = \bar{f} - \epsilon$, in the same order, and similarly $\hat{x}_j = \bar{x}_j + e$ and $\check{x}_j = \bar{x}_j - e$, with $j = 1, \ldots, n$. Let the complexity c of a sign pattern π be the number of nonzero

entries of π. The algorithm is divided into two phases, conceptually organized as follows (see Figure 3.23 for reference).

Computation of inconsistent patterns $\bar{\Pi}$ from data D

- Set $\bar{\Pi} = \emptyset$
- For all pairwise different data points $(\bar{x}', \bar{f}')$ and $(\bar{x}'', \bar{f}'')$ in D:

  If $\hat{f}'' < \check{f}'$

  - Define $\bar{\pi} = (\bar{\pi}_1, \ldots, \bar{\pi}_n)$ by $\bar{\pi}_j = \begin{cases} 1, & \check{x}'' > \hat{x}' \\ -1, & \hat{x}'' < \check{x}' \\ 0, & \text{otherwise} \end{cases}$, with $j = 1, \ldots, n$
  - Include $\bar{\pi}$ in $\bar{\Pi}$

- Return $\bar{\Pi}$

At this stage, a generic pattern π is inconsistent if and only if $\pi \sqsubseteq \bar{\pi}$ for some $\bar{\pi} \in \bar{\Pi}$ (Porreca et al., 2010a).

Computation of minimal consistent patterns Π^* from $\bar{\Pi}$

- Set $\Pi^* = \emptyset$
- For $c = 0, 1, \ldots, n$:
  - Enumerate all possible patterns π of complexity c
  - For every such π:

    If $\nexists\, \bar{\pi} \in \bar{\Pi}$ such that $\pi \sqsubseteq \bar{\pi}$ (π is consistent), and

    If $\nexists\, \pi^* \in \Pi^*$ such that $\pi^* \sqsubseteq \pi$ (π is minimal consistent), then

    Include π in Π^*
- Return P^*

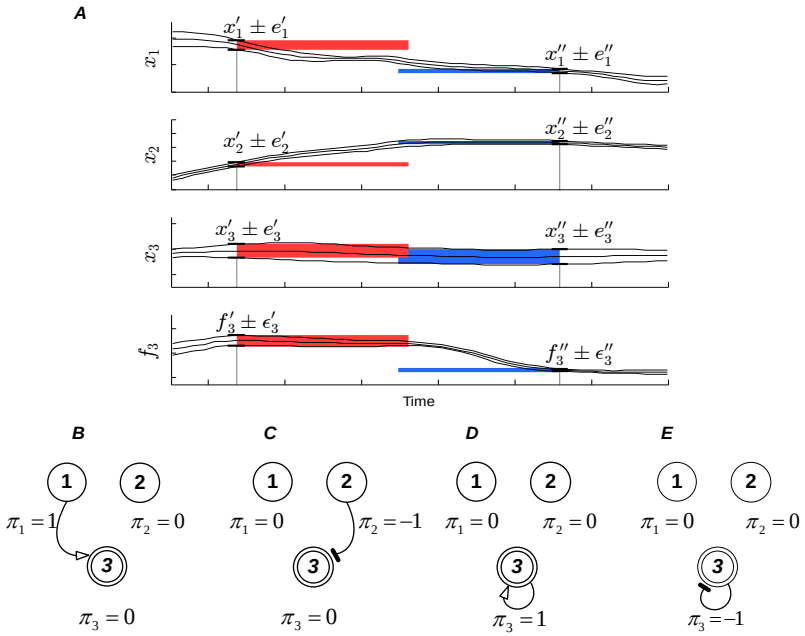

Figure 3.23: Computation of inconsistent and minimal consistent sign patterns from data. An example of the method of Section 3.2.3 is shown for the regulation of gene 3 in a hypothetical network with genes 1,2,3. *A:* From top to bottom, example time profiles and corresponding confidence intervals (thin black lines) for the concentrations of the proteins enconded by genes 1,2,3 and the synthesis rate of gene 3. Only the two data points (x', f'_3) and (x', f''_3) are considered in this example. Non-overlapping confidence intervals of f'_3 and f''_3 (reported next to each other for ease of comparison by the orange and blue shaded regions) imply $\hat{f}''_3 < \check{f}'_3$. Similarly, non-overlapping confidence intervals for x'_1, x''_1 and for x'_2, x''_2 imply $\hat{x}''_1 < \check{x}'_1$ ($\bar{\pi}_1 = -1$) and $\check{x}''_2 > \hat{x}'_2$ ($\bar{\pi}_2 = 1$), respectively, while confidence intervals for x'_3 and x''_3 overlap ($\bar{\pi}_3 = 0$). Whence, $\bar{\pi} = (-1, 1, 0)$. If this was the sign pattern of f_3, then $f_3(x)$ should increase for x'_1 decreasing to x''_1 and x'_2 increasing to x''_2 (x_3 is irrelevant in the hypothesis $\bar{\pi}_3 = 0$), whereas the observation says that $\hat{f}''_3 < \check{f}'_3$. Pattern $\bar{\pi} = (-1, 1, 0)$ is thus inconsistent with the data. *B - E:* Regulation patterns for gene 3, corresponding to the consistent sign patterns of f_3 deduced from the inconsistent patterns $\bar{\Pi} = \{\bar{\pi}\}$ obtained in *A*. Circles represent genes; directed arcs represent regulation of the target gene 3 by regulator j, with $j = 1, 2, 3$; arrow ends represent activation ($\pi_j = 1$), line ends represent inhibition ($\pi_j = -1$). Black arrows represent the minimal consistent sign patterns $(1, 0, 0)$ in *B*, $(0, -1, 0)$ in *C*, $(0, 0, 1)$ in *D*, and $(0, 0, -1)$ in *E*. Every consistent pattern is obtained from one of the cases *B* - *E* by turning the corresponding $\pi_j = 0$ into either $\pi_j = -1$ or $\pi_j = 1$.

At this stage, a generic pattern π is consistent if and only if $\pi^* \sqsubseteq \pi$ for some $\pi^* \in \Pi^*$ (Porreca et al., 2010a).

In practice, the above operations can be made computationally efficient. Notably, in our implementation, "for" loops and "if" tests are replaced by suitable algebraic and Boolean matrix operations in Matlab.

3.2.4 Derivation of regulation function of motility genes

We develop a kinetic model for the regulation of the expression of *tar* as a function of the total concentrations of FliA and FlgM (**see Figure 3.1 in Section 3.1.1**). The model is based on a quasi-equilibrium approximation of the mass-action kinetics for the formation of the FliA·FlgM complex, and a phenomenological Hill-type regulatory law of *tar* expression by free FliA.

Let $p_{A,free}$, $p_{M,free}$ and p_{AM} denote the concentrations of free FliA, free FlgM and FliA·FlgM, respectively, and let p_A, p_M denote total concentrations for FliA and FlgM. Assuming complex formation and dissociation are fast events relative to gene expression and protein degradation, we make the approximation

$$\frac{d}{dt}p_{AM} = k^+ \cdot p_{A,free} \cdot p_{M,free} - k^- \cdot p_{AM} \simeq 0,$$

with $k^- > 0$ and $k^+ > 0$. Using the facts that $p_A = p_{A,free} + p_{AM}$ and $p_M = p_{M,free} + p_{AM}$, substitution into the above to eliminate $p_{M,free}$ and p_{AM} from the equation yields

$$k^+ \cdot p_{A,free} \cdot \big(p_M - (p_A - p_{A,free})\big) - k^- \cdot (p_A - p_{A,free}) = 0,$$

which is a second-order polynomial equation in $p_{A,free}$. The solution of the equation that satisfies $0 \leq p_{A,free} \leq p_A$ is

$$p_{A,free}(p_A, p_M) = \frac{1}{2}\left(-(K + p_M - p_A) + \sqrt{(K + p_M - p_A)^2 + 4Kp_A}\right), \qquad (3.12)$$

with $K = k^-/k^+$, which is a function of the total concentrations p_A and p_M (Buchler and Louis, 2008). Only the free FliA molecules regulate the expression of the *tar* promoter, and we quantify the regulatory effect by the law

$$\frac{p^n_{A,free}}{p^n_{A,free} + \theta^n},$$

with $n \geq 1$. Multiplying by maximal synthesis rate k_1 and adding basal (unregulated) synthesis rate k_0 leads to the model we will be using to describe regulation of *fliA*-dependent genes

$$f(t) = k_0 + k_1 \frac{p_{A,free}^n}{p_{A,free}^n + \theta^n}, \tag{3.13}$$

Note that, in accordance with the expected regulatory pattern, the function $k_0 + k_1 \frac{p_{A,free}^n(p_A,p_M)}{p_{A,free}^n(p_A,p_M)+\theta^n}$ is increasing in p_A and decreasing in p_M. To verify this, it suffices to show that derivatives with respect to p_A and p_M are nonnegative and nonpositive, respectively. Ignoring k_0 and k_1 without loss of generality, the derivative of $p_{A,free}$ with respect to p_A can be written as

$$\frac{1}{2} + \frac{1}{2} \cdot \frac{-(p_M - p_A + K) + 2K}{\sqrt{(p_M - p_A + K)^2 + 4Kp_A}}.$$

This expression is obviously positive if $p_M - p_A + K \leq 0$. If instead $p_M - p_A + K > 0$, note that the expression is still positive if the square of the fraction,

$$\frac{\big((p_M - p_A + K) - 2K\big)^2}{(p_M - p_A + K)^2 + 4Kp_A},$$

is smaller than 1. But this is apparent since, under $p_M - p_A + K > 0$, the numerator is no bigger than $(p_M - p_A + K)^2$, whereas the denominator is no smaller than the same quantity. Similarly, the derivative of $p_{A,free}$ with respect to p_M can be written as

$$-\frac{1}{2} + \frac{1}{2} \cdot \frac{p_M - p_A + K}{\sqrt{(p_M - p_A + K)^2 + 4Kp_A}}.$$

This expression is obviously negative if $(p_M - p_A + K) \leq 0$. If instead $(p_M - p_A + K) > 0$, note that the square root is no smaller than $(p_M - p_A + K)$, hence the rightmost fraction is no bigger than 1, i.e. the overall expression is again negative.

The additional regulatory effect of the global physiological state of the cell is quantified via further multiplication by a function $f_{const}(t)$:

$$f(t) = f_{const}(t) \left[k_0 + k_1 \frac{p_{A,free}^n}{p_{A,free}^n + \theta^n} \right], \tag{3.14}$$

Monotonicity with respect to p_A and p_M remains unchanged. In addition, f is increasing in f_{const}. In all cases, the model depends on the (nonnegative) parameters k_0, k_1, n, θ, K.

3.2.5 Parameter estimation

The promoter activity models we have considered in the Section 3.1 have the form $f(t) = f\big(x(t), c\big)$, where c is a vector of unknown parameters and x is a vector of regressors. The specific form of $f\big(x(t), c\big)$ is given for *tar* in **Eq. 3.13 and 3.14**, and is analogous for *flgM*. The regressors take different forms in consecutive sections of this thesis, consisting either of the activities f_A and f_M of the *fliA* and *flgM* promoters ($x = (f_A, f_M)$) or the reconstructed concentrations p_A and p_M of FliA and FlgM ($x = (p_A, p_M)$). In all sections, $c = (k_0, k_1, n, \theta, K)$. The superscript symbol s indicates the experimental condition, where $s \in S = \{\Delta rpoS\text{-M9}, \Delta cpxR\text{-M9}, \Delta csgD\text{-M9}, \Delta csgD\text{-LB}, \text{WT-LB}\}$. Given measurements $\big(\bar{x}^s(t), \bar{f}^s(t)\big)$ of $\big(x(t), f(t)\big)$ (averages of 6-7 experimental replicates) at times $t \in T^s$ along with confidence intervals $(\bar{f}^s(t) \pm \epsilon^s(t))$ (computed from the same experimental replicates with ϵ^s equal to twice the standard error of the mean $\bar{f}^s$), we estimate c by solving the optimization problem

$$\min_{c \in C} Q(c), \qquad Q(c) = \sum_{s \in \S} \sum_{t \in s} \frac{1}{2\epsilon^s(t)} \left| \bar{f}^s(t) - f\big(\bar{x}(t), c\big) \right|^2 .$$

The solution is found in MATLAB using the numerical global search function `gs` with standard settings (interior-point method `fmincon` for local minimizations). For *tar* activity, the search is initialized at the values $(k_0, k_1, n, \theta, K) = (\hat{k}_0, \hat{k}_1, \hat{n}, \hat{\theta}, \hat{K})$ defined as $\hat{k}_0 = \min\{\bar{f}^s(t) : \ t \in T^s, s \in S\}$; $\hat{k}_1 = \max\{\bar{f}^s(t) - \hat{k}_0 : \ t \in T^s, s \in S\}$; $\hat{n} = 1$; $\hat{K} = \bar{\bar{x}}_1$, where the double bar stands for mean over $t \in T^s$ and $s \in S$; and, in view of Eq.3.12,

$$\hat{\theta} = \frac{1}{2}\left(-(\hat{K} + \bar{\bar{x}}_2 - \bar{\bar{x}}_1) + \sqrt{(\hat{K} + \bar{\bar{x}}_2 - \bar{\bar{x}}_1)^2 + 4\,\hat{K}\,\bar{\bar{x}}_1} \right).$$

The parameter search space C is given by the constraints $k_0 \geq 0$, $k_1 \geq 0$, $n \in [1, 3]$, $\theta \in [0, 10 \times \bar{p}_{A,free}]$, $K \in [0, 10 \times K_{max}]$, where $K_{max} = max\{x_2^s(t) : t \in T^s, s \in S\}$. For the estimation of the regulation function of *flgM*, the condition WT-LB is not available and hence excluded from the computation of $Q(c)$. Moreover, K is fixed for biological consistency, in this case, to the value inferred from the fitting of *tar* promoter activity.

3.2.6 Validation of reporter gene data using quantitative RT-PCR

We verified the reporter gene measurements by means of qRT-PCR in the WT-LB condition, following a previously described protocol (Lee et al., 2006).

According to Eq. 2.13 in Chapter 2, the ratio of the promoter activities f_1, f_2 of two genes is proportional to the ratio of the mRNA concentrations m_1, m_2, that is,

$$\frac{f_1(t)}{f_2(t)} = \frac{k_{p,1}}{k_{p,2}} \frac{m_1(t)}{m_2(t)}. \tag{3.15}$$

Measuring gene expression by qRT-PCR allows the relative abundances of the mRNA of a target gene to be quantified with respect to the mRNA of a reference gene (VanGuilder et al., 2008). This provides a direct way to verify if the relative promoter activities measured with reporter genes are confirmed by another, independent experimental method. We compared the promoter activity of *tar*, as an example of a motility gene, with the activity of the constitutive pRM promoter. The validation of the ratio f_{tar}/f_{pRM} was carried out by means of the WTpRM strain, a modified BW25113 strain carrying a natural copy of *tar* and a transcriptional fusion of the pRM promoter with a *gfp* gene inserted into the *intS* loci on the chromosome (Section 3.2.1). Quantitative RT-PCR was used to quantify the relative abundances of *tar* and *g*fp mRNA, using a standard qPCR protocol (Lee et al., 2006).
We took 5 μL samples at 11 time-points from cultures of the WTpRM strain, growing in a microplate under the conditions described in Section 3.2.1. Total mRNA was protected using the RNAprotect Bacteria Reagent kit (Quiagen) and then extracted using the RNeasy mini kit (Quiagen) according to the protocols of the manufacturer. The RNA samples were then treated using the turbo DNAse (Ambion) to avoid DNA contamination. Approximately 1 μg of RNA for each of the 11 time-points was reverse transcribed using SuperScript II Reverse Transcriptase (Invitrogen). The cDNA samples were diluted 10x into MESA Green qPCR Master Mix (Eurogentec), supplemented with primers for the *tar* and *gfp* genes. Quantitative PCR was performed in a StepOnePlus Real-Time PCR System (Applied Biosystems) according to the instructions of the manufacturer. Briefly, 5 μl reaction mixtures were incubated for 10 min at 95 ^{0}C and 40 PCR cycles (15 s at 95° C, 10 s at 62° C and 10 s at 70° C). PCRs were run in quadruplicate. Raw data were transformed into threshold cycle (C_T) values. PCR amplification efficiencies for *tar* and *gfp* were determined by constructing standard curves from serial dilutions (Lee et al., 2006).

The results were analyzed by means of a standard model for computing $m_1(t)/m_2(t)$ at the sample time-points t with respect to $m_1(t_0)/m_2(t_0)$, the same quantity at a ref-

erence time-point t_0 (Pfaffl, 2001):

$$q(t) = \frac{m_1(t)/m_1(t_0)}{m_2(t)/m_2(t_0)} = \frac{E_{gfp}^{\Delta C_T^{gfp}}}{E_{tar}^{\Delta C_T^{tar}}}, \tag{3.16}$$

where C_T^{gfp} and C_T^{tar} are the measured C_T values for *gfp* and *tar*, respectively, $\Delta C_T^{gfp}(t) = C_T^{gfp}(t) - C_T^{gfp}(t_0)$, $\Delta C_T^{tar}(t) = C_T^{tar}(t) - C_T^{tar}(t_0)$. As our reference time-point, we chose a measurement during exponential growth on glucose. As a consequence, the changes in mRNA abundance are relative to the mRNA abundance in exponential phase. The efficiencies were measured to be 109% for *gfp* ($E_{gfp} = 2.09$) and 105% for *tar* ($E_{tar} = 2.05$).

From Eq. 3.15 it follows that

$$q(t) = \frac{f_1(t)/f_1(t_0)}{f_2(t)/f_2(t_0)}. \tag{3.17}$$

The right-hand side of this equation can be computed from the measured promoter activities, as explained in Chapter 2. Figure 3.24 compares the value of $q(t)$ measured by means of reporter genes and qRT-PCR. There is a good qualitative and quantitative correspondence between the two independent methods (qRT-PCR and gene reporter genes) for measuring gene expression.

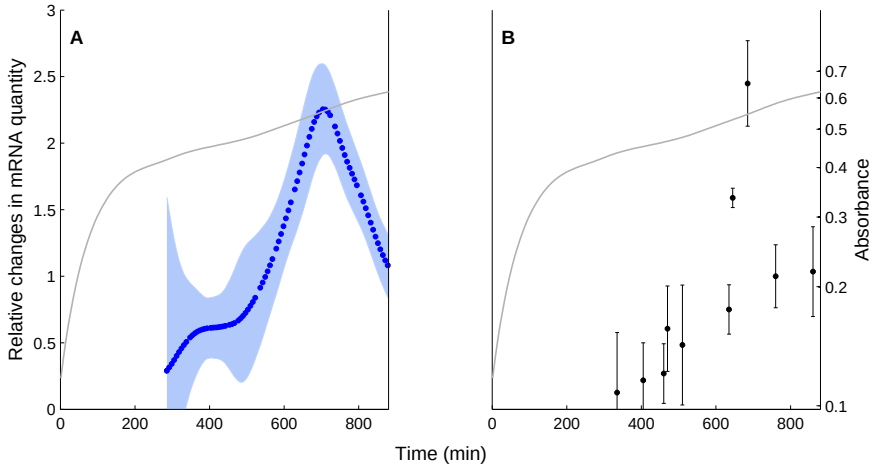

Figure 3.24: Validation of the reporter gene measurements using quantitative RT-PCR. *A*: The figure shows the promoter activity of *tar* with respect to global physiological effects (•, blue). The promoter activities were derived from corrected absorbance and fluorescence, measured using plasmids expressing the GFP reporter for the phage promoter pRM and the *tar* promoter. The shaded regions represent confidence intervals computed as ± the standard error of the mean of 5 replicates. *B*:The figure reports the expression of *tar* gene with respect to the mRNA quantity of pRM (WTpRM strain, Section 3.2.1) measured by qPCR (•, black). We have normalized with respect to the observed Tar mRNA quantity in exponential growth following Eq. 3.16. Expression of *tar* gene is observed to be maximal around 700 min. The results obtained by using the two independent techniques are in good agreement. The errorbars were computed from the standard error of the mean of 4 replicates. The absorbances have been also plotted on the figures (solid line, grey).

4. Conclusion

4.1 Summary of results

In this thesis, we develop methods for inferring genetic regulatory interactions and quantitative gene regulatory functions from gene expression profiles. The validation of the approach proposed in this work is carried out on a well-studied but complex gene regulatory network responsible for the motility of the bacterium *E. coli* . This chapter summarizes the main contributions of the thesis and puts our results into perspective relative to other methods for network inference.

While many problems persist in current methods for inferring gene regulatory networks from expression data, we believe that the solution should not only be sought in technical improvements of the algorithms themselves, but may come from a better understanding of the precise information on gene expression provided by the experimental data and their integration into appropriate modeling formalisms. The relation between the primary data and physiological quantities like the cellular concentrations of mRNA and protein is usually indirect and obscured by simplifications and assumptions that do not generalize beyond the specific situations for which they were designed. The main regulators of gene expression are proteins. Even though the concentrations of mRNA and proteins are weakly correlated at steady state, this is generally not the case when considering time-varying, dynamical expression data.

Reporter gene systems can yield gene expression profiles in complex *in vivo* experiments. Such dynamical data are well described by ordinary differential equations. We have therefore adapted an ODE modeling framework to convert (fluorescent) reporter gene data into biologically relevant quantities, such as promoter activities and protein concentrations. This approach is more appropriate for the inference of gene networks in the sense that it explicits the relation between experimental data and physiological quantities by means of mathematical (measurement) models of gene expression. Although an important source of the incomplete and sometimes spurious findings of classic inference algorithms, not many modeling frameworks for inference of gene regulatory networks focus on the distinction between these quantities.

Computing protein concentrations by means of the measurement models presented in this thesis supposes that we know the approximate values of the protein half-lives. With the exception of yeast, genome-wide studies of the stability of individual proteins in microorganisms are rare. It should be noted though that most proteins in *E. coli* are stable, with half-lives >10 h, so that the decay of protein concentrations is dominated by growth dilution, that is, $\mu \gg \gamma_p$. In other words, in many situations, in order to obtain a reasonable estimate of the effective protein half-life, it will be sufficient to perform the experiments in growth media supporting bacterial growth rates that results in doubling times well below 10 h.

Moreover, many methods for network inference rest on the common and tacit belief that the regulation of gene expression in bacteria is controlled uniquely by transcription factors and other specific regulators. In fact, the complex regulatory activity of the transcriptional and translational machinery of the cell, as well as other intrinsic global physiological effects, may induce major changes in gene expression over the course of an experiment . As Lovén et al. (2012) point out, many transcriptome studies assume that the total quantity of RNA is similar between different experimental conditions and use this quantity for normalization of the data. As a consequence, a global increase or decrease of transcriptional activity across conditions may lead to erroneous interpretation of the experimental data and the inference of spurious regulatory interactions. In this thesis, we have developed an improved model that, in addition to specific regulatory interactions, accounts for global regulatory effects of bacterial cells. We estimate the global physiological state of the cell from the activity of a constitutively expressed gene and whose expression only depends on the activity of the transcriptional and translational machinery.

Our analysis of reporter gene datasets thus makes it possible to account for the difference between mRNA and protein concentrations as well as for global physiological effects. We have validated the adequacy of the models to describe bacterial gene expression by comparing the deduced promoter activities with independent RT-qPCR measurements. We have shown that the inclusion of information about both global physiological effects and protein concentrations can improve the inference of regulatory interactions and the identification of quantitative regulation functions from time-series data. Compared to the classical inference approach, i.e., the inference of structure and

quantitative functions of gene regulatory networks from promoter activities, the inclusions of global regulatory effects significantly improves the prediction results. A further improvement is achieved by explicitly considering proteins as the regulators by taking into their half-lives.

To provide an integrated and straightforward inference approach, our method combines the above described experimental and computational models with a structural and parametric identification algorithm. The algorithm aims at inferring the structure of gene networks from time-series data by exploring the monotonicity properties of the network (or the model) and recovering only structures in good agreement with the data. For these structures, parameter estimation is performed to find the best prediction of both the (qualitative) model structure and the (quantitative) parameters. Generally, large search spaces reduce the performance, and may even compromise results of inference algorithms, the selection of data-consistent gene network models allows us to focus on precisely analyzing a small set of candidate models. Ultimately, this leads to results that are interpretable and relevant to the initial biological question.

The practical validation of our approach rests on the study of a gold standard biological system, the FliA-FlgM (motility) module in *E. coli*. Although atypical, this module possesses rich dynamics. The short half-lives of the FliA and FlgM proteins are time-varying and depend on the experimental conditions, inducing the time course of flagella synthesis. We investigated the capacity of our approach to infer from reporter gene data both the regulatory structure and the quantitative regulation function of two uniquely FliA-dependent motility genes (*flgM* and *tar*). When integrating information on the activity of the gene expression machinery and reconstructing protein concentrations from promoter activities, both the structure and the dynamics of the regulation of the *tar* and *flgM* promoters could be inferred successfully. For this analysis, we used available measurements of FliA and FlgM half-lives. This extended model contains the same number of parameters as the initial model and an improved fit between data and model is therefore not simply a consequence of increasing the degrees of freedom. Our results also confirmed the importance of global physiological effects and the active regulation of FliA and FlgM half-lives in predicting the activity of FliA-dependent promoters. When global physiological effects were ignored, or the FliA and FlgM half-lives were set to typical values of stable *E. coli* proteins, a sharp drop in the quality of the prediction of the gene regulation models was observed.

More generally, under which conditions does the inclusion of the above factors lead to better results and when can they be ignored? We performed a simulation study in which we systematically varied the relative contribution of global physiological effects to cross-condition variations in the expression of a target gene and the half-lives of the regulators. These results showed that increasing half-lives of the activating transcription factor and stronger variations of global physiological effects make it more difficult to obtain good fits when using promoter activities and data on specific regulators only, respectively. While these conclusions are not surprising, it is important to emphasize that in the system studied here, where FliA and FlgM have half-lives that are exceptionally short for bacterial proteins, a considerable improvement of the fit could be obtained. For regulatory proteins with more typical half-lives, the gain may therefore be even more important than observed here.

4.2 Perspectives

Our method to more fully exploit the information contained in time-series (population-averaged) data of the transcriptional response of bacterial cells to a changing environment depends on kinetic models of gene expression, relating the primary fluorescence and absorbance data to promoter activities and protein concentrations. In order to further improve the estimation of the biologically important quantities (promoter activity, etc.) from the primary data, we could take into account delays that are due to the maturation of GFP and the time for rounds of transcription and translation to complete. This refinement was not necessary here since the GFP reporter used in our study is fast-folding and the transcription and translation delays are very short with respect to the time-scale of the experiments.

In principle, it should be possible to apply the same approach to the inference of regulatory networks from high-throughput transcriptome data, such as DNA microarray and RNAseq data. The primary data would directly yield mRNA measurements, eliminating one step of the data treatment. However, since our algorithm relies on the interpretation of the dynamics of the system, we would need a relatively high sampling density. Furthermore, since we need to estimate the error on the change in mRNA or protein concentrations, many replicate experiments would have to be performed. Combining a high sampling density with numerous replicates may rapidly become very

costly. Future improvements, and the associated cost reduction, of sequencing techniques may soon make RNAseq data available for being analyzed by our algorithm.

We have found that an estimate of the protein concentration is crucial for improving the reliability of the network reconstruction. Since the measurement of mRNA is much easier than the direct measurement of protein concentrations, this implies that we need to know (or estimate) the protein half-lives. As pointed out above, in fast-growing cultures, the effective protein half-life is dominated by growth dilution. Experiments carried out in these conditions are directly amenable to analysis by our method. However, often this is not possible, for example, when measuring the transition from exponential growth to stationary phase. In this case, we would need to measure protein half-lives, for example using Western blots. However, such experiments are time and money consuming. An alternative could be to use translational fusions of the genes in the network to different flavors of fluorescent proteins. The stability of the protein would be directly correlated with the easily observable fluorescence signal. Control experiments would have to ascertain that the GFP-tag does not affect the half-life of the protein. The ideal solution, a direct observation of the proteins in the cell, for example by quantitative proteomics, remains too time and money consuming even more so than in the case of high-throughput transcriptomics.

The method we propose has the advantage that it can be used to monitor the dynamics of gene expression and global physiological effects in real time, without any additional preparation steps. However, reporter constructs have to be constructed for the genes of interest on plasmids or integrated into the chromosome. Such cloning tasks become increasingly automated and it seems possible that in the near future, many hundreds of such constructs can be assembled in parallel in a liquid handling robot. Efficient DNA assembly techniques, such as the Gibson assembly, will be optimized to streamline vector construction. We therefore anticipate that the reporter gene technology will be easily extended to other bacteria and certain eucaryotes, such as yeast.

Our approach to network reconstruction is based on the analysis of time-series data to first pinpoint possible network structures and then adjust model parameters in order to obtain a quantitative fit of the model to the experimental data. In the first step, any additional information limiting the possible network topologies could potentially be incorporated in order to reduce the search space of the algorithm. This

additional information can come from any of a number of sources: biological data that exclude a particular network structure, "classical" inference methods using steady-state data of the same system to limit the possible network topologies, or any other modeling approaches or measurements that fix certain parameters of the quantitative model. We therefore consider our algorithm a further improvement to the general problem of network inference, adding a new powerful tool that can be combined with existing methods to reliably deduce the underlying regulatory structure from time-series expression data.

A. Monitoring the expression of *flgA* promoter

The *flgM* gene can be expressed from two different promoters, its own specific class 3 *flgM* promoter and the class 2 *flgA* promoter from the same transcriptional unit (Figure A.1). Expression from the class 2 promoter is initiated by the master regulator of the flagellar cascade, the FlhDC heteromultimeric complex. When the Hook-Basal Body (HBB) structure is completed, the σ^{28} factor initiates transcription from class 3 promoters. It has been reported that the level of expression of the *flgM* gene from its class 2 promoter is only of about 20% of the total expression level (*Salmonella typhimurium*, Gillen and Hughes (1993), Karlinsey et al. (2000b)).

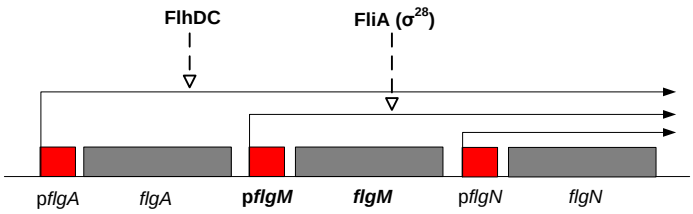

Figure A.1: The *flgAMN* operon.

We have tested whether the transcription from the *flgA* promoter has an important contribution to the expression of the *flgM* gene in *E. coli* by means of fluorescent reporter genes (the plasmid carying the *flgA* promoter has been taken from the plasmid library of Zaslaver et al. (2006)). The results show that the measured fluorescence signal representative for the activity of the *flgA* promoter is very low and are shown in Figure A.2.

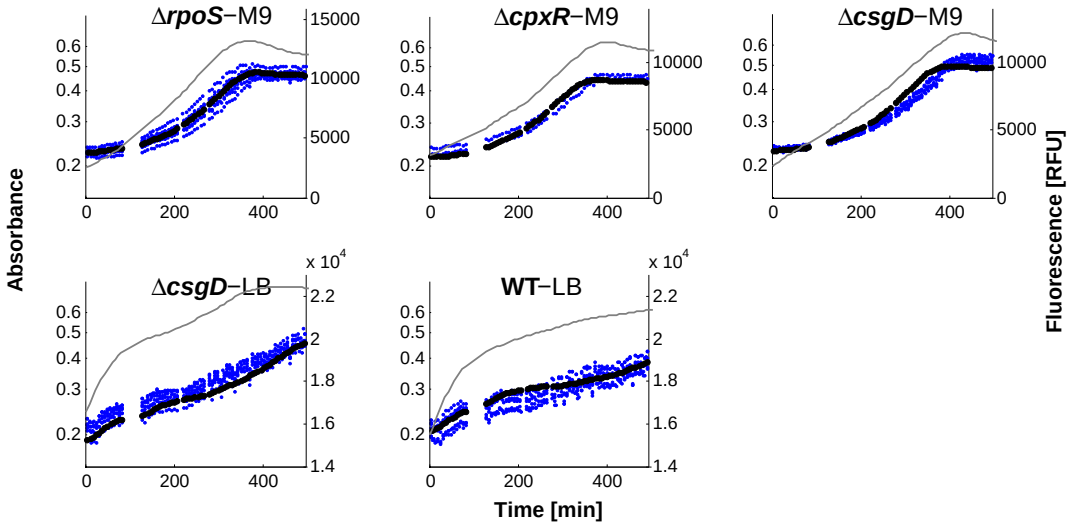

Figure A.2: Monitoring the expression of *flgA* promoter. The figure shows the fluorescence profiles (•, blue) corresponding to the activity of *flgA* with respect to the background fluorescence (•, black). The class 2 promoter expression has been observed in all the strains and growth media considered in this study (Δ*rpoS*, Δ*cpxR*, Δ*csgD*-M9, Δ*csgD*-LB and WT-LB). The fluorescence signal is not distinguishable from the background fluorescence in any of the conditions, thus gene expression from *flgA* promoter can be ignored. Absorbances (solid line, grey) show that growth conditions are similar with those in the experiments monitoring the expression from *flgM* promoter.

B. Additional information on identification of gene regulation functions from estimates of protein concentrations

In order to test if the known structure of the regulatory network of the FliA-FlgM (motility) module could be recovered from only protein concentrations of FliA and FlgM (over a range of plausible half-life values), we applied the sign pattern analysis (details in Chapter 3). The results for the structural inference are shown in Figure B.1 for *tar* and in Figure B.2 for *flgM*. A large number of combinations of half-lives are compatible with activation of FliA-dependent genes by FliA and with inhibition by FlgM and thus consistent with the known regulatory network. We then fitted the quantitative regulation functions (Eqs. 3.1-3.2) for both *tar* (Figure B.3) and *flgM* (Figure B.4). Except for the $\Delta csgD - M9$ condition, the model is not able to match the expression peaks.

We also tested if addition of global regulatory effects to protein concentrations of FliA and FlgM reconstructed for invalid half-lives (very stable half-lives, such as reporter protein half-lives) could improve the fit enough to match the expression peaks in all conditions (model of Eqs. 3.2-3.3). Although the fit improves a little bit quantitatively ($Q = 32$ vs. $Q = 36$) in the case of *tar* (Figure B.5), the model is not able to account for the expression dynamics in any of the conditions considered. Moreover, the model cannot obtain a good fit for *flgM* ($Q = 41$). The results are shown in Figure B.6.

B. ADDITIONAL INFORMATION ON IDENTIFICATION OF GENE REGULATION FUNCTIONS FROM ESTIMATES OF PROTEIN CONCENTRATIONS

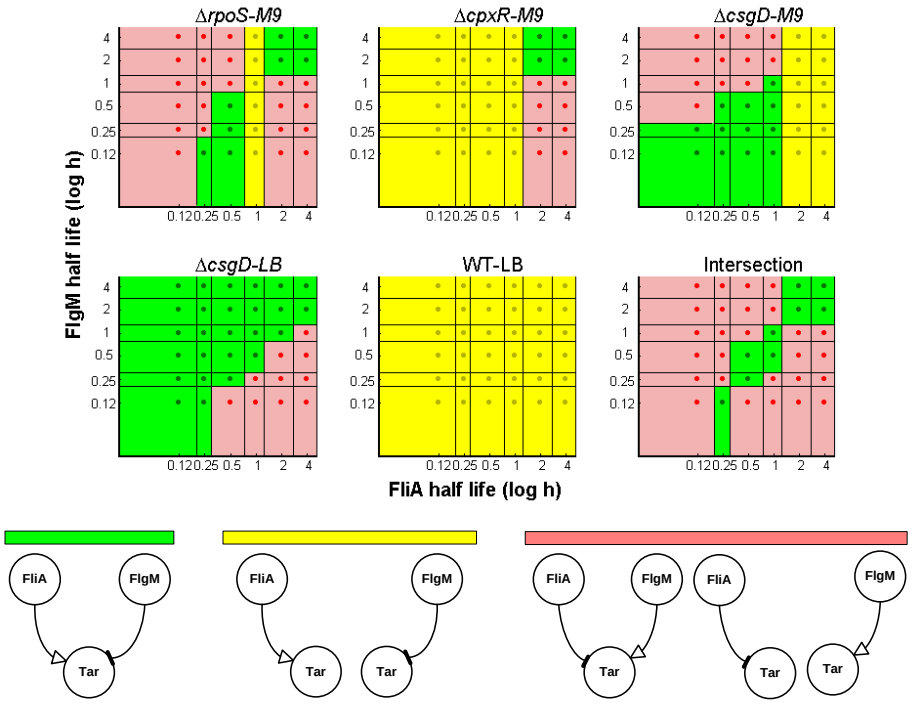

Figure B.1: Minimal patterns of regulatory interactions for *tar* over a range of physiologically realistic half-lives. The minimal regulatory patterns for the gene *tar* in the motility network of Figure 3.7 as a function of the half-lives of FliA and FlgM. The plots correspond to the five experimental conditions considered (Δ*rpoS*-M9, Δ*cpxR*-M9, Δ*csgD*-M9, Δ*csgD*-LB, and WT-LB) as well as the pooling of the data sets from all five conditions. The dot in the center of each region in the plots corresponds to a tested combination of half-lives of FliA and FlgM, and thus to specific protein concentration profiles computed from the kinetic model of gene expression (Section 3.2.4). The minimal regulatory patterns were obtained by applying the minimal sign pattern algorithm (Porreca et al., 2010a). The color codes represent the different categories of minimal signal patterns inferred. A region is colored green if the expected regulatory patterns is among the minimal sign patterns returned by the algorithm, and yellow if it is compatible with the returned sign patterns. A region is colored red if none of the returned sign patterns is consistent with the data. Two examples of inconsistent sign patterns are shown. The values of the half-lives are represented in `log`.

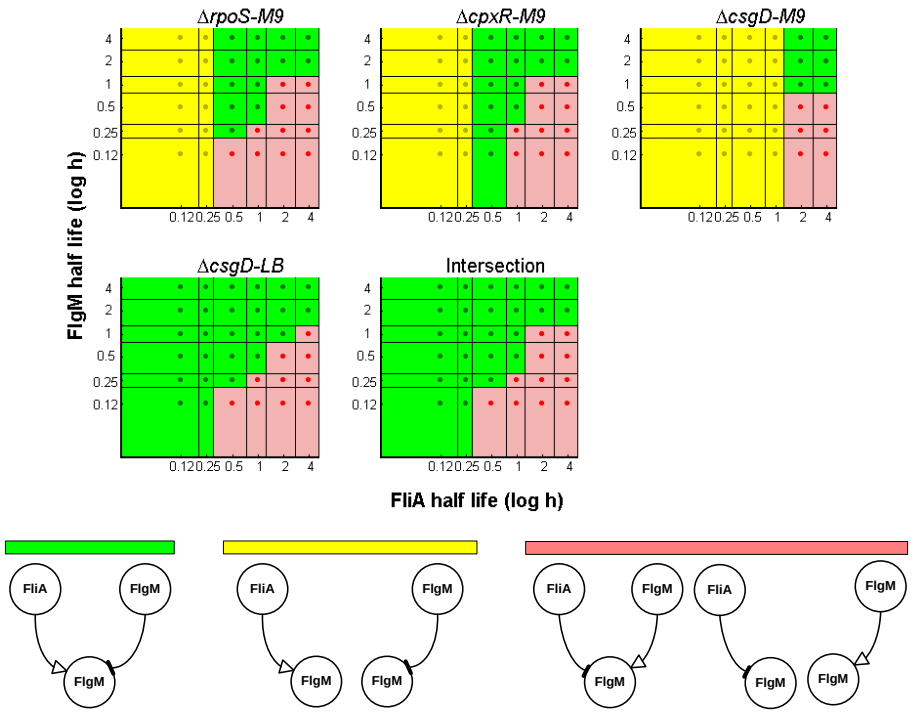

Figure B.2: Minimal patterns of regulatory interactions for *flgM* over a range of physiologically realistic half-lives. The minimal regulatory patterns for the gene *flgM* in the motility network of Figure 3.7 as a function of the half-lives of FliA and FlgM. Similarly to Figure 3.14, the plots correspond to the four of the experimental conditions considered (Δ*rpoS*-M9, Δ*cpxR*-M9, Δ*csgD*-M9, Δ*csgD*-LB) as well as the pooling of the data sets from all five conditions. The condition WT-LB was not used in the analysis of the regulation of the *flgM* promoter. The dot in the center of each region in the plots corresponds to a tested combination of half-lives of FliA and FlgM, and thus to specific protein concentration profiles computed from the kinetic model of gene expression (Section 3.2.4). The minimal regulatory patterns were obtained by applying the minimal sign pattern algorithm (Porreca et al., 2010a). The color codes represent the different categories of minimal signal patterns inferred. A region is colored green if the expected regulatory patterns is among the minimal sign patterns returned by the algorithm, and yellow if it is compatible with the returned sign patterns. A region is colored red if none of the returned sign patterns is consistent with the data. Two examples of inconsistent sign patterns are shown. The values of the half-lives are represented in `log`.

B. ADDITIONAL INFORMATION ON IDENTIFICATION OF GENE REGULATION FUNCTIONS FROM ESTIMATES OF PROTEIN CONCENTRATIONS

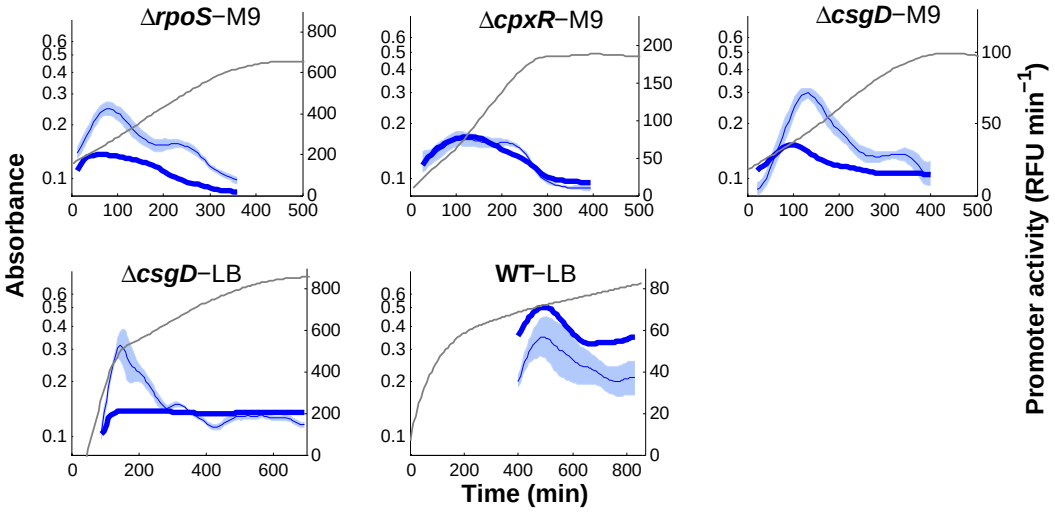

Figure B.3: Fits of regulation function of *tar* to reporter gene data when reconstructing protein concentrations from the reporter gene data and ignoring global physiological effects. The regulation function of Eqs. 3.1-3.2 was fit to the data using the promoter activity for *tar* (Figure 3.3) and concentrations of FliA and FlgM reconstructed from the activities of their promoters for physiologically realistic half-lives (Figure 3.12 and Figure 3.13). The parameters were estimated using a multistart global optimization algorithm (see Section 3.2.5 for details). The best fit is shown, for measured half-lives of FliA and FlgM of 30 min and 18 min, respectively (solid line, $Q = 36$, $(k_0, k_1, n, \theta, K) = (13.6, 206.8, 1.9, 3985, 223700)$).

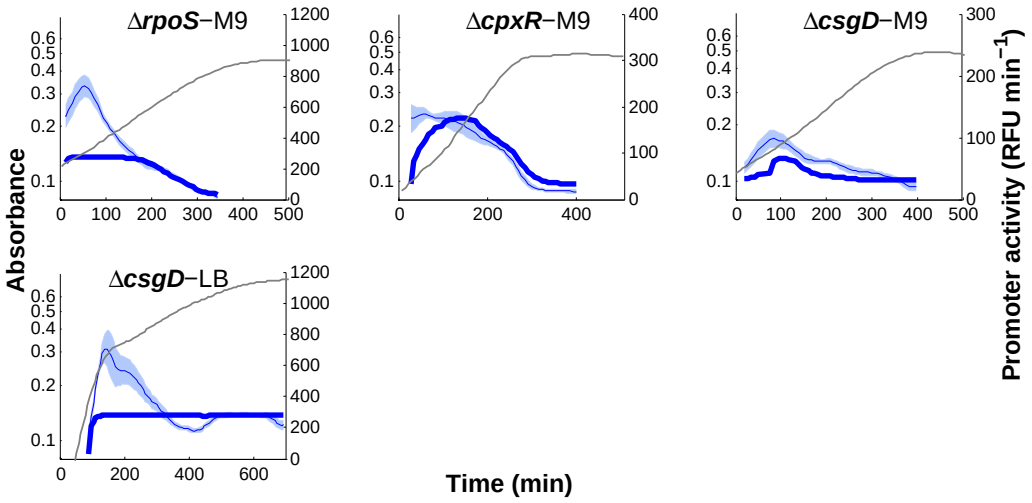

Figure B.4: Fits of regulation function of *flgM* to reporter gene data when reconstructing protein concentrations from the reporter gene data and ignoring global physiological effects. The regulation function of Eqs. 3.1-3.2 was fit to the data using the promoter activity for *flgM* (Figure 3.3) and concentrations of FliA and FlgM reconstructed from the activities of their promoters for physiologically realistic half-lives (Figure 3.12 and Figure 3.13). The parameters were estimated using a multistart global optimization algorithm (see Section 3.2.5 for details). The best fit is shown, for measured half-lives of FliA and FlgM of 30 min and 18 min, respectively (solid line, $Q = 27$, $(k_0, k_1, n, \theta, K) = (31.2, 246.8, 3, 2353, 223700)$).

B. ADDITIONAL INFORMATION ON IDENTIFICATION OF GENE REGULATION FUNCTIONS FROM ESTIMATES OF PROTEIN CONCENTRATIONS

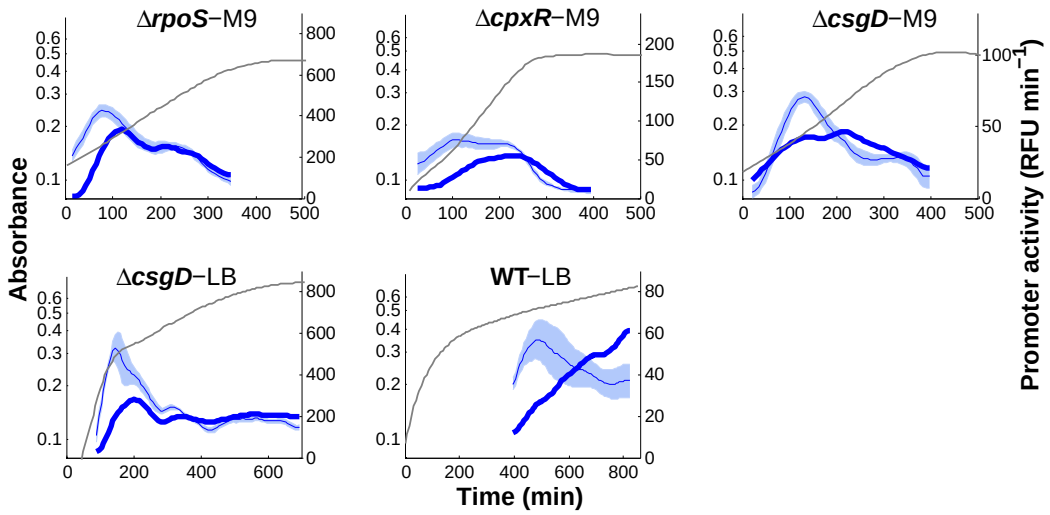

Figure B.5: Fits of regulation function of *tar* to reporter gene data when reconstructing protein concentrations from the reporter gene data for reporter half-lives and including global physiological effects. The regulation function of Eqs. 3.2-3.3 was fit to the data using the promoter activity for *tar* (Figure 3.3), concentrations of FliA and FlgM reconstructed from the activities of their promoters for reporter half-lives (Figure 3.12 and Figure 3.13), and the activity of the constitutively expressed pRM promoter quantifying global physiological effects (Figure 3.8). The parameters were estimated using a multistart global optimization algorithm (see Section 3.2.5 for details). The best fit is shown, for very stable half-lives of FliA and FlgM of 18 h (solid line, $Q = 32$, $(k_0, k_1, n, \theta, K) = (0.3, 4.5, 3, 83969, 27303)$).

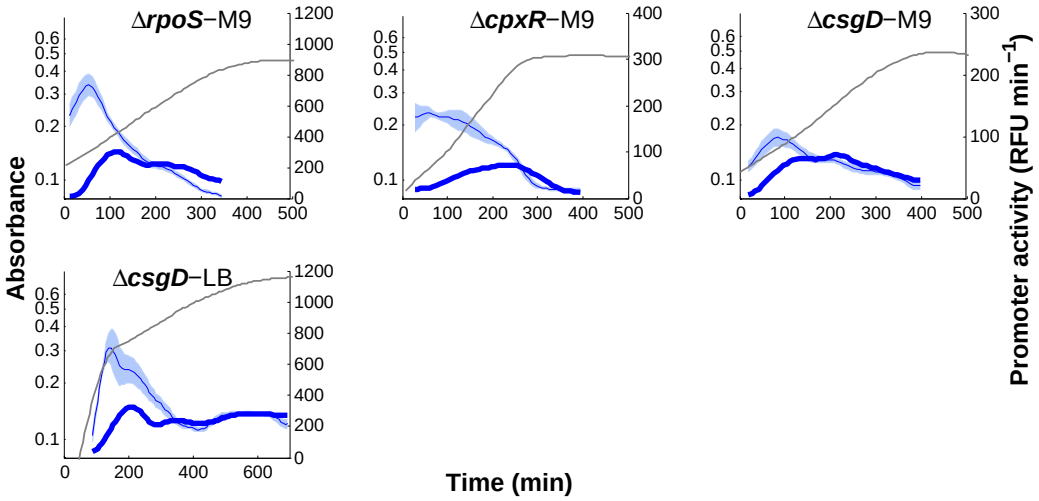

Figure B.6: Fits of regulation function of *flgM* to reporter gene data when reconstructing protein concentrations from the reporter gene data for reporter half-lives and including global physiological effects. The regulation function of Eqs. 3.2-3.3 was fit to the data using the promoter activity for *flgM* (Figure 3.3), concentrations of FliA and FlgM reconstructed from the activities of their promoters for reporter half-lives (Figure 3.12 and Figure 3.13), and the activity of the constitutively expressed pRM promoter quantifying global physiological effects (Figure 3.8). The parameters were estimated using a multistart global optimization algorithm (see Section 3.2.5 for details). The best fit is shown, for very stable half-lives of FliA and FlgM of 18 h (solid line, $Q = 41$, $(k_0, k_1, n, \theta, K) = (0.04, 11.4, 3, 24347, 27303)$).

B. ADDITIONAL INFORMATION ON IDENTIFICATION OF GENE REGULATION FUNCTIONS FROM ESTIMATES OF PROTEIN CONCENTRATIONS

C. Computation of active FliA

The active regulator in the FliA-FlgM module is free FliA, that is, FliA not bound to FlgM. The active concentration of FliA can be computed from the total concentration of FliA using Eq. 2 in the main text, given a value for the equilibrium constant K and possibly the half-lives of FliA and FlgM, estimated by fitting the model to the *tar* data. This has been done for all situations considered here: (i) replacing protein concentrations by promoter activities; (ii) replacing protein concentrations by promoter activities, while taking into account global physiological effects; (iii) computing protein concentrations for the reference half-lives of FliA and FlgM, while taking into account global physiological effects; (iv) computing protein concentrations for optimized half-lives of FliA and FlgM, while taking into account global physiological effects. The results are shown in Figures C.1-C.4. Notice that in some of the experimental conditions, FliA is only partially active when protein concentrations instead of promoter activities are used.

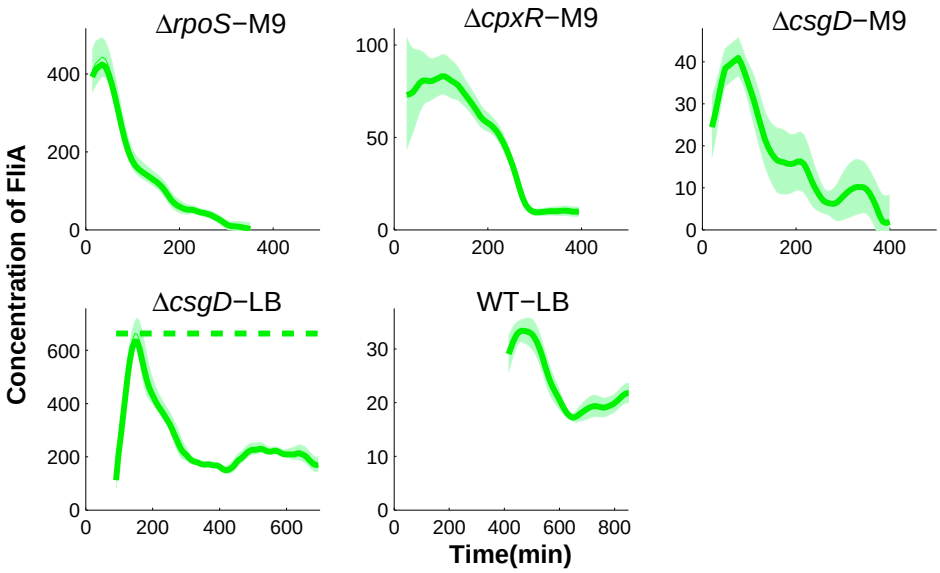

Figure C.1: Free and total concentration of FliA when using promoter activities. The concentration of free FliA (solid line, green) is computed by means of Eq. 3.2 in Chapter 3, for the optimal fit shown in Figure 3.5 in Chapter 3. The shaded regions represent the confidence intervals of total FliA and correspond to the mean of the promoter activities for 6 replicates ± twice the standard error of the mean. The threshold parameter θ is shown as a dashed green line.

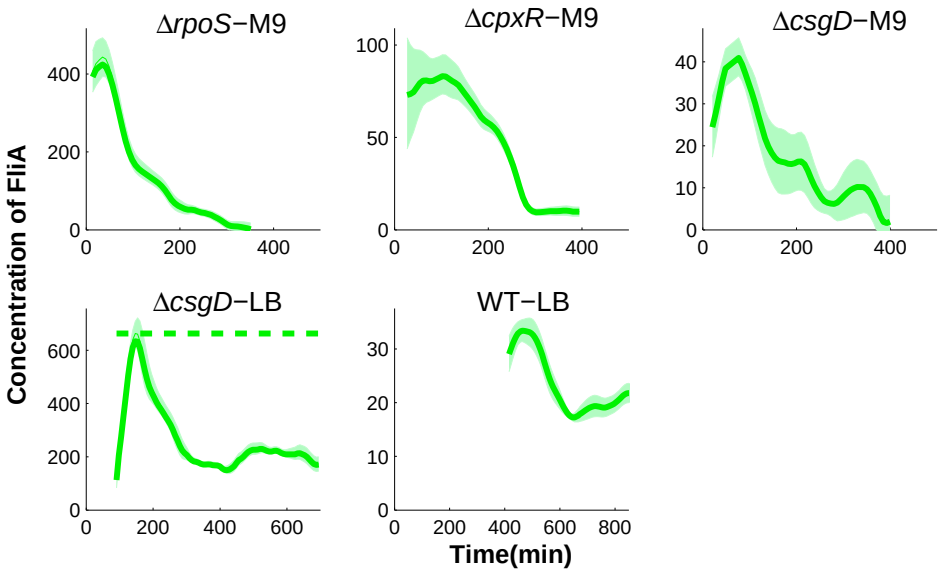

Figure C.2: Free and total concentration of FliA when using promoter activities and including global physiological effects. The concentration of free FliA (solid line, green) is computed by means of Eq. 3.2 in Chapter 3, for the optimal fit shown in Figure 3.10 in Chapter 3. The shaded regions represent the confidence intervals of total FliA and correspond to the mean of the promoter activities for 6 replicates ± twice the standard error of the mean. The threshold parameter θ is shown as a dashed green line.

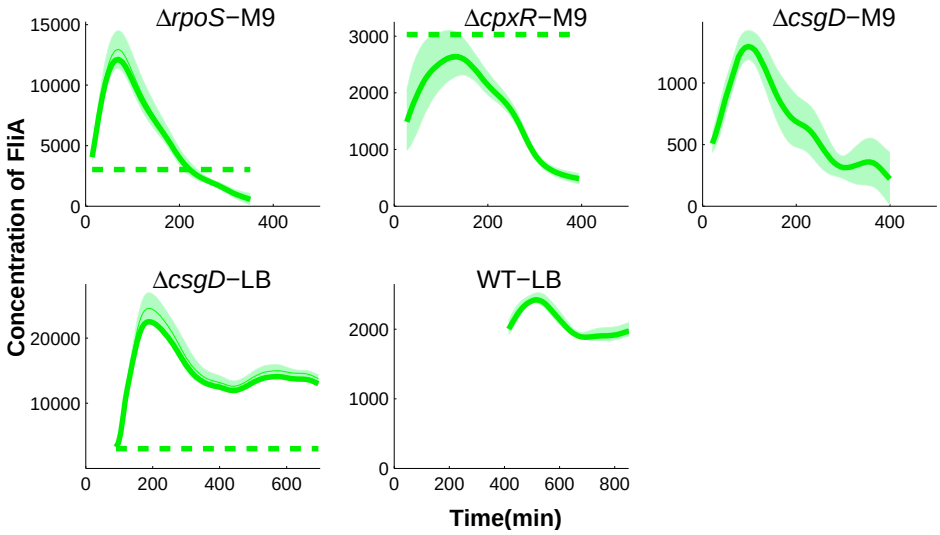

Figure C.3: Free and total concentration of FliA when using reconstructed protein concentrations for the measured reference half-lives, and including global physiological effects. The concentration of free FliA (solid line, green) is computed by means of Eq. 3.2 in Chapter 3, for the optimal fit shown in Figure 3.16 in Chapter 3. The shaded regions represent the confidence intervals of total FliA and correspond to the mean of the promoter activities for 6 replicates ± twice the standard error of the mean. The threshold parameter θ is shown as a dashed green line.

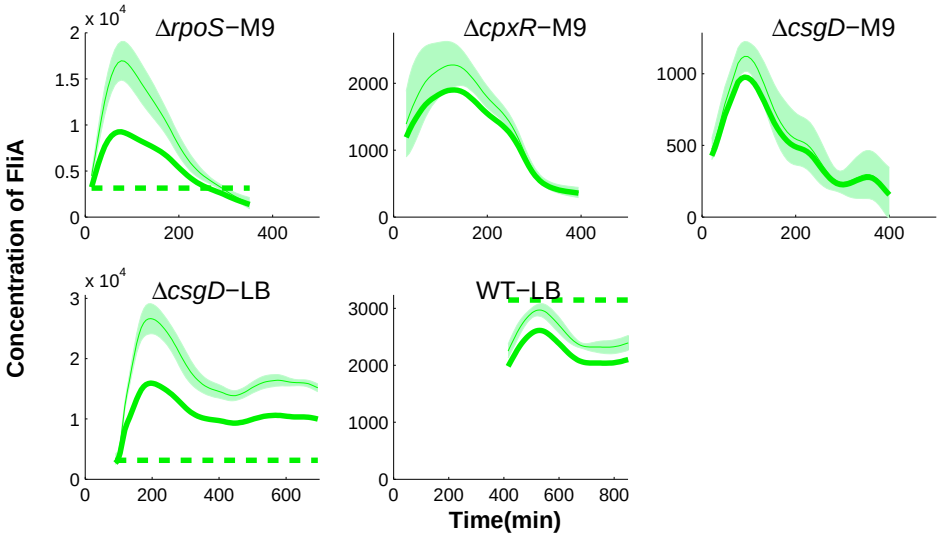

Figure C.4: Free and total concentration of FliA when using reconstructed protein concentrations for physiologically-realistic half-lives estimated from the data, and including global physiological effects for *tar* regulation function. The concentration of free FliA (solid line, green) is computed by means of Eq. 3.2 in Chapter 3, for the optimal fit shown in Figure 3.18 in Chapter 3. The shaded regions represent the confidence intervals of total FliA and correspond to the mean of the promoter activities for 6 replicates ± twice the standard error of the mean. The threshold parameter θ is shown as a dashed green line.

D. Parameter estimation

The kinetic model for the regulation of FliA-dependent genes (*tar* and *flgM*) is developed as a function of the total concentration of FliA and FlgM and ignores (Eq. D.1) or includes (Eq. D.2) global physiological effects:

$$f(t) = k_0 + k_1 \frac{p_{A,free}^n}{p_{A,free}^n + \theta^n}, \tag{D.1}$$

$$f(t) = f_{const}(t) \left[k_0 + k_1 \frac{p_{A,free}^n}{p_{A,free}^n + \theta^n} \right], \tag{D.2}$$

We analyse next the properties of the sigmoidal function (Eq. D.3) contained by the kinetic models above with respect to the parameter values estimated for θ.

$$f_{sig}(p_{A,free}, \theta, n) = \frac{p_{A,free}^n}{p_{A,free}^n + \theta^n}, \tag{D.3}$$

Figures D.1 and D.2 show explicitly how the estimated value of the θ parameter changes when using promoter activities of FliA and FlgM instead of their protein concentrations (model Eq. D.1), when including global physiological effects (model Eq. D.2), and when using protein concentrations and the global physiological effects (model Eq. D.2). As expected, when not using protein concentrations in the regulation function, neither for *tar* (Figure D.1-*A,B*) nor *flgM* (Figure D.2-*A,B*) the estimated value of θ is not in the range of FliA concentration.

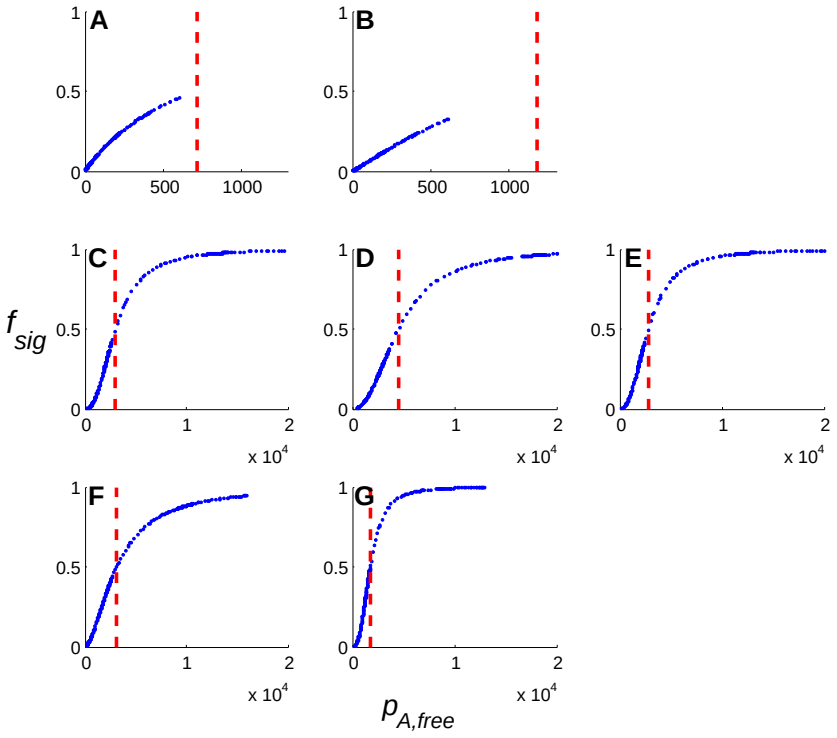

Figure D.1: Sigmoids for identified parameters for *tar* regulation. The sigmoidal functions (Eq. D.3) (•, blue) shown in Figures *A*-*G* are computed using identified parameters (k_0, k_1, n, θ, K) for *tar* regulation in the case when *A:* promoter activities for *fliA* and *flgM* replace protein concentrations of FliA and FlgM, respectively, *B:* promoter activities for *fliA* and *flgM* replace protein concentrations of FliA and FlgM, respectively, and global physiological effects are added, *C:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for physiologically realistic half-lives (30 min for FliA and 18 min for FlgM) and global physiological effects are added, *D:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for physiologically realistic half-lives (50 min for FliA and 27 min for FlgM) and global physiological effects are added, *E:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for physiologically realistic half-lives (27 min for FliA and 9 min for FlgM) and global physiological effects are added, *F:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for estimated half-lives from the data and global physiological effects are added; FliA half-lives values are $(50\ min, 24\ min, 24\ min, 35\ min, 45\ min)$ in ($\Delta rpoS$, $\Delta cpxR$, $\Delta csgD$-M9, $\Delta csgD$-LB, WT-LB) conditions, respectively and FlgM half-lives are equal to $(27\ min, 18\ min, 24\ min, 18\ min, 18\ min)$, *G:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for estimated half-lives from the data and global physiological effects are added; FliA $(40\ min, 40\ min, 24\ min, 1\ h, 45\ min)$ and FlgM $(11\ min, 27\ min, 13\ min, 24\ min, 18\ min)$ half-lives are similar to the previous case. The parameters θ of the sigmoids are shown in dashed, red lines.

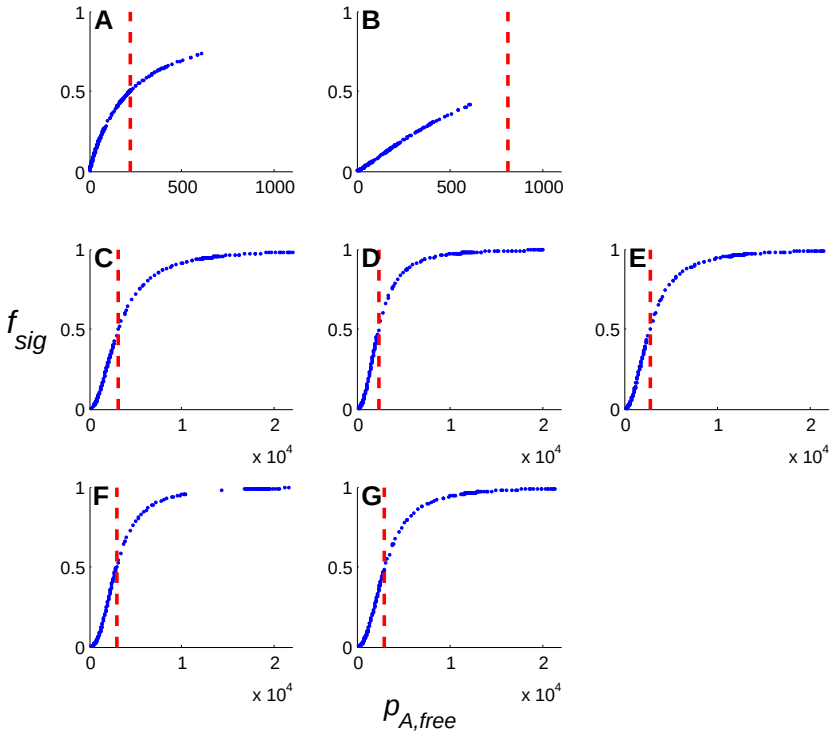

Figure D.2: Sigmoids for identified parameters for *flgM* regulation. The sigmoidal functions (Eq. D.3) (•, blue) shown in Figures *A*-*G* are computed using identified parameters (k_0, k_1, n, θ, K) for *flgM* regulation in the case when *A:* promoter activities for *fliA* and *flgM* replace protein concentrations of FliA and FlgM, respectively, *B:* promoter activities for *fliA* and *flgM* replace protein concentrations of FliA and FlgM, respectively, and global physiological effects are added, *C:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for physiologically realistic half-lives (30 min for FliA and 18 min for FlgM) and global physiological effects are added, *D:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for physiologically realistic half-lives (24 min for FliA and 27 min for FlgM) and global physiological effects are added, *E:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for physiologically realistic half-lives (27 min for FliA and 9 min for FlgM) and global physiological effects are added, *F:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for estimated half-lives from the data and global physiological effects are added; FliA half-lives values are $(24\ min, 35\ min, 27\ min, 50\ min)$ in the (Δ*rpoS*, Δ*cpxR*, Δ*csgD*-M9, Δ*csgD*-LB) conditions, respectively, and FlgM half-lives are equal to $(27\ min, 11\ min, 11\ min, 9\ min)$, *G:* protein concentrations of FliA and FlgM are reconstructed from the activities of their promoters for estimated half-lives from the data and global physiological effects are added; FliA $(24\ min, 35\ min, 27\ min, 27\ h)$ and FlgM $(18\ min, 13\ min, 20\ min, 27\ min)$ half-lives are similar to the previous case. The parameters θ of the sigmoids are shown in dashed, red lines.

E. Additional information on plasmid construction

In order to account for the global physiological effects we used the vector from the library developed at the Weizmann Institute (Zaslaver et al., 2006) and we constructed a reporter for the constitutive promoter pRM of the phage lambda. The pRM promoter region was cloned into the pUA66*gfp* plasmid backbone using the Gibson Assembly method (Gibson, 2011) and the primer sequences detailed in the table below (Table E.1).

Plasmid	Primer sequence
pUA66pRM-*gfp*	pRM-fw: GAGGC CCTTT CGTCT TCACC
	TCGAG CCTAT CACCG CCAGA
	pRM-re: TTCTT AAATC TAGAG GATCC
	GGTTT CTTTT TTGTG CTGAT
	gfp-fw: ATCAG CACAA AAAAG AAACC
	GGATC CTCTA GATTT AAGAA
	gfp-re: TCTGG CGGTG ATAGG CTCGA
	GGTGA AGACG AAAGG GCCTC

Table E.1: Primers used for the construction of pUA66pRM-*gfp* plasmid. The pUA66pRM-*gfp* plasmid was constructed with the Gibson Assembly method Gibson (2011). The pUA66*gfp* plasmid backbone was amplified using the primers gfp-fw and gfp-re. The pRM promoter region was amplified from the pZE1RM*gfp* plasmid Berthoumieux et al. (2013b) using the primers pRM-fw and pRM-re. pRM-fw and pRM-re contain the XhoI and BamHI restriction sites, respectively, allowing the insertion of the amplified DNA between these two sites on the pUA66*gfp* plasmid.

E. ADDITIONAL INFORMATION ON PLASMID CONSTRUCTION

References

J. Adler and B. Templeton. The effect of environmental conditions on the motility of escherichia coli. *J. Gen. Microbiol.*, 46(2):175–84, 1967.

T. Akutsu, S. Myiano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–34, 2000.

P.D. Aldridge, J.E. Karlinsey, C. Aldridge, C. Birchall, D. Thompson, and et al. The flagellar-specific transcription factor, $\sigma^2$8, is the type iii secretion chaperone for the flagellar-specific anti-$\sigma^2$8 factor flgm. *Genes Dev.*, 20(16):2315–2326, 2006.

D.B. Allison, X. Cui, G.P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews. Genetics*, 7(1):55–65, 2006.

U. Alon. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, 6(8):450–61, 2007.

R. Amato, A. Ciaramella, N. Deniskina, C. Del Mondo, D. di Bernardo, C. Donalek, G. Longo, G. Mangano, G. Miele, G. Raiconi, A. Staiano, and R. Tagliaferri. A multi-step approach to time series analysis and gene expression clustering. *Bioinformatics (Oxford, England)*, 22(5):589–96, 2006.

A. Ambesi and D. Bernardo. Computational Biology and Drug Discovery: From Single-Target to Network Drugs. *Curr. Bioinform.*, 1(1):3–13, 2006.

J. Aracena. Maximum number of fixed points in regulatory Boolean networks. *Bull. Math. Biol.*, 70 (5):1398–409, 2008.

T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K.A. Datsenko, M. Tomita, B.L. Wanner, and H. Mori. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, 2:2006.0008, 2006.

J.R. Banga. Optimization in computational systems biology. *BMC Syst. Biol.*, 2:47, 2008.

J.R. Banga, E. Balsa-Canto, C.G. Moles, and A.A. Alonso. Dynamic optimization of bioprocesses: efficient and robust numerical strategies. *J. Biotechnol.*, 117(4):407–19, 2005.

M. Bansal, G. Della Gatta, and D. di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–22, 2006.

M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, 3:78, 2007.

C. Barembruch and R. Hengge. Cellular levels and activity of the flagellar sigma factor flia of escherichia coli are controlled by flgm-modulated proteolysis. *Mol. Microbiol.*, 65(1):76–89, 2007.

K. Basso, A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, 37(4):382–90, 2005.

REFERENCES

M.A. Beer and S. Tavazoie. Predicting Gene Expression from Sequence. *Cell*, 117(2):185–198, 2004.

A. Bensimon, A.J.R. Heck, and R. Aebersold. Mass spectrometry-based proteomics and network biology. *Annual review of biochemistry*, 81:379–405, 2012.

H.C. Berg. *E. coli in motion.* Springer, 2004.

J.A. Bernstein, A.B. Khodursky, P.-H. Lin, S. Lin-Chao, and S.N. Cohen. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. USA*, 99(15):9697–9702, 2002.

R.M. Berry and J.P. Armitage. Microbiology. How bacteria change gear. *Science*, 320(5883):1599–600, 2008.

S. Berthoumieux, M. Brilli, D. Kahn, H. de Jong, and E. Cinquemani. On the identifiability of metabolic network models. *J. Math. Biol.*, 67(6-7):1795–832, 2013a.

S. Berthoumieux, H. de Jong, G. Baptist, C. Pinel, C. Ranquet, and et al. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.*, 9: 634, 2013b.

J.S. Bloom, Z. Khan, L. Kruglyak, M. Singh, and A.A. Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC genomics*, 10:221, 2009.

R. Bonneau, D.J. Reiss, P. Shannon, M. Facciotti, L. Hood, N.S. 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.

R. Bonneau, M.T. Facciotti, D.J. Reiss, A.K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M.H. Johnson, J.C. Bare, W. Longabaugh, M. Vuthoori, K. Whitehead, A. Madar, L. Suzuki, T. Mori, D-E. Chang, J. Diruggiero, C.H. Johnson, L. Hood, and N.S. Baliga. A predictive model for transcriptional control of physiology in a free living cell. *Cell*, 131(7):1354–65, 2007.

S. Bornholdt. Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society, Interface / the Royal Society*, 5 Suppl 1:S85–94, 2008.

S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge,UK, 2004.

H. Bremer and P.P. Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. In F.C. Neidhardt, R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter, and H.E. Umbarger, editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, pages 1553–69. ASM Press, Washington, DC, 2nd edition, 1996.

S.A. Brown, K.L. Palmer, and M. Whiteley. Revisiting the host as a growth medium. *Nat. Rev. Microbiol.*, 6(9):657–66, 2008.

N.E. Buchler and M. Louis. Molecular titration and ultrasensitivity in regulatory networks. *J. Mol. Biol.*, 384(5):1106–19, 2008.

R. Bundschuh, F. Hayot, and C. Jayaprakash. Fluctuations and slow variables in genetic networks. *Biophys. J.*, 84(3):1606–15, 2003.

S.A. Bustin and T. Nolan. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *Journal of biomolecular techniques : JBT*, 15(3):155–66, 2004.

S.A. Bustin, V. Benes, J. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G. Shipley, C.T. Wittwer, P. Schjerling, P.J. Day, M. Abreu, B. Aguado, J.F. Beaulieu, A. Beckers, S. Bogaert, J.A Browne, F. Carrasco-Ramiro, L. Ceelen, K. Ciborowski, P. Cornillie, S. Coulon, A. Cuypers, S. De Brouwer, L. De Ceuninck, J. De Craene, H. De Naeyer, W. De Spiegelaere, K. Deckers, A. Dheedene, K. Durinck, M. Ferreira-Teixeira, A. Fieuw, J.M. Gallup, S. Gonzalo-Flores, K. Goossens, F. Heindryckx, E. Herring, H. Hoenicka, and et. al. The need for transparency and good practices in the qPCR literature. *Nat. Methods*, 10(11):1063–7, 2013.

A.J. Butte and I.S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 418:29, 2000.

I. Cantone, L. Marucci, F. Iorio, M.A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M.P. Cosma. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172–81, 2009.

E. Cassuto, T. Lash, K.S. Sriprakash, and C.M. Radding. Role of exonuclease and beta protein of phage lambda in genetic recombination. v. recombination of lambda dna *in vitro*. *Proc. Natl. Acad. Sci. USA*, 68(7), 1971.

A.Y. Chen, Z. Deng, A.N. Billings, U.O.S. Seker, M.Y. Lu, R.J. Citorik, B. Zakeri, and T.K. Lu. Synthesis and patterning of tunable multiscale materials with engineered cells. *Nat. Mater.*, 2014.

F.F.V. Chevance and K.T. Hughes. Coordinating assembly of a bacterial macromolecular machine. *Nat. Rev. Microbiol.*, 6:455–65, 2008.

D.M. Chickering. *Learning from data: artificial intelligence and statistics.* Springer, New York, NY, 1996.

G.S. Chilcott and K.T. Hughes. Coupling of flagellar gene expression to flagellar assembly in *Salmonella enterica* serovar typhimurium and *Escherichia coli*. *Microbiol. Mol. Biol. Rev.*, 64(4):694–708, 2000.

J.P. Comet, M. Noual, A. Richard, J. Aracena, L. Calzone, J. Demongeot, M. Kaufman, A. Naldi, H. Snoussi, and D. Thieffry. On circuit functionality in boolean networks. *Bull. Math. Biol.*, 75(6): 906–19, 2013.

J.M. Conly, K. Stein, L. Worobetz, and S. Rutledge-Harding. The contribution of vitamin K2 (menaquinones) produced by the intestinal microflora to human nutritional requirements for vitamin K. *Am. J. Gastroenterol.*, 89(6):915–23, 1994.

J. Cox and M. Mann. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry*, 80:273–99, 2011.

REFERENCES

M.J. Daly, E.K. Gaidamakova, V.Y. Matrosova, A. Vasilenko, M. Zhai, A. Venkateswaran, M. Hess, and et al. Accumulation of Mn(II) in Deinococcus radiodurans facilitates gamma-radiation resistance. *Science.*, 306(5698):1025–8, 2004.

L.M. de Godoy, J.V. Olsen, J. Cox, M.L. Nielsen, N.C. Hubner, and et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217): 1251–1254, 2008.

H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9(1):67–103, 2002.

H. de Jong, C. Ranquet, D. Ropers, C. Pinel, and J. Geiselmann. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Syst. Biol.*, 4: 55, 2010.

R. de Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, 8:717–29, 2010.

P.P. Dennis, M. Ehrenberg, and H. Bremer. Control of rRNA synthesis in *Escherichia coli*: a systems biology approach. *Microbiol. Mol. Biol. Rev.*, 68(4):639–68, 2004.

P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–26, 2000.

Y. Dharmadi and R. Gonzalez. Dna microarrays: experimental issues, data analysis, and application to bacterial systems. *Biotechnol. Prog.*, 20(5):1309–24, 2004.

D. di Bernardo, M.J. Thompson, T.S. Gardner, S.E. Chobot, E.L. Eastwood, A.P. Wojtovich, S.J. Elliott, S.E. Schaus, and J.J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. biotechnology*, 23(3):377–83, 2005.

F.J. Doyle and J. Stelling. Systems interface biology. *J. R. Soc. Int.*, 3(10):603–16, 2006.

F.J. Doyle III and J. Stelling. Systems interface biology. *J. R. Soc. Interface*, 3(10):603–616, 2006.

O. Dudin, S. Lacour, and J. Geiselmann. Expression dynamics of rpos/crl-dependent genes in *Escherichia coli*. *Res. Microbiol.*, 164(8):838–47, 2013.

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–8, 1998.

J.J. Faith and S.T. Gardner. Reverse-engineering transcription control networks. *Phys. Life Rev.*, 2: 65–88, 2005.

J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowsk, G. Cottarel, S. Kasif, J.J. Collins, and S. T. Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):e8, 2007.

J.B. Fenn, M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

C.A. Floudas and C.E. Gounaris. A review of recent advances in global optimization. *Journal of Global Optimization*, 45(1):3–38, 2009.

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):102–5, 2003.

A.E. Gelfand and A.F. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *J. Amer. Stat. Assoc.*, 85:398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence*, 6(6):721–41, 1984.

L. Gerosa, K. Kochanowski, M. Heinemann, and U. Sauer. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Mol. Syst. Biol.*, 9:658, 2013.

D.G. Gibson. Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol.*, 498:349–61, 2011.

B. Giepmans, S. Adams, M. Ellisman, and R. Tsien. The fluorescent toolbox for assessing protein location and function. *Science.*, 312(5771):217–24, 2006.

K.L. Gillen and K.T. Hughes. Transcription from two promoters and autoregulation contribute to the control of expression of the salmonella typhimurium flagellar regulatory gene flgm. *J. Bacteriol.*, 175(21):7006–15, 1993.

H.S. Girgis, Y. Liu, W.S. Ryu, and S. Tavazoie. A comprehensive genetic characterization of bacterial motility. *PLoS Genet.*, 3(9):e154, 2007.

I. Golding, J. Paulsson, S.M. Zawilski, and E.C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

A. Greenfield, A. Madar, H. Ostrer, and R. Bonneau. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10):e13397, 2010.

J. Grefenstette, S. Kim, and S. Kauffman. An analysis of the class of gene regulatory functions implied by a biochemical model. *Bio Systems*, 84(2):81–90, 2006.

R. Guthke, U. Möller, M. Hoffmann, F. Thies, and S. Töpfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8): 1626–34, 2005.

J. Handl, J. Knowles, and D.B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–12, 2005.

M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: data integration in dynamic models-a review. *Bio Systems*, 96(1):86–103, 2009.

F. Hillenkamp, M. Karas, R.C. Beavis, and B.T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical chemistry*, 63(24):1193A–1203A, 1991.

REFERENCES

M.J. Holden and L. Wang. *Quantitative Real-Time PCR: fluorescent probe options and issues*. Springer Berlin Heidelberg, 2008.

Z. Huang, F. Senocak, A Jayaraman, and J. Hahn. Integrated modeling and experimental approach for determining transcription factor profiles from fluorescent reporter data. *BMC Syst. Biol.*, 2:64, 2008.

N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, 316(5824):593–7, 2007.

A.S. Jarrah, B. Raposa, and R. Laubenbacher. Nested Canalyzing, Unate Cascade, and Polynomial Functions. *Physica D. Nonlinear phenomena*, 233(2):167–174, 2007.

S. Kalir and U. Alon. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell*, 117(6):713–20, 2004.

S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, and et al. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, 292(5524):2080–83, 2001.

S. Kalir, S. Mangan, and U. Alon. A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*. *Mol. Syst. Biol.*, 1:2005.0006, 2005.

G. Karakousis, N. Ye, Z. Li, S.K. Chiu, G. Reddy, and C.M. Radding. The beta protein of phage lambda binds preferentially to an intermediate in dna renaturation. *J. Mol. Biol.*, 276(4), 1998.

J.E. Karlinsey, H.C Tsui, M.E. Winkler, and K.T. Hughes. Flk couples flgm translation to flagellar ring assembly in salmonella typhimurium. *J. Bacteriol.*, 180(20):5384–5397, 1998.

J.E. Karlinsey, J. Lonner, K.L. Brown, and K.T. Hughes. Translation/secretion coupling by type iii secretion systems. *Cell*, 102(4):487–497, 2000a.

J.E. Karlinsey, S. Tanaka, V. Bettenworth, S. Yamaguchi, W. Boosa, and et al. Completion of the hook-basal body complex of the salmonella typhimurium flagellum is coupled to flgm secretion and flic transcription. *Mol. Microbiol.*, 37(5):1220–1231, 2000b.

A.E. Karu, Y. Sakaki, H. Echols, and S. Linn. The gamma protein specified by bacteriophage lambda. Structure and inhibitory activity for the RecBC enzyme of *Escherichia coli*. *J. Biol. Chem.*, 250 (18):7377–87, 1975.

S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Random Boolean network models and the yeast transcriptional network. *Proc. Natl. Acad. Sci. USA*, 100:14796–9, 2003.

S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Genetic networks with canalyzing Boolean rules are always stable. *Proc. Natl. Acad. Sci. USA*, 101(49):17102–7, 2004.

S.A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22(3):437–67, 1969.

A.D. Keller. Model genetic circuits encoding autoregulatory transcription factors. *J. Theor. Biol.*, 172: 169–85, 1995.

M. Keller and K. Zengler. Tapping into microbial diversity. *Nat. Rev. Microbiol.*, 2(2):141–50, 2004.

L. Keren, O. Zackay, M. Lotan-Pompan, U. Barenholz, E. Dekel, and et al. Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol.*, 9:701, 2013.

I.M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A.M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Mu niz Rascado, Q. Ong, S. Paley, I. Schröder, A.G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R.P. Gunsalus, I. Paulsen, and P.D. Karp. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.*, 41(Database issue):D605–12, 2013.

E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application.* Wiley, Weinheim, Germany, 2005.

E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. *Gene expression models.* 2009.

S. Klumpp and T. Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc. Nat. Acad. Sci. USA*, 105(51):20245–50, 2008.

S. Klumpp, Z. Zhang, and T. Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139:1366–75, 2009.

K. Kutsukake, Y. Ohya, and T. Iino. Transcriptional analysis of the flagellar regulon of *Salmonella typhimurium*. *J. Bacteriol.*, 172(2):741–7, 1990.

K.L. Larrabee, J.O. Phillips, G.J. Williams, and A.R. Larrabee. The relative rates of protein synthesis and degradation in a growing culture of *Escherichia coli*. *J. Biol. Chem.*, 255(9):4125–30, 1980.

C. Lee, J. Kim, S.G. Shin, and S. Hwang. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. *J. Biotechnol.*, 123(3):273–80, 2006.

S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, pages 18–29, 1998.

S. Lin-Chao and H. Bremer. Effect of the bacterial growth rate on replication control of plasmid pBR322 in *Escherichia coli*. *Mol. Gen. Genet.*, 203(1):143–9, 1986.

D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–80, 1996.

D. Longo and J. Hasty. Dynamics of single-cell gene expression. *Mol. Syst. Biol.*, 2:64, 2006.

J. Lovén, D.A. Orlando, A.A. Sigova, C.Y. Lin, P.B. Rahl, and et al. Revisiting global gene expression analysis. *Cell*, 151(3):476–82, 2012.

REFERENCES

P. Lu, C. Vogel, R. Wang, X. Yao, and E.M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, 25(1): 117–24, 2007.

W. Luo, K.D. Hankenson, and P.J. Woolf. Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinformatics.*, 9(1):467, 2008.

T. MacCarthy, A. Pomiankowski, and R. Seymour. Using large-scale perturbations in gene network reconstruction. *BMC Bioinformatics.*, 6:11, 2005.

I. M. Mackay. *Real-Time PCR in microbiology: From diagnosis to characterization.* Caister Academic Press, Norfolk, UK, 2007.

R.M. Macnab. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, chapter Flagella and motility, page 123145. F.C. Neidhardt and R. Curtiss III and J.L. Ingraham and E.C.C. Lin and K.B. Low and et al., Washington, D.C., asm press edition, 1996a.

R.M. Macnab. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, chapter Chemotaxis, page 11031129. F.C. Neidhardt and R. Curtiss III and J.L. Ingraham and E.C.C. Lin and K.B. Low and et al., Washington, D.C., asm press edition, 1996b.

T. Maier, M. Guell, and L Serrano. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.*, 583(24):3966–73, 2009.

O. Maloe. *Biological Regulation and Development*, chapter Regulation of the protein synthesizing machinery, ribozomes, tRNA, factors and so on, pages 487–542. Springer US, 1979.

D. Marbach, R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA*, 107(14):6286–91, 2010.

D. Marbach, J.C. Costello, R. Kffner, N.M. Vega, R.J. Prill, D.M. Camacho, K. R. Allison, The DREAM5 Consortium, M. Kellis, J.J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nat. Meth.*, 9:796–804, 2012.

A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics.*, 7(Suppl 1):S7, 2006.

F. Markowetz and R. Spang. Inferring cellular networks–a review. *BMC Bioinformatics.*, 8 Suppl 6: S5, 2007.

E. Marshall. Getting the noise out of gene arrays. *Science*, 306(5696):630–1, 2004.

P.E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics & systems biology*, page 79879, 2007.

J.H. Miller. *Experiments in Molecular Genetics.* CSHL Press, Cold Spring Harbor, NY, 1972.

D.H. Milone, G. Stegmayer, M. Lopez, L. Kamenetzky, and F. Carrari. Improving clustering with metabolic pathway data. *BMC Bioinformatics*, 15:101, 2014.

A. Miró, C. Pozo, G. Guillén-Gosálbez, J.A. Egea, and L. Jiménez. Deterministic global optimization algorithm based on outer approximation for the parameter estimation of nonlinear dynamic biological systems. *BMC Bioinformatics.*, 13(1):90, 2012.

R.D. Mosteller, R.V. Goldstein, and K.R. Nishimoto. Metabolism of individual proteins in exponentially growing *Escherichia coli. J. Biol. Chem.*, 255(6):2524–2532, 1980.

K.C. Murphy. Lambda gam protein inhibits the helicase and chi-stimulated recombination activities of *Escherichia coli* recbcd enzyme. *J. Bacteriol.*, 173(18):5808–21, 1991.

D. Murrugarra and R. Laubenbacher. Regulatory patterns in molecular interaction networks. *J. Theor. Biol.*, 288:66–72, 2011.

Nir N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659): 799–805, 2004.

A.I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics*, 4:1419–40, 2005.

S. Nikolajewa, M. Friedel, and T. Wilhelm. Boolean networks with biologically relevant rules show ordered behavior. *Biosystems.*, 90(1):40–7.

M.K. Oh, L. Rohlin, K.C. Kao, and J.C. Liao. Global expression profiling of acetate-grown *Escherichia coli. J. Biolo. Chem.*, 277(15):13175–83, 2002.

A.B. Oppenheim, O. Kobiler, J. Stavans, D.L. Court, and S. Adhya. Switches in bacteriophage lambda development. *Annu. Rev. Genet.*, 39:409–29, 2005.

S.J. Park, S.Y. Lee, J. Cho, T.Y. Kim, J.W. Lee, J.H. Park, and M-J. Han. Global physiological understanding and metabolic engineering of microorganisms based on omics studies. *Applied microbiology and biotechnology*, 68(5):567–79, 2005.

J.M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717): 1965–9, 2005.

T.T. Perkins, R.A. Kingsley, M.C. Fookes, P.P. Gardner, K.D. James, L. Yu, S.A. Assefa, M. He, N.J. Croucher, D.J. Pickard, D.J. Maskell, J. Parkhill, J. Choudhary, N.R. Thomson, and G. Dougan. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS genetics*, 5(7):e1000569, 2009.

C. Pesavento, G. Becker, N. Sommerfeldt, A. Possling, N. Tschowri, and et al. Inverse regulatory coordination of motility and curli-mediated adhesion in *Escherichia coli. Genes Dev.*, 22(17):2434–46, 2008.

J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J.A. Schloss, V. Bonazzi, J.E. McEwen, K.A. Wetterstrand, C. Deal, C.C. Baker, V. Di Francesco, T.K. Howcroft, R.W. Karp, R.D. Lunsford,

REFERENCES

C.R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A.R. Little, H. Peavy, C. Pontzer, M. Portnoy, M.H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. The NIH Human Microbiome Project. *Genome Res.*, 19(12):2317–23, 2009.

M.W. Pfaffl. A new mathematical model for relative quantification in real-time rt-pcr. *Nucleic Acids Res.*, 29(9):e45, 2001.

P. Picotti and R. Aebersold. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods*, 9(6):555–66, 2012.

P. Picotti, O. Rinner, R. Stallmach, F. Dautel, T. Farrah, B. Domon, H. Wenschuh, and R. Aebersold. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods*, 7(1):43–6, 2010.

R. Porreca, E. Cinquemani, J. Lygeros, and G. Ferrari-Trecate. Identification of genetic network dynamics with unate structure. *Bioinformatics*, 26(9):1239–45, 2010a.

R. Porreca, E. Cinquemani, J. Lygeros, and G. Ferrari-Trecate. Structural identification of unate-like genetic network models from time-lapse protein concentration measurements. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 2529–34, 2010b.

S.L. Porter, G.H. Wadhams, and J.P. Armitage. Signal processing in complex chemotaxis pathways. *Nat. Rev. Microbiol.*, 9(3):153–65, 2011.

R.J. Prill, D. Marbach, J. Saez-Rodriguez, P.K. Sorger, L.G. Alexopoulos, X. Xue, N.D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PLoS One*, 5(2):e9202, 2010.

L. Raeymaekers. Dynamics of Boolean networks controlled by biologically meaningful functions. *J. Theor. Biol.*, 218(3):331–41, 2002.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics.*, 25(15):1923–9, 2009.

B. Regenberg, T. Grotkjaer, O. Winther, A. Fausbøll, M Akesson, C. Bro, L.K. Hansen, S. Brunak, and J. Nielsen. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in *Saccharomyces cerevisiae*. *Genome Biol.*, 7(11):R107, 2006.

D.J. Reiss, N.S. Baliga, and R. Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics.*, 7(1):280, 2006.

L. Reiter, O. Rinner, P. Picotti, R. Hüttenhain, M. Beck, M.Y. Brusniak, M.O. Hengartner, and R. Aebersold. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. methods*, 8(5):430–5, 2011.

J.J. Rice, Y. Tu, and G. Stolovitzky. Reconstructing biological networks using conditional correlation analysis. *Bioinformatics.*, 6(21):765–73, 2005.

M. Rodriguez-Fernandez, J.A. Egea, and J.R. Banga. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics*, 7:483, 2006.

M. Ronen, R. Rosenberg, B.I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA*, 99(16): 10555–60, 2002.

N. Rosenfeld, T.J. Perkins, U. Alon, M.B. Elowitz, and P.S. Swain. A fluctuation method to quantify in vivo fluorescence data. *Biophysical journal*, 91(2):759–66, 2006.

A. Roy and C.B. Post. Detection of long-range concerted motions in protein by a distance covariance. *J. Chem. Theory Comput.*, 8(9):3009–3014, 2012.

H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Mu niz Rascado, J.S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A.M. Huerta, C. Bonavides-Martínez, Y.I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. Del Moral-Chávez, A. Hernández-Alvarez, E. Morett, and J. Collado-Vides. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, 41(Database issue):D203–13, 2013.

M.S. Samoilov. *Reconstruction and functional analysis of general chemical reactions and reaction networks*. 1997.

N.A. Saunders and M.A. Lee. *Real-Time PCR: Advanced Technologies and Applications*. Caister Academic Press, Norfolk, UK, 2013.

M. Schaechter, J.L. Ingraham, and F.C. Neidhardt. *Microbe*. ASM Press, 2006.

M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70, 1995.

A. Schmidt, M. Claassen, and R. Aebersold. Directed mass spectrometry: towards hypothesis-driven proteomics. *Curr. opinion in chemical biology*, 13(5-6):510–7, 2009.

M. Scott and T. Hwa. Bacterial growth laws and their applications. *Curr. Opin. Biotechnol.*, 22(4): 559–65, 2011.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27: 379–423,623–656, 1948.

S.K. Sharan, L.C. Thomason, S.G. Kuznetsov, and D.L. Court. Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protocols*, 4(2), 2009.

I. Shmulevitch, E.R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–74, 2002.

E. Southern, K. Mir, and M. Shchepinov. Molecular interactions on microarrays. *Nature genetics*, 21 (1 Suppl):5–9, 1999.

REFERENCES

C. Southward and M. Surette. The dynamic microbe: Green fluorescent protein brings bacteria to life. *Mol. Microbiol.*, 45(5):1191–96, 2002.

P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, 9(12):3273–97, 1998.

J. Stelling. Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.*, 7(5):513–8, 2004.

R. Steuer, J. Kurths, C.O. Daub, J. Weise, and J. Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics.*, 18 Suppl 2:S231–40, 2002.

G. Storz and R. Hennge-Aronis. *Bacterial stress responses.* ASM Press, 2000.

J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003.

K. Suvarna, D. Stevenson, R. Meganathan, and M.E. Hudspeth. Menaquinone (vitamin K2) biosynthesis: localization and characterization of the menA gene from *Escherichia coli. J. Bacteriol.*, 180 (10):2782–7, 1998.

G. Szederkényi, J.R. Banga, and A.A. Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Syst. Biol.*, 5:177, 2011.

G.J. Székely and M.L. Rizzo. Brownian distance covariance. *Ann. Appl. Stat.*, 3(4):1236–1265, 2009.

Y. Taniguchi, P.J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X.S. Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991): 533–539, 2010.

J. Tegner, M.K.S. Yeung, J. Hasty, and J.J. Collins. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA*, 100(10):5944–9, 2003.

O. Tenaillon, D. Skurnik, B. Picard, and E. Denamur. The population genetics of commensal *Escherichia coli. Nat. Rev. Microbiol.*, 8(3):207–17, 2010.

R. Thomas. Boolean formalization of genetic control circuits. *J. Theor. Biol.*, 42(3):563–85, 1973.

P.J. Turnbaugh, R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, and J.I. Gordon. The human microbiome project. *Nature*, 449(7164):804–10, 2007.

A.H.M. van Vliet. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS microbiology letters*, 302(1):1–7, 2010.

H.D. VanGuilder, K.E. Vrana, and W.M. Freeman. Twenty-five years of quantitative pcr for gene expression analysis. *Biotechniques*, 44(5):619–26, 2008.

C. Vilas, E. Balsa-Canto, M.S. Garcia, J.R. Banga, and A.A. Alonso. Dynamic optimization of distributed biological systems using robust and efficient numerical techniques. *BMC Syst. Biol.*, 6:79, 2012.

A.F. Villaverde and J.R. Banga. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *J. R. Soc. Interface.*, 11(91):20130505, 2014.

A.F. Villaverde, J. Ross, and J.R. Banga. Reverse Engineering Cellular Networks with Information Theoretic Methods. *Cells*, 2(2):306–329, 2013.

X. Wang, B. Errede, and T. Elston. Mathematical analysis and quantification of fluorescent proteins as transcriptional reporters. *Biophys. J.*, 94(6):2017–26, 2008.

L.S. Waters and G. Storz. Regulatory RNAs in bacteria. *Cell*, 136(4):615–28, 2009.

A. Wepf, T. Glatter, A. Schmidt, R. Aebersold, and M. Gstaiger. Quantitative interaction proteomics using mass spectrometry. *Nat. Meth.*, 6(3):203–5, 2009.

A.K. White, M. VanInsberghe, O.I. Petriv, M. Hamidi, D. Sikorski, M.A. Marra, J. Piret, S. Aparicio, and C.L. Hansen. High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl. Acad. Sci. USA*, 108(34):13999–4004, 2011.

W.B. Whitman, D.C. Coleman, and W.J. Wiebe. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. USA*, 95(12):6578–83, 1998.

H-T. Yang, C-P. Hsu, and M-J. Hwang. An analytical rate expression for the kinetics of gene transcription mediated by dimeric transcription factors. *J. Biochem.*, 142(2):135–44, 2007.

J. Yu, V.A. Smith, P.P. Wang, A.J. Hartemink, and E.D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics.*, 20(18): 3594–603, 2004.

A. Zaslaver, A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, and et al. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods*, 3(8):623–8, 2006.

K. Zhao, M. Liu, and R.R. Burgess. Adaptation in bacterial flagellar and motility systems: from regulon members to "foraging"-like behavior in *E. coli* . *Nucleic Acids Res.*, 35(13):4441–52, 2007.

P. Zoppoli, S. Morganella, and M. Ceccarelli. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics.*, 11:154, 2010.