Study of the dynamics of catabolite repression : from mathematical models to experimental data

Valentin Zulkower

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques appliquées**

Arrêté ministérial : 7 aout 2006

Présentée par

Valentin Zulkower

Thèse dirigée par **Hidde de Jong**
et codirigée par **Johannes Geiselmann et Delphine Ropers**

préparée au sein de **l'Equipe-projet IBIS, INRIA Grenoble-Rhône-Alpes** et de **l'Ecole doctorale Mathematiques, Science et Technologies de l'Information, Informatique (MSTII)**

Etude de la dynamique des mécanismes de la répression catabolique

Des modèles mathématiques aux données expérimentales

Thèse soutenue publiquement le **3 mars 2015**,
devant le jury composé de :

Pr. Julio Rodriguez Banga
Professeur, CSIC, Vigo (Espagne), Rapporteur
Dr. Stefan Klumpp
Chercheur, Institut Max Planck, Potsdam (Allemagne), Rapporteur
Dr. Olivier Martin
Directeur de Recherche, INRA - UMR de Génétique Végétale , Président
Dr. Hidde de Jong
Directeur de Recherche, INRIA Grenoble-Rhône-Alpes, Directeur de thèse
Pr. Johannes Geiselmann
Professeur, Université Joseph Fourier, Grenoble, Co-directeur de thèse
Dr. Delphine Ropers
Chargée de Recherche, INRIA Grenoble-Rhône-Alpes, Co-encadrante de thèse

RÉSUMÉ EN FRANÇAIS

La répression catabolique désigne un mode de régulation très répandu chez les bactéries, par lequel les enzymes nécessaires à l'import et la digestion de certaines sources carbonées sont réprimées en présence d'une source carbonée avantageuse, par exemple le glucose dans le cas de la bactérie *E. coli*. Nous proposons une approche mathématique et expérimentale pour séparer et évaluer l'importance des différents mécanismes de la répression catabolique. En particulier, nous montrons que l'AMP cyclique et l'état physiologique de la cellule jouent tous deux un rôle important dans la régulation de gènes sujets à la répression catabolique. Nous présentons également des travaux méthodologiques réalisés dans le cadre de cette étude et contribuant à l'étude des réseaux de régulation génique en général. En particulier, nous étudions l'applicabilité de l'approximation quasi-stationnaire utilisée pour la réduction de modèles, et présentons des méthodes pour l'estimation robuste de taux de croissance, activité de promoteur, et concentration de proteines à partir de données bruitées provenant d'expériences avec gènes rapporteur.

RÉSUMÉ SUBSTANTIEL EN FRANÇAIS

La répression catabolique est un mode de régulation génique ubiquitaire chez les bactéries, faisant référence au fait qu'en présence d'une source carbonée préférée, par exemple le glucose chez *E. coli*, les enzymes nécessaires à l'import et la digestion de sources carbonées moins favorables sont réprimées. Plusieurs mécanismes ont été identifiés comme étant les médiateurs de la répression catabolique, en particulier l'*exclusion de l'inducteur* et la régulation transcriptionnelle par le métabolite AMP cyclique et sa protéine réceptrice, CRP. Cependant, l'importance de l'AMP cyclique dans la répression catabolique a récemment été remise en cause par plusieurs revues de la littérature, et son rôle fonctionnel pour la bactérie reste partiellement élucidé.

Bien que les mécanismes qui sous-tendent la répression catabolique soient étudiés depuis des dizaines d'années, la grande variété de souches bactériennes et milieux de cultures utilisés dans les différentes études rendent difficile de tirer des conclusions à partir des données disponibles. De plus, la plupart de

ces recherches ont porté sur des souches observées en croissance stationnaire, et ne prennent donc pas en considération les dynamiques de régulation très particulières des gènes régulés par l'AMPc entre différentes phases de croissance. L'influence de la concentration de CRP est également souvent négligée, et peu de modèles tiennent compte des effets physiologiques reflétant l'abondance et l'activité des ribosomes et de l'ARN polymérase.

Notre étude vise à quantifier, de manière systématique, l'importance de ces différents niveaux de régulation. Nous suivons, au moyen de gène rapporteurs, l'activité d'un gène exclusivement CRP-AMPc-dépendant conçu pour notre étude, ainsi que les concentrations de CRP et d'AMP cyclique. Nous mesurons aussi l'activité d'un promoteur constitutif, représentatif de l'état physiologique de la cellule. Nous formulons un modèle de régulation simple pour l'activité du promoteur synthétique, comprenant les variables CRP, AMPc, et l'état physiologique de la cellule. Nous testons ce modèle en comparant ses prédictions à des données recueillies sur plusieurs diauxies (glucose-glycerol, glucose-acétate, glucose-fructose, etc.) ainsi que certaines expériences montrant les limites de notre modèle, en particulier dans des conditions où l'addition d'AMP cyclique au milieu de culture ne lève pas la répression catabolique.

La comparaison des prédictions du modèle à des données expérimentales demande une analyse rigoureuse des mesures dynamiques d'expression génique. Nous avons donc mené une analyse quantitative des différences entre les prédictions découlant de modèles d'expression génique à deux étapes, et leurs version simplifiée à une étape, qui sera utilisée dans cette étude. Nous présentons également de nouvelles méthodes pour l'estimation de taux de croissance, d'activités de promoteurs, et de concentrations de protéines à partir de gènes rapporteurs, et montrons que ces méthodes sont moins biaisées et plus robustes que les méthodes existantes.

TITLE IN ENGLISH

A dynamical study of the mechanisms underlying carbon catabolite repression – From mathematical models to experimental data

---

ABSTRACT IN ENGLISH

Carbon Catabolite Repression (CCR) is a wide-spread mode of regulation in bacteria by which the enzymes necessary for the uptake and utilization of some carbon sources are repressed in presence of a preferred carbon source, e.g., glucose in the case of *Escherichia coli*. We propose a joint mathematical and experimental approach to separate and evaluate the importance of the different components of CCR. In particular, we show that both cyclic AMP and the global physiology of the cell play a major role in the regulation of the cAMP-dependent genes affected by CCR. We also present methodological improvements for the study of gene regulatory networks in general. In particular, we examine the applicability of the Quasi-Steady-State-Approximation to reduce mathematical gene expression models, and provide robust methods for the robust estimation of growth rate, promoter activity, and protein concentration from noisy kinetic reporter experiments.

LONG ABSTRACT IN ENGLISH

Carbon Catabolite Repression (CCR) is an ubiquitous mode of regulation in bacteria, referring to the observation that in the presence of a preferred carbon source, e.g., glucose in *E. coli*, the enzymes necessary for the uptake and utilization of less favorable carbon sources are repressed. Several molecular mechanisms have been identified as mediators of CCR, in particular inducer exclusion and transcription regulation by the signaling metabolite cyclic AMP (cAMP) and its receptor protein, CRP. Recent reviews of the literature, however, have questioned the importance of cAMP in CCR. The functional role of cAMP in the regulation of carbon metabolism in bacteria remains incompletely understood.

Even though the mechanisms underlying CCR have been studied for decades, the variety of bacterial strains and growth media used in the different studies make it difficult to draw conclusions from the available data. Furthermore, most of this research has focused on steady-state conditions, overlooking the singular dynamics of cAMP-dependent genes during phase transitions in diauxic growth. The influence of the concentration of CRP is also generally ignored, and few models account for global physiological effects reflecting the

abundance and activity of RNA polymerase and ribosome.

Our study aims at systematically quantifying the importance of these different levels of regulation. By means of reporter gene experiments we experimentally monitor the activity of an exclusively CRP-cAMP-dependent synthetic promoter designed for the study, along with the concentrations of CRP and cAMP. We also measure the activity of a constitutive promoter, representative of the global physiological state of the cell. We formulate a simple regulation model for the activity of the synthetic promoter, comprising the variables CRP, cAMP, and the global physiological state of the cell, and we test this model by comparing its predictions to dynamical data from several diauxic growth experiments (glucose-glycerol, glucose-acetate, glucose-fructose, etc.), as well as experiments showing the limits of the model, in particular conditions where adding cAMP to a glucose-containing medium does not relieve CCR.

The comparison of model predictions to experimental data requires a rigorous analysis of these dynamical gene expression measurements. We have therefore analyzed the differences of predictions derived from two-step models of gene expression (taking into account transcription and translation) and their simplified, one-step counterparts, which are used in this study. We also provide new methods for the estimation of growth rate, promoter activity and protein concentration from reporter gene experiments data, and show that our methods are more robust and less biased than the currently existing methods.

Contents

Summary of Chapter 1

In this chapter we present a State of the Art on carbon catabolite repression, in particular the different mathematical approaches and experimental techniques used for its study.

We detail the different mechanisms by which the *lac* operon of *E. coli* is repressed by glucose, and that some data from the literature can be quantitatively explained by a simple model in which this operon is regulated by both cyclic AMP and global modulations of cell machinery activity.

What are the relative importance of cAMP, global physiological effects, and other actors, in the expression of a cAMP-regulated gene ? Can simple model staging cAMP, its receptor protein CRP, and global modulations, predict the activity of a synthetic whose sole specific regulator is cAMP ?

We propose an approach based on the systematic observation of a collection of isogenic strains, during perturbation experiments designed so as to separate the effects of the different known mechanims of carbon catabolite repression. These experiments are used to calibrate a mathematical model whose parameters give insights ont the relative importance of cAMP and other actors in CCR (see Chapter 4).

Résumé du Chapitre 1

Introduction. Ce chapitre présente un état de l'art sur la répression catabolique, et notamment les modèles mathématiques et techniques expérimentales utilisés pour son étude.

Nous détaillons les différents mécanismes par lesquels la présence de glucose dans le milieu de culture réprime l'opéron *lac*, et montrons que certaines observations de la littérature peuvent être quantitativement expliquées par un modèle dans lequel la régulation de cet opéron dépend à la fois de l'AMP cyclique et d'effets globaux reflétant l'activité de la machinerie cellulaire.

Quelles sont les importances relatives de l'AMP cyclique, des effets *globaux* liés à la physiologie de la cellule, et d'autres acteurs, dans l'expression d'un gène AMPc-dépendent ? Un simple modèle prenant en compte les variations de concentration d'AMP cyclique, de sa protéine réceptrice CRP, et les modulations globales de l'expression génique dues à la physiologie cellulaire, peut-il prédire l'activité d'un gène synthétique dont le seul régulateur spécifique est l'AMP cyclique ?

Pour répondre à ces questions nous proposons une approche reposant sur l'observation systématique d'une collection de souches isogéniques, lors de cinétiques conçues pour perturber les cellules tout en séparant les effets de différents composants de la répression catabolique. Ces expériences permettent de calibrer un modèle mathématique dont les paramètres donneront une meilleure évaluation des influences respectives de l'AMP cyclique et des autres acteurs dans la répression catabolique (voire Chapitre 4).

Chapter 1

Introduction

1.1 Context

1.1.1 Carbon Catabolite Repression: hierarchy of nutrient utilization in bacteria

Bacteria are one of the most ubiquitous organisms on earth. Since their discovery in the 17th century, they have been found in very different environments, including the deep sea, volcanoes, and clouds (Brock et al., 1972; Bauer et al., 2002). Bacteria also live as hosts in higher organisms. The human gut, for instance, counts more bacteria than there are human cells in the entire body (Steinhoff, 2005). In this work, we will focus on *Escherichia coli*, a rod-shaped, 2-micrometer long bacterium that makes up 0.1% of the human gut flora. Due to its ability to grow rapidly in inexpensive and simple-to-prepare media, *E. coli* is a favorite model organism in microbiology and is commonly used for the industrial production of synthetic fuels or pharmaceutical molecules (Tsakraklides et al., 2012; Escalante and Calderón, 2010). *E. coli* is one of the most studied and best understood micro-organisms, even though the role of many of its circa 4000 genes is still not understood.

In their natural environment, bacteria compete with other species for food and space, and must adapt to the varying quality and availability of nutrients. In the example of *E. coli*, the host receives new nutrients after each meal,

and is low in nutrients between meals. As a consequence, bacteria have developed the capacity to adapt to different environments, such as the identity of the carbon source: *E. coli* can consume a range of sugars including glucose, mannose, maltose, lactose, xylose, etc. Each carbon source enters the cell through a dedicated transport system and is processed by specific enzymes of the cell. In particular, glucose enters the cell through a specific transport unit called phospho-transpherase system (PTS). Upon entrance, glucose takes one phosphate group from a phosphorylated enzyme EIIB of the cell, to enter the cell as Glucose-6-Phosphate (G6P) (Figure 1.1A). G6P is then processed through a chain of chemical reactions called glycolysis, that breaks it down to pyruvate and generates ATP and other co-factors such as NADH. Pyruvate subsequently enters the Krebs cycle to yield even more reducing equivalents and ATP.

The metabolite just upstream of pyruvate in the glycolytic pathway, phosphoenolpyruvate (PEP), donates the phosphate that will be transferred through a chain of membrane enzymes EI, Hpr, EIIA and EIIB, to form of a new phosphorylated enzyme EIIB. An important consequence of that mechanism is that the influx of glucose lowers the number of phosphorylated EIIB and EIIA enzymes in the cell, and that the concentration of phosphorylated EII rises quickly upon glucose exhaustion. We will see later in this section that these enzymes are involved in the cellular response to changes in carbon source. Lactose is metabolized by converting this disaccharide to glucose, which is phosphorylated and enters glycolysis. Lactose is imported through a specific transporter called *lactose permease*, and transformed by the enzyme β-galactosidase into one molecule of glucose and one molecule of galactose, which is in turn tranformed into glucose by a chain of enzymatic reactions (Figure 1.1B).

Carbon Catabolite Repression (CCR) denotes the fact that bacteria such as *E. coli*, in presence of several carbon sources, consume them sequentially, generally starting with the carbon source ensuring the highest growth rate. A typical example of CCR is the repression by glucose of the *lac* operon, a set of genes coding for the *lac* permease and β-galactosidase. In the presence of both glucose and lactose, bacteria will exhibit a *diauxic growth* (Monod, 1942) consisting in a fast growth phase on glucose, followed by a slower growth phase on lactose (Figure 1.2). This nutrient hierarchization enables the species to

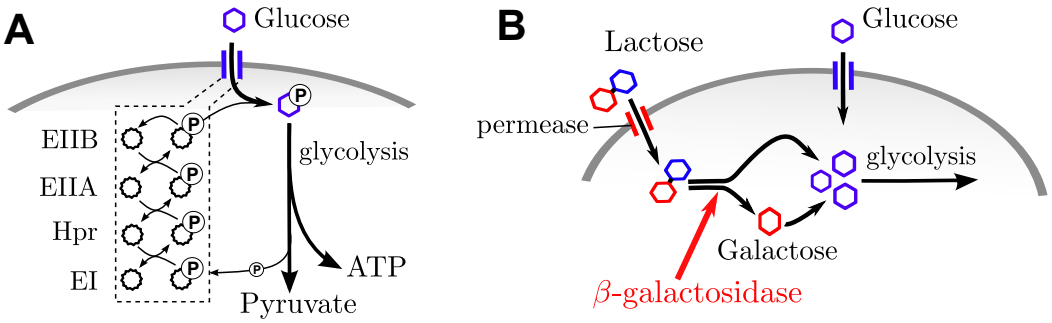


Figure 1.1: **A.** Schematic representation of the phospho-transpherase system (PTS). **B.** Transport and digestion of glucose and lactose in *E. coli*.

optimize its growth rate at all times and confers a clear advantage in the competition for food. For industrial purposes, however, it could be advantageous to abolish CCR to enable a bacteria to co-utilize several carbon sources at once and speed up the production of a product of interest (Tsakraklides et al., 2012). Achieving this goal is no simple task, as the biological mechanisms underlying CCR are still not fully understood even after decades of research. We will see in the next sections that even the much-studied regulation of the *lac* operon is still partly unexplained.

1.1.2 Regulation of the *lac* operon

The glucose-lactose diauxy observed in *E. coli* is a consequence of the low levels of β-galactosidase and lactose permease in bacteria in the presence of glucose, which prevent lactose import and digestion during this phase. In this section we review the different factors influencing the concentration of β-galactosidase, and the known mechanisms by which glucose impacts these factors.

The *lac* operon comprises the gene *lacZ*, coding for β-galactosidase, the gene *lacY* coding for Lactose Permease (also denoted LacY), and the gene *lacA*, coding for trans-acetylase (also denoted LacA). Lactose metabolism requires the first two genes, and we will focus on them from here on. The expression and regulation of these two genes is schematically represented in

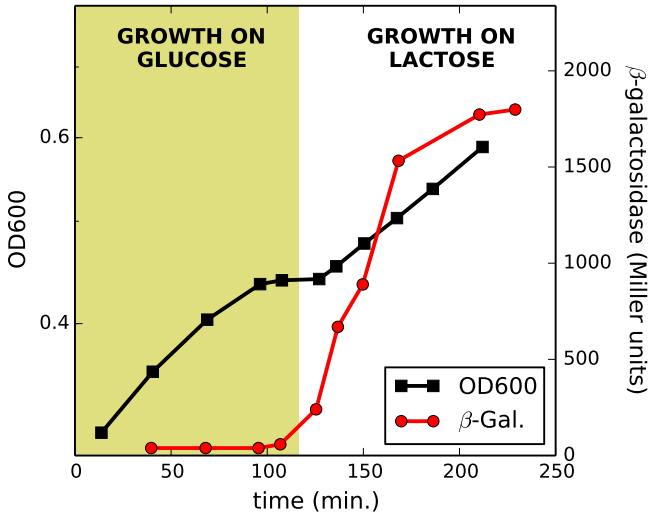


Figure 1.2: Diauxic growth of *E. coli* in the presence of glucose and lactose (Inada et al., 1996). The bacteria initially consume glucose (yellow shaded area) and grow fast. After a short lag, necessary for producing the enzymes responsible of lactose metabolism, such as β-galactosidase, the bacteria resume growth on this less-preferred carbon source.

Figure 1.3. Bacterial genes are expressed in two primary steps: transcription (synthesis of mRNA by RNA polymerase) and translation (synthesis of proteins by the ribosomes, from the information contained in the mRNA). In the case of the *lac* operon, the *lac* genes are included in an operon and therefore transcribed together to form a single mRNA molecule encoding β-galactosidase, LacY and LacA.

The proteins produced have a typical half-life of more than ten hours, which represents several bacterial generations. Each time a cell divides, the proteins pool is distributed over the two daughter cells. Therefore the intracellular concentration of proteins in bacteria is determined not only by their synthesis rate (quantity of protein synthesized per cell per minute), but also by the rate at which they are degraded, and diluted through population growth. We will assume that there is no specific degradation mechanism for the proteins we

study and that this rate remains constant over time. The variations in time of a protein concentration in a cell population, denoted $p(t)$ and expressed in moles per liter (M), is a function of the protein synthesis rate $s(t)$, which reflects the regulation of the gene and is expressed in M.min^{-1}, the degradation rate of the protein d_p (in min^{-1}), and the dilution rate (or growth rate) of the bacterial population, denoted $\mu(t)$ (in min^{-1}):

$$\frac{d}{dt}p(t) = s(t) - (\mu(t) + d_p)\, p(t). \tag{1.1}$$

At steady state, i.e. when the growth rate and all protein concentrations in the bacterial population are constant, we deduce from Equation 1.1 the relationship between the concentration p^* of a given protein, its degradation rate, its constant synthesis rate s^*, and the growth rate μ^*:

$$p^* = \frac{s^*}{\mu^* + d_p}. \tag{1.2}$$

This equation shows the interplay between the synthesis rates and the dilution rate in determining protein concentrations. Since β-galactosidase has a negligible degradation rate its concentration at steady-state is mainly determined by the rates of synthesis and dilution:

$$p^* \simeq \frac{s^*}{\mu^*}. \tag{1.3}$$

The protein synthesis rate $s(t)$ of the *lac* operon, for which we will propose a formula in Section 1.3.1, depends on the rates of transcription and translation. The *lac* operon is regulated at the level of the *lac* promoter (p*lac*), a DNA sequence upstream of the *lac* genes (see Figure 1.3). This region can be bound by proteins (called regulators) that will favor or disfavor the binding of RNA polymerase to the promoter. The structure of the *lac* promoter is discussed in more detail in Section 1.2. Transcription is repressed by protein LacI, which binds to specific DNA sites and blocks transcription. When LacI is absent or inactive, transcription is activated by the CRP-cAMP complex, formed by the metabolite cAMP and its repector protein CRP. This complex binds to a specific site of the *lac* promoter and facilitates binding of the RNA polymerase to the promoter.

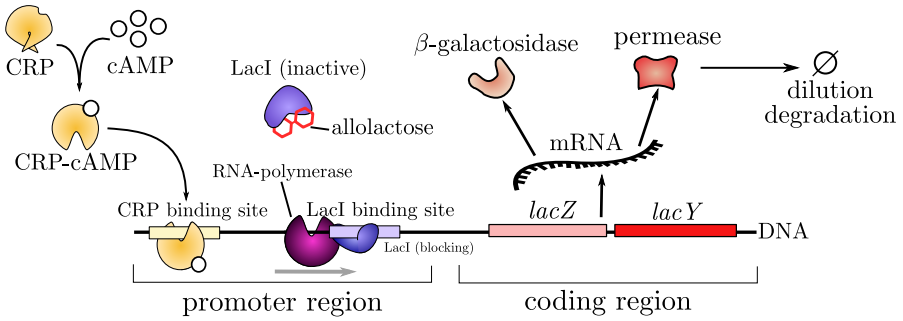


Figure 1.3: Regulation of the *lac* operon. The transcription of the operon is activated by the CRP-cAMP complex, that binds upstream of the promoter and facilitates recruitment of RNA polymerase. The binding site of LacI overlaps the promoter and the bound Lac repressor therefore sterically blocks RNA polymerase binding and transcription. When bound to allolactose, the repressor dissociates from the DNA.

The *lac* operon is regulated through different mechanisms represented schematically in Figure 1.4. The mechanism most largely accepted as the major factor of repression involves sensing of the presence of lactose by the *lac* repressor (Inada et al., 1996; Görke and Stülke, 2008a). After entry in the cell, lactose is transformed by β-galactosidase (which is always present at least at basal levels) into allolactose, which binds LacI and prevents it from repressing the *lac* operon. Therefore, lactose indirectly induces the production of β-galactosidase and *lac* permease responsible for its own utilization by *E. coli*, a phenomenon called *enzyme induction* (Figure 1.4A). It has been shown that either deleting the *lacI* gene or preventing it from binding to DNA using IPTG (an allolactose analog that passively enters the cell and neutralizes LacI) increases the activity of the *lac* operon several hundred folds (Kuo et al., 2003). In addition to the relief of repression, an activation function also contributes to enzyme induction. In presence of glucose, the intracellular concentration of cAMP is low. As explained earlier, the entry of glucose through the PTS lowers the pool of phosphorylated EIIA-P, an enzyme catalyzing cAMP synthesis by adenylate cyclase. In the absence of cAMP, CRP does not bind to its target site on the DNA and does not activate transcription of the *lac* operon.

Cyclic AMP has long been seen as a major mediator of CCR (Epstein et al., 1975; Kuo et al., 2003), not only in the case of the *lac* operon, but also for the many genes responsible for the metabolism of other carbon sources, such as mannose, maltose, or arabinose, which are not expressed in the absence of cAMP. However, we will see in the next section that several paradoxical observations have raised questions about the role of cAMP in CCR. Yet another mechanism, called inducer exclusion, prevents the expression of the *lac* operon in the presence of glucose. The influx of glucose through the PTS increases the pool of unphosphorylated enzyme EIIA, which inhibits the *lac* permease, thus preventing the uptake of lactose, and repressing the *lac* operon by lack of induction by lactose (Figure 1.4B).

The quality of the nutrients consumed by bacteria has a great influence on their gene expression machinery (ribosomes, RNA polymerases) and the energetic pool, and affects the expression of all genes at both the transcriptional and translational levels. *Global physiological effects* have regained interest in recent years, and it has been shown that the activities of many genes, regulated or not, are subject to these global modulations (Berthoumieux et al., 2013; Gerosa et al., 2013) However, their influence on cAMP-dependent genes has not yet been systematically studied. We show in Section 1.1.4 that taking into account the effects of cAMP and these global physiological effects can help explain paradoxical data from the literature, which suggests that both actors are equally important for the understanding of the regulation of catabolic genes.

1.1.3 Controversy over the role of Cyclic AMP in CCR

In my thesis, I will investigate in more details the roles of cAMP and of global physiological effects in CCR. The role of cyclic AMP as a mediator of Carbon Catabolite Repression has recently been debated over a body of seemingly contradictory data (Crasnier-mednansky et al., 2008; Görke and Stülke, 2008b; Narang, 2009a). In (Wanner et al., 1978), the authors show that adding cAMP to the growth medium does not substantially increase the activity of the *lac* operon in bacteria growing on glucose, even when LacI is inactivated by saturating concentrations of IPTG in the growth medium

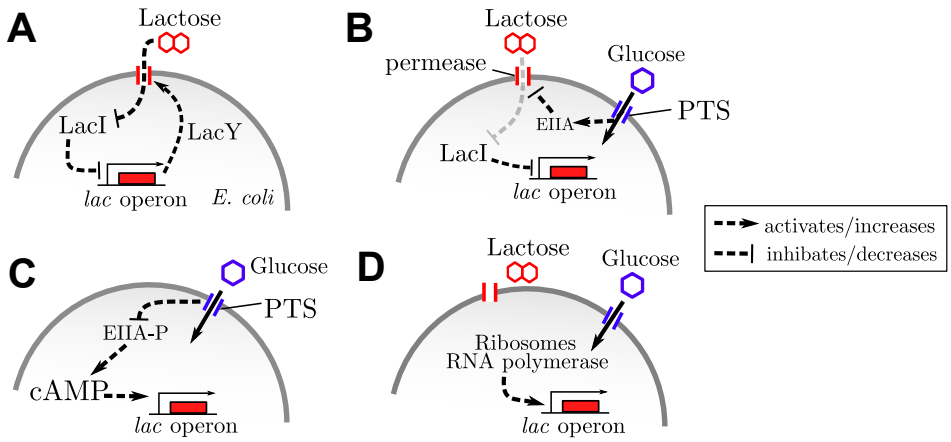


Figure 1.4: **Mechanisms of the regulation of the *lac* operon by glucose.** **A.** Enzyme induction by relief of repression. In the presence of lactose, LacI dissociates from its binding sites overlapping the *lac* promoter. **B.** Unphosphorylated EIIA inhibits the activity of the lactose transporter. **C.** Modulation of enzyme induction by the modulation of the concentration of cAMP. The presence of glucose prevents the accumulation of phosphorylated EIIA, necessary for activating adenylate cyclase. CRP is therefore inactive and does not activate the *lac* operon. **D.** The carbon source modulates the global physiology of the cell, in particular the growth rate and the concentration of ribosomes and RNA polymerase, thereby modulating gene expression.

and enzyme induction by activation (Figure 1.4C) is the major mechanism controlling the transcription of the *lac* operon. This result has been cited in (Narang, 2009b) to support the view that the lowering of cAMP levels by glucose is not a significant mechanism of CCR. We will study this phenomenon more thoroughly in Chapter 4 and argue that the cAMP added to the growth medium is most likely not imported in the cells during growth on glucose, which would explain the absence of an effect on the expression of *lac*.

It has also been claimed that glucose repression remains in a mutant strain expressing CRP*, an allele of CRP which can bind CRP-cAMP binding sites even in absence of cAMP (Dessein et al., 1978; Tagami et al., 1995). Even though the relevance of these results is questionable, as it has been shown

that strains expressing CRP* exhibit CRP*-mediated down-regulation of *crp** by glucose, they suggest that CCR could occur independently of cAMP, simply by modulation of the concentration of CRP.

These observations, together with the fact that CCR is also implemented through cAMP-independent mechanisms (like inducer exclusion in the case of the *lac* operon), have led T. Hwa and co-workers to the conclusion that *"the true physiological function of cAMP signalling in E. coli remains open nearly 50 years after its discovery"*, and that *"it is unclear to what extent cAMP signalling is intended for implementing CCR"* (You et al., 2013). Furthermore, cAMP is produced from ATP at a high energetic cost for the bacteria, which hints that it must have important functions, some of which still wait to be discovered: in (You et al., 2013), for instance, the authors postulate that cAMP is a sensor of the accumulation of metabolic precursors.

In the discussion of the data proving or disproving the role of cAMP in CCR, we need to carefully consider the different protocols used and the interpretation of the primary data. In the next section we discuss a recent study (Narang, 2009b), that claims that some data, presented as evidence of the role of cAMP, paradoxally tend to prove the contrary when interpreted rigorously. We will show, as a preliminary result, that these paradoxical data do not necessarily disprove the role of cAMP, and can be explained by taking into account the effects of global regulations.

1.1.4 Possible explanation of paradoxical data with a simple mathematical model

Most of the experimental data produced since the earliest work on Carbon Catabolite Repression have been obtained in steady-state experiments (Kuo et al., 2003; Dessein et al., 1978). In these studies cell cultures are grown in chemostats or are observed in mid-exponential phase of a batch culture. The bacteria are grown in minimal media supplemented either with different concentrations of glucose, or with different carbon sources (glucose, maltose, succinate, etc.), and IPTG is added to the growth medium in order to prevent LacI from repressing the *lac* operon. When the bacterial cultures reach a steady state growth, with a growth rate depending on the quality and quantity

of the carbon source, they are sampled to estimate the intracellular concentration of cyclic AMP, and the specific β-galactosidase activity (in Miller units per mg), which is proportional to the intracellular concentration of β-galactosidase. Typical results from such experiments are presented in Figure 1.5: panels A and B show data obtained with different sugars, while panels C and D show results for different glucose concentrations. We see that the carbon source has a strong impact on the activity of β-galactosidase (panel A) and that this activity correlates with the intracellular level of cyclic AMP (panel B). In cells grown with glucose as the sole carbon source, β-galactosidase activity decreases with growth rate (black dots in panel C), as does the intracellular cAMP level (in D). The authors see this as evidence of the role of cAMP in CCR. However, as pointed out in (Narang, 2009b), cAMP is not expected to directly influence the concentration of β-galactosidase, but rather its synthesis rate, as the CRP-cAMP complex is an activator of the transcription of the *lac* operon. This synthesis rate can be computed from the β-galactosidase concentration (or activity) and the growth rate μ with the following formula derived from Equation 1.3:

$$s = \mu \cdot [\beta\text{-gal.}] \tag{1.4}$$

Figure 1.6A shows the synthesis rate computed from the data in Figure 1.5C using Equation 1.4. While we expected synthesis rate to be correlated with the concentration of cAMP, we observe the opposite. This surprising anti-correlation of cyclic AMP concentration and β-galactosidase synthesis rates led Narang to the conclusion that *the 10-fold variation of the β-galactosidase activity with the cAMP level, attributed thus far to regulation of lac expression by cAMP, is primarily due to dilution.* He proposes that the CRP-cAMP levels in the cell are always at near-saturation level for activating the *lac* promoter – even during growth on glucose, where cAMP levels are low – and variations of the concentration of cAMP have therefore little effect on the transcription dynamics of *lac* operon.

Does cAMP have no importance in the regulation of the *lac* operon? Under this hypothesis, we would expect that the expression profile of the *lac* promoter ressembles the profile of a similar promoter that is independent of CRP. In (Kuo et al., 2003) the authors also study such a promoter (called *lacUV5*),

over a range of growth rates on glucose+IPTG. This promoter is a variant of the *lac* promoter with a higher affinity for RNA polymerase. Binding of RNA polymerase is therefore always maximal, in the presence or absence of CRP. They observe a very different expression profile (white dots in Figure 1.5C). The synthesis rate computed from these data, represented in Figure 1.6B, is approximately proportional to the growth rate. Since the experiments are done in presence of IPTG, which inactivates LacI, the expression of β-galactosidase is constitutive in the *lacUV5* strain, i.e. without any specific regulation, and its synthesis rate can be supposed to reflect the availability or activity of the gene expression machinery (ribosomes, RNA polymerases...). Such a proportionality between constitutive synthesis rate and growth rate has been more thoroughly studied in (Gerosa et al., 2013).

We can therefore extend the simple description and suppose that the synthesis rate of the cAMP-dependent p*lac* promoter of the wild-type strain depends on both the concentration of cAMP, and the activity of the expression machinery, which leads to the following regulation model:

$$s^*_{lac}(\mu) = k_l \cdot s^*_c(\mu^*) \cdot [\text{cAMP}]^*(\mu^*), \tag{1.5}$$

where $s^*_{lac}(\mu^*)$, denoting the growth-rate-dependent synthesis rate of β-galactosidase in the wild-type strain at steady state (Figure 1.6A), is the product of the intracellular cAMP level [cAMP](μ) (also Figure 1.6A) and $s_c(\mu)$, the constitutive synthesis rate measured for the p*lacUV5* promoter, which we use here as a proxy of the activity of the gene expression machinery. The proportionality constant k_l is estimated from the available data, and the resulting fit (Figure 1.6C) shows that the simple model of Equation 1.5 can predict the β-galactosidase synthesis rate in this experiment with very good accuracy ($r^2 = 0.94$). This analysis suggests that global regulations should be taken into account when studying the activity of cAMP-dependent genes, and at the same time confirms the important role of cAMP in CCR, as a factor counterbalancing the modulation of global physiological effects. While the expression of unregulated genes is 12 times lower in media with limiting carbon sources due to a shortage of energy and machinery (Figure 1.6C), the expression of a cAMP-dependent promoter is only 3-4 times lower, due to the activating effect of cAMP that partially compensates the global decrease of gene expression at

low growth rates. Physiologically, the increase of the cAMP concentration is responsible for an increase of β-galactosidase concentration at lower growth rates (black dots in Figure 1.5), whereas the concentration of a protein encoded by a cAMP-independent gene remains relatively constant (white dots). This explanation still holds that glucose has a negative influence, through cAMP, on the pool of β-galactosidase.

More generally, this analysis shows how a simple mathematical model could be used to propose a possible explanation of the available data and infer a putative relationship between different biological variables, an approach that has been increasingly used in the study of CCR, as we will see in the next section.

1.1.5 Mathematical approaches to the study of Carbon Catabolite Repression

Different mathematical formalisms have been used to model the interactions between genes, metabolites and environmental factors as part of a broader discipline called Systems Biology (de Jong, 2002; Alon, 2007; Klipp et al., 2009). In this section we will focus on the mathematical modeling of Carbon Catabolite Repression. A specificity of this subject is that changes in the nutrient source are accompanied by an important modification of bacterial physiology, from the composition of the membrane to the conformation of the DNA and the concentration of hundreds of metabolites in the cytoplasm. This multitude of effects makes it difficult to understand which ones of these changes constitute a signal and which ones are a response to this signal, which interactions are direct and significant and which ones are marginal or indirect.

The different mathematical models proposed to study Carbon Catabolite Repression differ in their scale (typically, the number of genes staged in the model), and granularity (level of detail). A first class of models aims at quantitatively characterizing a small system, typically made up of a single gene, by modelling its response as a function of different biological variables and parameters. In (Kuhlman et al., 2007) the authors hypothesize, based on biological evidence and previous models, that in a strain lacking cAMP and growing on glucose, the β-galactosidase activity in response to external cAMP

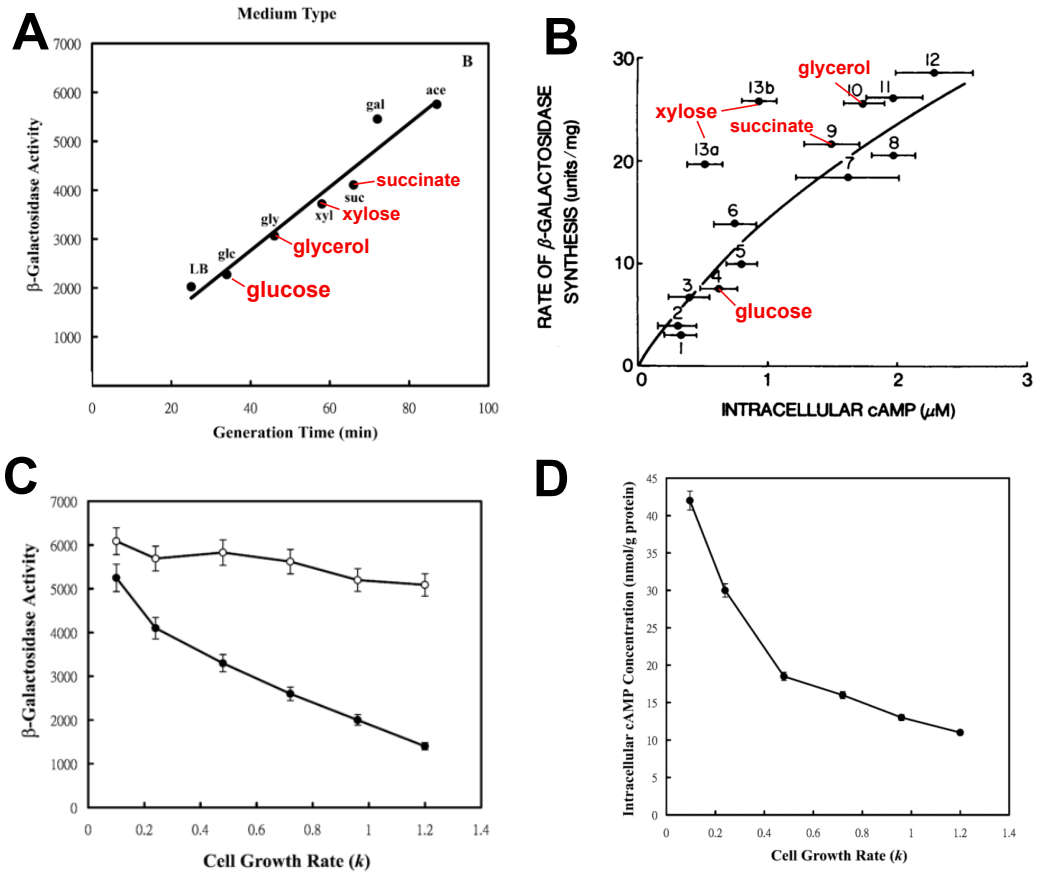


Figure 1.5: Literature data on the relationship between growth rate, β-galactosidase activity and cyclic AMP concentration. **A.** Activity of β-galactosidase and concentration of cAMP for cell cultures growing on various carbon sources (Epstein et al., 1975) **B.** β-galactosidase activity and generation time of cell cultures growing on various substrates. **C.** β-galactosidase activity in a wild-type *E. coli* (black dots) and a *lacUV5* mutant (white dots) grown in continuous cultures on minimal medium with various concentrations of glucose. **D.** Internal cAMP concentration measured in the populations of panel C (Kuo et al., 2003).

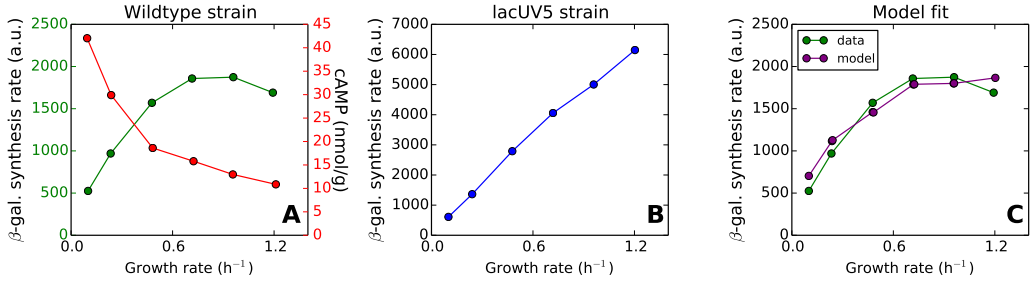


Figure 1.6: Synthesis rate of β-galactosidase as a function of the growth rate in cells growing on glucose + IPTG. **A.** Synthesis of β-galactosidase in wild-type *E. coli* (computed from data in Figure 1.5C) and intracellular concentration of cAMP (as in Figure 1.5D) **B.** Synthesis of β-galactosidase in the *lacUV5* strain (computed from data in Figure 1.5C).**C.** Comparison of the model of Equation 1.5 and the observed β-galactosidase synthesis rate of panel A.

concentration can be modelled by

$$\alpha_{cAMP} = \frac{1 + f_{cAMP}([cAMP_{ext}]/C_{cAMP})^{m_{cAMP}}}{1 + ([cAMP_{ext}]/C_{cAMP})^{m_{cAMP}}}. \tag{1.6}$$

In Equation 1.6, f_{cAMP} represents the fold change between the lowest and highest values of β-galactosidase activity, C_{cAMP} represents a threshold of external cAMP from which β-galactosidase is expressed at half its maximal value, m_{cAMP} expresses a possible non-linearity in the reponse to external cAMP.

The calibration of this model from observations of α_{cAMP} at different cAMP concentrations yields $m_{cAMP} \simeq 1$, which indicates weak or no cooperativity. The authors also find that in a strain lacking the cAMP-phosphodiesterase (PDE) the activation threshold C_{cAMP} was lower, which shows that PDE somehow isolates the cell from the influence of external cAMP.

In another study, (Kaplan et al., 2008) measure a set of catabolic genes whose response depends on cAMP and another inducer (for instance arabinose or maltose). They hypothesize that the response of these genes to external cAMP and inducer could be modelled as the product of two independent

responses:

$$\alpha_{cAMP,Ind.} = \left(\frac{[cAMP]^{m_{cAMP}}}{C_{cAMP} + [cAMP]^{m_{cAMP}}}\right)\left(\frac{[Ind]^{m_{Ind}}}{C_{Ind} + [Ind]^{m_{Ind}}}\right), \qquad (1.7)$$

and they calibrate this model for each gene over a range of cAMP and inducer concentrations. Surprisingly, they observe that this model cannot fit well the reponse of some genes to external cAMP, which is non-monotonous (i.e., β-galactosidase activity is observed to decrease at high concentrations of external cAMP). Their study also shows that, while different genes involved in the metabolism of arabinose had a similar response, genes of the galactose and maltose regulons have heterogeneous response profiles. This study shows how difficult it can be to predict even the shape of a gene expression response. However, it should be kept in mind that the promoters studied here are embedded in a complex cellular regulatory network and the regulation of transcription is not the simple response to two signals, cAMP and the respective inducer.

Such small-scale models can be used as building blocks of larger ones to study the interactions in a system of several genes and/or metabolites, using differential-equations-based dynamical models (Bettenbrock, 2005). A difficulty in handling large models is that they often require large amounts of experimental data to be properly calibrated and validated. In practice, this generally means pooling together data and parameters from different studies, possibly acquired in different conditions, and using different strains. These models have been used to analyse the dynamics of gene regulation and they can help to understand which biological variables and parameters are essential for explaining a given phenomenon, or singular bacterial behaviours. For instance, it has been noticed that within a population growing on a mix of glucose and lactose, a minor fraction of the cells use lactose as a preferred source. In (Ozbudak et al., 2004), the authors remark that the expression of *lacY* is partly auto-amplified: *lacY* codes for the *lac* permease, which favorizes the entry of lactose in the cell, and therefore indirectly the expression of *lacY* (see Figure 1.4A). In these conditions, a bistable behaviour is to be expected: cultures grown on lactose for an extended period will continue to prefer lactose even in presence of glucose, at least for some time. Interestingly, their model manages to accurately predict the population response in different

growth media with variable concentrations of TMG (a non-metabolized lactose analog).

While the dynamics of inducer exclusion seem to have been well captured by the current models (see (Santillán and Mackey, 2008) for a review), mathematical approaches have until now failed to properly explain the role of cyclic AMP, as several models staging different actors have been proposed and could not be discriminated. For instance, in (Narang and Pilyugin, 2007) the authors propose a model to explain nutrient preferences in which dilution through growth is the main factor driving CCR: to each nutrient source corresponds a specific growth rate which ensures that enzymes for the digestion of other nutrients remain below their activation threshold. In another study (Zhuang et al., 2011), the authors remark that the ATP yield and growth rate of bacteria could be accurately predicted over a range of different glucose uptake rates using Flux Balance Analysis. Their model introduces an *occupancy constraint* imposing that the number of transporters in a bacterium is limited by the surface area of its membrane. This limitation of the membrane obliges bacteria to optimize the nature and relative proportion of its different transporters (the authors speak of *the economics of cell membrane*) and leads to the conclusion that *the occupancy constraint* [not gene regulation] *may be a fundamental governing constraint of cellular metabolism and physiology*.

To date, no model has been proposed to specifically clarify the role of cAMP and the gene machinery in the regulation of the *lac* operon and, despite the many studies on CCR, the body of data gathered is insufficient to build, validate and discriminate the different models. Some key biological variables playing a role in CCR have been little observed, partly due to technological limitations. Few studies focus on the variations of the concentration of CRP, even though it has been shown that CRP can be a limiting factor in the regulation of the *lac* operon (Ishizuka et al., 1993). And, although constitutive promoters have been studied for decades, they have not been commonly used as proxies of the activity of the gene machinery until very recently (Klumpp et al., 2009; Berthoumieux et al., 2013; Gerosa et al., 2013). Finally, most of the available data come from steady-state experiments, which are much less informative than perturbation experiments, such as up- or downshift of nutrients, that trigger a global re-organisation of cellular metabolism (Kao

et al., 2005). In particular, cAMP has interesting dynamics during the growth transition from glucose to a poorer carbon source, characterized by a rapid and transient increase of the concentration of cAMP, yielding valuable information about the regulation of the *lac* operon.

In conclusion, some of the most basic questions regarding carbon catabolite repression are still unanswered despite the large amount of available data. Elucidating the roles of cAMP and global regulations will require new, carefully planned experiments, where the protocols and observed variables are chosen so as to best calibrate and validate a tailor-made regulation model in order to assess the relative importance of the different regulatory factors controlling CCR.

1.2 Problem statement

The recent observations that the availability of energy and the abundance and activity of gene machinery may play a significant role in the expression of all bacterial genes could shed a new light on Carbon Catabolite Repression, and in particular on our understanding of the role of cyclic AMP. The question becomes: how can we dissect and evaluate the influence of CRP-cAMP and global physiology on the activity of a cAMP-dependent promoter of *E. coli*?

As our biological model, we will observe the activity of a synthetic promoter, p*lac**, obtained by deleting the LacI binding sites of the *lac* operon, leaving CRP-cAMP as the only known specific regulator (Oehler and Amouyal, 1994; Müller et al., 1996) (Figure 1.7). To which extent and in which conditions can CRP-cAMP and physiological effects explain the observed dynamical response of this promoter to different stimuli? And what is the contribution of each factor to the physiological response, i.e., the final concentration of β-galactosidase in the cell?

We have seen in Section 1.1.3 that steady-state activities of the *lac* operon can be explained by a simple model staging both global regulations and cAMP. But does a similar model also explain the dynamics of p*lac** activity during growth transitions? In this thesis we propose and calibrate a dynamical model describing quantitatively the regulations of a cAMP-dependent promoter over

time. We assess to what extent this model is sufficient for explaining the observations over a range of experiments, as well as the seemingly contradictory results of the literature, or whether a third important actor of CCR is still missing.

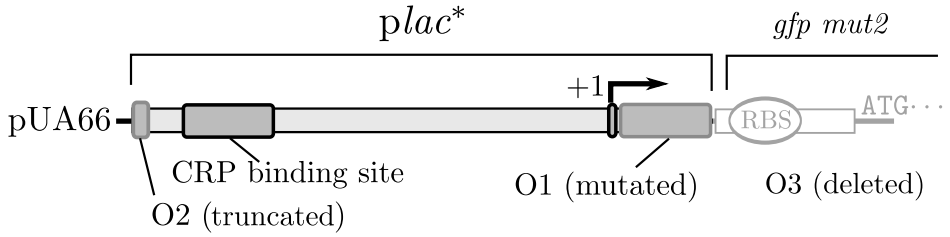


Figure 1.7: Synthetic cAMP-dependent promoter constructed for the study. O1, O2, O3 denote LacI binding sites, as in (Oehler and Amouyal, 1994). These sites are either deleted, partially deleted, or mutated, so as to disable any influence of LacI on the promoter.The plasmids are derivatives of pUA66 and the activity of the promoter is assessed by measuring the fluorescence of the *gfp-mut2* gene replacing *lacZ*.

1.3 Approach

One difficulty in separating the effects of cAMP from global physiological effects is that both vary during growth transitions, along with many other biological variables. Measuring the relevant biological variables and separating their effects on the p*lac** promoter requires therefore a systematic study with a dedicated model and carefully chosen experiments. We will formalize our biological question with a simple mathematical model, which will be be calibrated by means of perturbation experiments on a collection of isogenic strains carrying expression and reporter plasmids.

1.3.1 Mathematical formulation of the problem

Assuming that CRP-cAMP and global physiological effects act independently on the activity of the p*lac** promoter, denoted $a_{plac^*}(t)$ (in M.min^{-1}), this

activity can be represented as the product of independent factors:

$$a_{plac^*}(t) = f_a \, a_c(t) \, a_g(t), \tag{1.8}$$

where f_a denotes the maximal activity of the plac^* promoter, and $a_c(t), a_g(t)$, both varying between 0 and 1, denote the modulations of the promoter activity due to CRP-cAMP and global physiological effects, respectively.

Assuming that the concentration of cAMP is always much larger than the concentration of CRP-cAMP complex, the concentration of CRP-cAMP depends on the concentrations of intracellular cAMP and CRP as follows:

$$[\text{CRP-cAMP}](t) = [\text{CRP}](t) \frac{[\text{cAMP}](t)^{m_c}}{K_c^{m_c} + [\text{cAMP}](t)^{m_c}}, \tag{1.9}$$

where K_c is the dissociation constant of cAMP and CRP. The influence of an activator on the activity of a gene is generally represented by a Hill function. The influence $a_c(t)$ of the CRP-cAMP concentration on the activity of the plac^* promoter, is therefore modelled as follows:

$$a_c(t) = \frac{[\text{CRP-cAMP}](t)^{m_a}}{C_c^{m_a} + [\text{CRP-cAMP}](t)^{m_a}}. \tag{1.10}$$

In this equation C_c represents the activation threshold of the plac^* promoter (value of [CRP-cAMP] at which $a_c(t)$ is at half its theoretical maximum). This corresponds also to the dissociation constant of CRP-cAMP for its binding site upstream of the plac^* promoter. Coefficient m_a represents the cooperativity between CRP-cAMP complexes in binding the promoter; the results from (Kuhlman et al., 2007) mentioned in Section 1.1.5 $m_a \simeq 1$ and $m_c \simeq 1$.

The influence in time of global physiological effects, denoted $a_g(t)$, is estimated by observing the time-dependent activity of the constitutive promoter pRM, $a_{pRM}(t)$ (in M.min^{-1}; (Berthoumieux et al., 2013)). As different genes may have different sensitivities to global regulations (Gerosa et al., 2013), we will express $a_g(t)$ as follows:

$$a_g(t) = \frac{a_{pRM}(t)}{K_{pRM} + a_{pRM}(t)}. \tag{1.11}$$

Everything combined, the model from Equation 1.8 becomes:

$$a_{plac^*}(t) = f_a \left(\frac{[\text{CRP-cAMP}](t)}{C_c + [\text{CRP-cAMP}](t)} \right) \left(\frac{a_{pRM}(t)}{K_{pRM} + a_{pRM}(t)} \right). \tag{1.12}$$

While very simple, this model will enable us, once tested and calibrated with experimental data, to assess the importance of each variable in the regulation of plac^*. If the estimated dissociation constant K_c is below the concentrations of cAMP observed during a given experiment, we can conclude that [CRP-cAMP]$\simeq$[CRP], meaning that the variations of [CRP-cAMP] are in fact driven by variations of CRP, not cAMP. In the same way, if the estimate of C_c is low, $a_c(t) \simeq 1$, i.e. variations of [CRP-cAMP] have no effect on the activity of plac^*, we can conclude that CRP-cAMP is always at saturating concentrations with respect to the promoter (a hypothesis formulated in (Narang, 2009b)). Finally, a low value of K_{pRM} would indicate that global physiological effects play no role in the regulation of plac^* ($a_g(t) \simeq 1$), while a high value would mean that the activity of plac^* is sensitive to global physiological effects ($a_g(t) \simeq a_{pRM}(t)/K_{pRM}$).

Because the concentrations of CRP and cAMP, as well as global regulations, vary abruptly during growth transitions, these coefficients cannot be identified from a single kinetic experiment. Their estimation requires a set of separating experiments, using isogenic strains and specific plasmids, that will allow us to separate the different variables and create physiological conditions in which this model can be further simplified and more easily calibrated.

1.3.2 Systematic study of isogenic strains

Despite the abundant literature on Carbon Catabolite Repression, the study of (Kuo et al., 2003) presented in section Section 1.1.4 is the only work with measurements, in the same conditions, of the concentration of intracellular cAMP, cAMP-dependent β-galactosidase activity, and constitutive β-galactosidase activity. More data would be required to calibrate our model of Equation 1.12, but it was not possible to gather a coherent dataset from the literature because the studies often differ by the protocols, growth media, and strains used.

We have therefore conducted a systematic study of isogenic strains (represented in Figure 1.8), obtained by gene deletion or plasmid insertion from the same original BW25113 *E. coli* strain. This set of strains allows us to monitor the different variables of our model, and create scenarii in which the activity of plac^* will only depend on one or two of its regulators. In wild-type

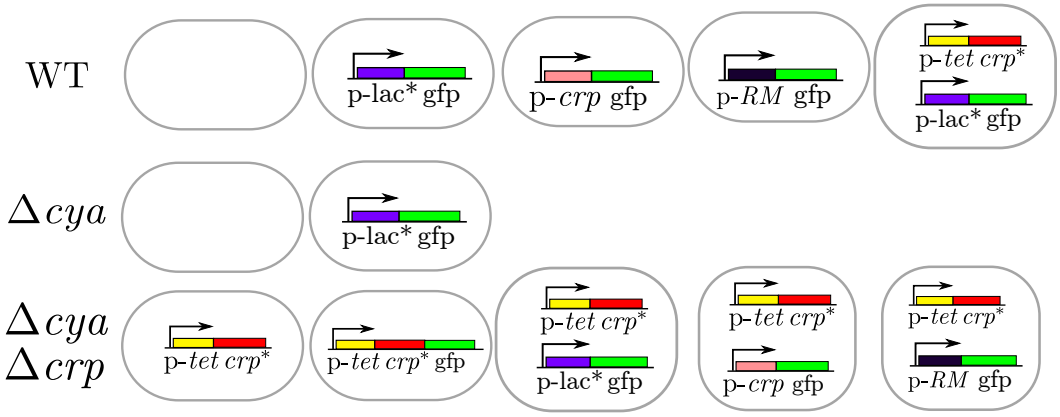


Figure 1.8: Isogenic strains used in this thesis. Each line represents a collection of strains obtained by insertion of reporter or expression plasmids into the same mother strain, whose genotype is indicated in the first column. The first strain represented on each line does not carry any reporter plasmid, and will be used as a control of the bacterial autofluorescence. The p*tet* promoter of the expression plasmids can be induced by adding ATc to the growth medium. The plasmids used in this study are presented in more detail in section Section C2 of the appendix.

strains, all components of the model are present: CRP, cAMP, and global regulation. In Δ*cya* strains, where the gene coding for adenylate cyclase is deleted, cAMP is absent, unless added to the growth medium. This enables us to observe the activity of p*lac** in absence of cAMP or in the presence of defined concentrations of externally added cAMP. Finally, Δ*cya* Δ*crp* strains produce neither CRP nor cAMP. These strains carry a gene coding for CRP* (which can directly bind to and activate p*lac** without any need for cAMP) inserted downstream of a p*tet* promoter and located on an expression plasmid. Adding defined quantities of the inducer ATc (anhydro-tetracylcine) to the medium allows a controlled expression of the cAMP-independent variant of CRP, CRP*. Under these conditions the activity of p*lac** should depend solely on the concentration of CRP*, and global regulations. In cells growing exponentially on a single carbon source, we can expect that cell physiology varies little, and the variations in the activity of p*lac** will mainly reflect the

varying concentration of CRP*. When CRP* concentrations are nearly constant, however, the variations of $a_{plac^*}(t)$ should be mainly influenced by global variations of the gene expression machinery. Thus we can separate and observe the relative importance of each effect on the activity of the plac^* promoter.

These strains will be continuously monitored in a series of microplate experiments in well-controlled conditions and growth media. We will see in the next section how the data gathered from these experiments are analyzed to obtain estimated profiles of the relevant biological signals.

1.3.3 Estimation of biological signals from dynamic experiments

The mathematical model proposed in Section 1.3.1 involves promoter activities (a_{plac^*}, a_{pRM}) and concentrations of molecules ([CRP],[cAMP]). In order to compare model prediction with experiments we have to estimate these quantities from the measured time course of the expression of reporter genes. However these estimations are not trivial as they involve mathematical measurement models and require pre-processing the data (noise filtering, background correction, etc.). To this end, we have developed new methods of parameter estimation from noisy experimental measurements. We will also compare these new methods to other existing solutions.

The activity of a_{plac^*}, a_{pRM}, as well as [CRP](t) will be estimated by means of fluorescent reporter genes in microplate experiments with frequent measurements of the total fluorescence of the bacterial population. The fluorescence data thus provide an estimate of the quantity $R(t)$ of fluorescent proteins in the population. At the same time we measure the absorbance, which provides an estimate of the volume $V(t)$ of the bacterial population, up to a proportionality factor α. As will be explained more thoroughly in Chapters 2 and 3 of this thesis, the promoter activity $a(t)$ of a gene of interest can be estimated from the obversations of $R(t)$ and $V(t)$ using a measurement model of the form

$$\frac{d}{dt}R(t) = \alpha V(t)a(t) - d_R R(t) \tag{1.13}$$

where d_R denotes the degradation rate of the reporter. While these experiments yield data with good precision and high temporal resolution, one must

be careful to avoid any artifacts that could lead to a biased estimation of $a(t)$. We use a low-copy plasmid whose copy number per cell has been found to be constant in different conditions, and a long-lived fluorescent protein allele (*gfp-mut2*, half-life $\simeq$ 24h) with negligible problems of bleaching (Zaslaver et al., 2006). Some studies of the *lac* operon, such as the ones discussed in Section 1.1.3, rely on measurements of β-galactosidase activity, which is proportional to the number of β-galactosidase enzymes that were present in the cells from which the proteins were extracted. This technique is unintrusive compared to reporter gene experiments, as the strain does not need to carry a reporter plasmid, but it requires sampling the bacterial population and perform an enzymatic assay for every time point. Only the real-time reporter gene experiments in an automated microplate reader offer the necessary time-resolution for measuring transients during growth transitions.

Constitutive (i.e., unregulated) promoters have long been used in microbiology, where they are generally created by removing or inactivating the regulators of a known promoter. For instance we have seen in Section 1.1.4 that p*lacUV5* was made constitutive by adding IPTG to the growth medium, thereby inactivating the repressor LacI. Constitutive genes have only recently been used as indicators of the cell physiology. In (Gerosa et al., 2013) the authors show that the expression profile of constitutive genes is highly correlated with the profile of the growth rate (and thus with the global physiology of the cell). The authors use genes whose known repressor gene has been deleted, a technique that enabled the authors to test *constitutive* genes with different expression strength, but requires using mutant strains. In our study, we will use the pRM promoter from phage λ, which does not possess any known regulator in *E. coli*, and has already been used as an indicator of cell physiology in (Berthoumieux et al., 2013).

While CRP is an important transcription factor of *E. coli*, its concentration is rarely measured in the CCR studies. It is commonly assumed that the variations of [CRP-cAMP] are driven by variations of cAMP. We will explicitly verify this assumption in our study by monitoring (indirectly) the intracellular concentration of CRP using reporter genes (as explained in Chapter 3). Other possible techniques for estimating the concentration of CRP include mRNA quantification using DNA microarrays or qRT-PCR(Enjalbert et al., 2013).

Such measurements can be misleading, as we will see that the transcription rate of *crp* can decrease but still lead to an increase of [CRP] due to a decreased rate of dilution. Antibody-based Western blots have also been used to measure the concentration of CRP in the cell (Inada et al., 1996). While they yield a more direct estimation of CRP levels in the bacterial population, they do not allow for measurements as precise and frequent as fluorescent reporter genes (see (de Jong et al., 2010) for a comparison of the two techniques). Finally, CRP concentrations can be estimated using mass-spectrometry-based proteomics techniques, which offers good accuracy but currently weak time resolution.

Estimating the concentration of intracellular cAMP is a difficult task, in part because cAMP production can rise and fall several folds in a matter of minutes upon addition or depletion of glucose. While there is no method to date for reliably measuring the intracellular cAMP concentration directly, it is possible to quantify the concentration of cAMP in the growth medium. We measure external cAMP using a competitive ELISA assay with a specific anti-cAMP antibody line in (Berthoumieux et al., 2013), where luminescence intensities measured in microplate can be related to extracellular cAMP concentrations by means of a calibration curve. By measuring the cAMP concentration at several points in time during an experiment, we obtain a time profile of the accumulation rate of cAMP in the growth medium. The rate of export of cAMP from the cell has been shown to be proportional to the intracellular cAMP concentration (see Figure 1.9). This leads to the following equation, relating the observed external cAMP concentration to the internal cAMP level and the volume $V(t)$ of the bacterial population, estimated from the absorbance:

$$\frac{d}{dt}[cAMP_{ext.}](t) = C_e V(t)[cAMP_{int.}](t) - d_{cAMP}[cAMP_{ext.}](t). \qquad (1.14)$$

In Equation 1.14 C_e (min^{-1}) is an export rate and d_{cAMP} (min^{-1}) the degradation rate of cAMP in the medium, which is assumed to be constant. This equation, which is similar to Equation 1.13 for the estimation of promoter activity, enables us to estimate the profile of $[cAMP_{int.}](t)$ from the observables $[cAMP_{ext.}](t)$ and $V(t)$ with a method similar to the one developped in

Chapter 3. Some authors also measured intracellular cAMP by removing the bacteria from their medium using fast filtering techniques, lysing the washed bacteria and quantifying the released cAMP (Kuo et al., 2003). However, as the concentration of cAMP inside bacteria can vary considerably over short periods of time (Epstein et al., 1975), artifacts can arise from contamination with extracellular cAMP (Pastan and Sankar, 1976). This method also requires to sample large volumes in order to obtain cAMP concentrations measurable with a standard assay. As a final note, the use of FRET (Fluorescence resonance energy transfer) techniques could provide precise, dynamical, and non-intrusive quantification of the cAMP concentration *in vivo* in the future (Sekar and Periasamy, 2003; Odaka et al., 2014).

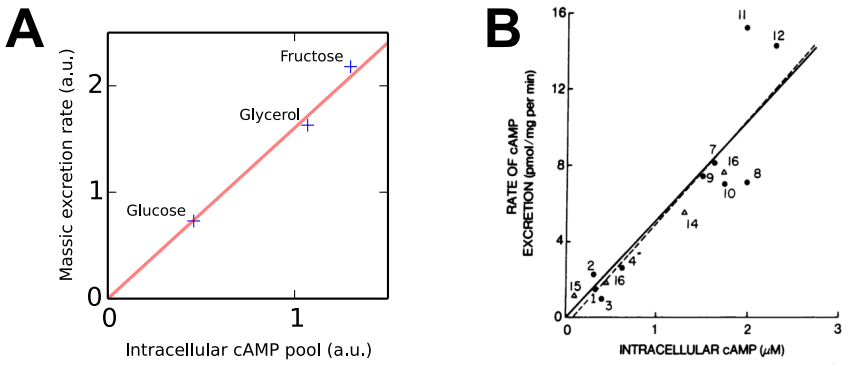


Figure 1.9: Relationship between intracellular cAMP and per-cell acculmulation rate of cAMP in the growth medium. Panel A shows the data from (Fraser and Yamazaki, 1979) for glucose, fructose and glycerol, which will be used in our study, and correspond to numbers 4, 7, and 10 respectively in panel B, from (Epstein et al., 1975). Both graphs show a linear relationship between intracellular cAMP concentrations and the rate of accumulation of cAMP in the growth medium.

1.4 Organisation of the thesis

The chapters of this thesis are organized as follows:

Chapter 2 - One-step and two-step models of gene expression This chapter briefly presents a mathematical and experimental approach justifying the common use of one-step differential equations when modeling gene expression with ordinary differential equations. The results in this chapter are the subject of an article submitted for publication.

Chapter 3 - Estimation of dynamic biological signals by means of linear inverse methods. This chapter focuses on the estimation of growth rate, promoter activity, and protein concentration from dynamical absorbance and reporter fluorescence data, which are the primary data of our study of CCR. We present new mathematical methods for the treatment of data and show that these novel techniques compare favorably with the existing methods in terms of robustness to noise, precision, and the ability to capture fast changes during growth transitions. The results in this chapter are the subject of an article submitted for publication.

Chapter 4 - Dissection of the regulations of a CRP-cAMP dependent promoter In this chapter we present and discuss the results of our study of CCR. The results in this chapter are the subject of an article in preparation.

Summary of chapter 2

The analysis of large regulatory networks requires gene expression to be modeled in a simple and accurate way. Most models describe the synthesis of proteins as either two sequential steps (transcription and translation) or as a single, lumped step. While one-step models are easier to handle, their validity in practice has been rarely studied.

We define the model reduction error as the relative difference between the predicted protein concentration profiles in one-step and two-step Ordinary Differential Equation models. We experimentally quantify the model reduction error for the genes *crp* and *acs*, involved in carbon metabolism in *Escherichia coli*, by means of fluorescent reporter genes. Although the two genes have quite distinct dynamics during growth transitions, the one-step model was able to reproduce the predictions of the two-step model with an error on the order of a few percent for both genes (0.5-4%).

In bigger models these errors can amplified, notably through non-linearities in the gene interactions. These results are consistent with our mathematical analysis, which provides simple formulas for predicting the upper bound of the model reduction error in a system of one gene or in a regulation chain. Our study shows that for biologically plausible parameters this error will be small compared to the observed biological variability. It also provides simple rules to understand which genes of a network model should or should not be modeled with one step.

Résumé du Chapitre 2

Modèles d'expression génique à une et deux étapes. L'analyse de grands réseaux de régulation génique nécessite de modéliser l'expression génique de manière à la fois simple et numériquement précise. Les modèles dynamiques utilisés décrivent généralement la synthèse de protéines comme un processus à deux étapes (transcription et traduction), ou regroupent ces deux phénomènes en un modèle simplifié à une seule étape. Si les modèles à une étape peuvent être préférés pour leur simplicité (et en particulier leur moins grand nombre de paramètres), leur validité est rarement testée en pratique.

Nous définissons *l'erreur de réduction* comme la différence relative entre les prédictions de concentration de protéines de modèles basés sur des équations différentielles à une et deux étapes respectivement. Nous quantifions cette erreur expérimentalement dans le cas des gènes *crp* et *acs* impliqués dans le métabolisme carboné d'*Escherichia coli*, à l'aide de gènes rapporteurs pour suivre l'activité génique de ces gènes au cours du temps.

Bien que ces deux gènes aient des dynamiques assez distinctes au moment des transitions de phase, le modèle simplifié à une étape s'est avéré capable de produire les mêmes prédictions que le modèle à une étape pour ces deux gènes, avec une erreur de seulement quelques pour-cents ($0.5 - 4\%$). Dans le cas de modèles impliquant plusieurs gènes ces erreurs peuvent être amplifiées.

Ces résultats sont consistants avec notre analyse mathématique, qui fournit des formules simples pour les bornes de l'erreur de réduction dans un système à un gène ou dans le cas d'une chaine de régulations. Notre étude montre que, pour tous paramètres biologiquement acceptables, l'erreur de réduction sera presque toujours très faible, ce qui justifie l'emploi de modèles à une étape dans notre étude de la répression catabolique. Nous donnons également quelques règles simples pour déterminer quels gènes d'un modèle devraient voire leur expression modélisée en une seule étape afin de réduire le modèle.

Chapter 2

One-Step and Two-Step Models of Gene Expression in Bacteria

2.1 Motivation

Protein synthesis is a complex process involving many biochemical reactions and intermediate products. Most models of gene expression do not capture the full complexity of this process, but rather distinguish two prime steps: transcription and translation (Kremling, 2007; Bolouri, 2008; Mehra et al., 2003). Transcription denotes the synthesis of messenger RNA (mRNA), while translation is the process by which proteins are produced from the information contained in the mRNA. Such two-step gene expression models are usually the building blocks of larger models describing the interactions of several genes, mRNAs, proteins, and metabolites. To make the network simpler to analyze and easier to handle computationally, it may be worthwhile to simplify even further this two-step model, and lump the entire gene expression process into a single step. This simplifying assumption underlies approximate models like piecewise-linear differential equation (PLDE) and logical models (Thomas, 1973; Glass and Kauffman, 1973).

We will focus on Ordinary Differential Equations (ODE) models, for which both two-step or one-step models of gene expression can be used. Examples of well-characterized gene regulatory networks for which ODE models have been

developed are the network controlling the early development of the Drosophila embryo (von Dassow et al., 2000), the circadian clock in mammals and plants (Goldbeter, 1996; Locke et al., 2005), and the expression of the *lac* operon in *E. coli* (Santillán and Mackey, 2001). In the context of ODEs, the reduction of two-step models to one-step models is usually based on the assumption that the mRNA concentrations are in quasi-steady state, in the sense that, on the time-scale of variations in the protein concentrations, they adapt almost instantaneously to changes in the promoter activity. This assumption, which makes it possible to overlook the variations of the mRNA concentration, and write the variations of the protein concentration directly as a function of the promoter activity, is known as the quasi-steady-state approximation (QSSA). Simplifications based on similar arguments have been extensively studied in enzyme kinetics, notably in the context of the reduction of mass-action models to the Michaelis-Menten rate law (Borghans et al., 1996; Chen et al., 2010; Roussel and Fraser, 2001; Segel and Slemrod, 1989). Gene expression models have been less studied from this point of view. Under which conditions is it justified to simplify two-step models to one-step models, and to which extent will this simplification influence the model predictions? In this paper we address the above two questions with both an analytic and and experimental approach. We characterize the difference over time of the predictions of the one-step and the two-step model, which we will refer to as the model reduction error. We provide mathematical propositions to easily predict upper bounds for the model reduction error, and show how the time profile of this error can be estimated from gene expression data obtained by means of fluorescent reporter genes. As an application, we determine the model reduction error for two genes involved in carbon metabolism of the enterobacterium *E. coli*. The gene *crp* codes for the pleiotropic transcription factor Crp (Gosset et al., 2004a; Kolb et al., 1993a), whereas *acs* encodes the enzyme acetyl-CoA synthetase (Acs), catalyzing an important step in acetate metabolism. The transcription of *acs* is regulated by the complex formed by Crp and the signalling metabolite cyclic AMP (cAMP), alongside other transcription factors (Wolfe, 2005). The population-level expression dynamics of these genes are measured in batch conditions during a typical growth-phase transition, when the externally-supplied carbon source glucose is depleted. While the activity of

the *crp* promoter exhibits slow variations, on the order of the doubling time of the cells, the *acs* promoter is induced within minutes after the growth arrest.

The reduction of two-step to one-step models is usually based on the assumption that mRNA half-life is much smaller than protein half-life (Polynikis et al., 2009). Our theoretical study of the model reduction error partly confirms this, but more interestingly, indicates that slow variation of the promoter activity is a sufficient condition for model reduction, independently of the protein half-life. The experimental data show that, for physiologically-relevant parameter sets, the relative error induced by the model reduction of a gene will remain below the observed biological variability, both for genes with fast and slowly-varying promoter activities. In networks of several genes, however, the model reduction is more problematic, since the error made on one gene can cause large perturbations as it is propagated to the model elements downstream of that gene. Our study also shows that reducing the models of several genes at once may quickly lead to large prediction errors. The theoretical criteria developed in this paper provide guidelines for the choice of one-step or two-step models when describing the dynamics of gene regulatory networks, in bacteria and higher organisms.

2.2 Methods and materials

2.2.1 Strains and growth conditions

The *E. coli* strain used in this study is the strain BW25113 (Baba et al., 2006). The strain was transformed with low-copy pUA66 plasmids from the Alon collection (Zaslaver et al., 2006), bearing a *gfp* reporter gene and a *kan* resistance marker. In particular, we used plasmids bearing a transcriptional fusion of the *crp* and *acs* promoter regions with the *gfp* reporter and a promoterless vector for background correction (see below). The reporter gene encodes a stable and fast-folding version of the GFP reporter (GFPmut2).

Glycerol stocks (-80°C) of the above-mentioned reporter strains were grown overnight (about 15 h) at 37°C, with shaking at 200 rpm, in M9 minimal medium (Miller, 1972) supplemented with 0.3% glucose and mineral trace elements. For plasmid-carrying strains, the growth medium was supplemented

with 100 μg ml^{-1} kanamycin. The overnight cultures were strongly diluted (1500-7000 fold) into a 96-well microplate, so as to obtain an adjusted initial OD_{600} of 0.001. The wells of the microplate contain M9 minimal medium supplemented with 0.3% glucose, mineral trace elements, and 1.2% of the buffering agent HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) for maintaining physiological pH levels in the growth medium. No antibiotics were added at this stage. The wells were covered with 60 μl of mineral oil to avoid evaporation. The microplate cultures were then grown for about 24 h at 37°C, with agitation at regular intervals, in the Fusion microplate reader (Perkin Elmer).

2.2.2 Fluorescent reporter gene measurements

The strains growing in the wells of the microplate express a fluorescent reporter of the genes *crp* and *acs*. During a typical experimental run, we acquire about 120 readings each of absorbance (600 nm) and fluorescence (485-520 nm). From the measured signal we remove the background signals of absorbance and fluorescence measured on wells containing growth medium only and strains carrying a promoterless reporter plasmid, respectively. The treatment of these data is described in details in Section A7 of the Supplementary Information (SI). We computed the promoter activities and model reduction error from the resulting absorbance and fluorescence signals, as described in Section 2.3.4.

2.3 Results

2.3.1 One-step and two-step models of gene expression

Two-step models describe gene expression as two coupled processes: mRNA synthesis by transcription of DNA, followed by protein synthesis by translation of mRNA (Fig. 3.1). The processes can be mathematically described as follows. Let $m(t)$ [μM] and $p(t)$ [μM] denote the time-varying concentrations of mRNA and protein, respectively, with time $t \in \mathbb{R}_+$. The two-step model is a system

of two ODEs for $m(t)$ and $p(t)$:

$$\frac{d}{dt}m(t) = \kappa_m\, f(t) - (\gamma_m + \mu(t))\, m(t), \quad m(0) = m_0, \tag{2.1a}$$

$$\frac{d}{dt}p(t) = \kappa_p\, m(t) - (\gamma_p + \mu(t))\, p(t), \quad p(0) = p_0, \tag{2.1b}$$

with κ_m [M min^{-1}] the maximum synthesis rate of mRNA, and κ_p [min^{-1}] the synthesis rate of protein per unit mRNA. The function $f(t) : \mathbb{R}_+ \to [0, 1]$ describes the modulation over time of the rate of mRNA synthesis by transcriptional regulators. Accordingly, $\kappa_m f(t)$ is referred to as the promoter activity. mRNA and protein are degraded in a first-order reaction with degradation constants γ_m [min^{-1}] and γ_p [min^{-1}], respectively, and diluted through the growth of the cell population, with growth rate $\mu(t)$ [min^{-1}]. The degradation constants are related to the half-lives of mRNA and protein, denoted by $\tau_{m,1/2}$ and $\tau_{p,1/2}$ respectively, as follows: $\tau_{m,1/2} = \ln 2/\gamma_m$ and $\tau_{p,1/2} = \ln 2/\gamma_p$. The growth rate of the cell population can be expressed in terms of the volume $V(t)$ [L] of the population:

$$\mu(t) = \frac{1}{V(t)}\frac{d}{dt}V(t), \tag{2.2}$$

or alternatively in terms of the biomass, which is proportional to the volume over a large range of growth rates (Volkmer and Heinemann, 2011). We refer to $m(t)$ and $p(t)$ as the state variables of the model and $\mu(t)$ and $f(t)$ as the (time-varying) input variables.

The typical half-life of mRNA in bacteria (on the order of a few minutes (Bernstein et al., 2002)) is small compared to the time-scale of other phenomena like cell division (on the order of tens of minutes in rich media to hours in minimal media (Schaechter et al., 1958; Andersen and von Meyenburg, 1980)) or protein degradation (on the order of hours for almost all proteins (Mosteller et al., 1980; Larrabee et al., 1980b)). As a consequence, it is usually assumed that mRNA concentrations relax to their steady state much faster than the other variables do. That is, the mRNA concentration is always considered at quasi-steady state with respect to the protein concentration (QSSA, quasi-steady state approximation). This translates in terms of equations as

$$\frac{d}{dt}m(t) = 0 = \kappa_m\, f(t) - (\gamma_m + \mu(t))\, m(t),$$

from which follows that

$$m(t) = \frac{\kappa_m\, f(t)}{\gamma_m + \mu(t)}.$$

We can simplify this equation using the above-mentioned observation that the growth rate is usually small compared to the degradation constant of the mRNA, and write $m(t) = \kappa_m\, f(t)/\gamma_m$. Reinjecting this equality into the first equation leads to an approximate system of a single ODE describing the dynamics of the protein concentration with the same regulatory input $f(t)$:

$$\frac{d}{dt}\hat{p}(t) = \frac{\kappa_m\, \kappa_p}{\gamma_m}\, f(t) - (\gamma_p + \mu(t))\, \hat{p}(t), \quad \hat{p}(0) = \hat{p}_0. \tag{2.3}$$

In Eq. 2.3 the transcription and translation processes are lumped into one step. As a consequence, the ratio $\kappa_m\kappa_p/\gamma_m$ can be treated as a single phenomenological synthesis parameter, thus significantly reducing the number of parameters of the model. An advantage of Eq. 2.3 is that the fast variable $m(t)$ is no longer explicitly considered, so that the model will be easier to solve numerically (reduction of stiffness). The one-step model of gene expression is schematically compared with the two-step model in Fig. 3.1.

A mathematical basis generally invoked for the application of the QSSA is Tykhonov's theorem for dynamical systems (Heinrich and Schuster, 1996; Khalil, 2001). However, this theorem only gives a limit behavior of the system when some scaling parameter converges to infinity. Moreover, it usually does not consider input variables, which vary on a particular time-scale themselves. This raises the question how well the one-step model approximates the two-step model in a particular context, as defined by specific half-lives of mRNA and protein, specific initial conditions, and a specific promoter activity and population growth rate. In our study we will focus on the relative model reduction error, which provides a measure of the quality of the approximation:

$$\Delta(t) = \frac{\mid p(t) - \hat{p}(t) \mid}{p(t)}. \tag{2.4}$$

We will show that this error can be estimated through reporter gene experiments, and that its upper bound is linked in a simple way to the biological parameters.

Intensive variables (concentrations)		Extensive variables (total amounts in the population volume)	
		acs* and *crp	
$m(t)$	mRNA concentration	$M(t) = m(t)V(t)$	Total amount of mRNA in the population
$p(t)$	Protein concentration	$P(t) = p(t)V(t)$	Total amount of protein in the population
$f(t)$	Promoter activity	$F(t) = f(t)V(t)$	Total activity of promoters in the population
		***gfp* reporter**	
$n(t)$	mRNA concentration	$N(t) = m(t)V(t)$	Total amount of mRNA in the population
$q(t)$	Concentration of unfolded protein	$Q(t) = q(t)V(t)$	Total amount of unfolded proteins in the population
$r(t)$	Concentration of folded protein	$R(t) = r(t)V(t)$	Total amount of folded proteins in the population (proportional to the emitted fluorescence)

Table 2.1: **Intensive and extensive variables used in the models**, with units in terms of concentrations and total amounts summed over the cell population volume, respectively.

2.3.2 Computation of the model reduction error from extensive variables

In order to better understand which parameters and which signals are relevant for the applicability of the reduction of two-step to one-step models, we propose a reformulation of Equation 2.1 and Equation 2.3 , based on a change of variables.

The system described in Equation 2.1 undergoes two distinct external perturbations, $f(t)$ and $\mu(t)$, reflecting the influences of transcriptional regulation and changes in the cell population volume, respectively. It is possible to aggregate these effects by introducing the following variables:

$$M(t) = V(t)\, m(t), \quad P(t) = V(t)\, p(t), \quad F(t) = V(t)\, f(t), \quad \hat{P}(t) = V(t)\, \hat{p}(t). \tag{2.5}$$

$M(t)$, $P(t)$, and $\hat{P}(t)$, expressed in mole units, represent the amounts of mRNA and protein summed over the volume of the cell population, while $F(t)$ is the cumulative activity of all promoters of the gene in the cell population, and has the unit mole min^{-1}. This change of variables allows the two-step model to be rewritten as follows (Section A1):

$$\frac{d}{dt}M(t) = \kappa_m F(t) - \gamma_m M(t), \tag{2.6a}$$

$$\frac{d}{dt}P(t) = \kappa_p M(t) - \gamma_p P(t). \tag{2.6b}$$

while the reduced model becomes

$$\frac{d}{dt}\hat{P}(t) = \frac{\kappa_m \kappa_p}{\gamma_m} F(t) - \gamma_p \hat{P}(t). \tag{2.7}$$

Notice that the growth-dilution terms have disappeared from the reformulated models, as we do no longer consider concentrations but total amounts of molecules in a (possibly) expanding volume. Table 2.1 illustrates the conceptual change underlying the reformulation of the model equations. Instead of focusing on a unit volume and considering the rates of production, degradation and dilution of the mRNA and protein molecules within this volume, we consider the amount of molecules produced and degraded within the total volume of the cell population. In physical terms, the reformulation of the model implies a change from intensive to extensive variables.

In general a steady state in terms of the total amount of mRNA and protein, summed over the volume of the cell population, is not equivalent to a steady state in terms of the mRNA and protein concentrations. Strictly speaking, steady states for the systems with intensive variables (Equation 2.1 and Equation 2.3) are equivalent to steady states of the systems with extensive variables (Equation 2.6 and Equation 2.7), if and only if the growth rate is 0. However, when the volume does not vary much, the two notions of steady state are close.

Conveniently, the model reduction error defined in Equation 2.4 can be written as a function of the extensive variables as well:

$$\Delta(t) = \frac{\mid P(t) - \hat{P}(t) \mid}{P(t)}. \tag{2.8}$$

In what follows, we study $\Delta(t)$ by means of the new systems of Equation 2.6 and Equation 2.7 . In this reformulation the relative error $\Delta(t)$ does not depend on $f(t)$ and $\mu(t)$ separately, but is affected by their joint influence on

the variable $F(t)$. Besides greatly simplifying the analysis of the systems, this comes with interesting practical implications, as experimentally measuring $f(t)$ and $\mu(t)$ would require accurate measurements of the volume of the cell population during a reporter experiment, while $F(t)$ can be directly inferred from the measured population-level fluorescence signals (Section 2.3.4).

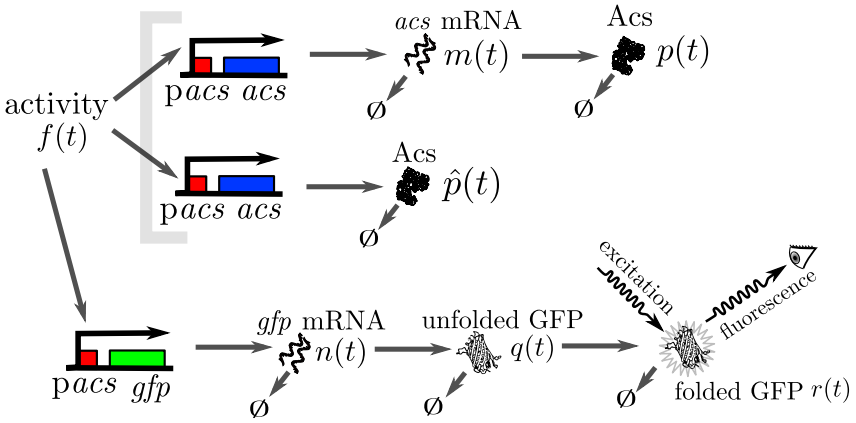


Figure 2.1: **Two-step, one-step, and reporter models for the *acs* gene**. The two-step model (top) includes transcription and translation processes, whereas the one-step model (center) directly links protein synthesis to the activity of the promoter of the gene. In the reporter model (bottom), the *gfp* gene has the same activity $f(t)$ as *acs*, as it has the same promoter region p*acs*. The promoter activity modulates the production of a fluorescent protein whose variations can be observed and used to estimate $f(t)$.

2.3.3 Theoretical upper bounds for the model reduction error

The following two properties give upper bounds for the model reduction error $\Delta(t)$ in the case of slow and fast variations of $F(t)$, respectively. The first property states that the error bound depends on the (maximum) relative change of the cumulated promoter activity $F(t)$ in the time-interval under consideration (see Section A2 of the SI for the proof).

Proposition 1. *Assume that the systems of Eq. 2.6 and Eq. 2.7 verify*

$$\frac{d}{dt}M(0) = \frac{d}{dt}P(0) = \frac{d}{dt}\hat{P}(0) = 0.$$

Then for all $t > 0$*,*

$$\Delta(t) < \frac{1}{\gamma_m} \sup_{s<t} \left| \frac{1}{F(s)} \frac{d}{ds} F(s) \right|, \tag{2.9}$$

or equivalently, in terms of the original variables,

$$\Delta(t) < \frac{1}{\gamma_m} \sup_{s<t} \left| \mu(s) + \frac{1}{f(s)} \frac{d}{ds} f(s) \right|. \tag{2.10}$$

Note that in Proposition 1 the production constants κ_m, κ_p appear to play no role in the formation of the approximation error: this can be considered a general rule as long as we neglect the role of the initial conditions of the systems (Section A1.3).

The second property states that, when the promoter activity $f(t)$ suddenly steps up or down from an initial steady-state level, the upper bound of the error linearly depends on the activity fold change (see Section A3 for the proof).

Proposition 2. *Assume that the systems of Eq. 1 and Eq. 3 are such that*

$$\frac{d}{dt}m(0) = \frac{d}{dt}p(0) = \frac{d}{dt}\hat{p}(0) = 0.$$

If at $t = 0$ *the promoter activity changes by a relative factor* $\chi > -1$ *after the initial state, i.e.,* $f(t) = (1 + \chi)f(0)$*,* $t > 0$*, then under the assumption that the growth rate is constant (*$\mu(t) = \mu$*), and bearing in mind that* $\gamma_m > \gamma_p$*, it holds for all* $t > 0$ *that*

$$\Delta(t) \leq |\chi| \frac{\gamma_p + \mu}{\gamma_m - \gamma_p} (1 + \frac{\mu}{\gamma_m}) + \frac{\mu}{\gamma_m}. \tag{2.11}$$

For $\gamma_m \gg \mu, \gamma_p$, Eq. 2.11 simplifies to

$$\Delta(t) \leq |\chi| \frac{\gamma_p + 2\mu}{\gamma_m}. \tag{2.12}$$

The two propositions provide simple rules of thumbs to predict upper bounds for the model reduction error in various situations. Proposition 1

shows that the relative variations of the promoter activity and the volume contribute to the error $\Delta(t)$ in an additive way. Moreover, they are of equal importance, independently of the parameters characterizing the synthesis and degradation processes. At steady state, when the promoter activity $f(t)$ is constant, the model reduction error is bounded by μ/γ_m. For a typical growth rate of 0.01 min^{-1} (doubling time of 70 min) and a typical mRNA degradation rate of 0.5 min^{-1} (half-life of 1.5 min), we are ensured that the model reduction error will stay below 2%. If the promoter activity exhibits rapid variations, however, Proposition 1 becomes little informative, as it will provide a rough estimation of the upper bound. In this case, a long half-life of the proteins will give high amounts of protein (or equivalently, concentrations), and slow variations of these amounts (or concentrations), thus leading to small relative errors. Proposition 2 demonstrates the antagonistic actions of the degradation rates γ_m and γ_p on the model reduction error in the extreme case of a step change in promoter activity. Were the promoter activity to instantly double or vanish (in both cases $|\chi| = 1$), then for a typical protein degradation rate of 0.01 min^{-1} (half-life of 70 min) the approximation error would stay below 4-6%. This also ensures that high-frequency perturbations of small amplitude (*e.g.*, biological noise), which occur on a fast time-scale, have a very limited impact on the approximation error Δt.

In conclusion, Proposition 1 shows that having slow relative variations of $F(t)$ (as compared to the characteristic time-scale of the mRNA dynamics, determined by γ_m) is a sufficient condition for a low error $\Delta(t)$. Were these conditions not fulfilled, Proposition 2 shows that a small ratio γ_p/γ_m will buffer the effect of fast variations of $F(t)$ (or $f(t)$). These two last considerations are supported by a qualitative analysis of the systems based on filter theory (Section A4). We will assess the predictive power of these propositions in the next sections, where we use experimentally measured promoter activities to obtain a profile of $\Delta(t)$ for two genes of *E. coli*.

2.3.4 Experimental quantification of the model reduction error

We have shown in the previous previous section that the relative error $\Delta(t)$ mainly depends on γ_p, γ_m, and the variations of the cumulative promoter activity $F(t)$. We will show in this section how the temporal profile of the promoter activity can be inferred from reporter gene experiments. This method will be used in the next section to estimate the the $\Delta(t)$ profile for the two *E. coli* genes.

Current reporter gene technologies, based on Green Fluorescent Proteins (GFPs) and other fluorescent and luminescent reporter proteins, provide an excellent means to measure promoter activities *in vivo* and in real time ((Giepmans et al., 2006; Tsien, 1998), Fig. 3.1). The underlying principle of the technology is to fuse the promoter region and possibly (part of) the coding region of a gene of interest to a reporter gene. The expression of the reporter gene generates a visible signal (fluorescence) that is easy to capture and reflects the expression of a gene of interest. Measurement models allow biologically-relevant quantities occurring in the models, such as $f(t)$, $F(t)$, $p(t)$, and $P(t)$, to be reconstructed from the primary fluorescent data (see (de Jong et al., 2010; Finkenstädt et al., 2008; Huang et al., 2008) and references therein). In particular, measurement models separate the signal of interest from background noise and correct for systematic biases, such as different half-lives of reporter and host proteins, folding times, photo-bleaching effects, *etc.*

In earlier work (de Jong et al., 2010), we developed and validated a measurement model for the interpretation of fluorescence data. This model is a variant of the two-step model, taking into account that the fluorescent activity of GFP in response to light excitation depends on post-translational modifications, notably the folding of the protein to an appropriate conformation, including the autocatalytic formation of the chromophore (Tsien, 1998). This maturation process gives rise to an additional reaction step from inactive to active GFP (Fig. 3.1). As a consequence, the state variables of the measurement model are *gfp* mRNA ($n(t)$ [μM]), inactive GFP ($q(t)$ [μM]), and active GFP ($r(t)$ [μM]). The synthesis parameters κ_n and κ_q correspond to κ_m and κ_p in the gene expression model of Eq. 2.1, respectively, while κ_l [min^{-1}] is

the GFP folding constant. The degradation constants of reporter mRNA and of the folded and unfolded reporter proteins are denoted by γ_n, γ_r, and γ_q, respectively. By construction, the promoter regions of the reporter gene and the gene of interest are the same, so it is natural to assume that both have the same promoter activity $f(t)$.

$$\frac{d}{dt}n(t) = \kappa_n\, f(t) - (\gamma_n + \mu(t))\, n(t), \tag{2.13a}$$

$$\frac{d}{dt}q(t) = \kappa_q\, n(t) - (\gamma_q + \kappa_r + \mu(t))\, q(t), \tag{2.13b}$$

$$\frac{d}{dt}r(t) = \kappa_r\, q(t) - (\gamma_r + \mu(t))\, r(t). \tag{2.13c}$$

While the degradation constants of the folded and unfolded GFP are identical ($\gamma_q = \gamma_r$), the degradation constants of the reporter mRNA and reporter protein are different from the degradation constants for the products of the gene of interest, that is, $\gamma_q \neq \gamma_p$ and $\gamma_n \neq \gamma_m$. Following a change of variables, replacing like in Section 2.3.2 concentrations by total amounts of molecules in the cell population volume, we obtain the system

$$\frac{d}{dt}N(t) = \kappa_n F(t) - \gamma_n N(t), \tag{2.14a}$$

$$\frac{d}{dt}Q(t) = \kappa_q N(t) - (\gamma_r + \kappa_r)Q(t), \tag{2.14b}$$

$$\frac{d}{dt}R(t) = \kappa_r Q(t) - \gamma_r R(t). \tag{2.14c}$$

A summary of the model variables and their biological meaning is given in Table 2.1.

In the model of Eq. 2.14, $R(t)$ represents the total amount of fluorescent protein in the cell population volume. This variable can be assumed proportional to the experimentally observed reporter fluorescence $I(t)$. Eq. 2.14 enables us to express $F(t)$ directly as a function of $R(t)$ and its derivatives, so that the profile of $F(t)$ can be inferred, up to a multiplicative constant, from the fluorescence data (see Section A5 for the proof):

$$F_e(t) \stackrel{\text{def.}}{=} I(t) + a\frac{d}{dt}I(t) + b\frac{d^2}{dt^2}I(t) + c\frac{d^3}{dt^3}I(t) \propto F(t), \tag{2.15}$$

where

$$a = \frac{1}{\gamma_n} + \frac{1}{\gamma_r} + \frac{1}{\gamma_r + \kappa_r}, \quad b = \frac{1}{\gamma_n\gamma_r} + \frac{1}{\gamma_n(\gamma_r + \kappa_r)} + \frac{1}{\gamma_r(\gamma_r + \kappa_r)},$$
$$c = \frac{1}{\gamma_r\gamma_n(\gamma_r + \kappa_r)}.$$

Because Eq. 2.6 and Eq. 2.7 are linear systems, scaling the input $F(t)$ with an unknown multiplicative constant will scale the outputs $P(t)$ and $\hat{P}(t)$ by the same factor, and thus not change the relative error. Therefore, using F_e instead of $F(t)$ will lead to the same result. F_e can be computed from the fluorescence data in different ways, using smoothing splines, sliding windows, or a direct inversion method (Bansal et al., 2012). Since all of these methods yield, up to small differences, the same profiles for $F(t)$ and $\Delta(t)$ (see Section A8), only the results obtained from the splines method are reported below. We verified that using different degrees of smoothing when treating the experimental data did not impact the results, which is a consequence of the filtering behavior of the gene expression system, as explained in Section A4.

To infer the model reduction error $\Delta(t)$ we injected the expression of $F_e(t)$ in the two-step and one-step systems given by Eq. 2.6 and Eq. 2.7, which we supposed to be initially at steady state with initial value $F_e(0)$ for all variables. We numerically solved the equations and thus obtained profiles of $P(t)$ and $\hat{P}(t)$, respectively, from which we computed $\Delta(t)$ according to Eq. 2.8. The next section presents the results obtained for the genes *acs* and *crp* in *E. coli*.

2.3.5 Model reduction error for selected *E. coli* genes

Microorganisms like the enterobacterium *E. coli* use glucose and other carbon sources for growth. External carbon sources are taken up by the cell and converted by central carbon metabolism into precursors for the synthesis of macromolecular constituents of the cell as well as the energy (ATP) required for biosynthetic functions (Gottschalk, 1986). *E. coli* has intricate regulatory mechanisms, on both the metabolic and genetic level, to adapt the functioning

of metabolism to the availability of different carbon sources in the environment. For instance, when a preferred (rich) carbon source like glucose is depleted, it continues its growth on less preferred (poorer) carbon sources like acetate. A change in carbon source is accompanied by a profound reorganization of the metabolic flux distribution and of the expression of the genes encoding enzymes of the metabolic reactions (Oh et al., 2002).

We consider here two genes that are typical for the kind of expression patterns encountered during growth transitions. The gene *crp* encodes the transcription factor Crp that is a pleiotropic regulator of the cell (Kolb et al., 1993b). The complex Crp·cAMP regulates the transcription of hundreds of genes in *E. coli*, many of them enzymes catalyzing reaction steps in carbon metabolism (Gosset et al., 2004a; Gutierrez-Ríos et al., 2007). The protein concentration has been shown to vary little during the transition phase (Berthoumieux et al., 2013; Kuhlman et al., 2007). The second gene considered here is *acs*, which is notably regulated by the Crp·cAMP complex, together with other transcription factors (Wolfe, 2005). It encodes the enzyme acetyl-CoA synthetase (Acs), which converts acetate to acetyl-CoA, a critical step in acetate assimilation. The gene *acs* is expressed at a weak basal level in the presence of glucose, but is strongly induced upon its depletion (Fig. 2.3A).

Based on the theoretical results of Section 2.3.3, one expects that a rapid and strong change in promoter activity leads to a (transient) increase of the model reduction error. On the other hand, weak variations of promoter activities are expected to keep the model reduction error low. We tested the above predictions by means of a reporter gene experiment.

Batch cultures of *E. coli* were grown in a microplate at 37 °C in M9 minimal medium supplemented with glucose, as described in Section 2.2.1. To measure the expression of the genes of interest, we used strains transformed with reporter plasmids carrying a transcriptional fusion of a *gfp* reporter gene and the promoter region of *crp* and *acs*, respectively (Zaslaver et al., 2006). By means of an automated microplate reader, we monitored in real time and *in vivo* the absorbance of the culture and the emitted fluorescence. The absorbance is not strictly necessary for our study, as we have shown that the variable of interest, $\Delta(t)$, can be reconstructed from $I(t)$ alone. However, the absorbance is used as a measure for the population volume, in order to compute the variables

$p(t)$, $\hat{p}(t)$, and $m(t)$, and to synchronize datasets from different wells on the microplate.

Fig. 2.2 and Fig. 2.3 show the results obtained with the reporters of the *crp* and *acs* reporters, respectively. For each gene we applied the methods described in Section 2.3.4 to infer the cumulative promoter activity $F(t)$, reconstruct the profiles of the total amount of protein $P(t)$, and compute $\Delta(t)$. We also provide the profiles of $f(t)$, $p(t)$, and $\hat{p}(t)$ for each gene, obtained by dividing $F(t)$, $P(t)$ and $\hat{P}(t)$ by the absorbance. The computation of the promoter activities of *crp* and *acs* requires the degradation and folding constants of the reporter system to be known. The GFP used in our study was shown to have a half-life of about 17 h in our conditions ($\gamma_q = 0.0007$ min^{-1}) and a folding rate κ_r=0.3 min^{-1}. For *acs* and *crp* we took half-lives measured in our conditions for the gene *fis* as reference values for γ_p and γ_m (about 100 min and 1.2 min, respectively, corresponding to $\gamma_p = 0.0065$ min^{-1} and $\gamma_m = 0.56$ min^{-1}).

The promoter activities of *crp* and *acs* are seen to react in quite distinct ways to the depletion of glucose. Fig. 2.2 shows that well before the growth arrest, the promoter activity $f(t)$ of *crp* starts to decrease and reaches a two-fold lower stationary level. The expression of *acs* is negligible while glucose is present, resulting in a fluorescence signal close to the background level (Section A7). However, when glucose is exhausted the transcription of *acs* is strongly, but transiently induced under the effect of Crp·cAMP (Fig. 2.3*A*-*B*).

One can see in Fig. 2.2 that the relative error stays below 1% in the case of the gene *crp*. The growth of the total promoter activity, $(1/F)dF/dt$, is equal to 0.013 during exponential phase. Proposition 1 then predicts an error inferior to $(1/\gamma_m)(1/F)dF/dt = 2.3\%$, which is verified. The approximation error is much larger in the case of the gene *acs*, but only transiently. In this case the promoter activity profile could be assimilated to a step function, like in Proposition 2. However, this step has a very large amplitude as the promoter activity starts from a negligible level. Therefore, Proposition 2 predicts a rough and not very informative upper bound for $\Delta(t)$, far above the maximum obtained (4%). In Fig. 2.2C and Fig. 2.3C the predictions of the one-step and two-step models are almost superimposed: the difference induced by the model reduction is much smaller than the variability of the predictions observed on

biological replicates, which is on the order of 10%.

For longer-lived RNAs or shorter-lived proteins, larger errors can be expected. Fig. 2.4 shows the maximal approximation errors obtained when performing the computations leading to Fig. 2.2D and Fig.2.3D using different values of γ_p and γ_m. The inverse relationship between the maximal error and γ_m, which was put forward in Propositions 1 and 2, is striking for both *acs* and *crp*. Moreover, $\Delta(t)$ appears to be much more sensible to γ_m than to γ_p in the parameter region considered. Even in the most disadvantageous (and unlikely) case that $\gamma_m = \gamma_p = 0.1$, the approximation error remains below 10%.

2.3.6 Propagation of the approximation error

When gene expression systems such as described in Eq. 2.1 are part of bigger models involving several genes and proteins, an error in the prediction of the protein concentration $p(t)$ can affect the prediction of other proteins regulated by p. Even though the model reduction error will generally stay under 4% for a system of a single gene, as shown in Sec. 2.3.4, it is possible that this error will be amplified throughout the gene network, due to non-linearities in the interactions between the genes.

To study the propagation of the error we consider a simple system in which a gene g_1 coding for a protein p_1 is the activator of another gene g_2 coding for a protein p_2. How will a model reduction error on p_1 affect the predictions on p_2 ? And how will the predictions on p_2 differ if we model both genes using one step, or two steps ? We assume that the concentration $p_1(t)$ is driven by Eq. 2.1, while the concentration of $p_2(t)$ is driven by a similar system depending on $p_1(t)$:

$$\frac{d}{dt}m_2(t) = \kappa'_m\, f_2\left(p_1(t)\right) - \left(\gamma'_m + \mu(t)\right) m_2(t), \quad m_2(0) = m_{2,0}, \tag{2.16a}$$

$$\frac{d}{dt}p_2(t) = \kappa_2\, m_2(t) - \left(\gamma_2 + \mu(t)\right) p_2(t), \quad p_2(0) = p_{2,0}. \tag{2.16b}$$

The function f_2 describes the activation of g_2 by p_1, and is classically modelled using a Hill function (Kuhlman et al., 2007):

$$f_2\left(p_1(t)\right) = \frac{p_1(t)^a}{K^a + p_1(t)^a}, \tag{2.17}$$

where a is a cooperativity constant and K represents the saturation threshold of the promoter of g_2 with respect to the protein p_1. A large value of a indicates a higher non-linearity in the action of protein p_1 on the gene g_2.

To understand how the parameters K and a influence the propagation of the model reduction error we compared the outputs of three models with different degrees of model reduction (as illustrated in Fig. 2.5). In the *full model*, the gene expression systems of p_1 and p_2 are given by a two-step models as in Eq. 2.1 and Eq. 2.16. The promoter activity of p_1, as well as the degradation rates γ_m, γ_p of the two genes, were chosen as for the experimentally observed gene *acs* (Fig. 2.3). We denote $p_2(t)$ the output of the simulation. In a second model, we reduced the expression model of g_1 to one-step (as in Eq. 2.3), which led to an approximate output $\hat{p_2}(t)$ in the two-genes system. In a third model we simultaneously reduced the models of genes g_1 and g_2 to one step, and we denote $\mathring{p_2}(t)$ the output of the simulations. We led simulations of the three models for different values of K and a and after each round of simulations we computed the maximal relative differences between $p_2(t)$ and $\hat{p_2}(t)$ (Fig. 2.6A) and between $p_2(t)$ and $\mathring{p_2}(t)$ (Fig. 2.6B). Thus, Fig. 2.6A illustrates the propagation to another gene of a model reduction error on one gene, while Fig. 2.6B shows the effect of multiple model reductions on the output of a gene cascade.

We see on Fig. 2.6B that, in case of important non-linearities (i.e. for large values of a), an amplification of the model reduction error can be expected. This is mainly due to the fact that, in the particular experiment used for simulations, p_1 takes small values and is therefore in the non-linear region of f_2. A mathematical analysis (reported in Sec. A6) shows that the condition

$$\min p_1(t) \geq K \sqrt[a]{a-1} \tag{2.18}$$

is sufficient to ensure the non amplification of the model reduction error. This condition means that the activity of gene g_2 is at least at $(1 - 1/a)$ of its theoretical maximum, and can be generalized to $\min p_1(t) \geq 4K/3$. We conclude that there are two categories of genes which will not amplify the model

reduction error made on their regulator: genes whose regulator has a weak cooperativity constant ($a \leq 1$) and genes that are always activated at a high level.

Applying a double-model reduction (Fig. 2.6B), where both the regulator and the regulated genes are modelled with one step, led to a doubling of the error in our example. An approximately three-fold multiplication can be observed for a cascade of three genes (data not shown). Fig. 2.6C shows that, in addition to being larger than the error on p_1, the propagated error on p_2 is also more persistent in time. Thus, although the model reduction results in acceptable errors for systems of a single gene, as shown in the previous sections, multiple model reductions can cause important accuracy losses in gene networks, and one should select with care the genes of the network which will have their model reduced.

2.4 Discussion

Mathematical models of cellular networks are composed of sub-models describing the synthesis and degradation of the products of individual genes (Karr et al., 2012). In this paper we have compared classical two-step models of gene expression, which explicitly distinguish transcription and translation and which are themselves reduced versions of models describing individual reaction steps in detail (Kremling, 2007; Morozova et al., 2012), with one-step models lumping gene expression into a single step. This reduction is often motivated by the distinct time-scales of the dynamics of mRNA and protein concentrations, as a consequence of the large difference in half-life (minutes for mRNA, hours for protein in bacteria). We have defined a measure of the time-varying error incurred when approximating the two-step model by a one-step model, and shown that this error depended mainly on the half-life of the mRNA and the variation rate of the gene's promoter activity. Moreover, we have shown how the model reduction error can be experimentally quantified using fluorescent reporter gene data. This error was computed for two typical genes in *E. coli*: the gene *crp*, encoding a global transcription regulator, and the gene *acs*, encoding an enzyme of central metabolism.

Probably the most interesting observation of this paper is that the model reduction error is largely negligible over different growth phases, exponential growth on glucose and growth arrest after glucose depletion. For *crp* the error remains below 1%, while for *acs* it transiently rises to 4%, following the rapid induction of this gene after growth arrest. However, even this transient peak is well below the variability of experimental replicates (10%). This conclusion is robust to different treatments of the primary fluorescent data and it remains valid over a range of physiological values of the degradation constants (half-lives) of mRNA and protein.

We presented theoretical results that give some deeper insights into the factors that influence the magnitude of the model reduction error. For a system that is initially in a non-growing steady state, we have shown that that the model reduction error depends on the sum of the rate of change of the promoter activity and the growth rate of the cell population (Proposition 1). If these rates are small in comparison with the mRNA half-life, the model reduction error will remain acceptably low, as illustrated for the gene *crp*. While a rapid change in promoter activity acts as a perturbation driving the mRNA concentration away from its quasi-steady-state value, thus increasing the model reduction error, a short mRNA half-life favors a rapid return to the quasi-steady state and avoids the accumulation of the error. In a different context, Turányi *et al.* arrive at a similar conclusion when analyzing transient errors occurring before the system approaches the slow manifold (Turanyi et al., 1993). Proposition 2 predicts that after a rapid perturbation of the promoter activity, for instance the accumulation of cAMP following the exhaustion of glucose in the growth medium, which induces the transcription of *acs*, the model reduction error is bounded by the activity fold change and the ratio $(\gamma_p + 2\mu)/\gamma_m$. Intuitively, like for Proposition 1, a higher activity fold change drives the mRNA concentration away from its quasi-steady-state value, while a shorter mRNA half-life curbs this effect. Notice that Proposition 1 only applies when initially the systems of extensive variables, given by Eq. 2.6 and Eq. 2.7, are at steady state. However, since the dependence of $P(t)$ and $\hat{P}(t)$ on the initial conditions quickly becomes negligible with time (Section A4), Proposition 1 can be used even when the initial steady-state conditions are not strictly satisfied.

The possibility to estimate the model reduction error *a posteriori* from fluorescence data allows one to assess the appropriateness of using one-step or two-step models for specific genes in a network model, given a predefined error bound. If no experimental data are available but the degradation constants of the proteins and mRNAs are (approximately) known, as well as upper bounds on changes in promoter activity, then the Propositions 1 and 2 can provide a (conservative) *a priori* estimation of the model reduction error.

The results of this paper suggest that, for a wide range of bacterial genes, one-step models can safely replace two-step models. This may be beneficial in practice, especially when dealing with networks having a large number of genes. The approximation significantly reduces the number of parameters to estimate and the time to simulate the system. Although replacing two-step models by one-step models introduces a negligible error locally, this can have a non-negligible effect on the global network dynamics. The extent to which the errors are amplified will depend on the network structure, the non-linearities, and the parameter values (Polynikis et al., 2009). We showed that, in the special case of a regulatory cascade without feedback, replacing the two-step model by a one-step model for a single upstream gene will amplify the model reduction error in a way that depends on the parameters describing the interaction between the two genes. In particular, we provided simple conditions under which the model of a regulator gene can be simplified without risk of error amplification in the genes it regulates. When applying the model reduction to both the regulator and the regulated genes, the effect on the model reductions is quasi-additive. Therefore, model reduction should be applied selectively, and not systematically to all genes of the network. In large gene networks, for a given experiment, some modules of genes may exhibit slow variations (Tournier and Chaves, 2009), and be therefore particularly fit for model reduction. Using the conditions provided in Sec. 2.3.6, one can first model a gene network using simple one-step expression models, and then use the parameters and dynamics estimated from this model in order to select which genes should be modelled with two steps. Note, however, that these conditions do not cover the cases of more complex network topologies like feedback loops, which are known to lead to an explosion of the approximation error through the network (Chen et al., 2010; Pedraza and van Oudenaarden,

2005).

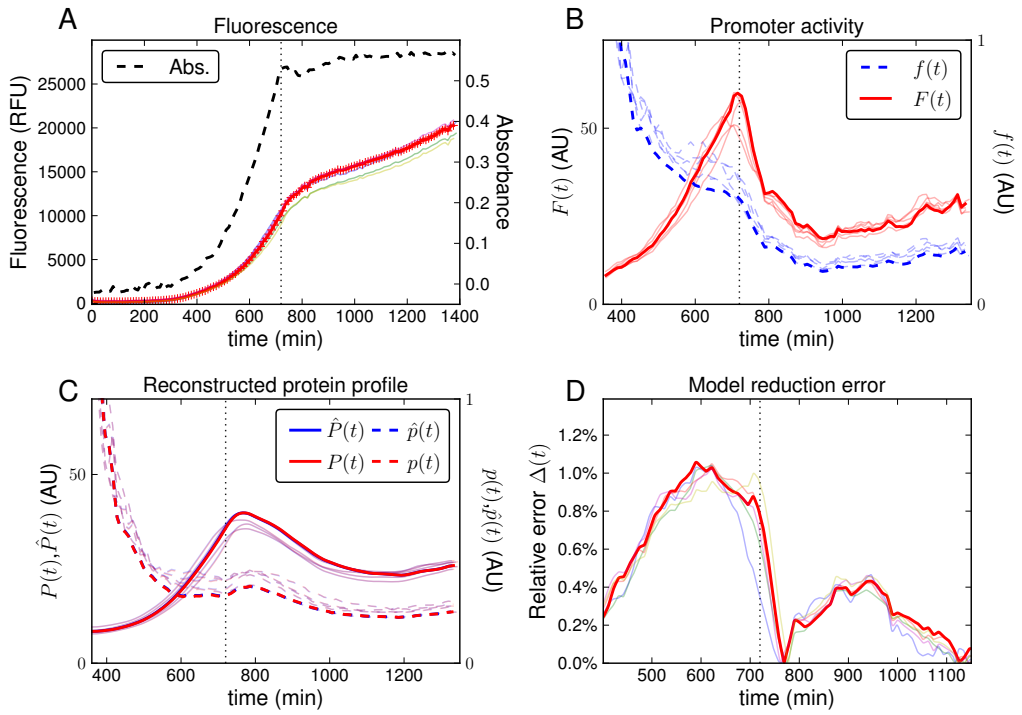


Figure 2.2: **Experimental estimation of the approximation error for the gene *crp*.** The profiles shown are computed from data obtained for five different wells on the microplate. The bold profile corresponds to one of these replicates (the same in the four panels). **A.** Observed fluorescence, after background substraction (cf. Section 2.2.2). We provide the absorbance from one well as a measure of the population volume. **B.** Promoter activities $F(t)$ and $f(t)$, computed from the data in A, according to Eq. 2.15 . **C.** Estimated profiles of the total amount of Crp in the cell population volume ($P(t)$) and the protein concentration ($p(t)$). The profiles are reconstructed from the promoter activities presented in B. Note that the reconstructed profiles for the different replicates overlap to the point of being almost indistinguishable. **D.** Relative error $\Delta(t)$ computed for the different replicates using the reconstructed Crp profiles from C.

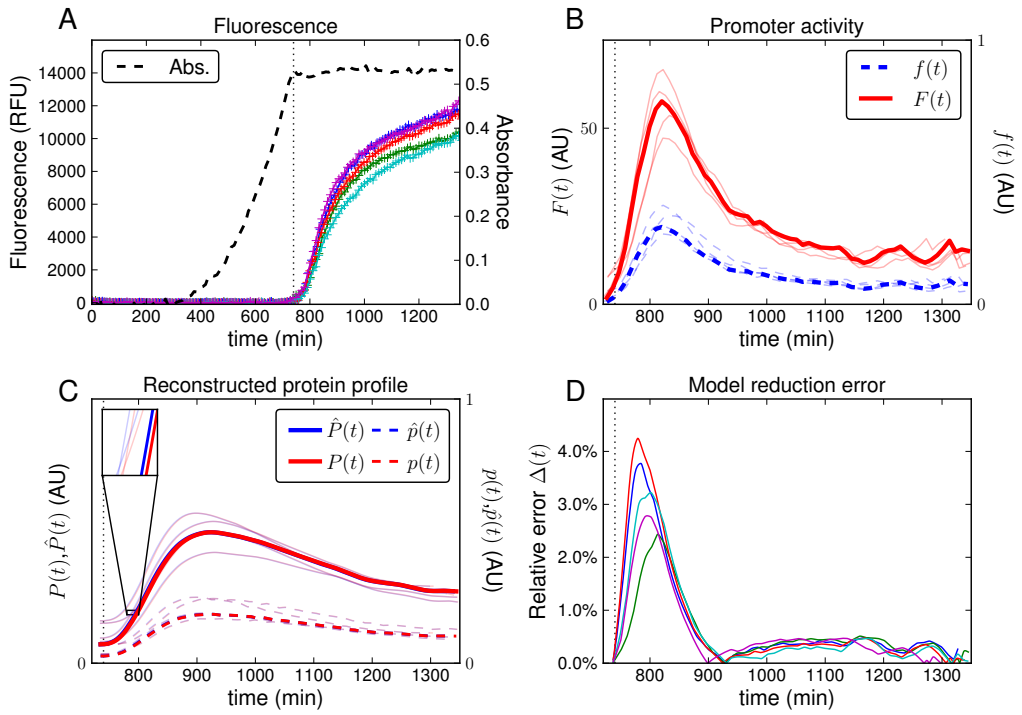


Figure 2.3: **Experimental estimation of the approximation error for the gene *acs*.** The profiles shown are computed from data obtained for five different wells on the microplate. The bold profile corresponds to one of these replicates (the same in the four panels). **A.** Observed fluorescence, after background substraction (cf. Section 2.2.2). We provide the absorbance from one well as a measure of the population volume. **B.** Promoter activities $f(t)$ and $F(t)$. The activities are computed from the data in A. **C.** Estimated profiles of the total amount of Acs in the cell population volume ($P(t)$) and the protein concentration ($p(t)$). The profiles are reconstructed from the promoter activities presented in B. Note that the reconstructed profiles for the different replicates overlap to the point of being almost indistinguishable. **D.** Relative error $\Delta(t)$ computed for the different replicates using the reconstructed Acs profiles from C.

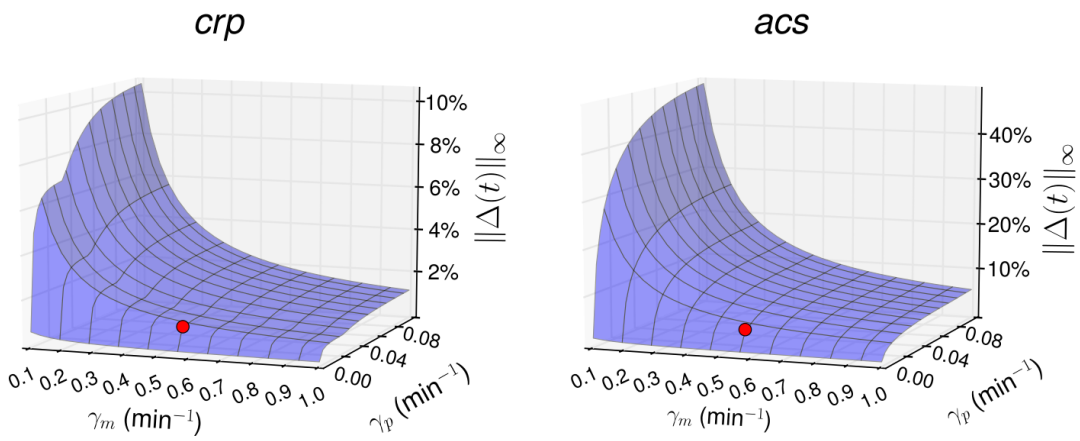


Figure 2.4: **Model reduction error as a function of degradation parameters γ_m and γ_p.** The two-step and one-step models were simulated as described in Section 2.3.4 using the estimated promoter activity of *crp* (Fig. 2.2) and *acs* (Fig. 2.3). The plots show the maximal value $\| \Delta \|_\infty$ of the model reduction error over the simulation interval, for different values of γ_p and γ_m. The red dots indicate the parameter values used for the computation of $\Delta(t)$ in Fig. 2.2 and Fig. 2.3.

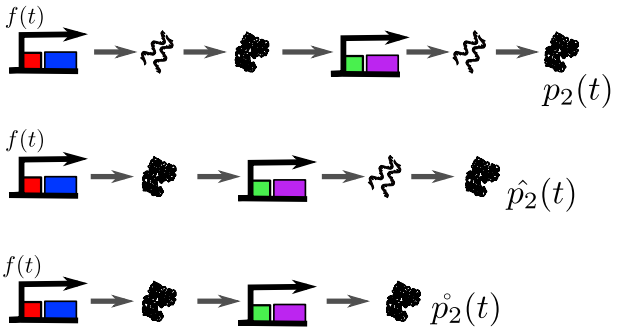


Figure 2.5: Systems of two genes with no model reduction, one model reduction applied to the first gene, and model reductions applied to both genes.

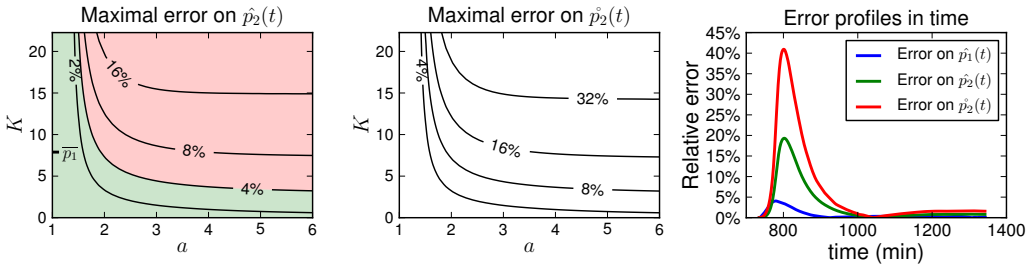


Figure 2.6: **Propagation of the model reduction error in a two-gene system.** A two-gene cascade was simulated using one or two model reductions, as explained in Sec. 2.3.6, the maximal difference over time to the unreduced model was plotted for different values of the parameters a and K. **A.** Maximal relative difference over time between $p_2(t)$ and $\hat{p}_2(t)$. The red region indicates the parameter values for which the model reduction error on $\hat{p}_1(t)$ (4%, see Fig. 2.3) results in an even larger error in $\hat{p}_2(t)$. The values of K were chosen of the same order of magnitude as the mean value of $p_1(t)$, denoted $\overline{p_1}$. **B.** Maximal relative difference over time between $p_2(t)$ and $\mathring{p}_2(t)$. **C.** Time profiles of the relative errors $\hat{p}_1(t)$, $\hat{p}_2(t)$ and $\mathring{p}_2(t)$ for the parameter values $a = 5, K = 20$.

Summary of Chapter 3

Time-series observations from reporter gene experiments are commonly used for inferring and analyzing dynamical models of regulatory networks. The robust estimation of promoter activities and protein concentrations from primary data is a difficult problem due to measurement noise and the indirect relation between the measurements and quantities of biological interest.

We propose a general approach based on regularized linear inversion to solve a range of estimation problems in the analysis of reporter gene data, notably the inference of growth rate, promoter activity, and protein concentration profiles. We evaluate the validity of the approach using in-silico simulation studies, and observe that the methods are more robust and less biased than indirect approaches usually encountered in the experimental literature based on smoothing and subsequent processing of the primary data. We apply the methods to the analysis of fluorescent reporter gene data acquired in kinetic experiments with Escherichia coli. The methods are capable of reliably reconstructing time-course profiles of growth rate, promoter activity, and protein concentration from weak and noisy signals at low population volumes. Moreover, they capture critical features of those profiles, notably rapid changes in gene expression during growth transitions.

The methods described in this paper are made available as a Python package (LGPL licence) and also accessible through a web interface.

Résumé du Chapitre 3

Reconstruction robuste de profiles d'expression génique à partir de données de gènes rapporteur, par inversion linéaire. Les modèles

dynamiques de réseaux de régulations géniques sont couramment calibrés et validés à l'aide de séries temporelles d'absorbance et de fluorescence obtenues lors d'expériences avec gènes rapporteurs.

Dans ce chapitre nous montrons comment certaines variables biologiques d'intérêt (taux de croissance, taux de synthèse protéique, concentration de protéine) peuvent être estimées à partir de ces données, par des méthodes d'inversions linéaires.

Nous testons les méthodes proposées sur des données simulées et montrons qu'elles permettent de correctement estimer les variations de ces signaux biologiques. En particulier, elles sont robustes au bruit de mesure, peu biaisées, et capturent bien les variations rapides lors des transitions de phase, tout en ne nécessitant quasiment aucun réglage de paramètre. Elles sont donc bien adaptées à une utilisation *de routine* pour le traitement de grands jeux de données provenant, par exemple, d'expériences en micro-plaques. Nous appliquons ces méthodes à l'étude de l'activité de différents gènes d'*E. coli* durant la réponse à une augmentation puis diminution de la quantité de nutriments dans le milieu de culture. Ces méthodes seront également mises en œuvre dans notre étude de la Répression Catabolique au Chapitre 4.

Chapter 3

Robust reconstruction of gene expression profiles from reporter gene data using linear inversion

3.1 Introduction

Over the past decade a variety of new experimental technologies have become available for measuring gene expression over time. They provide valuable information for the construction and validation of models of gene regulatory networks, involving tasks like parameter estimation, hypothesis testing, and model selection (Villaverde and Banga, 2014; Bansal et al., 2007; de Smet and Marchal, 2010). A critical step in the exploitation of the experimental data is the estimation of biologically relevant quantities, in particular promoter activities, mRNA concentrations or protein concentrations, from the primary data provided by the measurement instruments. This requires data analysis procedures that are unbiased and robust to measurement noise.

Fluorescent reporter genes have become widely used for monitoring gene expression in bacteria at high temporal resolution in a non-intrusive way (Chudakov et al., 2010; Giepmans et al., 2006). The underlying principle is the fusion of a natural gene of interest and/or the promoter region driving its expression with a gene encoding a fluorescent protein (Figure 3.1). A bacterial

strain carrying the resulting reporter gene, either on the chromosome or on a plasmid, emits a fluorescence signal proportional to the amount of reporter protein in the cell. When reporter strains are grown in a microplate, the fluorescence as well as the absorbance (optical density) of the culture can be automatically measured every few minutes, in a highly parallelized way. The resulting data contain information on population-level gene expression that is highly valuable for applications such as the inference and analysis of regulatory networks in bacterial cells (Berthoumieux et al., 2013; Gerosa et al., 2013; Keren et al., 2013; Ronen et al., 2002; Stefan et al., 2015).

The extraction of useful information from reporter gene data is not easy to achieve though, since it is often buried in noise, especially at low population volumes. Moreover, the fluorescence and absorbance measurements are only indirectly related to promoter activities and protein concentrations, requiring dynamical models of the expression of reporter genes for their interpretation. Several methods have been proposed to process the fluorescence and absorbance signals and estimate time-varying promoter activities and protein concentrations from the data (Aïchaoui et al., 2012; Bansal et al., 2012; de Jong et al., 2010; Finkenstädt et al., 2008; Leveau and Lindow, 2001; Lichten et al., 2014; Porreca et al., 2010; Ronen et al., 2002; Wang et al., 2008). The methods differ in the scope of the estimation problems considered, some being restricted to the inference of promoter activities and others also considering mRNA and protein concentrations. In addition, the approaches used to estimate these quantities from the primary data are quite different. Some methods are indirect, in the sense that they smoothen the data first and reconstruct the profiles of interest *via* the measurement model only in a second step. This results in a propagation of estimation errors that is difficult to control. Other methods state a regularized data fitting problem directly in terms of the quantities of interest, thus proceeding in a single and better controlled optimization step.

In this paper we propose a general, comprehensive approach towards the reconstruction of gene expression profiles from reporter gene data and solve the estimation problems it comprises in a mathematically sound and practical manner. We formulate the estimation problems in the classical framework of regularized linear inversion (Bertero, 1989; Wahba, 1990; de Nicolao et al.,

1997), which gives access to a range of powerful tools for robust estimation. Contrary to the related work of (Bansal et al., 2012) and (Porreca et al., 2010), we consider not only the inference of promoter activities, but also of growth rates and protein concentrations. Moreover, no restrictions are imposed that limit the practical applicability of the approach. We propose efficient procedures for the implementation of the methods and show by means of an *in-silico* simulation study under realistic conditions that they perform better than the indirect approaches usually encountered in the experimental literature. The algorithms have been implemented in a Python package and are also accessible through a web application.

Our linear inversion methods have been tested on fluorescent reporter gene data acquired in experiments with the model bacterium *Escherichia coli*. These experiments aim at quantifying the dynamics of gene expression during growth transitions induced by carbon upshifts and carbon depletion. We show that linear inversion succeeds in robustly reconstructing growth rate, promoter activity and protein concentration over the entire duration of the experiment, in particular in the beginning of the experiment when the population density and thus the signal-to-noise ratio are low. Moreover, we show that our methods reliably capture rapid changes in gene expression during the growth transitions, when promoter activities may change ten to hundred-fold within a dozen of minutes (Baptist et al., 2013; Enjalbert et al., 2013; Kao et al., 2005). Reconstructing these transient gene expression profiles from the data is highly important for increasing our understanding of the functioning of the underlying regulatory networks, but is particularly difficult to achieve.

To the best of our knowledge, the methods and computer tools presented in this paper provide the most comprehensive solution for analyzing reporter gene data available to date. Although the application has focused on fluorescent reporter gene data, the methods are directly applicable to the analysis of luminescent reporter gene data or other time-series gene expression data sets. In addition, the gene expression models underlying the methods are valid not only for bacteria but also for higher organisms.

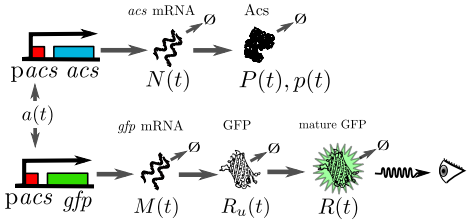


Figure 3.1: **Expression of the gene *acs* in *Escherichia coli* and the associated reporter gene p*acs-gfp*.** *acs* and *gfp* mRNA are transcribed, and translated into the proteins Acs and GFP (Green Fluorescent Protein), respectively. Both mRNA and protein are degraded. Moreover, GFP is transformed into a mature form in which it emits fluorescence when excited. Since *acs* and its reporter have the same promoter region, the transcriptional regulation of the two genes is identical. The variables are as defined in Equation 3.16 and Equation 3.17 .

3.2 Linear inversion methods

In this section we review properties of linear ordinary differential equations (ODEs) and linear relationships between different outputs driven by the same input. This theoretical framework enables us to estimate growth rate, promoter activity and reporter concentration using simple linear inversions in Section 3.3.

3.2.1 Inversion of a linear ODE system

We consider the following linear ODE model with input $u(t) \in \mathbb{R}$ and output $y(t) \in \mathbb{R}$:

$$\begin{cases} \dfrac{d}{dt}\mathbf{x}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)u(t), \\ y(t) = \mathbf{C}(t)\mathbf{x}(t), \\ \mathbf{x}(0) = \mathbf{x}_0. \end{cases} \tag{3.1}$$

In this system, $\mathbf{x}(t) \in \mathbb{R}^n$ is a vector of state variables, and $\mathbf{A}(t), \mathbf{B}(t), \mathbf{C}(t)$ are known time-varying matrices with dimensions $n \times n$, $n \times 1$, $1 \times n$, respectively. Given a set of noisy observations $(\tilde{y}(t_i))_{1 \leq i \leq N_y}$ of $y(t_i)$, we wish to estimate the

unknown input $u(t)$ and initial conditions $\mathbf{x}_0$. The solution of Equation 3.1 at time t with input u and initial conditions $\mathbf{x}_0$ can be formulated explicitly as:

$$y(t, u, \mathbf{x}_0) = \mathbf{C}(t) \left(\mathbf{\Phi}(t, 0)\, \mathbf{x}_0 + \int_0^t \mathbf{\Phi}(t, \tau)\, \mathbf{B}(\tau)\, u(\tau)\, d\tau \right), \tag{3.2}$$

where $\mathbf{\Phi}(t, \tau)$ is the state transition matrix (Chen, 1970). Notice that in this equation $y(t, u, \mathbf{x}_0)$ depends linearly on both the signal u and the initial conditions $\mathbf{x}_0$, making the estimation of these variables from $(\tilde{y}(t_i))_{1 \leq i \leq N_y}$ a *linear inversion problem* (Bertero, 1989; Wahba, 1990; de Nicolao et al., 1997).

Under the classical assumption of Gaussian i.i.d. measurement noise, the maximum likelihood solution of this problem can equivalently be written as

$$\text{Find } (\hat{u}, \hat{\mathbf{x}}_0) = \underset{(u, \mathbf{x}_0)}{\operatorname{argmin}} \operatorname{Err}(u, \mathbf{x}_0), \tag{3.3}$$

where

$$\operatorname{Err}(u, \mathbf{x}_0) = \sum_{i=1}^{N_y} \left(y(t_i, u, \mathbf{x}_0) - \tilde{y}(t_i) \right)^2 .$$

Without further assumptions on $u(t)$ this problem is ill-posed, *i.e.* there are infinitely many equivalent solutions $(\hat{u}(t), \hat{\mathbf{x}}_0)$, although these solutions may present biologically unrealistic values or variations. The problem must therefore be *regularized* by formulating new assumptions that lead to a unique, acceptable solution.

To this end, we discretize the time space of the input into N_u intervals of the form $[\tau_j, \tau_{j+1}[$, of equal length $\delta\tau$. We assume that on this grid of time intervals the input $u(t)$ is sufficiently well approximated by a piecewise constant input $(u(\tau_j))_{1 \leq j \leq N_u}$:

$$u(t) = \sum_{j=1}^{N_u} u(\tau_j) \mathbf{1}_{[\tau_j, \tau_{j+1}[}(t). \tag{3.4}$$

Because the output $y(t)$ depends linearly on u, the values $(y(t_i, u, \mathbf{x}_0))_{1 \leq i \leq N_y}$

depend linearly on $(u(\tau_j))_{1\leq j\leq N_u}$. If we define the following vectors:

$$\mathbf{u} = \begin{pmatrix} u(\tau_1) \\ u(\tau_2) \\ \vdots \\ u(\tau_{N_u}) \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y(t_1, u, \mathbf{x}_0) \\ y(t_2, u, \mathbf{x}_0) \\ \vdots \\ y(t_{N_y}, u, \mathbf{x}_0) \end{pmatrix}, \tilde{\mathbf{y}} = \begin{pmatrix} \tilde{y}(t_1) \\ \tilde{y}(t_2) \\ \vdots \\ \tilde{y}(t_{N_y}) \end{pmatrix},$$

and $\mathbf{w} = \begin{pmatrix} \mathbf{x}_0 & \mathbf{u} \end{pmatrix}^T$, then there exists an *observation matrix* $\mathbf{H_w}$ with dimension $(N_u + n) \times N_y$, such that

$$\mathbf{H_w}\,\mathbf{w} = \mathbf{y}. \tag{3.5}$$

The matrix $\mathbf{H_w}$ can be written as the juxtaposition of two matrices:

$$\mathbf{H_w} = \begin{pmatrix} \mathbf{H_{x_0}} & \mathbf{H_u} \end{pmatrix},$$

where $\mathbf{H_{x_0}}$ is a $N_y \times n$ matrix describing the influence of the initial conditions on $\mathbf{y}$, and $\mathbf{H_u}$ a $N_y \times N_u$ matrix describing the influence of $\mathbf{u}$ on $\mathbf{y}$. The computation of $\mathbf{H_w}$ can be generally performed using a numerical ODE solver, as explained in Section B3. However, for the cases of interest described in Section 3.3, we provide more effective formulas for the computation of $\mathbf{H_w}$.

Our inversion problem now writes as a multivariate linear regression problem:

$$\text{Find } \begin{pmatrix} \hat{\mathbf{x}}_0 \\ \hat{\mathbf{u}} \end{pmatrix} = \hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \ \|\mathbf{H_w}\mathbf{w} - \tilde{\mathbf{y}}\|_2^2. \tag{3.6}$$

This problem may also be ill-posed, in particular when $N_u > N_y$. *Tikhonov regularization* on the first derivative consists in introducing a penalty on the successive variations of u, whose importance is modulated by a *regularization parameter* $\lambda \geq 0$:

$$\text{Find } \hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \ \|\mathbf{H_w}\mathbf{w} - \tilde{\mathbf{y}}\|_2^2 + \lambda \sum_{j=1}^{N_u - 1} \left(\mathbf{u}_{j+1} - \mathbf{u}_j\right)^2, \tag{3.7}$$

where $\mathbf{u}_j$ and $\mathbf{u}_{j+1}$ denote the jth and $(j+1)$th element of $\mathbf{u}$, respectively. Practically, this penalty is implemented by introducing a new $(N_u+n)\times 1$-vector $\mathbf{v}=\mathbf{L_w}\mathbf{w}$ and a new $(N_u+n)\times(N_u+n)$-matrix $\mathbf{H_v}=\mathbf{H_w}\mathbf{L_w}^{-1}$, where $\mathbf{L_w}$ is a matrix of the form

$$\mathbf{L_w}=\begin{pmatrix}\omega\mathbf{I}_n & \mathbf{0}\\ \mathbf{0} & \mathbf{L_u}\end{pmatrix}.$$

In the formulation above, $\mathbf{I}_n$ is the $n\times n$ identity matrix, and $\omega\in\mathbb{R}$ a small but non-zero number ensuring that the values of $\mathbf{x}_0$ contribute negligibly to the penalty term while keeping $\mathbf{L_w}$ invertible. $\mathbf{L_u}$ is the $N_u\times N_u$ discrete differentiation matrix

$$\mathbf{L_u}=\begin{pmatrix}\epsilon & & & \mathbf{0}\\ -1 & 1 & & \\ & \ddots & \ddots & \\ \mathbf{0} & & -1 & 1\end{pmatrix}.$$

In (Bansal et al., 2012) the parameter ϵ is chosen equal to 1, but this results in a biased estimation of $\mathbf{u}_0$ as ϵ represents a penalty on this parameter. Section B1 discusses how to find an appropriate value for ϵ, typically $0<\epsilon\ll 1$.

The inversion problem of Equation 3.7 can be reformulated in matrix form as

$$\text{Find}\quad\begin{pmatrix}\hat{\mathbf{x}}_0\\ \hat{\mathbf{u}}\end{pmatrix}=\mathbf{L_w}^{-1}\hat{\mathbf{v}},\quad\text{where} \tag{3.8}$$

$$\hat{\mathbf{v}}=\underset{\mathbf{v}}{\operatorname{argmin}}\ \|\mathbf{H_v}\mathbf{v}-\tilde{\mathbf{y}}\|_2^2+\lambda\|\mathbf{v}\|_2^2. \tag{3.9}$$

For λ large enough, this problem admits a unique solution (Hoerl and Kennard, 1970):

$$\begin{pmatrix}\hat{\mathbf{x}}_0\\ \hat{\mathbf{u}}\end{pmatrix}=\hat{\mathbf{w}}=\mathbf{L_w}^{-1}(\mathbf{H_v}^T\mathbf{H_v}+\lambda\mathbf{I})^{-1}\mathbf{H_v}^T\tilde{\mathbf{y}}. \tag{3.10}$$

The regularization parameter λ can be set arbitrarily. However, λ too large will lead to over-smoothed estimates of $u(t)$, whereas λ too small will lead

to under-smoothed (unstable) estimates of $u(t)$. Many techniques have been proposed to automatically select a proper λ to regularize a given problem. In this article the choice of λ will always be based on *generalized cross-validation (GCV)* (Golub et al., 1979), a fast procedure which aims at maximizing the predictive power of the resulting estimate of u. It is also straightforward to deal with additional linear constraints in the problem of Eqs 3.8-3.9, for instance to ensure that the estimated input $\hat{\mathbf{u}}$ is always positive (see Section B2 for technical details).

3.2.2 Linear inversion involving ODE systems with identical input

We now consider two linear ODE systems, defined as in Equation 3.1, sharing the same input $u(t)$, but having different variables $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$, possibly different parameters, different initial conditions $\mathbf{x}_{0,1}$ and $\mathbf{x}_{0,2}$, and different outputs y_1 and y_2. The goal is to estimate the profile y_1 from observations of y_2. This case will be found useful in Section 3.3.4 for computing protein concentrations. We have seen in the previous section that there is a linear relationship between u and y_1 on one hand, and between u and y_2 on the other hand. This gives rise to linear inversion problems defined by the observation matrices $\mathbf{H}_{\mathbf{w}_1}$ and $\mathbf{H}_{\mathbf{w}_2}$, respectively (Figure 3.2):

$$\mathbf{H}_{\mathbf{w}_1}\begin{pmatrix}\mathbf{x}_{0,1}\\ \mathbf{u}\end{pmatrix} = \mathbf{y}_1 \quad \text{and} \quad \mathbf{H}_{\mathbf{w}_2}\begin{pmatrix}\mathbf{x}_{0,2}\\ \mathbf{u}\end{pmatrix} = \mathbf{y}_2.$$

When $\mathbf{H}_{\mathbf{w}_1}$ is invertible (which can be enforced in many cases since the times τ_i for the discretization of $\mathbf{u}$ can be chosen arbitrarily), it is possible to relate $\mathbf{y}_1$ to $\mathbf{y}_2$ through a chain of linear transformations:

$$\begin{pmatrix}\mathbf{x}_{0,2}\\ \mathbf{y}_1\end{pmatrix} \xrightarrow{\mathbf{H}_1} \begin{pmatrix}\mathbf{x}_{0,2}\\ \mathbf{x}_{0,1}\\ \mathbf{u}\end{pmatrix} \xrightarrow{\mathbf{H}_2} \begin{pmatrix}\mathbf{x}_{0,2}\\ \mathbf{u}\end{pmatrix} \xrightarrow{\mathbf{H}_{\mathbf{w}_2}} \mathbf{y}_2,$$

where matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are defined as follows:

$$\mathbf{H}_1 = \begin{pmatrix} \mathbf{I}_{n_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{\mathbf{w}_1}^{-1} \end{pmatrix}, \quad \mathbf{H}_2 = \begin{pmatrix} \mathbf{I}_{n_2} & \mathbf{0}_{n_2 \times n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{N_\mathbf{u} \times n_1} & \mathbf{I}_{N_\mathbf{u}} \end{pmatrix},$$

and n_1, n_2 are the lengths of vectors $\mathbf{x}_1(t), \mathbf{x}_2(t)$.

By lumping this chain into a single transformation matrix $\mathbf{H}_\mathbf{y} = \mathbf{H}_{\mathbf{w}_2}\mathbf{H}_2\mathbf{H}_1$ we obtain

$$\mathbf{H}_\mathbf{y} \begin{pmatrix} \mathbf{x}_{0,2} \\ \mathbf{y}_1 \end{pmatrix} = \mathbf{y}_2,$$

and $\mathbf{y}_1$ can be estimated from observations of $\mathbf{y}_2$ using Tikhonov regularization with generalized cross-validation, as explained in the previous section.

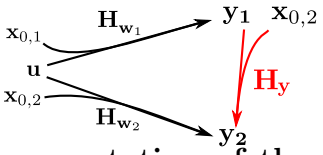


Figure 3.2: **Schematic representation of the linear relationships between variables $\mathbf{u}$, $\mathbf{x}_{0,1}$, $\mathbf{x}_{0,2}$, $\mathbf{y}_1$ and $\mathbf{y}_2$.** Arrows indicate the linear relationships derived in Section 3.2.2.

3.3 Estimation of gene expression profiles from fluorescent reporter gene data

In this section, we will show how recurring problems in the analysis of reporter gene data, the estimation of growth rate, promoter activity, and protein concentration, can be mapped to the linear inversion problems formulated in the previous section. We apply the resulting methods to the analysis of fluorescent and absorbance signals measured in population-level experiments in *E. coli*, in conditions involving strong time-varying changes in growth rate and gene expression.

3.3.1 Fluorescent reporter gene experiments in *E. coli*

Changes in the environment trigger responses on different levels in bacterial cells, typically affecting intracellular metabolite pools within seconds and, on a longer time-scale, protein concentrations and physical parameters like cell size. The regulatory networks controlling these adaptations are complex and only partially understood.

In this paper we consider four genes playing a key role in the adaptation of *E. coli* to perturbations due to the sudden availability or depletion of carbon sources in the medium. These genes are *fis*, encoding a global regulator responsible in particular for activating ribosomal RNA transcription (Bradley et al., 2007); *gyrA*, coding for DNA gyrase which negatively supercoils DNA (Travers and Muskhelishvili, 2005); *crp*, whose product regulates the transcription of hundred of genes when activated by the secondary messenger cyclic AMP (Gosset et al., 2004a); and *acs*, encoding an enzyme required for acetate consumption (Wolfe, 2005). We used reporter strains obtained by transforming the *E. coli* wild-type strain with reporter plasmids carrying a transcriptional fusion of the promoter region of the aboves genes with a *gfp* reporter gene. The reporter gene for *acs* codes for GFPmut2, a reporter with a long half-live (19 h), whereas the other reporter genes code for GFPmut3, with a short half-live of 1 h (see Section B4 for details on the plasmids and strains used in this study).

Overnight stationary-phase cultures of the reporter strains were diluted into the wells of a microplate containing minimal medium with glucose. The bacteria were observed in a microplate reader up until a few hours after glucose exhaustion (see Section B4 for details on the experimental conditions). The carbon upshift provokes a strong activation of the expression of many genes, while growth arrest following glucose exhaustion triggers the activation of so-called catabolite genes, responsible for the assimilation of secondary carbon sources, such as acetate secreted during rapid growth on glucose (Baptist et al., 2013; Enjalbert et al., 2013; Kao et al., 2005). The absorbance (600 nm) and fluorescence (485/520 nm) of the growing bacterial cultures was measured for each of the 96 wells, typically one measurement per minute per well.

The absorbance (optical density) measurements are usually assumed pro-

portional to the volume $V(t) \in \mathbb{R}_+$ of the growing cell population. More precisely, for measurements made at time-points t_i, we have the following measurement model:

$$\tilde{V}(t_i) = \alpha V(t_i) + \nu_i, \tag{3.11}$$

where $\tilde{V}(t_i)$ represents the absorbance measurement at t_i, $\alpha \in \mathbb{R}_+$ an unknown proportionality coefficient, and ν_i measurement noise. Similarly, the fluorescence measurements, after background correction (de Jong et al., 2010; Lichten et al., 2014), can be assumed proportional to the total quantity of active (mature) fluorescent protein $R(t)$ in the growing cell population:

$$\tilde{R}(t_i) = \beta R(t_i) + \nu'_i, \tag{3.12}$$

where $\tilde{R}(t_i)$ represents the fluorescence measurement at t_i, $\beta \in \mathbb{R}_+$ an unknown proportionality coefficient, and ν'_i measurement noise.

The absorbance and fluorescent measurements that will be used in the remainder of this paper are shown in the top row of Figure 3.3. The data illustrate some of the difficulties encountered in the analysis, namely weak signals in the beginning of the experiment, when the volume of the cell population is low, and rapid changes during growth transitions.

3.3.2 Estimation of growth rate

The exhaustion of glucose in the medium around 500 min is followed by immediate growth arrest, causing a break in the absorbance curves (Figure 3.3). In order to sharply distinguish the growth phases, it is important to precisely estimate the *growth rate* of the population, defined by

$$\mu(t) = \frac{1}{V(t)} \frac{d}{dt} V(t). \tag{3.13}$$

It is possible to compute $\mu(t)$ from the absorbance measurements $\tilde{V}(t_i)$ of the volume $V(t)$, by smoothing interpolation, subsequent differentiation, and numerical resolution of Equation 3.13. However, this method is unstable when the signal-to-noise ratio is low, especially in the early stages of the experiment.

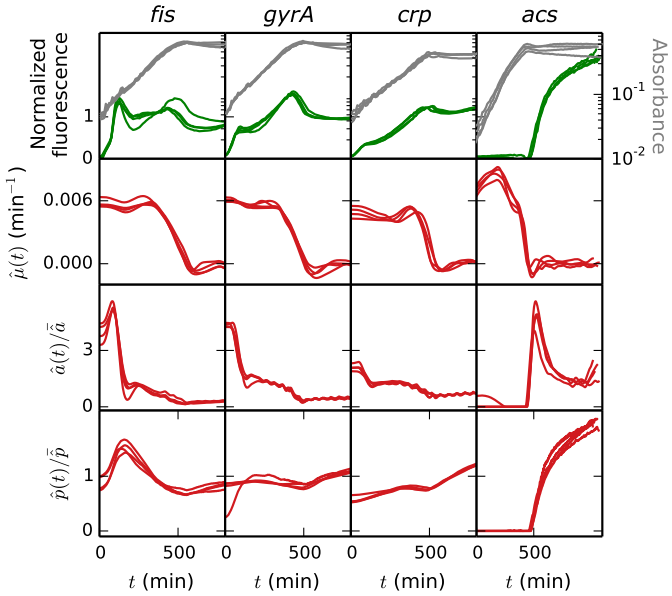


Figure 3.3: **Fluorescence and absorbance data obtained from reporter gene experiments in *E. coli* and estimations of growth rate, promoter activity, and protein concentration from these data.** The measured fluorescence and absorbance signals are shown in the top row. The estimations of growth rate, promoter activity, and protein concentration are denoted by $\hat{\mu}(t)$, $\hat{a}(t)$, and $\hat{p}(t)$, respectively. The fluorescence signal, $\hat{a}(t)$, and $\hat{p}(t)$ have been divided by their mean as they have different orders of magnitude for genes *fis*, *gyrA*, *crp*, and *acs*. For each signal, four replicates are shown, corresponding to different wells of the microplate.

As an alternative approach, we formulate the problem as a linear inversion problem. We first rewrite Equation 3.13 as

$$\frac{d}{dt}(\alpha\, V)(t) = \alpha\, V(t)\, \mu(t) \simeq \tilde{V}(t)\, \mu(t), \tag{3.14}$$

where $\tilde{V}(t)$ is an interpolated version of the measurements $\tilde{V}(t_i)$. Replacing the volume by the experimentally measured absorbance signal has the advantage of bringing the equation into the form of Equation 3.1, with

$$u(t) = \mu(t), \;\; \mathbf{x}(t) = \alpha V(t), \;\; y(t) = \alpha V(t)$$
$$\mathbf{A}(t) = 0, \quad \mathbf{B}(t) = \tilde{V}(t), \;\; \mathbf{C}(t) = 1.$$

That is, the growth rate is the input and the volume the output of a linear system, so that the growth rate can be estimated by linear inversion from the absorbance measurements.

The observation matrix $\mathbf{H_w}$ for this system can be computed as explained in Section B3.1. Solving the problem of Eqs 3.8-3.9 by regularization, we obtain the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}_0$ of the growth rate and the initial volume, repectively. The growth rate estimations $\hat{\mu}(t)$ are shown in the second row of Figure 3.3. As can be seen, upon glucose exhaustion the growth rate steeply drops from its maximum value to 0 within approximately one hour.

As our method relies on penalizing successive variations of $\mu(t)$, the question arises whether this entails a strong bias. In particular, how well estimated are the timing of the transition between the two growth phases, and the values of the growth rate during each phase? We tested the method on simulated data similar to the measurements in Figure 3.3, notably with equivalent sampling densities and signal-to-noise ratios. The results are presented in Section B5. They show that our estimation method is able to recover different growth-rate profiles, with very small bias and moderate variance.

For comparison we also estimated the growth rate with the indirect method described after Equation 3.13. In particular, we smoothed the absorbance measurements $\tilde{V}(t_i)$ by means of smoothing splines in order to estimate the volume and its derivative, and computed an estimate of $\mu(t)$ by means of Equation 3.13. For simulated data shown in Figure 3.4A, the results in panel B show that at the beginning of the experiment, when the absorbance signal is low, the growth rate estimation is highly unstable. Additional numerical experiments, shown in Section B5, indicate that increasing the smoothing parameter to reduce the variance of the estimates introduces a strong bias on the estimate, reflecting the well-known variance-bias trade-off. We conclude that the proposed linear inversion method performs better than this indirect approach.

Notice that the estimation of growth rate and initial volume also leads to a denoised estimation of the population volume by Equation 3.5:

$$\widehat{\boldsymbol{\alpha}\,\mathbf{V}} = \mathbf{H}_{\mathbf{w}} \begin{pmatrix} \widehat{\boldsymbol{\alpha}\,\mathbf{V}}_0 & \hat{\boldsymbol{\mu}} \end{pmatrix}, \tag{3.15}$$

which will be used in the next sections for the estimation of promoter activity and protein concentration.

3.3.3 Estimation of promoter activity

The interest of the use of reporter genes is that, by construction, they provide information on the expression of a gene of interest. We will focus on transcriptional fusions here, where the reporter gene and the gene of interest share the same promoter region and their promoter activities can be considered identical, possibly up to a multiplicative constant (Figure 3.1).

The relation between promoter activity and observed fluorescence and absorbance signals is indirect and models of the gene expression process are needed to interpret the primary data. Several models have been proposed in the literature (see the references in the introduction), but here we follow with some modifications the model used in (de Jong et al., 2010). The expression of a fluorescent reporter gene is modelled as a three-step process involving transcription, translation, and maturation of the fluorescent protein (Figure 3.1). The variables of the model are $M(t), R_u(t), R(t) \in \mathbb{R}_+$, denoting the total quantity of *gfp* mRNA in a growing cell population and the total quantity of immature and mature GFP, respectively (in mmol). In comparison with most other models, we consider total quantities of molecules and not concentrations. This has the advantage of simplifying the estimation of promoter activities, since it omits terms due to growth dilution from the equations.

The rate of transcription drives the dynamics of gene expression and is defined as $k_m\, a(t)\, V(t)$, representing the total amount of mRNA produced per time unit in the growing cell population (usually expressed in mmol $\cdot$ mL^{-1} $\cdot$ min^{-1}). We will call $a(t) \in \mathbb{R}_+$, which is a dimensionless quantity scaled between 0 and 1, the *promoter activity*, whereas $k_m \in \mathbb{R}_+$ represents the maximum transcription rate. With the constants $d_M, k_U, d_R, k_R \in \mathbb{R}_+$

(min^{-1}), characterizing the degradation, translation, and maturation steps, we obtain the following ODE system:

$$\begin{cases} \dfrac{d}{dt}M(t) = k_M\, a(t)\, V(t) - d_M\, M(t), \\ \dfrac{d}{dt}R_u(t) = k_U\, M(t) - (d_R + k_R)\, R_u(t), \\ \dfrac{d}{dt}R(t) = k_R\, R_u(t) - d_R\, R(t). \end{cases} \tag{3.16}$$

Notice that the transcription rate is modulated by the volume of the growing cell population, which can be replaced by its estimate from Equation 3.15. As a consequence, the first equation of the model writes

$$\frac{d}{dt}M(t) = k'_M\, a(t)\, \widehat{\alpha V}(t) - d_M\, M(t),$$

where $k'_M = k_M/\alpha$. The resulting gene expression model can be easily brought into the form of Equation 3.1:

$$\mathbf{x}(t) = \begin{pmatrix} M(t) \\ R_u(t) \\ R(t) \end{pmatrix}, \quad \mathbf{A}(t) = \begin{pmatrix} -d_M & 0 & 0 \\ k_U & -(d_R + k_R) & 0 \\ 0 & k_R & -d_R \end{pmatrix},$$
$$\mathbf{B}(t) = \begin{pmatrix} k'_M\, \widehat{\alpha V}(t) \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{C}(t) = \begin{pmatrix} 0 \\ 0 \\ \beta \end{pmatrix}, \quad u(t) = a(t),$$
$$y(t) = \beta R(t).$$

This allows the promoter activity $a(t)$ as well as the initial conditions M_0, R_{u_0}, R_0 to be estimated from the measured fluorescence signal $\tilde{R}(t_i)$. Whereas the degradation constants d_R, d_M and the maturation constant k_R are usually available, the other parameters are generally not known precisely. However, we prove in Section B6 that setting β, k_U, k'_M to 1 still allows the time-varying profile of $a(t)$ to be estimated, up to some unknown multiplicative coefficient (as usual in the literature).

Section B3.2 provides an efficient procedure for computing the observation matrix $\mathbf{H_w}$ for the above system. The efficiency and accuracy of the computation of the observation matrix can be further increased when *gfp* mRNA is unstable and the maturation time is fast (*i.e.* for k_R and d_M large compared to d_R), which is the case for the reporter genes used in this study (de Jong et al., 2010). This makes it possible to lump the gene expression model of Equation 3.16 into a single step and thus simplifiy the regularized regression problem.

The linear inversion method for computing the promoter activity was applied to the reporter gene data in Figure 3.3, resulting in the estimates shown in the third row. We notice a sharp peak in the promoter activity of *fis* right after the nutrient upshift, which is consistent with previous reports (Azam et al., 1999), and the same behavior is observed for *gyrA*. Whereas the activity of the *crp* promoter shows little variation, consistent with the observation that the Crp concentration does not change much across growth phases (Kuhlman et al., 2007), upon glucose exhaustion the activity of *acs* shows a dramatic increase, in large part due to sudden accumulation of cyclic AMP in the cell (Berthoumieux et al., 2013; Wolfe, 2005). These examples illustrate that the method correctly infers known fast changes in gene expression from the data, while avoiding overfitting outside the transition region.

Like in Section 3.3.2, we used *in-silico* benchmarks resembling the actual data to further evaluate the ability of the method to reconstruct promoter activities, in particular the timing of the peak and its amplitude. The results in Figure 3.4C and Section B5 show that the method is stable even when the signal-to-noise ratio is low, and manages to capture the rapid variations in promoter activity with high precision. Like in Section 3.3.2, we remark that the method is robust, but nevertheless introduces some bias in extreme cases. However, this bias is much smaller than that obtained by an indirect method analogous in spirit to that outlined in the previous section (Figure 3.4C).

The estimation of the promoter activity is interesting in its own right, as it gives insight into changes in the transcriptional activity of specific genes during growth transitions. However, it can also be the first step towards the estimation of the concentration of the regulators of a gene (Bansal et al., 2012; Finkenstädt et al., 2008) or of the concentration of the protein encoded by the

gene of interest (de Jong et al., 2010). In the next section, we will develop a direct linear inversion method for addressing the latter problem.

3.3.4 Estimation of protein concentration

The expression of a gene of interest involves the same steps as the expression of the reporter gene, without the maturation step (Figure 3.1). As explained in the previous section, in the case of transcriptional fusions the promoter activities $a(t)$ are the same for the reporter gene and the gene of interest. However the other parameters describing mRNA and protein synthesis and degradation may be different.

In order to model the expression of the gene of interest, we introduce new variables describing the total amount of mRNA and protein for the gene of interest, denoted by $N(t)$ and $P(t)$ (mmol), respectively, and new kinetic constants k_N, d_N, k_P, d_P (min^{-1}). The concentration of the protein of interest is given by $p(t) = P(t)/V(t)$. This results in the following ODE system:

$$\begin{cases} \dfrac{d}{dt}N(t) = k_N\, V(t)\, a(t) - d_N\, N(t), \\ \dfrac{d}{dt}P(t) = k_P\, N(t) - d_P\, P(t), \\ p(t) = P(t)/V(t). \end{cases} \tag{3.17}$$

Introducing the variables $N'(t) = \alpha N(t)$ and $P'(t) = \alpha P(t)$, the system of Equation 3.17 becomes

$$\begin{cases} \dfrac{d}{dt}N'(t) = k_N\, \alpha\, V(t)\, a(t) - d_N\, N'(t), \\ \dfrac{d}{dt}P'(t) = k_P\, N'(t) - d_P\, P'(t) \\ p(t) = P'(t)/\left(\alpha\, V(t)\right). \end{cases} \tag{3.18}$$

Like in the previous section, $\alpha V(t)$ can be replaced by the experimentally measured absorbance signal $\widehat{\alpha V}(t)$, yielding:

$$\mathbf{x}(t) = \begin{pmatrix} N'(t) \\ P'(t) \end{pmatrix}, \quad \mathbf{A}(t) = \begin{pmatrix} -d_N & 0 \\ k_P & -d_P \end{pmatrix}, \quad u(t) = a(t),$$

$$\mathbf{B}(t) = \begin{pmatrix} k'_N \widehat{\alpha V}(t) \\ 0 \end{pmatrix}, \quad \mathbf{C}(t) = \begin{pmatrix} 0 \\ 0 \\ 1/\widehat{\alpha V}(t) \end{pmatrix}, \quad y(t) = p(t).$$

For the estimation of the protein concentration $p(t)$, the scheme outlined in Figure 3.2 applies, with $u(t) = a(t)$, $y_1(t) = p(t)$, and $\tilde{y}_2(t) = \tilde{R}(t)$. This allows an estimate of $p(t)$ to be obtained from the experimental measurement of $R(t)$ as explained in Section 3.2.2. When the gene expression model in Equation 3.18 can be reduced to a single step, the observation matrix of the problem can be computed in an efficient way as explained in Section B3.4.

The protein concentrations estimated from the *E. coli* reporter gene data by means of the above method are shown in the bottom row of Figure 3.3. The degradation constant of Fis was measured ($d_P = 0.0065$ min^{-1}; (de Jong et al., 2010)), whereas the other proteins were assumed to be long-lived ($d_P = 0.001$ min^{-1}), like most proteins in *E. coli* (Larrabee et al., 1980b). We observe that the Fis concentration transiently increases after the nutrient upshift, which is consistent with the role of Fis in activating the synthesis of stable RNAs necessary for growth (Dennis et al., 2004). The concentration of Crp is stable during growth on glucose and somewhat increases after glucose exhaustion, as expected from the fact that Crp activates catabolic genes needed for growth on poor carbon sources (Gosset et al., 2004b). Interestingly, this accumulation cannot be simply inferred by looking at the fluorescence data, which show no increase after glucose exhaustion. It illustrates the importance of taking into account different half-lives for reporter proteins and proteins of interest.

We also tested this method on the simulated data. The results are reported in Figure 3.4D and show that linear inversion is more stable and introduces less bias than other approaches, notably indirect approaches based on the estimation of $a(t)$ and numerical integration of Equation 3.18 using this estimate. Another advantage of the linear inversion method is that it does not need an estimate of the initial conditions, which are often unknown. In conclusion, our

direct method allows rich information on gene expression to be inferred from the absorbance and fluorescence data under reasonable assumptions.

3.4 Software for applying linear inversion methods

As they rely on few assumptions and require virtually no hand-tuning, the linear inversion methods developed in this paper are suitable for routine treatment of reporter data gene obtained in microplate experiments, which generate a huge quantity of measurements (typically $10^4 - 10^5$ data points). The linear inversion methods were implemented in the Python package WellFARE, relying on the scientific Python libraries NumPy and Scipy (Jones et al., 2001). In addition, the package provides utilities for parsing data files and removing possible outliers from the absorbance and fluorescence signals. The WellFARE package is available under an LGPL license, but has also been integrated into a web application called WellInverter, which provides a graphical user interface allowing access to the linear inversion methods through a web browser (Figure 3.5). The user can upload data files by means of WellInverter, remove outliers and subtract background, and launch the procedures for computing growth rates, promoter activities, and protein concentrations (Section B7).

3.5 Discussion

The inference of meaningful gene expression profiles from indirect experimental data is a key step in the analysis of dynamical models in systems biology. As reporter genes tend to become ubiquitous, it is important to develop reliable methods for the automated treatment of the large amounts of data becoming available. We have shown that the estimation of growth rate, promoter activity, and protein concentration from reporter gene data can be expressed as linear inversion problems using ODE-based measurement models and we proposed efficient procedures to compute the observation matrices solving these problems. The methods thus obtained were used to study the expression dynamics of several genes of *E. coli* during growth transitions, where they

confirmed their ability to handle critical issues in reporter gene data analysis: low signal-to-noise ratios and rapid changes in gene expression in response to environmental perturbations. The validity of these estimation procedures was reinforced by tests on simulated data, which showed that the methods are robust and little biased.

Several methods for the analysis of reporter gene data have been proposed in the literature, all of which are implicitly or explicitly based on the same or very similar measurement models for interpreting the data. The major differences between the approaches lie in the information they extract from the data and in the way the profiles are computed from the primary data. The basic idea underlying the linear inversion methods presented here is that they are direct, in the sense that they perform regularization on the quantity to be estimated, rather than by plugging empirically smoothed versions of the data into the measurement models (de Jong et al., 2010; Ronen et al., 2002). Our results show that this improves the robustness of the estimation process. In comparison with (Bansal et al., 2012), we extend the linear inversion methods to growth rates and protein concentrations, thus more fully exploiting the information contained in reporter gene data. Moreover, we improve the practical applicability of the approach in that we do not need to make assumptions that are often not realistic, such as zero initial promoter activities, constant growth rate, and direct measurements of reporter concentrations. The linear inversion methods remain tractable when improving estimation through the addition of linear constraints (*e.g.*, to ensure positive promoter activities and protein concentrations), the consideration of uncertainty on the data, or the use of different regularizations (L_1 regularization or regularization on the second derivative) (de Nicolao et al., 1997).

The methods described in this paper are made available as a Python package and can also be accessed through a user-friendly web application. Other tools for the analysis of reporter gene data are WellReader (Boyer et al., 2010) and BasyLICA (Aïchaoui et al., 2012). While the Matlab program WellReader uses the indirect approaches from (de Jong et al., 2010), BasyLICA is based on the use of Kalman filters, which also directly estimate quantities of interest from the reporter gene data by a Bayesian approach. In comparison with BasyLICA, WellInverter estimates not only promoter activities but also protein

concentrations from the data. In addition, WellInverter uses regularization based on generalized cross-validation to avoid hand-tuning.

The generality of the techniques used in this paper suggests that they could be applied to a much wider range of problems. A necessary condition for the application of the linear inversion methods is that the measured data is linearly related to the biological quantity of interest. Notice that this does not exclude time-varying parameters in Equation 3.16 or Equation 3.17, for instance a time-varying degradation constant of the protein, due to a change in half-live after a growth transition (Hengge-Aronis, 2002). As long as the time-varying parameters are known, for example when their profile has been measured, the inversion problem remains linear. To some extent, this even allows nonlinear estimation problems to be handled in our framework, as illustrated by the growth-rate estimation in Section 3.3.2.

The methods proposed in this paper provide the most general and comprehensive treatment of the reconstruction of gene expression profiles form reporter gene data available today, based on a solid mathematical foundation and supported by user-friendly computer tools. The approach directly carry over to luminescent reporter genes and may also apply to time-series data obtained by completely different experimental technologies, like DNA microarrays, RNA-Seq or quantitative proteomics. While we validated and illustrated the methods by means of reporter gene data from bacterial kinetics, the measurement models of Equation 3.16 and Equation 3.17 are sufficiently general to apply to higher organisms as well.

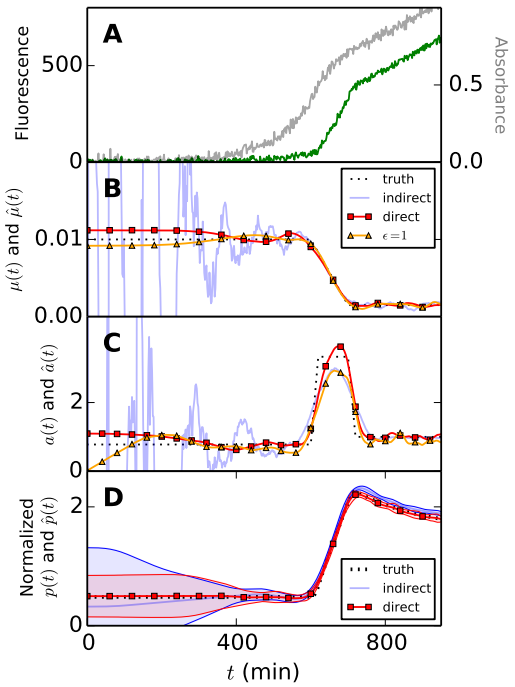


Figure 3.4: **Comparison of different methods for the estimation of growth rate and promoter activity from reporter gene data.** In particular, we compare indirect approaches based on plugging empirically smoothed versions of the data into measurement models with the direct linear inversion methods proposed here (including a variant in which ϵ is set to 1). Additional examples can be found in Section B5. **A.** Simulated noisy absorbance and fluorescence data. **B.** Estimates of the growth rate $\mu(t)$ obtained with the different methods. **C.** Estimates of promoter activity $a(t)$. **D.** Estimates of protein concentration $p(t)$, using the direct method developed in this paper, and an indirect method consisting in the estimation of $a(t)$, followed by numerical solution of Equation 3.17. Solid lines and shaded regions represent the mean $\pm$ one standard deviation over 100 simulations. The direct methods perform better than the indirect methods in that they yield estimates with less bias and lower variance. The use of $\epsilon = 1$ may introduce a bias.

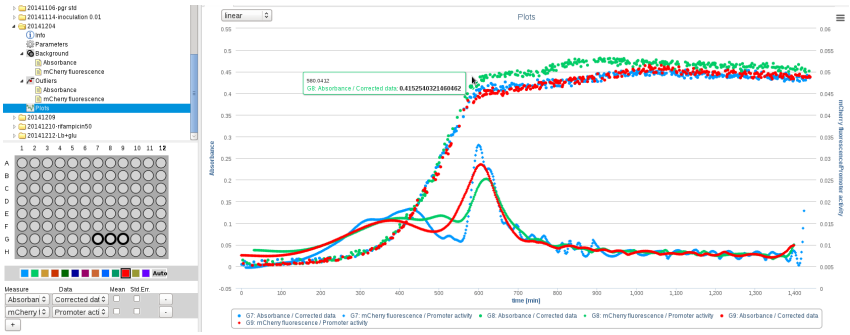


Figure 3.5: **Screenshot of the web application WellInverter.** WellInverter allows reporter gene experiments to be analyzed online through a web-based platform. The screenshot shows background-corrected absorbance data and estimated promoter activities for three different wells (G7, G8 and G9).

Summary of Chapter 4

In this chapter we present and discuss the results of our study of CCR, introduced in Chapter 1. We show that in a strain expressing *crp**, the variations of our synthetic p*lac** promoter are well explained by a model involving only CRP* and global physiological effects. By adding cyclic AMP to this model we were able to explain the activity profile of p*lac** during glucose-glycerol and a glucose-acetate diauxies. Our model fails to predict what happens when cAMP is added in large excess to the growth medium during growth on glucose. We solve this discrepancy by showing that cAMP import is likely repressed in presence of glucose

Once calibrated, our model enables us to assess the relative contribution of each regulator over the course of a given experiment (cAMP, CRP, and global physiological effects) by comparing the observed p*lac** activity and β-galactosidase concentration to the hypothetical scenarios where these different factors are kept constant over the time course of the experiment. We found that while cAMP exhibits dramatic variations during growth transitions, its contribution to the activity of p*lac** was actually of the same order of magnitude as CRP.

Résumé du Chapitre 4

Dans ce chapitre nous présentons et discutons les résultats de notre étude de la répression catabolique, introduite au Chapitre 1. Nous montrons que dans une souche exprimant *crp**, les variations de notre promoter synthétique p*lac** sont bien expliquées par un modèle mettant en jeu uniquement CRP* et des effets physiologiques globaux. En amendant dans ce modèle pour prendre en

compte l'action d'AMP cyclique, nous avons pu expliquer le profile d'activité du promoteur plac^* observé durant des diauxies glucose-glycerol and glucose-acetate. Nous observons que notre modèle ne permet pas de prédire l'effet de l'ajout d'AMP cyclique en large excès dans le milieu de culture lors d'une croissance sur glucose. Nous montrons que cela peut-être expliqué par le fait que l'import d'AMP cyclique est réprimé en présence de glucose.

Le modèle une fois calibré permet d'évaluer, pour expérience donnée, les contributions relatives de chaque acteur (AMPc, CRP, et effets physiologiques globaux) en comparant l'activité de plac^* et la concentration de β-galactosidase observées à leurs valeurs hypothétiques dans le cas ou chacun de ces ateurs resterait constant dans le temps. Nous observons que bien que l'AMPc varie dramatiquement entre deux phases de croissance, sa contribution à l'activité du promoteur plac^* reste du même ordre d'importance que la contribution de CRP.

Chapter 4

Regulation dynamics of a CRP-cAMP dependant promoter in *E. coli*

4.1 Motivation

We have seen in the introduction of this thesis that the role of cAMP in *E. coli*'s Carbon Catabolite Repression, and in particular its role in the regulation of the *lac* operon, remains controversial (Crasnier-mednansky et al., 2008; Narang, 2009a). We showed in Section 1.1.4 that the activity (in number of proteins produced per cell per minute) of a cAMP-dependent promoter can be well explained by the product of cAMP-related effects and the influence of global factors. We therefore postulated the following regulation model for the synthetic p*lac** promoter, designed so as to be regulated by the CRP-cAMP complex only.:

$$a_{plac^*}(t) = f_a \cdot \left(\frac{[\text{CRP-cAMP}](t)}{C_c + [\text{CRP-cAMP}](t)} \right) \cdot \left(\frac{a_{pRM}(t)}{K_{pRM} + a_{pRM}(t)} \right), \qquad (4.1)$$

, where $a_{plac^*}(t)$ denotes the activity of p*lac** at time t. This promoter is inserted upstream of a gene coding for a fluorescent protein (*gfp-mut2*), allowing us to infer promoter activity from fluorescence and absorbance measurements as explained in Chapter 3. The promoter activity $a_{pRM}(t)$ of a constitutive

reporter gene (pRM-*gfp*) is used as a proxy of the physiological state of the cell.

In a wild-type *E. coli* strain, changes in the activity and concentration of the gene expression machinery occur during growth transitions, simultaneously with important modulations of the cAMP and CRP pools. To calibrate the model of Equation 4.1 and ensure that each parameter can be identified we will need to separate the effects of the different regulators, i.e. use a set of specifically designed strains to create conditions in which influences due to the gene expression machinery dominate, and conditions in which the specific regulations of p*lac** determine the activity of this promoter. We will verify that the parameter estimates obtained in these conditions can also be used to explain p*lac** activity during growth transitions in wild-type strains. Finally, we will confront the calibrated model with paradoxical observations in which the addition of cAMP to the growth medium has no effect on the activity of p*lac** during growth on glucose.

4.2 Results

4.2.1 Regulation of a CRP-cAMP dependent promoter by CRP*

In a strain where the CRP-cAMP complex is replaced by a functionally analogous protein, CRP*, the model of Equation 4.1 becomes:

$$a_{plac^*}(t) = f_a\left(\frac{[\text{CRP}^*\,]}{C_c + [\text{CRP}^*\,]}\right)\left(\frac{a_{pRM}(t)}{K_{pRM} + a_{pRM}(t)}\right), \tag{4.2}$$

where we assume that the affinity constant C_c for the p*lac** promoter was the same for CRP-cAMP and CRP*. We thus obtain a simpler model, independent of fluctuations of the concentration of cAMP, which can therefore be more easily calibrated from experimental observations.

Mutant *E. coli* strains producing neither cAMP nor CRP (Δcya Δcrp genotype), were transformed with different expression and reporter plasmids in order to induce the synthesis of CRP* in the cells in a controlled way, and simultaneously monitor the variations of $a_{plac^*}(t)$ and $a_{pRM}(t)$ (Figure 4.1).

A first strain was obtained by transforming Δcya Δcrp cells with a plasmid carrying crp^* downstream of a ptet promoter, which allows crp^* transcriptionto be tuned by modulating the concentration of the inducer anhydro-tetracycline (ATc) in the growth medium. In order to monitor crp^* synthesis and estimate the time-varying profile of intracellular CRP* concentration, other strains were transformed with a plasmid carrying a ptet-crp^*-gfp construction, where the genes crp^* and gfp are transcribed together, as a consequence of which the synthesis rate of gfp can be assumed equal to the synthesis rate of CRP*. The strain carrying the ptet-crp^* induction plasmid was also transformed with reporter plasmids plac^*-gfp or pRM-gfp to monitor the time course of $a_{plac^*}(t)$ and $a_{pRM}(t)$ over time.

Bacteria were grown overnight on glucose, then diluted 1/20 in minimal media with different ATc concentrations and nutrient compositions (either 0.3% glycerol, or 0.3% glucose + fructose). Their growth and fluorescence were monitored in a microplate reader in order to infer the activities of the different promoters over time (dotted lines in Figure 4.2). The nutrient compositions were chosen so as to yield different plac^* activity profiles (panels G and H). The ptet-crp^*-gfp construction makes it possible to estimate a synthesis rate profile of CRP*, denoted $a_{crp^*}(t)$. We see in panels A and B of Figure 4.2 that $a_{crp^*}(t)$ increases at the beginning of the experiment due to the production of CRP*. The subsequent decrease in activity is due to the fact that ATc is degraded in the growth medium (Politi et al., 2014), and the initial ATc concentrations were chosen through trial and error to be slightly above the ptet activation threshold. The concentration of CRP* resulting from this synthesis can be computed using the following one-step expression model:

$$\frac{d}{dt}[\text{CRP}^*\,](t) = a_{crp^*}(t) - (d_{CRP} - \mu(t))\,[\text{CRP}^*\,](t), \tag{4.3}$$

where parameter d_{CRP} is the degradation rate of CRP (which is assumed equal to the degradation rate of CRP* since the two proteins are very similar). As d_{CRP} has never been measured (the only reported value, obtained by fitting a large mathematical model to experimental data, is 0.01 min^{-1}), it was estimated using dedicated experiments, reported in Section C1 of the SI. These experiments show that during growth on acetate, in conditions where the

variations of $a_{plac^*}(t)$ depend essentially on the variations of [CRP*], Equation 4.3 only yields [CRP*] profiles matching the experimental data when $d_{CRP} \simeq 0.0011$ min^{-1}, which indicates that CRP$^{(*)}$ is a stable protein (half-life of circa 11h).

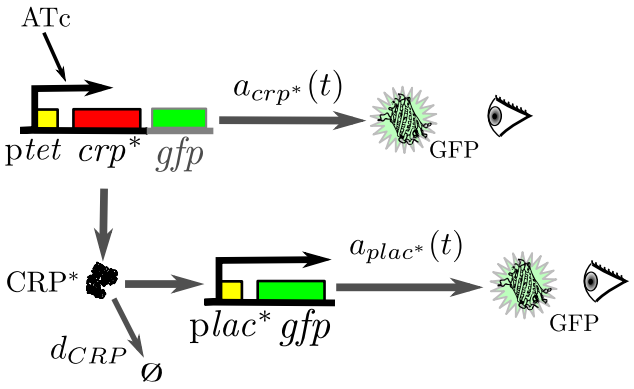


Figure 4.1: **Induction and reporter plasmids used in the Δ*cya* Δ*crp* *E. coli* mutant experiments.** Note that the strains carry at most one reporter gene, so strains carrying a p*tet-crp***-gfp* construction do not carry the p*lac***-gfp* reporter plasmid.

Using this value of d_{CRP} the profiles of the intracellular CRP* concentration in the different conditions could be estimated from the of profiles of *crp** activity (panels E and F of Figure 4.2). We see that in all conditions this concentration increases during the first hours of the experiment, then decreases as CRP* is not produced anymore, and is diluted through population growth (the decrease due to degradation being negligible). These profiles enable us to calibrate our model. We adjust the values of C_c and K_{pRM} such that, in all conditions, the profile of a_{plac^*}, given the profiles of a_{pRM} and [CRP*], best fits the experimental data. The best fit gives $C_c = 48000$ (arbitrary units) and $K_{pRM} = 25$ (a.u.), as shown in the first plot of Figure 4.3, and gives very good predictions of $a_{plac^*}(t)$ (solid lines in panels G and H of Figure 4.2). The fact that parameters C_c and K_{pRM} are (most of the time) above the observed values of, respectively, $a_{pRM}(t)$ and [CRP*](t) shows that p*lac** activity is sensitive to both these variables. In particular, this establishes that p*lac** expression

is controlled by global regulations (which can be clearly seen around $t = 400$ minutes in Figure 4.2F).

These results, which tend to prove the role of CRP and global effects in the regulation of a CRP-cAMP-dependent gene, also indicate that the simple model of Equation 4.2 is sufficient, in the sense that it captures well the observed activity of p*lac**, in a quantitative way, and no additional regulation mechanism needs to be invoked. It can therefore be expected that the full model of Equation 4.1 will be sufficient to explain the activity of p*lac** in a wild-type strain.

4.2.2 Regulation of a cAMP-regulated promoter during diauxic shifts

Catabolic genes are known to be transiently over-expressed during growth transition towards a poorer carbon source (Wolfe, 2005; Inada et al., 1996; Berthoumieux et al., 2013). Can the model of Equation 4.1, partly calibrated in the previous section, quantitatively explain these variations of activity?

Bacterial strains carrying reporter plasmids p*crp-gfp*, pRM-*gfp* and p*lac**-*gfp* were grown overnight on glucose, then rediluted (factor 1/1000) in fresh minimal medium containing either 0.2% glucose or 0.2% glucose and 0.2% glycerol. Their growth and fluorescence were monitored in the microplate reader. The resulting promoter activities are represented in Figure 4.4. During the experiments, wells with strains containing p*lac** cells were sampled and assayed for cAMP concentration in the medium. The intracellular cAMP concentration was inferred from these measurements as explained in Section 1.3.3 (black lines in Figure 4.4). The original data are presented in more detail in Appendix C. We see that gene expression profiles have similar shapes in the two diauxies (growth on glucose as the sole carbon source leads to a glucose-acetate diauxie because acetate is secreted during fast growth on glucose), but the amplitudes of the variations differ: the overshoot of the p*lac* activity is at least twice as large in the case of the glucose-glycerol diauxie, and the activity of $p_{RM}(t)$ falls more rapidly in absence of glycerol (- 50 % between the values at $t = 550$ min and $t = 650$ min) while it is maintained longer on glycerol (- 10 % between $t = 550$ min and $t = 650$ min). This difference may be ex-

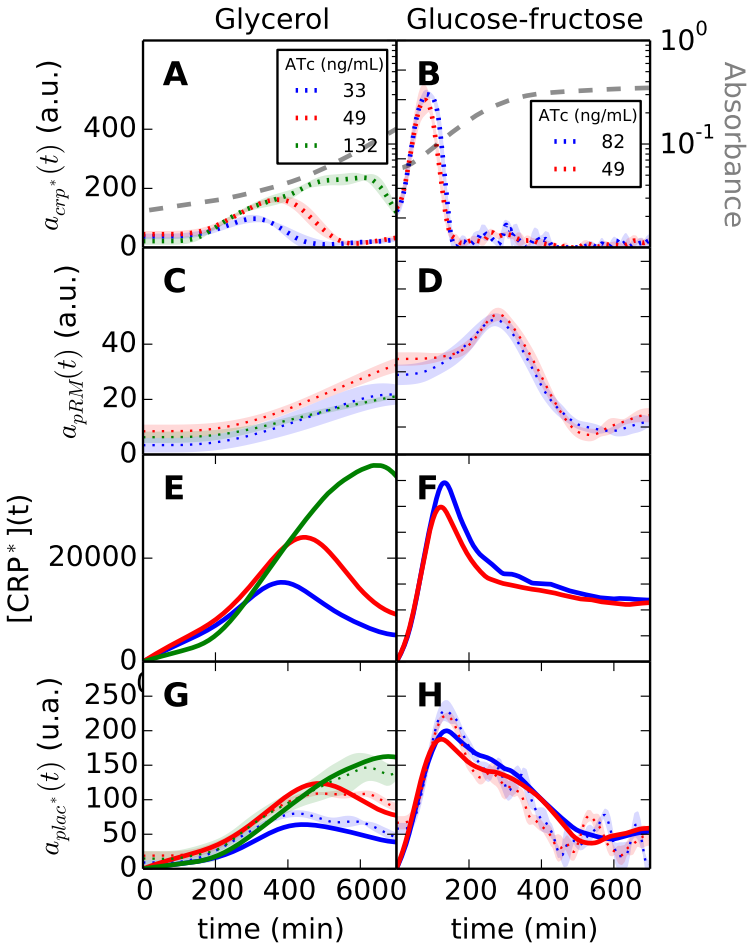


Figure 4.2: **Regulation of p*lac** by CRP* in a Δ*cya* Δ*crp* strain.** Dotted lines and shaded areas indicate the mean ± one standard deviation of at least 4 promoter activity profiles observed in different microplate wells. All observations in one column come from the same microplate. The production of *crp** in all strain was induced at $t = 0$ minutes by rediluting the cells in a fresh medium containing a concentration of ATc indicated in panels A and B. **A,B.** Promoter activity of the p*tet-crp***-gfp* gene, induced by ATc at $t = 0$ minutes. **C,D.** Activity of pRM. **E,F.** Concentrations of CRP*, estimated from the activities in panels A and B. **G,H.** The activity of p*lac** is predicted from $a_{pRM}(t)$ (panels C,D) and [CRP](t) (panels E,F) using the model of Equation 4.2. The predictions (solid lines) are compared to the observed data (dotted lines).

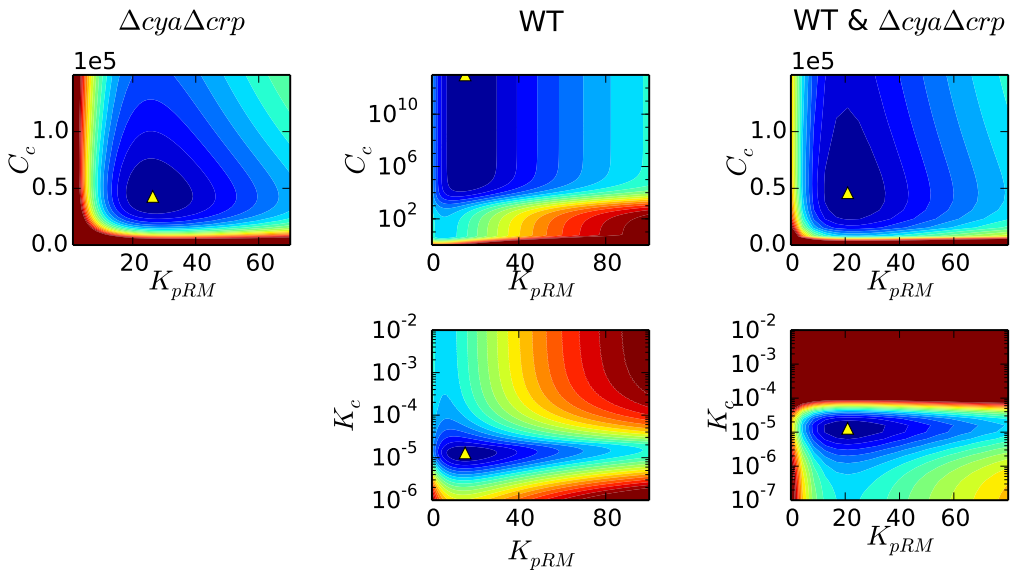


Figure 4.3: **Calibration of the p*lac** regulation model on observed data.** The parameters C_c, K_{pRM}, and K_c are varied and the goodness of fit is calculated for each parameter set. and The columns show calibration results using different datasets. The left panel uses only the data from Δcya Δcrp strains represented in Figure 4.2. The column in the middle is based on the data from diauxic shifts observations presented in Figure 4.4. The right column uses data from both types of experiments. In all plots, triangles indicate the best fit in the sense of the minimal squared error between model prediction and observed data. Dark blue and dark red regions indicate parameters yielding a squared error less than 5%, or more than twice larger than the error of the best fit, respectively.

plained by the fact that glycerol is a relatively rich carbon source, maintaining a relatively high activity of machinery. In both diauxies we also observe a decrease of $a_{crp}(t)$ during growth transitions, potentially due to global effects (Berthoumieux et al., 2013).

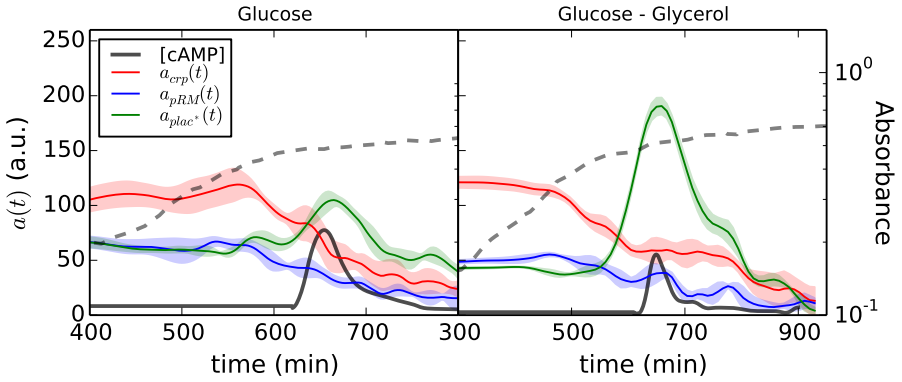


Figure 4.4: **Promoter activities and cAMP concentration in *E. coli* during diauxic shifts.** The promoter activities were measured from absorbance and fluorescence data as explained in Chapter 3. The curves indicate the mean and one standard deviation to the mean of at least 4 replicate wells. Cyclic AMP concentrations (in arbitrary unit, scale not represented) were deduced from external measurements of cAMP concentration using an inversion method, as explained in Chapter 1. The dashed line shows a representative absorbance profile on a logarithmic scale.

These experiments alone lead to a very poor calibration of the model of Equation 4.1 (middle column in Figure 4.3). In particular, these data do not provide any reliable estimate for the saturation threshold C_c of plac^* with respect to CRP-cAMP, and they only yield a very rough estimate of K_{pRM}, which does not even clearly allow to rule out the case $K_{pRM} = 0$. In other words, these data do not provide evidence that global regulations have an effect on plac^* (as was the case with the observation on mutants in the last section). However, pooling these observations with those of Figure 4.2 enables us to fully calibrate the model (right column in Figure 4.3). The best fit corresponds to biological parameters close to the ones found with Δcrp

Δcya mutants ($K_{pRM} = 21, C_c = 42000$ a.u.), meaning that the sensibility of the plac^* promoter to global modulations and to CRP-cAMP (or CRP*) is conserved between the different conditions. The calibrated model appears to explain well the activity of plac^* in both the wild-type strain and the mutant strains (Figure 4.5).

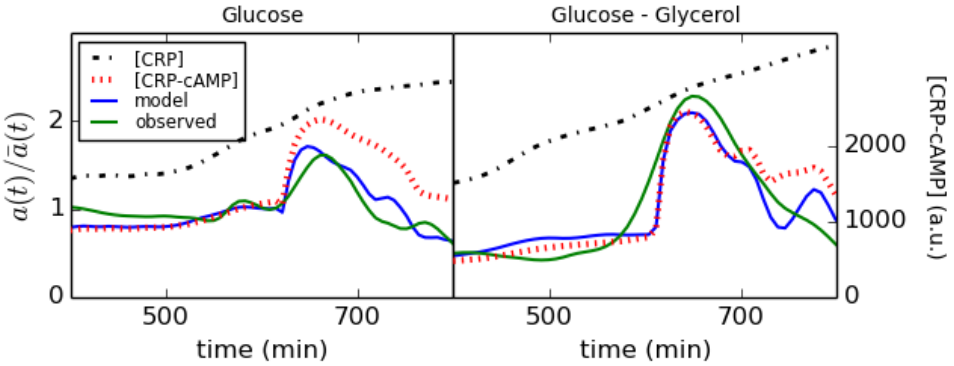


Figure 4.5: **Predictions of the calibrated model.** CRP concentrations were estimated from the crp^* promoter activities shown in Figure 4.4, using Equation 4.4. CRP-cAMP concentrations were computed from CRP and cAMP concentrations and the dissociation constant K_c estimated in Section 4.2.2. Blue lines are predictions from the model for $a_{plac^*}(t)$, computed from the CRP-cAMP estimations and $a_{pRM}(t)$ profiles. The predictions of the model are compared to the experimentally observed plac^* activities (green lines, same as in Figure 4.4). All promoter activities, predicted, and observed, are normalized by their means to accentuate the similarity in fold change between the model predictions and the observed data.

Parameters K_c, C_c and K_{pRM} characterize the sensibility of the plac^* promoter to the different variables. We observe that the estimated value of the CRP-cAMP dissociation constant K_c is of the same order of magnitude as cAMP concentrations during growth on glucose, which was also the conclusion drawn from in-vitro experimental data in (Narang, 2009b). Higher cAMP concentrations will saturate CRP: at the apex of the transient cAMP peak, almost 100% of the CRP molecules are bound by cyclic AMP, as can be seen in both plots of Figure 4.5. Thus, the dramatic increase of cAMP concentra-

tion at transitions (by a factor of 20) results in relatively mild variation of the CRP-cAMP concentration and p*lac** activity (by a factor of three in the glucose-glycerol diauxy).

The concentration of CRP in the cells, shown in Figure 4.5, was estimated from $a_{crp}(t)$ using an expression model similar to Equation 4.3:

$$\frac{d}{dt}[\text{CRP}](t) = \alpha a_{crp}(t) - (d_{CRP} + \mu(t))\,[\text{CRP}](t), \tag{4.4}$$

where $\alpha = 1/5$ is a correction factor reflecting the fact that the plasmid carrying the p*crp-gfp* reporter is present in approximately 5 copies per cells. This correction makes it possible to compare the obtained CRP concentration (in arbitrary units) to the CRP* concentration found in Section 4.2.1, where we assumed that the transcription rate of gene *gfp* equals that of *crp** in the p*tet-crp***-gfp* construction. A somehow paradoxical consequence of the stability of CRP (small d_{CRP}) is that even though the synthesis rate of *crp* decreases upon glucose exhaustion (Figure 4.4), the growth arrest causes CRP concentrations to actually increase: for both diauxies, [CRP] increases by approximately 50% between $t = 500$ min. and $t = 700$ min. The affinity constant C_c of the CRP-cAMP complex for the p*lac** promoter was found to be much higher than the observed CRP-cAMP concentration, which means that the p*lac** promoter is fully sensitive to the variations of the CRP-cAMP concentration. Since this concentration is proportional to the concentration of CRP, we conclude that the increase of CRP concentration after exhaustion of glucose increases the activity of p*lac* by the same amplitude.

We also find $K_{pRM} = 21$. Since the values of $a_{pRM}(t)$ measured in both diauxies are most of the time well above this value, we conclude that the modulations of p*lac** activity by global effects during growth transitions are less pronounced than the (already mild) variations of pRM activity (shown in Figure 4.4).

In order to understand the functional role of each regulator, we calculated their relative importance for setting the concentration of β-galactosidase (or any other cAMP-induced enzyme) in the cells, as this concentration determines the capacity of the cell to grow on secondary carbon sources. We can assume that the activity profiles of the p*lac** promoter observed here is equivalent to the activity of the *lac* operon in presence of IPTG. We have seen

in Section 1.1.3 that this concentration was related to the activity of the *lac* promoter as follows:

$$\frac{d}{dt}[\beta\text{-Gal.}](t) = a_{lac}(t) - \mu(t)[\beta\text{-Gal.}](t), \tag{4.5}$$

which enables us to obtain an estimation of [β-Gal](t) using $a_{plac^*}(t)$ as a (proportional) estimate of $a_{lac}(t)$. Table 4.1 shows the importance of the different variables in determining the activity of plac^* and therefore the intracellular β-galactosidase concentration after glucose exhaustion: we used the calibrated model to predict what would be the height of the plac^* activity peak at growth transition ($a_{plac^*}(t = 650)$) and the concentration of β-galactosidase after growth transition, at $t = 750$, if one of the variables [cAMP], [CRP], or a_{pRM}, were to be kept constant after $t = 500$ minutes, i.e., if it played no role in determining the value of $a_{plac^*}(t)$. These simulated experiments give an estimate of the fold change in plac activity and β-galactosidase concentration due to each actor, relative to the hypothetical situation where this actor does not play any role, all other things being equal. As expected, we see that cAMP stands out as the main cause of the peak of a_{plac^*}. Variations of [CRP] were found to have a smaller yet comparable importance, while the action of global regulations was found to be marginal in these experiments. In terms of influence on the β-galactosidase concentration after the growth transition, however, all these regulations seem to have a very moderate effects. In particular, the important overshoot of cAMP concentration only increases β-galactosidase activity by 24% compared to the scenario in which intra-cellular cAMP concentration would stay constant.

These results would need to be confirmed with other diauxies. We present in Appendix C a study of two additional diauxies on glucose-xylose, and glucose-fructose. The model calibrated in this section predicts relatively well the response of plac^* on xylose. However, the present calibration failed to correctly explain the particular dynamics observed on fructose. This discrepancy may be due to the poor quality of the cAMP concentrations measurements for this diauxie.

	Glucose		Glucose-Glycerol	
	$a_{plac^*}(t = 650\text{min.})$	$[\beta\text{-Gal.}]_{t=750\text{min.}}$	$a_{plac^*}(t = 650\text{min.})$	$[\beta\text{-Gal.}]_{t=750}$
[cAMP]	+37%	+8%	+62%	+24%
[CRP]	+33%	+13%	+24%	+13%
a_{pRM}	-14%	-9%	-16%	-14%

Table 4.1: **Effect of the different regulators on the activity of a CRP-cAMP dependent gene and β-galactosidase concentrations after the transition.**

4.2.3 Effect of external cAMP on the activity of a cAMP-regulated promoter

In (Wanner et al., 1978) the authors remark that adding cAMP to the growth medium does not have a large impact on the steady-state activity of the *lac* operon (Figure 4.6A). This is interpreted in (Narang, 2009b) to mean that cAMP levels on glucose, while lower than on other carbon sources, are nonetheless at near-saturation levels with respect to CRP. Our observation of the glucose-glycerol diauxy, as well as our calibrated model, corroborate this interpretation. However, our data suggest that a significant increase of the intracellular cAMP concentration could potentially amplify p*lac** activity by at least 3-fold, which is not observed in Wanner's data.

It has been suggested in (Ishizuka et al., 1994) that the CRP-cAMP complex down-regulates the expression of *crp*. Thus, higher intracellular cAMP concentrations could result in lower CRP concentrations and overall unchanged CRP-cAMP concentrations. This hypothesis is supported by the observation in (Ishizuka et al., 1993) that adding external cAMP has the expected up-regulatory effect on the *lac* promoter, provided that *crp* is over-expressed.

Another, even simpler explanation of the lack of any effect of external cAMP on the activity of the *lac* operon is that cAMP does not enter the cell. This hypothesis is rejected by the authors of (Wanner et al., 1978), who point out that cAMP addition to the growth medium has a clear impact in

some conditions, such as growth rate recovery on some sugars in Δcya cells. However this does not rule out the possibility that cAMP import (or absence of import) depends on the carbon source used by the bacteria, and in particular that cAMP import is blocked on rich carbon sources.

Figure 4.6B presents a series of experiments in which the activity of our plac^* promoter was observed during growth on glucose and until a few hours after glucose exhaustion, without cAMP into the medium, and with the addition of 2mM cAMP in the medium. We observe that in a wild-type strain, adding cAMP in the medium has no effect during growth on glucose: the only notable difference occurs during growth transition, where a transiently $\sim$30% higher activity can be observed for cells growing in presence of external cAMP. This observation suggests that the cells are oblivious to external cAMP during growth on glucose. In a strain lacking cAMP (Δcya) placed in a cAMP supplemented medium, it could be expected that the external cAMP, by entering the cell, would re-establish the original phenotype. Instead, we observe that the activity of plac^* remains negligible during growth on glucose and the promoter is suddenly activated upon glucose exhaustion, which is further evidence of the absence of cAMP import during growth on glucose. A possible reason is that glucose inhibits the cAMP import system, which would be coherent with the fact that cAMP is not stricly needed by *E. coli* during this growth phase. Levels of phosphodiesterase (which catalyzes cAMP degradation) could also play a role in screening external cAMP (Kuhlman et al., 2007). As an additional control we observed that inducing CRP* in the cells at the beginning of the experiment resulted in a 2-fold increase of plac^* activity on glucose, confirming that the inaction of external cAMP is a problem specific to cAMP import, and not due to a repression of the plac^* promoter during growth on glucose.

In conclusion, the data presented in (Wanner et al., 1978) should be interpreted with care, and not be seen as an indication of the sensitivity of the *lac* promoter to cAMP, as the addition of 5mM external cAMP may in fact result in different (and unknown) intracellular concentrations of cAMP, depending on the growth substrate.

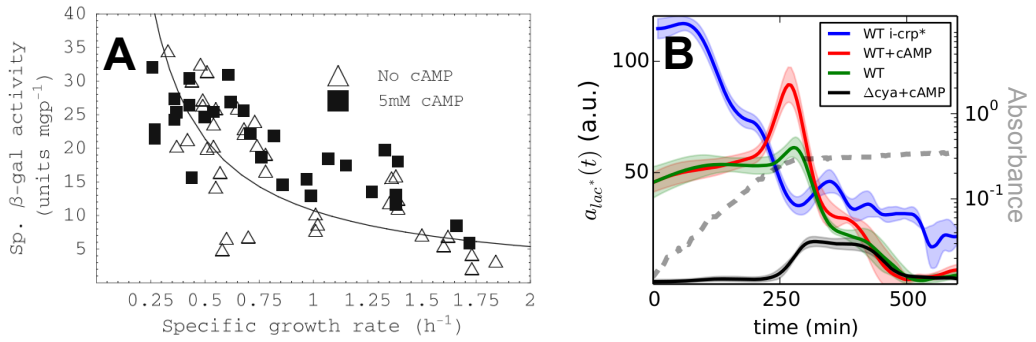


Figure 4.6: **Effect of cAMP addition to the medium on the regulation of p*lac**. A.** Steady-state observations of β-galactosidase activity in an IPTG-supplemented medium, on a range of subtrates (glucose, glycerol, acetate, etc.) with and without addition of 5mM cAMP (Wanner et al., 1978). **B.** Activity of p*lac* in *E. coli* growing on glucose. The curves from different experiments (indicating the mean and standard deviation from at least 4 replicates) are synchronized with respect to the entry into stationnary phase.

4.3 Discussion

While the *lac* operon has been studied for decades and in many different conditions, most studies only consider steady-state experiments, and overlook the very particular dynamics of the cAMP pool and cAMP-dependent gene activities during transitions. We found that the overshoot of cAMP was responsible for an up to 30% increase in enzyme concentration a few hours after glucose exhaustion. While this increase may seem mild compared to the hundred-fold modulations of gene activity caused by LacI repression, it contributes to maintain high enzymatic levels in the cells hours after glucose exhaustion (as enzymes like β-galactosidase are very stable). The observation that the expression of cAMP-dependent genes is also sensitive to the cells' physiology could explain the observed cAMP overproduction at early stationary phase: as the activity of the gene expression machinery decreases shortly after glucose exhaustion, it is interesting for the bacteria to maximize the production of enzymes before energy exhaustion. Using our calibrated model, we found

that, had the peak of cAMP concentration occured two hours later in the glucose-glycerol diauxy, the resulting peak of β-galactosidase synthesis would have been 30% smaller, due to the lower value of $a_{pRM}(t)$ in this time interval. This suggests that the reason of why we found that global regulations had a weak influence on the peak of cAMP in our experiments is that the peak occurs soon enough to avoid the global decrease of all gene expression following glucose exhaustion.

While CRP is necessary for the regulation of CRP-cAMP-dependent promoters, it is seldom considered as a modulator of gene activity, and the variations of its concentration during a diauxy had never been precisely quantified. We have shown here that the activity of p*lac** was fully sensitive to the concentration of CRP. We have established that CRP is stable, and, as a consequence accumulates in the cells after glucose exhaustion, thus increasing p*lac* activity by up to 50% a few hours after phase transition. This suggests that one possible mechanism by which glucose could lower p*lac** activity is by lowering CRP concentrations through dilution.

The pivotal role of cyclic AMP in Carbon Catabolite Repression has been postulated decades ago. Yet, on carbon limited cultures, the action of cAMP on catabolic genes appears mild compared to the action of specific regulators, such as LacI in the case of the *lac* operon. It is still unclear whether cAMP is *intended* for increasing the activity of these genes, or if it has another, more important functional role (Görke and Stülke, 2008a; You et al., 2013). A difficulty in the study of the role of cAMP is the estimation of intracellular cAMP concentrations, as the assays are expensive, difficult to perform in a normal laboratory setting, and can only give extracellular concentrations for a limited number of time points. The intracellular concentration must then be deduced, using a measurement model. Furthermore cAMP import, export and degradation in the cell are largely not understood. We have shown that by creating biological conditions in which p*lac** expression was independent of cAMP it was possible to estimate key biological parameters indicating the sensitivity of this promoter to (seemingly) all other regulators but cAMP, and that these parameters could be used to estimate the contribution of each actor, including cAMP, to the activity of the p*lac** promoter and β-galactosidase concentration in a wild-type strain. In this sense, our study shows the interest

of carefully planned perturbation experiments in order to characterize a small regulatory system.

Chapter 5

Conclusion

The inference and analysis of bacterial gene regulatory networks is a difficult task. Many mathematical formalisms and methods have been developed to model regulatory interactions and extract relevant information from experimental observations, but sometimes the available data are simply not informative enough to reach a conclusion. This is the case for carbon catabolite repression, for which the different studies have, until now, failed to produce a complete and coherent picture. One reason is that steady-state experiments provide only scarce data, and the data from different studies cannot be pooled due to differences in protocols and strains.

Dynamic perturbation experiments, in which we observe bacteria as they adapt to changes in their environment, provide a richer source of information than steady-state experiments. They enable us to observe, during a single experiment, a multitude of transition states that could not be observed during steady-state experiments. Since bacteria re-adapt in a matter of minutes, this approach has greatly benefited from the development of high-throughput technologies. In particular, fluorescent reporter genes offer a convenient and non-invasive way to monitor the expression of a gene of interest with much precision and high time resolution (on the order of minutes). These new technologies call for a reassessment of the mechanisms underlying CCR.

In this thesis I have developed mathematical methods for the inference of regulatory networks and the estimation of biological signals, with an application to the dynamical study of growth transitions in *E. coli*. We first provide

simple rules of thumbs for the reduction of gene expression models from two steps (transcription, translation) to one step. Such a reduction makes the model simpler to analyze and to identify (as it has less parameters) and faster to simulate. However, not taking into account transcription leads to a simplification error. We provide theoretical and experimentally-measured upper bounds for this error, and provide guidelines for the reduction of larger models involving several genes.

A second contribution is the development of robust methods for the analysis of fluorescent reporter gene data. We have shown how the relevant biological variables can be linked in a linear way to the observed data, and we have developed estimation procedures that were rigorously tested for bias and robustness. They constitute, to our knowledge, the most comprehensive and practical approaches to extract biological signals from reporter gene data acquired in kinetic experiments in a microplate reader.

We carried out a systematic characterization of the regulation of a cAMP-dependent promoter in *E. coli*. Through carefully designed experiments we could show that this promoter is affected by global regulations, and we evaluated the contribution of each actor (cAMP, CRP, and global effects) to the activity of this promoter during growth transitions. The kinetic parameters estimated in this study were shown to be conserved across experimental conditions. Therefore they can be used in larger models of CCR involving several catabolic genes and their regulators. We also showed that CRP is a limiting factor in the expression of our CRP-cAMP-dependent promoter, and that this key regulator accumulates in the cells upon growth arrest, thus having a significant positive role in the activity of the CRP-cAMP promoter. Notice that the accumulation of CRP in the cell is not due to any particular up-regulation mechanism, but simply to growth arrest. Since growth arrest influences the concentration of all proteins, it can be seen as a global post-transcriptional regulation signal. Finally, we enhanced our understanding of why adding cAMP in large excess to the growth medium does not affect the activity of a cAMP-dependent gene. The analysis of our dynamical experiments show that cAMP does not enter bacteria in the presence of glucose.

Conclusion en Français

L'inférence et l'analyse de réseaux de régulation génique est une tache difficile. De nombreux formalismes et méthodes mathématiques ont été développés pour modéliser les intéractions entre gènes et extraire des informations pertinentes des données expérimentales. Mais il arrive que les données expérimentales ne soient simplement pas suffisament informatives pour mener à une conclusion. C'est le cas pour la répression catabolique, dont les différentes études n'ont pas pu, jusqu'à présent, donner une image complète et cohérente. Une des raisons est que les expériences en régimes stationnaire produisent peu de données, et les données provenant de différentes études ne peuvent être utilisées ensemble de par les différences de souches et de protocoles.

Les expériences de perturbation dynamiques, où l'on observe des bactéries durant leur adaptation à des changements environementaux, fournissent une meilleur source d'information que les expériences en régime stationnaire. Elles permettent d'observer, sur une seule expérience, une multitude d'états transitoires qui seraient inobservable dans une expérience en régime stationnaire. Comme les bactéries s'adaptent à un nouvel environement en l'espace de quelques minutes seulement, cette approache a grandement bénéficié du récent développement de technologies à haut débit. En particulier, les gènes rapporteurs fournissent un moyen simple et non-invasif de suivre l'expression d'un gène d'intérêt durant une expérience, avec une grand présolution et résolution temporelle (de l'ordre de la minute). Ces nouvelles technologies invitent à une réévaluation des méchanismes de la répression catabolique.

Dans cette thèse nous avons développé plusieurs méthodes mathématiques pour l'inférence de réseaux de régulation géniques et l'estimation de signaux biologiques, que nous avons appliqué à l'étude dynamique des transitions chez

E. coli. Nous avons énoncé des règles pratiques pour la réduction de modèles d'expression génique de deux étapes (transcription, traduction) à une étape. Une telle réduction rend l'analyse du modèle plus simple (puisqu'il y a moins de paramètres) et sa simulation plus rapide. Cependant, ne pas prendre en compte l'étape de transcription conduit à une *erreur de simplification*. Nous donnons des bornes supérieures, théoriques et mesurées expérimentallement, pour cette erreur, ainsi que des instructions pour la réduction de modèles d'expression génique dans des plus grands réseaux de gènes.

Une deuxième contribution de cette thèse est le développement de méthodes robustes pour l'analyse de données de gène rapporteur fluorescent. Nous avons montré comment certaines variables biologiques d'intérêt étaient reliées linérairement aux données expérimentales, et avons développé des procédures d'estimations dont nous avons rigoureusement évalué la robusteté et d'éventuel biais. Ces procédures constituent, à notre connaissance, l'approche la plus complète et pratique pour extraire des signaux biologiques de données de gène rapporteur fluorescent acquises durant des cinétiques en lecteur de microplaque.

Nous avons caractérisé de façon systématique les régulations d'un promoteur cAMP-dépendant chez *E. coli*. Grâce à des expériences soigneusement conçues nous avons pu montrer que ce promoteur subissait l'influence d'effets physiologiques globaux, et nous avons évalué la contribution de chaque acteur (cAMP, CRP, et les effets physiologiques globaux) à l'activité de ce promoter durant des transitions vers l'état stationnaire. Nous avons que les paramètres cinétiques estimés dans cette étude sont conservé à travers les différents scénarii expérimentaux. Ils peuvent donc être utilisés dans des modèles plus larges de la répression catabolique, impliquant plusieurs gènes cataboliques et leurs régulateurs. Nous avons également montré que CRP était un facteur limitant dans l'expression de notre promoteur régulé par CRP-AMPc, et que ce régulateur clé s'accumule dans les bactéries lors d'un arrêt de croissance, ayant donc un important effet positif fur l'activité d'un gènes régulés pas CRP-cAMP. Remarquons que cette accumulation de CRP n'est pas due à un régulateur particulier, mais simplement à l'arrêt de croissance. Comme cet arrêt affecte la concentration de toutes les proteines, il peut être vu comme un signal global de régulation au niveau post-traductionel. Finalement, nous avons contribué

à la compréhension de l'absence d'effet de l'AMP cyclique lorsque celui-ci est ajouté au milieu de culture, en montrant par des expériences en dynamique que l'AMP cyclique n'est pas importé par des bactéries en présence de glucose

Appendix A

Supplementary Information on Chapter 2

A1 Reformulation of the models

The following proposition explains the transition between the equation system expressed in terms of concentrations and the system formulated in terms of extensive variables.

Proposition 1. *Eq. 2.1 and Eq. 2.6 in main text are equivalent.*

Proof. The computations below show the equivalence of Eq. 2.1a and Eq. 2.6a. The computations for Eq. 2.1b and Eq. 2.6b are similar.

$$\begin{aligned}
\frac{d}{dt}m(t) &= \kappa_m\, f(t) - (\gamma_m + \mu(t))\, m(t) \\
\frac{d}{dt}\left(\frac{M(t)}{V(t)}\right) &= \kappa_m\, \frac{F(t)}{V(t)} - (\gamma_m + \mu(t))\, \frac{M(t)}{V(t)} \\
\frac{1}{V(t)}\,\frac{d}{dt}M(t) - \frac{M(t)\,\mu(t)}{V(t)} &= \kappa_m\, \frac{F(t)}{V(t)} - (\gamma_m + \mu(t))\, \frac{M(t)}{V(t)} \\
\frac{d}{dt}M(t) &= \kappa_m\, F(t) - \gamma_m\, M(t).
\end{aligned}$$

□

A1.1 Further simplification of the equation systems

In this section we present a simpler yet equivalent formulation of Eq. 2.6 and Eq. 2.7. The equation systems thus obtained will be used as a basis for the proofs given in the next sections. We introduce the following dimensionless variables, which can be seen as *normalized* versions of the extensive variables F, M, and P:

$$\mathcal{F}(t) = \frac{F(t)}{F(0)}, \quad \mathcal{M}(t) = \frac{\gamma_m}{\kappa_m F(0)} M(t),$$
$$\mathcal{P}(t) = \frac{\gamma_m \gamma_p}{\kappa_m \kappa_p F(0)} P(t), \quad \hat{\mathcal{P}}(t) = \frac{\gamma_m \gamma_p}{\kappa_m \kappa_p F(0)} \hat{P}(t).$$

Note that by definition, $\mathcal{F}(0) = 1$. When substituting the above variables into Eq. 2.6, we obtain the following two-step model:

$$\frac{d}{dt}\mathcal{M}(t) = \gamma_m \left(\mathcal{F}(t) - \mathcal{M}(t)\right), \tag{A1a}$$
$$\frac{d}{dt}\mathcal{P}(t) = \gamma_p \left(\mathcal{M}(t) - \mathcal{P}(t)\right). \tag{A1b}$$

The one-step model of Eq. 2.7 transforms into

$$\frac{d}{dt}\hat{\mathcal{P}}(t) = \gamma_p(\mathcal{F}(t) - \hat{\mathcal{P}}(t)), \tag{A2}$$

which, incidentally, happens to be the model that one would obtain by directly reducing the system of Eq. A1 using the QSSA. This second reformulation also conserves the relative error:

$$\Delta(t) = \frac{\mid \mathcal{P}(t) - \hat{\mathcal{P}}(t) \mid}{\mathcal{P}(t)}, \tag{A3}$$

and we will use Eq. A1 and Eq. A2 as a basis for further discussion in this appendix.

A1.2 Influence of the parameters κ_m and κ_p

The production constants κ_m and κ_p do not play any role in the dynamics of Eq. A1 and Eq. A2. This can be explained by the fact that κ_m and κ_p influence the amplitude of the variations of M and P, but not the time-scale on which these variations occur. Moreover, under the assumptions of Propositions 1 and 2,

$$\frac{d}{dt}M(0) = \frac{d}{dt}P(0) = \frac{d}{dt}\hat{P}(0) = 0,$$

we have $\mathcal{M}(0) = \mathcal{P}(0) = 1$, independently of the values of κ_m and κ_p. Under different initial conditions, κ_m and κ_p appear in the definition of the initial conditions of $\mathcal{M}$, $\mathcal{P}$ and $\hat{\mathcal{P}}$, and thus influence $\Delta(t)$, but we show in the next section that the initial conditions play a limited role in the determination of $\Delta(t)$. This explains that κ_m and κ_p are not further discussed in our study.

A1.3 Influence of the initial conditions

The analytical solution of Eq. A2 [1] is

$$\hat{\mathcal{P}}(t) = \hat{\mathcal{P}}(0)e^{-\gamma_p t} + \gamma_p e^{-\gamma_p t}\int_0^t e^{\gamma_p x}\mathcal{F}(x)dx.$$

We can see here that the initial condition $\hat{\mathcal{P}}(0)$ is vanishing with a characteristic time γ_p. This means that, if the protein production rate is not null, then the initial conditions do not impact the value of $\hat{\mathcal{P}}$ after a few intervals of duration γ_p. In many cases where the population is growing with growth rate μ (i.e., the volume verifies $V(t) = e^{\mu t}$), the quantity of protein grows as well, and is grossly proportional to $e^{\mu t}$. Therefore the relative weight of the initial condition $\hat{\mathcal{P}}(0)$ in $\hat{\mathcal{P}}(t)$, given by $\hat{\mathcal{P}}(0)e^{-\gamma_p t}/\hat{\mathcal{P}}(t)$, will decrease in $e^{-(\gamma_p+\mu)t}$. This vanishing effect of the initial conditions hold for Eq. A1, and follows

[1] The general equation $dy(t)/dt = a(t)y(t) + b(t)$ has the solution:

$$y(t) = y(0)e^{\int_0^t a(x)dx} + \int_0^t e^{\int_x^t a(u)du}b(x)dx.$$

from the fact that these equation systems are linear filters (see Section A4 for further discussion).

A2 Proof of Proposition 1

We have seen in the previous section that the model reduction error $\Delta(t)$ can be expressed as a function of the reformulated variables $\mathcal{P}$ and $\hat{\mathcal{P}}$. Moreover, as the reformulated variable $\mathcal{F}$ is proportional to F, we have for all t,

$$\frac{1}{F(t)}\frac{d}{dt}F(t) = \frac{1}{\mathcal{F}(t)}\frac{d}{dt}\mathcal{F}(t).$$

As a consequence, proving

$$\Delta(t) \leq \frac{1}{\gamma_m}\sup_{s\leq t}\left|\frac{1}{F(s)}\frac{d}{dt}F(s)\right|$$

under the conditions

$$\frac{d}{dt}P(0) = \frac{d}{dt}M(0) = \frac{d}{dt}\hat{\mathcal{P}}(0) = 0$$

is equivalent to proving the formula written in terms of the new variables

$$\frac{|\mathcal{P}(t) - \hat{\mathcal{P}}(t)|}{\mathcal{P}(t)} \leq \frac{1}{\gamma_m}\sup_{s\leq t}\left|\frac{1}{\mathcal{F}(s)}\frac{d}{dt}\mathcal{F}(s)\right| \tag{A4}$$

under the conditions

$$\frac{d}{dt}\mathcal{P}(0) = \frac{d}{dt}\mathcal{M}(0) = \frac{d}{dt}\hat{\mathcal{P}}(0) = 0.$$

We will prove this in two steps. First we will prove that

$$\frac{|\mathcal{P}(t) - \hat{\mathcal{P}}(t)|}{\mathcal{P}(t)} \leq \sup_{s\leq t}\left|\frac{\mathcal{F}(s) - \mathcal{M}(s)}{\mathcal{M}(s)}\right|, \tag{A5}$$

and then we will prove that

$$\sup_{s\leq t}\left|\frac{\mathcal{F}(s) - \mathcal{M}(s)}{\mathcal{M}(s)}\right| \leq \frac{1}{\gamma_m}\sup_{s\leq t}\left|\frac{1}{\mathcal{F}(s)}\frac{d}{dt}\mathcal{F}(s)\right|. \tag{A6}$$

To prove Eq. A5 we denote $\delta(t) = \mathcal{P}(t) - \hat{\mathcal{P}}(t)$. By subtracting Eq. A2 from Eq. A1b the differential equation

$$\frac{d}{dt}\delta(t) = \gamma_p\big((\mathcal{M}(t) - \mathcal{F}(t)) - \delta(t)\big)$$

is obtained, whose analytic solution writes (see footnote 1)

$$\delta(t) = \underbrace{\delta(0)}_{0} e^{-\gamma_p t} + \gamma_p e^{-\gamma_p t}\int_0^t (\mathcal{M}(u) - \mathcal{F}(u))e^{\gamma_p u}du. \tag{A7}$$

We now find an upper bound for the integral. It is clear that, for all $u \in [0, t]$

$$\frac{|\mathcal{M}(u) - \mathcal{F}(u)|}{\mathcal{M}(u)} \leq \sup_{s\leq t}\left|\frac{\mathcal{M}(s) - \mathcal{F}(s)}{\mathcal{M}(s)}\right|$$

or written otherwise

$$|\mathcal{M}(u) - \mathcal{F}(u)| \leq \mathcal{M}(u)\sup_{s\leq t}\left|\frac{\mathcal{M}(s) - \mathcal{F}(s)}{\mathcal{M}(s)}\right|.$$

By injecting this into Eq. A7 we obtain

$$|\delta(t)| \leq \left(\sup_{s\leq t}\left|\frac{\mathcal{F}(s) - \mathcal{M}(s)}{\mathcal{M}(s)}\right|\right)\gamma_p e^{-\gamma_p t}\int_0^t \mathcal{M}(u)e^{\gamma_p u}du. \tag{A8}$$

Now, by solving Eq. A1b we obtain

$$\mathcal{P}(t) = \mathcal{P}(0)e^{-\gamma_p t} + \gamma_p e^{-\gamma_p t}\int_0^t \mathcal{M}(u)e^{\gamma_p u}du \geq \gamma_p e^{-\gamma_p t}\int_0^t \mathcal{M}(u)e^{\gamma_p u}du,$$

and therefore by dividing both sides of Eq. A8 by $\mathcal{P}(t)$ we obtain

$$\begin{aligned}\frac{|\delta(t)|}{\mathcal{P}(t)} &\leq \left(\sup_{s\leq t}\left|\frac{\mathcal{F}(s) - \mathcal{M}(s)}{\mathcal{M}(s)}\right|\right)\frac{\gamma_p e^{-\gamma_p t}\int_0^t \mathcal{M}(u)e^{\gamma_p u}du}{\mathcal{P}(t)} \\ &\leq \sup_{s\leq t}\left|\frac{\mathcal{F}(s) - \mathcal{M}(s)}{\mathcal{M}(s)}\right|.\end{aligned}$$

which was the first point to prove.

The proof of the second point is quite similar. We denote

$$\delta_{\mathcal{M}}(t) = \mathcal{M}(t) - \mathcal{F}(t).$$

By subtracting $\frac{d}{dt}\mathcal{F}(t)$ from both sides of Eq. A1a we obtain

$$\frac{d}{dt}\delta_{\mathcal{M}}(t) = -\frac{d}{dt}\mathcal{F}(t) - \gamma_m \delta_{\mathcal{M}}(t),$$

whose analytic solution (see footnote 1): is

$$\delta_{\mathcal{M}}(t) = \underbrace{\delta_{\mathcal{M}}(0)}_{0} e^{-\gamma_m t} + e^{-\gamma_m t} \int_0^t e^{\gamma_m u} \frac{d}{dt}\mathcal{F}(u)du.$$

So we have

$$|\delta_{\mathcal{M}}(t)| \leq \left(\sup_{s \leq t} \left| \frac{1}{\mathcal{F}(s)} \frac{d}{dt}\mathcal{F}(s) \right| \right) e^{-\gamma_m t} \int_0^t e^{\gamma_m u} \mathcal{F}(u)du.$$

Now, by solving Eq. A1a we obtain

$$\mathcal{M}(t) = \mathcal{M}(0)e^{-\gamma_m t} + \gamma_n e^{-\gamma_m t} \int_0^t e^{\gamma_m u}\mathcal{F}(u)du \geq \gamma_m e^{-\gamma_m t} \int_0^t e^{\gamma_m u}\mathcal{F}(u)du.$$

So by dividing each side by $\mathcal{M}(t)$ in the previous equation we obtain

$$\frac{|\delta_{\mathcal{M}}(t)|}{\mathcal{M}(t)} \leq \frac{1}{\gamma_m} \sup_{s \leq t} \left| \frac{1}{\mathcal{F}(s)} \frac{d}{dt}\mathcal{F}(s) \right|.$$

This proves Eq. A6. Together with Eq. A5, this proves Proposition 1 in the main text.

A3 Proof of Proposition 2

Because the extensive variables $\mathcal{F}$, $\mathcal{M}$, $\mathcal{P}$ and $\hat{\mathcal{P}}$ are more suitable for mathematically analyzing the gene expression system, we first present a lemma similar to Proposition 2 in the main text, but formulated in terms of these variables. We then provide a lemma that build on the result of this first

lemma, but is expressed in terms of intensive variables (molecular concentrations). Finally, we prove Proposition 2 by some adjustments of the second lemma.

A3.1 A lemma on the extensive variables

Lemma 1. *Under the conditions*

$$\frac{d}{dt}\mathcal{P}(0) = \frac{d}{dt}\mathcal{M}(0) = \frac{d}{dt}\hat{\mathcal{P}}(0) = 0,$$

and assuming that the function $\mathcal{F}$ is of the form

$$\mathcal{F}(t) = \begin{cases} 1 & , \quad t \leq 0, \\ 1+\chi & , \quad \textit{otherwise,} \end{cases}$$

where $\chi >= -1$, and assuming $\gamma_m > \gamma_p$, we have

$$\Delta(t) < \frac{\gamma_p}{\gamma_m - \gamma_p}|\chi|. \tag{A9}$$

Proof. Due to the initial steady-state hypothesis, $\mathcal{P}(0) = \mathcal{M}(0) = \hat{\mathcal{P}}(0) = 1$. For all $t > 0$, Eq. A1 writes

$$\frac{d}{dt}\mathcal{M}(t) = \gamma_m(1 + \chi - \mathcal{M}(t)), \tag{A10a}$$

$$\frac{d}{dt}\mathcal{P}(t) = \gamma_p\left(\mathcal{M}(t) - \mathcal{P}(t)\right). \tag{A10b}$$

Eq. A10a leads to

$$\mathcal{M}(t) = 1 + \chi - \chi e^{-\gamma_m t},$$

and Eq. A10b leads to

$$
\begin{aligned}
\mathcal{P}(t) &= \mathcal{P}(0)e^{-\gamma_p t} + \gamma_p \int_0^t e^{\gamma_p(u-t)}\mathcal{M}(u)du \\
&= e^{-\gamma_p t} + \gamma_p e^{-\gamma_p t}\int_0^t e^{\gamma_p u}\big(1+\chi-\chi e^{-\gamma_m u}\big)du \\
&= e^{-\gamma_p t} + \gamma_p e^{-\gamma_p t}\int_0^t e^{\gamma_p u}\big(1+\chi\big)du - \gamma_p e^{-\gamma_p t}\int_0^t \chi e^{(\gamma_p-\gamma_m)u}du \\
&= e^{-\gamma_p t} + \gamma_p\big(1+\chi\big)e^{-\gamma_p t}\int_0^t e^{\gamma_p u}du - \gamma_p\chi e^{-\gamma_p t}\int_0^t e^{(\gamma_p-\gamma_m)u}du \\
&= e^{-\gamma_p t} + \big(1+\chi\big)e^{-\gamma_p t}(e^{\gamma_p t}-1) - \frac{\gamma_p}{\gamma_p-\gamma_m}\chi e^{-\gamma_p t}(e^{(\gamma_p-\gamma_m)t}-1) \\
&= 1+\chi-\chi e^{-\gamma_p t} - \frac{\gamma_p}{\gamma_p-\gamma_m}\chi(e^{-\gamma_m t}-e^{-\gamma_p t}) \\
&= 1+\chi-\chi e^{-\gamma_p t} + \chi\frac{\gamma_p}{\gamma_m-\gamma_p}(e^{-\gamma_m t}-e^{-\gamma_p t})
\end{aligned}
$$

The reduced system writes

$$\frac{d}{dt}\hat{\mathcal{P}}(t) = \gamma_p\big(1+\chi-\hat{\mathcal{P}}(t)\big),$$

from which we deduce

$$\hat{\mathcal{P}}(t) = 1+\chi-\chi e^{-\gamma_p t}.$$

So, with the definition of $\Delta(t)$ in Eq. A3, and taking into account that $\gamma_m > \gamma_p$, we have

$$\Delta(t) = |\chi|\frac{\gamma_p}{\gamma_m-\gamma_p}\ \frac{e^{-\gamma_p t}-e^{-\gamma_m t}}{1+\chi-\chi e^{-\gamma_p t}+\chi\frac{\gamma_p}{\gamma_m-\gamma_p}(e^{-\gamma_m t}-e^{-\gamma_p t})}.$$

For $\chi = 0$, Lemma 1 is obviously true. In what follows we consider the case $\chi \neq 0$. To simplify this formula we define $z = e^{-\gamma_p t}$ and $\alpha = \gamma_m/\gamma_p$. Note that when t varies between 0 and $+\infty$, z varies between 1 and 0, and that by hypothesis $\alpha > 1$. Moreover,

$$e^{-\gamma_m t} = e^{-(\gamma_m/\gamma_p)\gamma_p t} = (e^{-\gamma_p t})^{\gamma_m/\gamma_p} = z^\alpha$$

and

$$\frac{\gamma_p}{\gamma_m - \gamma_p} = \frac{1}{\gamma_m/\gamma_p - 1} = \frac{1}{\alpha - 1}.$$

We now rewrite $\Delta(t)$ as a function of the variable z:

$$\begin{aligned}
\Delta(z) &= |\chi|\big(\frac{1}{\alpha-1}\big)\frac{z - z^\alpha}{1 + \chi - z\chi + \chi\frac{1}{\alpha-1}(z^\alpha - z)} \\
&= \frac{|\chi|}{\chi}\big(\frac{1}{\alpha-1}\big)\frac{z - z^\alpha}{\chi^{-1} + 1 - z + \frac{1}{\alpha-1}(z^\alpha - z)} \\
&= \frac{|\chi|}{\chi}\frac{z - z^\alpha}{(\alpha-1)(\chi^{-1}+1) - (\alpha-1)z + (z^\alpha - z)} \\
&= \frac{|\chi|}{\chi}\frac{z - z^\alpha}{(\alpha-1)(\chi^{-1}+1) - \alpha z + z^\alpha} \stackrel{def.}{=} \frac{|\chi|}{\chi}\frac{num(z)}{den(z)}
\end{aligned}$$

If $\chi > 0$, then differentiating the functions num and den and looking for global maxima and minima of these functions leads to the following inequalities for all $z \in [0, 1]$:

$$\begin{aligned}
0 < num(z) < (\alpha-1)\alpha^{-\frac{\alpha}{\alpha-1}}, \\
0 < \chi^{-1}(\alpha-1) < den(z),
\end{aligned}$$

from which we deduce that

$$\sup_{t\geq 0}\Delta(t) = \sup_{z\in[0,1]}\Delta(z) \leq |\chi|\alpha^{-\frac{\alpha}{\alpha-1}}. \tag{A11}$$

Since $\alpha > 1$, this inequality can be relaxed to :

$$\Delta(t) \leq |\chi|\alpha^{-\frac{\alpha}{\alpha-1}} < |\chi|\alpha^{-1} < |\chi|(\alpha-1)^{-1} = \frac{\gamma_p}{\gamma_m - \gamma_p}|\chi|,$$

which proves Lemma 1 for $\chi > 0$.

We now consider the case $\chi \in [-1, 0[$. Notice that in this case $|\chi|/\chi < 0$, so that proving Lemma 1 comes to proving that, for all $z \in [0, 1]$,

$$\frac{\chi}{\alpha-1} \leq \frac{num(z)}{den(z)} \leq 0. \tag{A12}$$

Since

$$\frac{|\chi|}{\chi}\frac{num(z)}{den(z)} = \Delta(z) \geq 0,$$

we have

$$\frac{num(z)}{den(z)} \leq 0,$$

In what follows we prove the second part of Eq. A12. For any $z \in [0,1]$, the function

$$g : u \longmapsto \frac{z-u}{(1+\chi^{-1})(\alpha-1)-\alpha z+u}$$

is an increasing function of u, since

$$\begin{aligned} g'(u) &= \frac{-\left((1+\chi^{-1})(\alpha-1)-\alpha z+u\right)-(z-u)}{\left((1+\chi^{-1})(\alpha-1)-\alpha z+u\right)^2} \\ &= \frac{(\alpha-1)(-1-\chi^{-1})+(\alpha-1)z}{\left((1+\chi^{-1})(\alpha-1)-\alpha z+u\right)^2} \\ &= \frac{(\alpha-1)(z-1-\chi^{-1})}{\left((1+\chi^{-1})(\alpha-1)-\alpha z+u\right)^2} \geq \frac{(\alpha-1)z}{\left((1+\chi^{-1})(\alpha-1)-\alpha z+u\right)^2} \geq 0. \end{aligned}$$

This indicates that, for all $z \in [0,1]$, $g(z^\alpha) > g(0)$, which translates into

$$\frac{z}{(1+\chi^{-1})(\alpha-1)-\alpha z} < \frac{z-z^\alpha}{(1+\chi^{-1})(\alpha-1)-\alpha z+z^\alpha} = \frac{num(z)}{den(z)}. \quad \text{(A13)}$$

Now, notice that the function

$$h : u \longmapsto \frac{u}{(1+\chi^{-1})(\alpha-1)-\alpha u}$$

is a decreasing function of u since

$$h'(u) = \frac{(1+\chi^{-1})(\alpha-1)}{\left((1+\chi^{-1})(\alpha-1)-\alpha u\right)^2} < 0.$$

So, for all $z \in [0,1]$, $h(z) \geq h(1)$, i.e.

$$\frac{z}{(1+\chi^{-1})(\alpha-1)-\alpha z} \geq \frac{1}{(1+\chi^{-1})(\alpha-1)-\alpha} = \frac{\chi}{\alpha-1-\chi} > \frac{\chi}{\alpha-1}. \tag{A14}$$

Equating Eq. A13 and Eq. A14 yields

$$\frac{num(z)}{den(z)} \geq \frac{z}{(1+\chi^{-1})(\alpha-1)-\alpha z} \geq \frac{1}{(1+\chi^{-1})(\alpha-1)-\alpha} = \frac{\chi}{\alpha-1-\chi} > \frac{\chi}{\alpha-1},$$

which proves the lemma.

□

Note that, although we came to the upper bound of Eq. A9 through a series of successive upper bounds, it appears to be almost optimal for moderate values of χ and large values of γ_m/γ_p, as shown in Fig. A1.

A3.2 A lemma on intensive variables

The lemma proven in this section considers concentrations, and it is therefore closer to Proposition 2 in the main text. However, it considers a variable $\hat{p}_2$ that is slightly different from the variable $\hat{p}$ in the main text, for reasons that will become clear in the next section, where Proposition 2 is proven.

Suppose that μ is constant, so that Eq. 1 in the main text can be written as

$$\frac{d}{dt}m(t) = \kappa_m\, f(t) - \gamma'_m\, m(t), \quad m(0) = m_0, \tag{A15a}$$

$$\frac{d}{dt}p(t) = \kappa_p\, m(t) - \gamma'_p\, p(t), \quad p(0) = p_0. \tag{A15b}$$

where

$$\gamma'_m = \gamma_m + \mu, \quad \gamma'_p = \gamma_p + \mu.$$

We then introduce the variable $\hat{p}_2(t)$, defined by the equation

$$\frac{d}{dt}\hat{p}_2(t) = \frac{\kappa_m\kappa_p}{\gamma'_m}\, f(t) - \gamma'_p\, \hat{p}_2(t), \quad \hat{p}_2(0) = \hat{p}_{2,0}. \tag{A16}$$

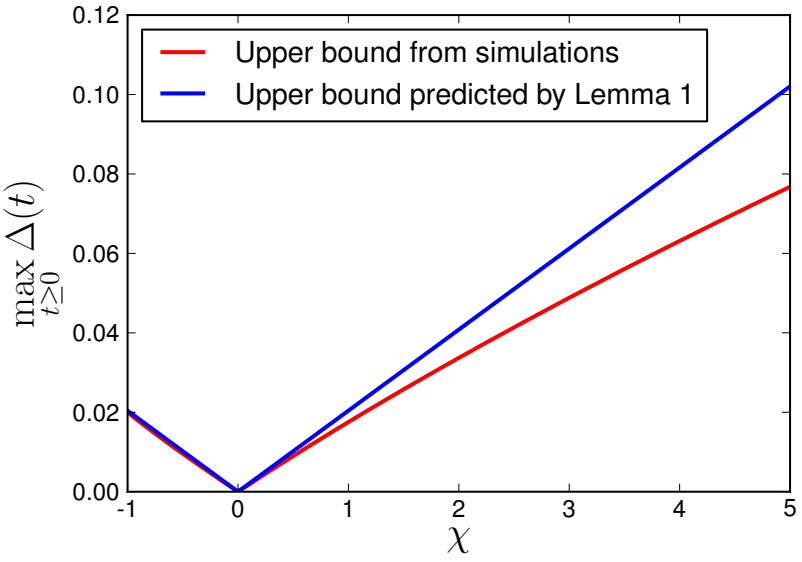


Figure A1: **Predicted and actual upper bounds of the model reduction error.** Eq. A1 and Eq. A2 were computationally simulated in the conditions specified in Lemma 1, using various values of χ, in order to compute the upper bound of $\Delta(t)$. For this example we used $\gamma_p = 0.01$ and $\gamma_m = 0.5$. The result is compared with the upper bound predicted by the same lemma.

Lemma 2. *Under the conditions*

$$\frac{d}{dt}p(0) = \frac{d}{dt}m(0) = \frac{d}{dt}\hat{p}_2(0) = 0,$$

we have for all $t \geq 0$,

$$\frac{|\hat{p}_2(t) - p(t)|}{p(t)} < |\chi| \frac{\gamma'_p}{\gamma'_m - \gamma'_p}. \tag{A17}$$

Proof. Eq. A15a and Eq. A16 are of the same form as the systems of extensive variables in Eq. 6 and Eq. 7 of the main text. Therefore, the proof of lemma 2 is in all points similar to that of Lemma 1. □

A3.3 Proof of Proposition 2

With μ constant, and using γ_p' as defined above, the differential equation defining $\hat{p}$ writes

$$\frac{d}{dt}\hat{p}(t) = \frac{\kappa_m \kappa_p}{\gamma_m} f(t) - \gamma_p' \hat{p}(t). \tag{A18}$$

We see that $\hat{p}(t)$ and $\hat{p}_2(t)$ are defined by similar equations whose synthesis rates differ by a factor γ_m'/γ_m. As Eq. A16 and Eq. A18 describe linear differential equations, they have the following property (which can be proven by solving the systems analytically). Under the condition

$$\frac{d}{dt}\hat{p}(0) = \frac{d}{dt}\hat{p}_2(0) = 0,$$

we have

$$\frac{\hat{p}(t)}{\hat{p}_2(t)} = \frac{\gamma_m'}{\gamma_m} = 1 + \frac{\mu}{\gamma_m},$$

which leads to

$$\frac{\hat{p}(t) - \hat{p}_2(t)}{\hat{p}_2(t)} = \frac{\mu}{\gamma_m}.$$

Finally, the following computations use Lemma 2 to prove Proposition 2:

$$
\begin{aligned}
\frac{|p(t)-\hat{p}(t)|}{p(t)} &= \frac{|p(t)-\hat{p}_2(t)+\hat{p}_2(t)-\hat{p}(t)|}{p(t)} \\
&= \left|\frac{p(t)-\hat{p}_2(t)}{p(t)} + \frac{\hat{p}_2(t)-\hat{p}(t)}{p(t)}\right| \\
&= \left|\frac{p(t)-\hat{p}_2(t)}{p(t)} + \frac{\hat{p}_2(t)}{p(t)}\frac{\hat{p}_2(t)-\hat{p}(t)}{\hat{p}_2(t)}\right| \\
&= \left|\frac{p(t)-\hat{p}_2(t)}{p(t)} + \left(1-\frac{p(t)-\hat{p}_2(t)}{p(t)}\right)\left(-\frac{\mu}{\gamma_m}\right)\right| \\
&= \left|\frac{p(t)-\hat{p}_2(t)}{p(t)}(1+\frac{\mu}{\gamma_m}) - \frac{\mu}{\gamma_m}\right| \\
&\leq \left|\frac{p(t)-\hat{p}_2(t)}{p(t)}\right|(1+\frac{\mu}{\gamma_m}) + \frac{\mu}{\gamma_m} \\
&\leq |\chi|\frac{\gamma_p'}{\gamma_m'-\gamma_p'}\left(1+\frac{\mu}{\gamma_m}\right) + \frac{\mu}{\gamma_m} \\
&= |\chi|\frac{\gamma_p+\mu}{\gamma_m-\gamma_p}\left(1+\frac{\mu}{\gamma_m}\right) + \frac{\mu}{\gamma_m}.
\end{aligned}
$$

A4 A filter-theoretical view of the model reduction error

Equation systems A1 and A2 describe linear low-pass filters: if $\mathcal{F}(t)$ is a sinusoidal signal of the form $\mathcal{F}(t) = A\cos(\omega t)$, with A a constant, then the variable $\mathcal{M}(t)$, once a stationary behavior is reached, possibly after a transition period, will be a sinusoidal signal of amplitude $AH(\omega)$, where

$$H(\omega) = \frac{1}{\sqrt{(\frac{\omega}{\gamma_m})^2+1}}.$$

This shows that the system is a low-pass filter with cut-off angular frequency (CAF) $\omega_c = \gamma_m$, meaning that all components of the spectrum of the input signal $\mathcal{F}$ which correspond to an angular frequency larger than γ_m will

be severely damped in the output $\mathcal{M}$. In a schematic view, the spectrum of $\mathcal{M}$ can be seen as a truncated version of the spectrum of $\mathcal{F}$.

The same considerations hold for the systems represented by Eq. A1 and Eq. A2. Here, one can see $\mathcal{P}(t)$ as the image of $\mathcal{F}(t)$ after passing the signal through two low-pass band filters with CAFs γ_m and γ_p, according to the two-step system of Eq. A1, while in the one-step system of Eq. A2, $\hat{\mathcal{P}}(t)$ is the image of $\mathcal{F}(t)$ after filtering once with a CAF γ_p (as illustrated in Fig. A2).

Let us now study different situations, illustrated in Fig. A3 :

- If γ_m lies above the spectrum of $\mathcal{F}(t)$ (Fig. A3A), then $\mathcal{M}(t)$ is a non-filtered version of $\mathcal{F}(t)$, *i.e.* $\mathcal{M}(t) \simeq \mathcal{F}(t)$. As a consequence, $\mathcal{P} \simeq \hat{\mathcal{P}}$, independently of the value of γ_p, since they are the image of roughly the same signal passing through a filter with CAF γ_p.

- If γ_m lies in the spectrum of $\mathcal{F}(t)$, then $\mathcal{M}(t)$ is a filtered version of $\mathcal{F}(t)$.

  - If $\gamma_p < \gamma_m$ (Fig. A3B), then the spectrum of $\mathcal{M}(t)$ is further truncated at CAF γ_p to give $\hat{\mathcal{P}}$. Both $\mathcal{P}$ and $\hat{\mathcal{P}}$ are then filtered responses to the signal $\mathcal{F}$ with the same CAF γ_p. Hence, we have $\mathcal{P} \simeq \hat{\mathcal{P}}$.

  - If $\gamma_p > \gamma_m$ (Fig. A3C), then $\mathcal{P}$ and $\hat{\mathcal{P}}$ are *in fine* filtered responses to the signal $\mathcal{F}$ with CAFs γ_m and γ_p, respectively, and they may differ significantly. The difference can be estimated using the part of the spectrum of $\mathcal{F}$ that lies between γ_m and γ_p.

From these considerations we can conclude, qualitatively, that the model reduction will be valid if the characteristic time constant of $\mathcal{F}$ is small compared to γ_m, or if $\gamma_p \ll \gamma_m$. This is consistent with Propositions 1 and 2 in the main text.

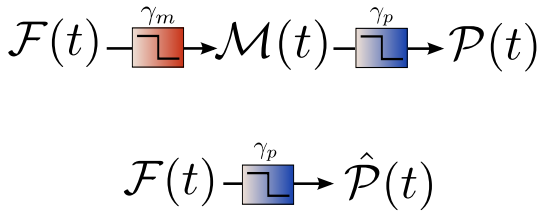


Figure A2: **Filter-theoretical representation of the two-step and the one-step models.** In the two-step model, the output $\mathcal{P}(t)$ is the result of filtering $\mathcal{F}(t)$ twice, through filters with cut-off frequencies γ_m and γ_p, respectively. In the one-step system, the signal $\hat{\mathcal{P}}(t)$ is the filtered response to $\mathcal{F}(t)$ involving a single filter with cut-off frequency γ_p.

A5 Proof of Eq. 2.15 in main text

Eq. 2.14c yields

$$Q(t) = \frac{1}{\kappa_r}\left(\frac{d}{dt}R(t) + \gamma_r R(t)\right),$$

and therefore

$$\frac{d}{dt}Q(t) = \frac{1}{\kappa_r}\left(\frac{d^2}{dt^2}R(t) + \gamma_r \frac{d}{dt}R(t)\right).$$

In the same way, we have by Eq. 2.14b

$$\begin{aligned} N(t) &= \frac{1}{\kappa_q}\left(\frac{d}{dt}Q(t) + (\gamma_r + \kappa_r)Q(t)\right) \\ &= \frac{1}{\kappa_r\kappa_q}\left(\frac{d^2}{dt^2}R(t) + (2\gamma_r + \kappa_r)\frac{d}{dt}R(t) + \gamma_r(\gamma_r + \kappa_r)R(t)\right), \end{aligned}$$

and

$$\frac{d}{dt}N(t) = \frac{1}{\kappa_r\kappa_q}\left(\frac{d^3}{dt^3}R(t) + (2\gamma_r + \kappa_r)\frac{d^2}{dt^2}R(t) + \gamma_r(\gamma_r + \kappa_r)\frac{d}{dt}R(t)\right).$$

Finally, Eq. 2.14a gives

$$\begin{aligned} F(t) &= \frac{1}{\kappa_n}\left(\frac{d}{dt}N(t) + \gamma_n N(t)\right) \\ &= \frac{\gamma_r(\gamma_r + \kappa_r)\gamma_n}{\kappa_n\kappa_q\kappa_r}\left(R(t) + a\frac{d}{dt}R(t) + b\frac{d^2}{dt^2}R(t) + c\frac{d^3}{dt^3}R(t)\right), \end{aligned}$$

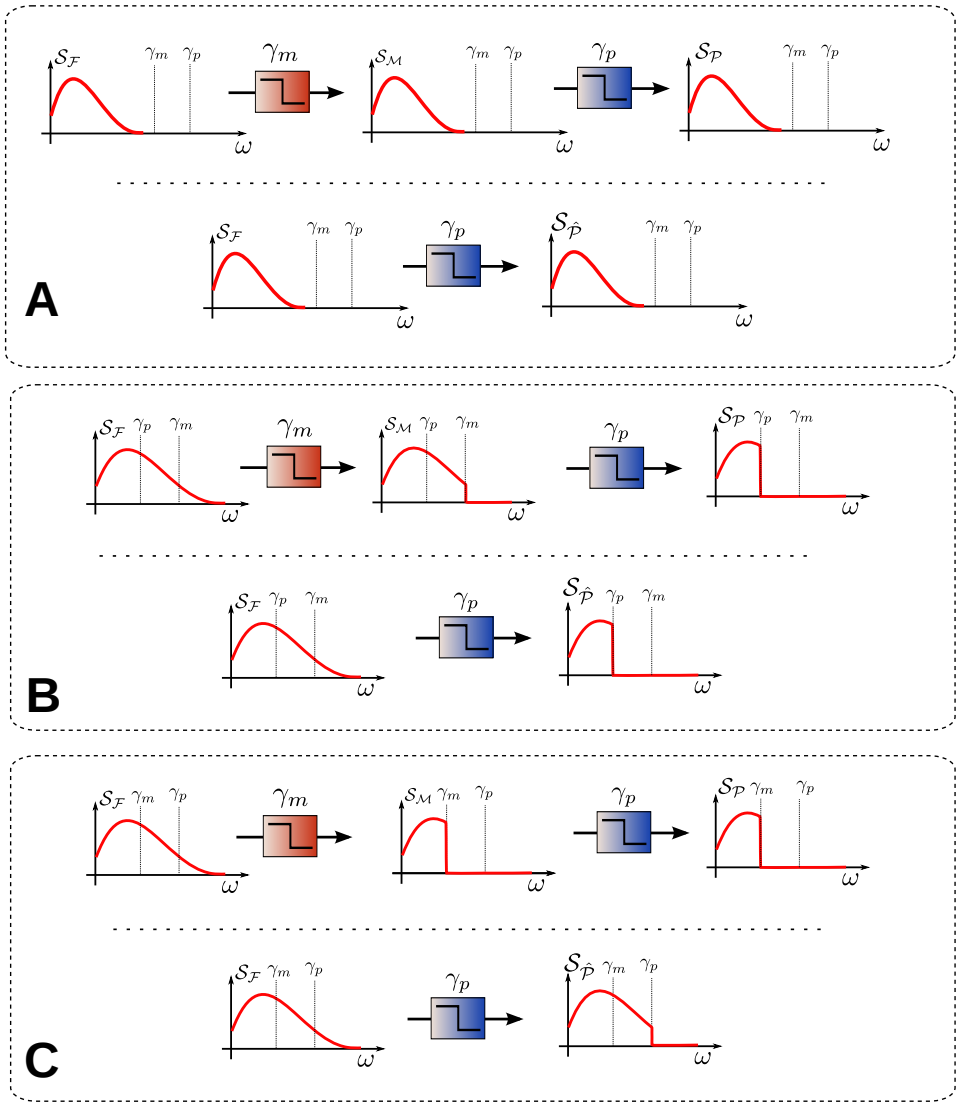


Figure A3: **Output spectra of the one-step and two-step filters**. $\mathcal{S}_{\mathcal{F}}$, $\mathcal{S}_{\mathcal{M}}$, $\mathcal{S}_{\mathcal{P}}$, and $\mathcal{S}_{\hat{\mathcal{P}}}$ denote the spectra of $\mathcal{F}$, $\mathcal{M}$, $\mathcal{P}$, and $\hat{\mathcal{P}}$ respectively. **A:** The spectrum of $\mathcal{F}$ lies entirely below γ_m, in which case $\mathcal{P}$ and $\hat{\mathcal{P}}$ have similar spectra whatever the value of γ_p. **B:** The spectrum of $\mathcal{F}$ is cut above γ_p in both the one-step and two-step models, so that $\mathcal{P}$ and $\hat{\mathcal{P}}$ have similar spectra, independently of γ_p. **C**: $\gamma_p > \gamma_m$. In this last case, the ouputs $\mathcal{P}$ and $\hat{\mathcal{P}}$ differ.

where a, b and c are defined like in the main text. Assuming that $R(t) = kI(t)$ (where the constant k is unknown) and denoting

$$F_e(t) = \frac{\kappa_n \kappa_q \kappa_r}{\gamma_r(\gamma_r + \kappa_r)\gamma_n k} F(t),$$

it follows that F_e is proportional to F and verifies Eq. 2.15 in the main text.

A6 Propagation of the model reduction error

In this section we prove our claim from Sec. 2.3.6 that if protein p_1 regulates the activity of the gene coding for protein p_2 through the regulation function

$$f_2\left(p_1(t)\right) = \frac{p_1(t)^a}{K^a + p_1(t)^a}, \tag{A19}$$

then the condition

$$\min_{t\geq 0} p_1(t) \geq K\sqrt[a]{a-1} \tag{A20}$$

is sufficient for ensuring that the error on protein p_1 will result in a not-larger error on the p_1-dependant protein p_2.

We will first show that under this condition a variation of $x\%$ in $p_1(t)$ will result in a variation of at most $x\%$ in $f_2\left(p_1(t)\right)$ (Lemma 3). We will then show that a relative variation of at most $x\%$ in f_2 will lead to a variation of at most $x\%$ in the prediction of the protein concentration $p_2(t)$ (Lemma 4). At the end of the section we reformulate this condition and loosen it to $(\min p_1(t) > 4K/3)$.

For convenience we define a class of regulation functions f which *conserve the relative error* on p:

Definition 1. *We denote by $\mathcal{C}$ the class of functions $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ defined for $p > 0$, such that for all p_1, p_2*

$$\frac{|f(p_1) - f(p_2)|}{f(p_1)} \leq \frac{|p_1 - p_2|}{p_1}.$$

We will prove that the functions in $\mathcal{C}$ ensure that the relative error on p_1 will not be amplified on its target p_2. But we first show that the Hill function in Eq. A19 belongs to $\mathcal{C}$ under the condition that for all t, $p_1(t) > K\sqrt[a]{a-1}$.

Lemma 3. *The restriction of f_2 (as defined in Eq. A19) to the interval $p > K\sqrt[a]{1-a}$ belongs to the class $\mathcal{C}$.*

Proof. We consider x, y with $0 < x \leq y$ and we will show that the lemma is true for the couples (x, y) and (y, x), i.e.

$$\frac{|f_2(x) - f_2(y)|}{f_2(x)} = \frac{f_2(y) - f_2(x)}{f_2(x)} \leq \frac{y - x}{x} = \frac{|x - y|}{x}, \tag{A21}$$

and

$$\frac{|f_2(y) - f_2(x)|}{f_2(y)} = \frac{f_2(y) - f_2(x)}{f_2(y)} \leq \frac{y-x}{y} = \frac{|y-x|}{y} \tag{A22}$$

For Eq. A21, we first notice that

$$\frac{f_2(y) - f_2(x)}{f_2(x)} = \frac{K^a(y^a - x^a)}{x^a(K^a + y^a)} = \frac{K^a \int_x^y a u^{a-1} du}{x^a(K^a + y^a)} < \frac{K^a a y^{a-1}(y-x)}{x^a(K^a + y^a)} = \frac{K^a a y^{a-1}}{x^{a-1}(K^a + y^a)} \frac{y-x}{x}.$$

We denote $\phi(y) = \frac{K^a a y^{a-1}}{x^{a-1}(K^a + y^a)}$. It is clear from the above computations that Eq. A21 is verified when $\phi(y) \leq 1$. If $a < 1$, this will always be the case, as

$$\phi(y) = a \frac{K^a}{(K^a + y^a)} \left(\frac{y}{x}\right)^{a-1} < 1 * 1 * 1 = 1.$$

For $a > 1$ we have

$$\frac{d\phi}{dy}(y) = \frac{1}{x^{a-1}} \frac{a K^a y^{a-2} \left(a K^a - K^a - y^a\right)}{k^{2a} + 2K^a y^a + y^{2a}},$$

which shows that ϕ is maximal for $y_{max} = k(a-1)^{1/a}$ and its value is then

$$\phi(y_{max}) = \left(\frac{K \sqrt[a]{a-1}}{x}\right)^{a-1},$$

therefore $\phi(y) < 1$ for all y, if $x \geq K \sqrt[a]{a-1}$.

For Eq. A22,

$$\frac{f_2(y) - f_2(x)}{f_2(y)} = \frac{K^a(y^a - x^a)}{y^a(K^a + x^a)} = \frac{K^a \int_x^y a u^{a-1} du}{y^a(K^a + x^a)} < \frac{K^a a y^{a-1}(y-x)}{y^a(K^a + x^a)} = \frac{K^a a}{(K^a + x^a)} \frac{y-x}{y},$$

so Eq. A22 is verified when

$$\frac{K^a a}{(K^a + x^a)} \leq 1$$

which also leads to $x \geq K \sqrt[a]{a-1}$ □

Note that a relative error on $f_2(t)$ is equivalent to the same relative error on its extensive equivalent $F_2(t) = V(t)f_2(t)$ or on the its standardized version $\mathcal{F}_2(t) = F_2(t)/F_2(0)$. The same way, a relative error on p_2 is equivalent to a relative error on $P_2(t) = V(t)p_2(t)$ or $\mathcal{P}_2(t)$. Therefore, we will show that a relative prediction error on $\mathcal{F}_2(t)$ will lead to a non-larger relative prediction error on $\mathcal{P}_2(t)$ (and therefore on $p_2(t)$).

Lemma 4. *We consider the two-step expression model of the protein p_2, expressed with global standardized variables:*

$$\frac{d}{dt}\mathcal{M}_2(t)(t) = \gamma_m \left(\mathcal{F}_2(t) - \mathcal{M}_2(t)(t)\right), \tag{A23}$$

$$\frac{d}{dt}\mathcal{P}_2(t) = \gamma_p \left(\mathcal{M}_2(t) - \mathcal{P}_2(t)\right), \tag{A24}$$

$$\mathcal{P}_2(0) = \mathcal{M}_2(0) = \mathcal{F}_2(0), \tag{A25}$$

where $\mathcal{M}_2$ represents the standardized total quantity of mRNA and $\mathcal{P}_2$ the standardized total quantity of protein p_2.

We call $\mathcal{P}_{2,a}(t)$ and $\mathcal{P}_{2,b}(t)$ the profiles of $\mathcal{P}_2(t)$ obtained using two different input functions for $F_2(t)$: $\mathcal{F}_{2,a}(t)$ and $\mathcal{F}_{2,b}(t)$. Then we have

$$\sup_{t\geq 0} \frac{|\mathcal{P}_{2,a}(t) - \mathcal{P}_{2,b}(t)|}{\mathcal{P}_{2,a}(t)} < \sup_{t\geq 0} \frac{|\mathcal{F}_{2,a}(t) - \mathcal{F}_{2,b}(t)|}{F_{2,a}(t)}.$$

Proof. We will only need to show that

$$\sup_{t\geq 0} \frac{|\mathcal{M}_{2,a}(t) - \mathcal{M}_{2,b}(t)|}{\mathcal{M}_{2,a}(t)} < \sup_{t\geq 0} \frac{|\mathcal{F}_{2,a}(t) - \mathcal{F}_{2,b}(t)|}{\mathcal{F}_{2,a}(t)}. \tag{A26}$$

Since $\mathcal{P}_2$ depends of $\mathcal{M}_2$ the same way that $\mathcal{M}_2$ depends on $\mathcal{F}_2$, the proof will be exactly of the same form for the inequality

$$\sup_{t\geq 0} \frac{|\mathcal{F}_{2,a}(t) - \mathcal{F}_{2,b}(t)|}{\mathcal{F}_{2,a}(t)} < \sup_{t\geq 0} \frac{|\mathcal{M}_{2,a}(t) - \mathcal{M}_{2,b}(t)|}{\mathcal{M}_{2,a}(t)}, \tag{A27}$$

and therefore will be omitted. Eq. A26 and A27 together prove the lemma.

The variables $\mathcal{M}_{2,a}$ and $\mathcal{M}_{2,b}$ verify the following systems :

$$\frac{d}{dt}\mathcal{M}_{2,a}(t) = \gamma_m\big(\mathcal{F}_{2,a}(t) - \mathcal{M}_{2,a}(t)\big) \tag{A28}$$

$$\frac{d}{dt}\mathcal{M}_{2,b}(t) = \gamma_m\big(\mathcal{F}_{2,b}(t) - \mathcal{M}_{2,b}(t)\big) \tag{A29}$$

We define

$$\delta M_2(t) = \mathcal{M}_{2,a}(t) - \mathcal{M}_{2,b}(t)$$
$$\delta\mathcal{F}_2(t) = \mathcal{F}_{2,a}(t) - \mathcal{F}_{2,b}(t)$$

By substracting Eq. A29 to Eq. A28 we obtain the differential equation

$$\frac{d}{dt}\delta\mathcal{M}_2(t) = \gamma_m\big(\delta\mathcal{F}_2(t) - \delta\mathcal{M}_2(t)\big) \tag{A30}$$

By analytically solving Eq. A30 and A28 we obtain

$$\frac{\mathcal{M}_{2,a}(t) - \mathcal{M}_{2,b}(t)}{\mathcal{M}_{2,a}(t)} = \frac{\delta\mathcal{M}_2(t)(t)}{\mathcal{M}_{2,a}(t)} = \frac{\delta\mathcal{M}_2(0)e^{-\gamma t} + \gamma e^{-\gamma t}\int_0^t \delta\mathcal{F}_2(u)e^{\gamma u}du}{\mathcal{F}_{2,a}(0)e^{-\gamma t} + \gamma e^{-\gamma t}\int_0^t \mathcal{F}_{2,a}(u)e^{\gamma u}du} \tag{A31}$$

We have

$$|\delta\mathcal{M}_2(0)| = |\mathcal{M}_{2,a}(0) - \mathcal{M}_{2,b}(0)| = |\mathcal{F}_{2,a}(0) - \mathcal{F}_{2,b}(0)| \leq \mathcal{F}_{2,a}(0)\left(\sup_{t\geq 0}\frac{|\mathcal{F}_{2,a}(t) - F_2(t)|}{\mathcal{F}_{2,a}(t)}\right),$$

and for all $u < t$

$$-\mathcal{F}_{2,a}(u)\left(\sup_{t\geq 0}\frac{|\mathcal{F}_{2,a}(t) - \mathcal{F}_{2,b}(t)|}{\mathcal{F}_{2,a}(t)}\right) \leq \delta\mathcal{F}_2(t) \leq \mathcal{F}_{2,a}(u)\left(\sup_{t\geq 0}\frac{|\mathcal{F}_{2,a}(t) - F_2(t)|}{\mathcal{F}_{2,a}(t)}\right)$$

therefore, from Eq. A31 follows

$$\frac{|\mathcal{M}_{2,a}(t) - \mathcal{M}_{2,b}(t)|}{\mathcal{M}_{2,a}(t)} \leq \sup_{t\geq 0}\frac{|\mathcal{F}_{2,a}(t) - \mathcal{F}_{2,b}(t)|}{\mathcal{F}_{2,a}(t)}$$

which proves the result. □

A6.1 Corollaries

Notice that

$$f_2(K\sqrt[a]{a-1}) = \frac{K^a(a-1)}{K^a(a-1)+K^a} = \frac{a-1}{a} = 1 - 1/a.$$

Therefore, the condition

$$\min_{t\geq 0} p_1(t) \geq K\sqrt[a]{a-1}$$

can be reformaluted as

$$\min_{t\geq 0} f_2(t) \geq 1 - 1/a,$$

which means thatthe activity of the gene coding for p_2 remains above $1 - 1/a$ of its highest possible value at all t.

Moreover, a quick analysis of the expression $\sqrt[a]{a-1}$ shows that this expression admits a maximum when a varies between 0 and $+\infty$. It seems that this maximum has no simple expression. Numerically we found it to be approximately 1.32110, i.e. smaller that 4/3. Therefore the condition on $p_1(t)$ can be generalized for all values of a. We obtain the quick tule of thumbs

$$\min_{t\geq 0} p_1(t) \geq 4K/3.$$

A7 Background correction of the data

In this part we discuss in more detail the processing of the data which led to the corrected fluorescence curves reported in Fig. 2.2 and Fig. 2.3 of the main text.

A non-negligible part of the experimentally measured fluorescence signal is not due to the GFP fluorescence. Thus, the measured fluorescence does not give direct information on the quantity of GFP proteins inside a cell population. Among the components of the signal that are not related to the quantity of GFP inside the cells, some can be removed using control wells:

- A small, uncorrelated noise is due to perturbations at the instrument level. We found that a part of this noise was the same for all wells of the

plate and could therefore be removed by subtracting from each well of the plate the fluorescence signals measured on a well containing only growth medium and no bacteria. This operation considerably reduces the noise but does not affect much the value of the fluorescence intensities, so that the result of this operation will still be called *measured fluorescence* in what follows.

- The natural fluorescence of the bacterial population (autofluorescence), which is mainly due to the production of fluorescent proteins, like flavins, by the cells.

The autofluorescence is generally non-negligible and its removal demands special care. We show in Fig. A4 and Fig. A5 the fluorescence and absorbance measured on replicate wells containing wild-type *E. coli* and a strain containing a promoterless plasmid, respectively, as explained in the main text. No noticeable difference exists between the strains (in other experiments, the data perfectly superimpose) and we found that both could be used for autofluorescence subtraction, giving approximately the same results. It appears that the fluorescent curves obtained from replicate wells differ by a small multiplicative factor, possibly due to small differences in culture volume. These curves, when normalized by the area under the curve, exhibit good reproducibility.

Fig. A6 and Fig. A7 show the absorbance and fluorescence obtained for strains carrying reporter plasmids of the genes *acs* and *crp*. The signal in the strain carrying a p*crp-gfp* reporter plasmid is well above the autofluorescence background. Removing the autofluorescence from each well gives the corrected fluorescence signals reported in Fig. 2.2 of the main text. For the strain transformed with the p*acs-gfp* reporter plasmid, the observed fluorescence signal does not differ from the autofluorescence signal during growth on glucose, which indicates that the expression of *acs* is negligible during this period. The induction of the gene upon glucose exhaustion is clearly noticeable, as the fluorescence signal separates from the autofluorescence signal at $t = 700$ min. For these wells, the removal of autofluorescence is more delicate and requires special care, such as synchronization of the curves by means of the measured absorbance signals. This results in the corrected fluorescence curves of Fig. 2.3 in the main text.

A8 Using an inversion method to evaluate the promoter activity

In the main text we showed that the time profile of the synthesis rate of GFP can be computed from the fluorescence signal. To this end, the fluorescence signal needs to be approximated through regression splines, which allows the time-derivatives of the signal to be computed. An inconvenience of this approach is that, due to noise, the resulting evaluation of the promoter activity may exhibit biologically irrelevant, typically negative values. In (Bansal et al., 2012) the authors describe another method, very similar in principle, but which enables one to directly formulate constraints on the promoter activity. The main idea is to optimize a parametric time profile of the promoter activity, so that the resulting fluorescence, predicted using the GFP expression model described in the main text, will optimally fit the data. We applied this method, with some adjustments, notably to allow non-null initial values for the variables, and found that it led to the same results as the regression spline methods, as shown in Fig. A8 and Fig. A9.

A9 Software

All numerical computations in the main text and the Supplementary Information have been performed using the Python scientific computing library `Scipy` (Jones et al., 2001). In addition, our implementation of the inversion method described in section A8 makes use of the Python optimization package `cvxopt` (Andersen et al., 2012) (see also Chapter 3 of this thesis).

B1 Sensitivity of the estimation results to the value of the parameter ϵ

Section 3.2.1 of the main text describes the practical implementation of the regularization parameter λ, involving the introduction of a parameter ϵ in the discrete differentiation matrix $\mathbf{L_u}$. When the value of ϵ is not carefully cho-

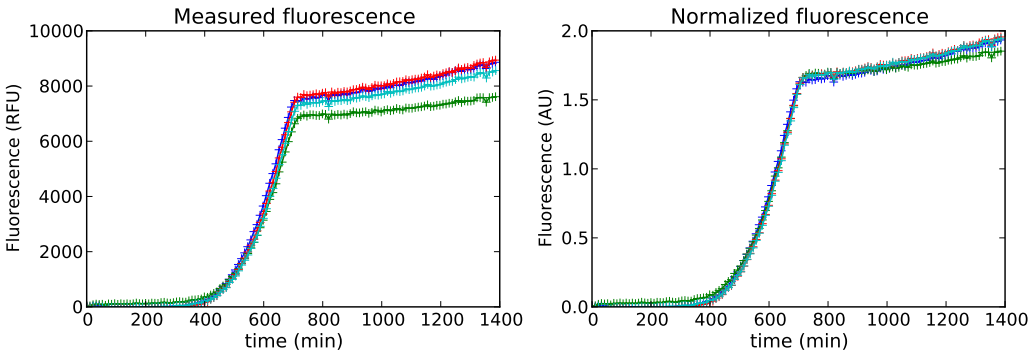


Figure A4: **Fluorescence signal measured for the *E. coli* BW25113 strain.** Shown is the measured fluorescence (equal to the autofluorescence of the cells), before (left) and after (right) normalization by the area under the curve.

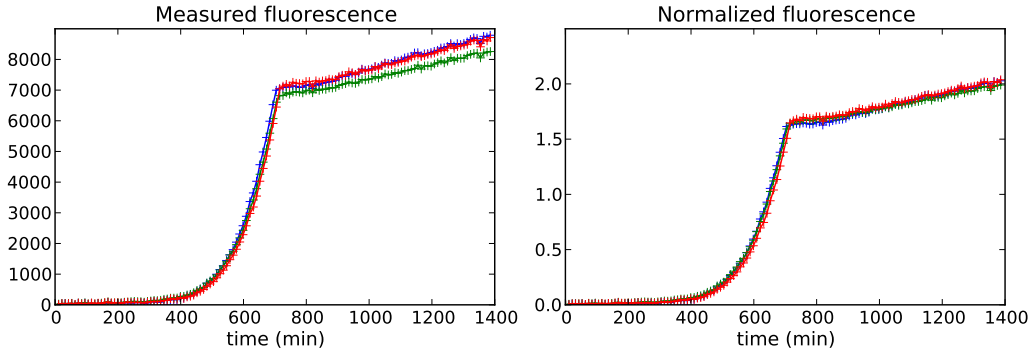


Figure A5: **Fluorescence signal measured for the BW25113 strain carrying a promoterless reporter plasmid**. Shown is the measured fluorescence (equal to the autofluorescence of the cells), before (left) and after (right) normalization by the area under the curve.

sen, it introduces an unwanted penalty on the initial values of the variable to estimate. This may modify the range of values of λ for which the matrix is invertible and thus influence the estimation results. In this section we investigate the sensitivity of the estimation results to the value of ϵ and suggest how to proceed in finding an appropriate value.

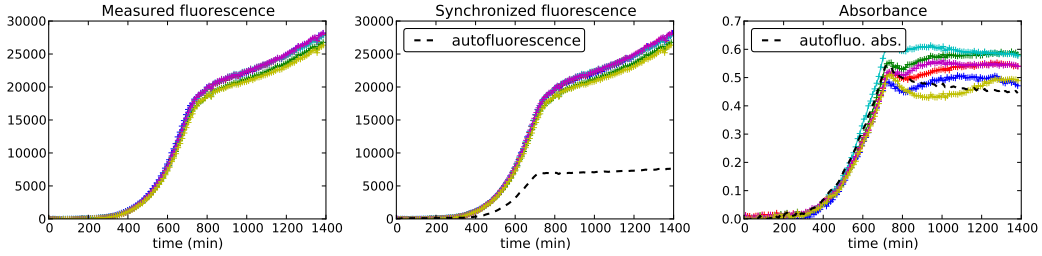


Figure A6: **Fluorescence and absorbance signals measured for the BW25113 strain carrying a p*crp-gfp* reporter plasmid**. The autofluorescence and absorbance measured on a strain carrying a promoterless plasmid are provided for comparison. The fluorescence curves in the middle plot have been obtained by shifting the fluorescence curves in the left plot over a time interval obtained by synchronizing the absorbance curves (right plot).

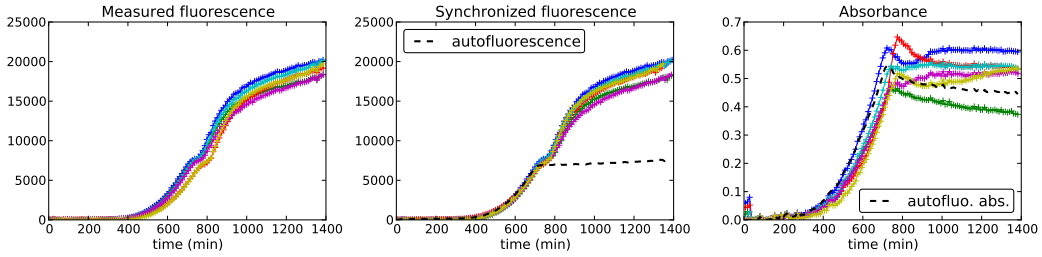


Figure A7: **Fluorescence and absorbance signals measured for the BW25113 strain carrying a p*acs-gfp* reporter plasmid**. See caption of Fig. A6.

As an illustration, we solved the growth rate estimation problem presented in Figure B1 for several different values of ϵ. The simulated data (in green in panel A) represent the bacterial population volume obtained from the growth rate (dashed line in the same panel), with added noise having the same properties as in the reporter gene data set in Figure 3.3. For $\epsilon = 1$, we observe as expected that the estimation of μ_0 is negatively biased. The penalization parameter $\lambda = \lambda_{gcv}$, chosen by generalized cross-validation, minimizes the error ErrReg(λ) associated with the regularized problem, defined as the right-hand side of Equation 3.9 in the main text (dashed vertical line in panel F). For

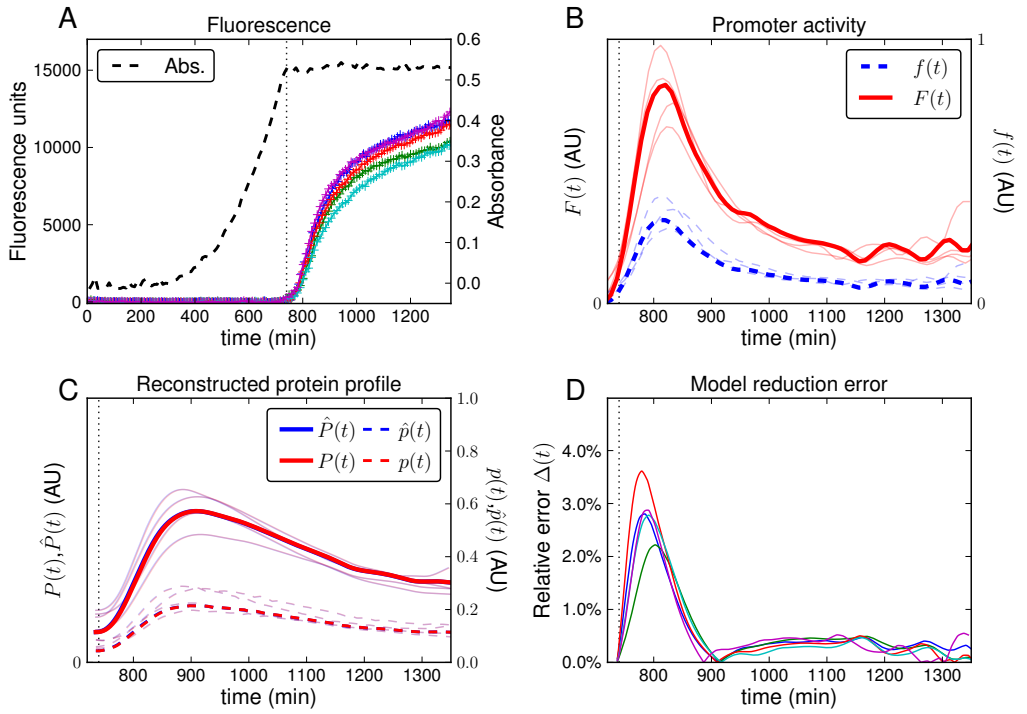


Figure A8: **Experimental estimation of the approximation error for the gene *acs***. The promoter activities have been obtained using an inversion method, as described in Section A8. See caption of Fig. 2.3 in the main text for details.

ϵ between 10^{-2} and 10^{-5}, the estimation is unbiased and corresponds well to the real input (panels B-C). The estimation is not sensible to ϵ in this interval and the value of λ_{gcv} is of the same order of magnitude as for the case $\epsilon = 1$. However, for even lower values of ϵ (panels D-E), the same value of λ_{gcv} makes the problem ill-posed (hence the peaks in ErrReg(λ) in panels I-J). As a consequence, the value of λ_{gcv} is aberrant and the resulting estimations are off the mark.

Throughout the paper we used $\epsilon = 10^{-5}$ and verified in each instance the appropriateness of this choice by a sensitivity analysis of the type shown in Figure B1. This procedure is recommended more generally.

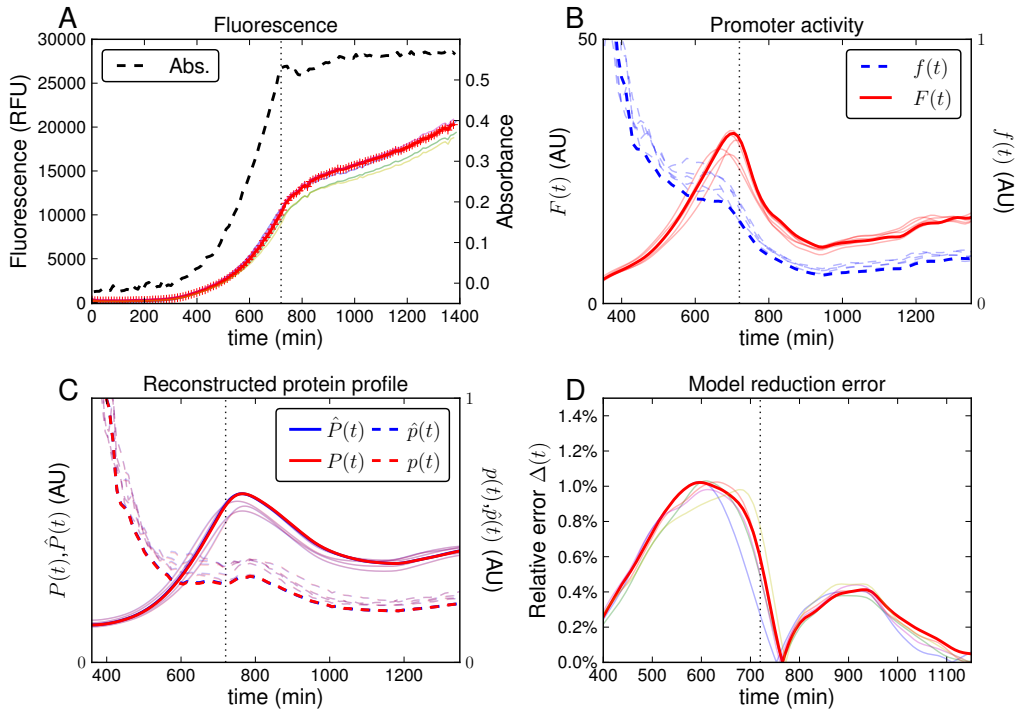


Figure A9: **Experimental estimation of the approximation error for the gene *crp***. The promoter activities have been obtained using an inversion method, as described in Section A8. See caption of Fig. 2.2 in the main text for details.

B2 Linear inversion problems with linear constraints

It can be useful to impose constraints on the values of the estimated $\mathbf{w}$, to ensure that they are comprised between certain bounds. Equation 3.7 in the main text can be reformulated as follows, using the definition of the discrete derivation matrix $\mathbf{L_w}$ in the same section:

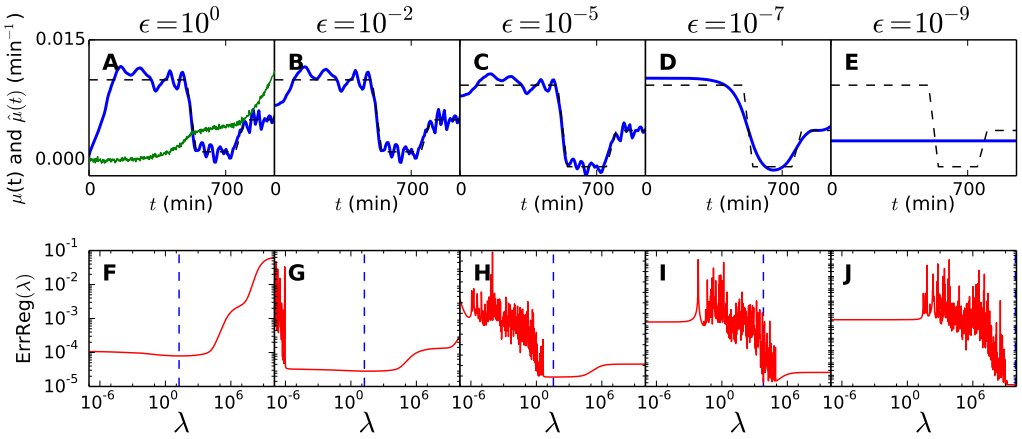


Figure B1: **Effect of parameter ϵ on growth rate estimation from *in-silico* data.** The simulated data (in green) for known input $\mu(t)$ (dashed line) shown in **A**. The estimation results (solid line) for different values of ϵ are shown in **A**-**E**. The corresponding error profile for the regularized problem, ErrReg(λ), and the minimal value λ_{gcv} for different choices of ϵ are shown in **F**-**J**.

$$
\begin{aligned}
\hat{\mathbf{w}} &= \underset{\mathbf{w}}{\operatorname{argmin}} \ \|\mathbf{H_w w} - \tilde{\mathbf{y}}\|_2^2 + \lambda \|\mathbf{L_w w}\|_2^2 \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \ (\mathbf{H_w w} - \tilde{\mathbf{y}})^T (\mathbf{H_w w} - \tilde{\mathbf{y}}) + \lambda (\mathbf{L_w w})^T (\mathbf{L_w w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \ \mathbf{w}^T \mathbf{H}_\mathbf{w}^T \mathbf{H_w w} - 2\tilde{\mathbf{y}}^T \mathbf{H_w w} + \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} + \lambda \mathbf{w}^T \mathbf{L}_\mathbf{w}^T \mathbf{L_w w} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \ \mathbf{w}^T (\mathbf{H}_\mathbf{w}^T \mathbf{H_w} + \lambda \mathbf{L}_\mathbf{w}^T \mathbf{L_w}) \mathbf{w} - 2\tilde{\mathbf{y}}^T \mathbf{H_w w}.
\end{aligned}
$$

To this quadratic minimization problem we can add a set of linear constraints of the form

$$\mathbf{G}_1 \mathbf{w} = \mathbf{c}, \tag{B1}$$

$$\mathbf{G}_2 \mathbf{w} \leq \mathbf{0}, \tag{B2}$$

where $\mathbf{G}_1, \mathbf{G}_2$ are constant matrices and $\mathbf{c}$ a constant vector. In this paper we want to ensure that the initial conditions (which represent quantities of molecules or volumes) and the input variable (which represents a growth rate, promoter activity, or protein concentration in Section 3.3 of the main text) is positive. This corresponds to setting

$$(\mathbf{G}_1, \mathbf{c}, \mathbf{G}_2) = (\mathbf{0}, \mathbf{0}, \mathbf{I}).$$

Several solvers have been proposed for the general quadratic programming problem, and in particular for the special case of ensuring positive solutions. In this paper we used the solver `cvxopt.solvers.qp` from the Python module `cvxopt`, which is well adapted to large-scale problems (Andersen et al., 2012). Notice that it is not possible to use generalized cross-validation on the constrained problem. Therefore, we first used GCV on the unconstrained problem to select the regularization parameter λ, and then solved the constrained problem for that particular value of λ.

B3 Computation of observation matrices

The computation of the observation matrix $\mathbf{H_w} = \begin{pmatrix} \mathbf{H_{x_0}} & \mathbf{H_u} \end{pmatrix}$ defined in Section 3.2.1 of the main text can be achieved in a straightforward way. The jth column of $\mathbf{H_{x_0}}$ is the vector of values obtained by solving Equation 3.1 at times $(t_i)_{1 \leq i \leq N_y}$, using $u(t) = 0$ and

$$\mathbf{x}_0 = (0, \cdots, 0, \underbrace{1}_{\mathbf{x}_0[j]}, 0, \cdots, 0).$$

The jth column of matrix $\mathbf{H_u}$ is obtained by solving the same system with $\mathbf{x}_0 = \mathbf{0}$ and

$$u(t) = \mathbf{1}_{[\tau_j, \tau_{j+1}[}(t).$$

The computation of $\mathbf{H_w}$ can be performed using a numerical differential equation solver, but this is usually time-consuming because of the large number of ODE integrations required $(n + N_u)$. An alternative is to use the explicit solution of Equation 3.2 and exploit the specific form that Equation 3.1 takes

when estimating growth rate, promoter activity, and protein concentration. The latter approach will be further developed in the remainder of this section.

B3.1 Explicit formula for the observation matrix for growth rate estimation

In section Section 3.3.2 we proposed the following model as a basis for the estimation of $\mu(t)$:

$$\frac{d}{dt}(\alpha\, V)(t) = \tilde{V}(t)\, \mu(t). \tag{B3}$$

This model admits the following general solution:

$$\alpha\, V(t) = \alpha\, V(0) + \int_0^t \tilde{V}(\sigma)\, \mu(\sigma)\, d\sigma. \tag{B4}$$

The observation matrix is of the form $\mathbf{H_w} = \begin{pmatrix} \mathbf{H_{x_0}} & \mathbf{H_u} \end{pmatrix}$. $\mathbf{H_{x_0}}$ has dimensions $1 \times N_y$ and its values are obtained by computing Equation B4 at the different observation times $(t_i)_{1 \leq i \leq N_y}$, with $\alpha\, V(0) = 1$ and $\mu(t) = 0$ for all t. Therefore, we have

$$\mathbf{H_{x_0}} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The element of $\mathbf{H_u}$ at position $[i, j]$ is computed by evaluating Equation B4 at time t_i with $\alpha\, V(0) = 0$ and

$$\mu(t) = \mathbf{1}_{[\tau_j, \tau_{j+1}[}(t).$$

This leads to

$$\mathbf{H_u}[i, j] = \int_0^{t_i} \tilde{V}(\sigma)\, \mathbf{1}_{[\tau_j, \tau_{j+1}[}(\sigma)\, d\sigma = \begin{cases} 0 & \text{if } t_i < \tau_j, \\ \int_{\tau_j}^{\min(t_i, \tau_{j+1})} \tilde{V}(\sigma)\, d\sigma & \text{otherwise.} \end{cases}$$

The size of the intervals $[\tau_i, \tau_{i+1}[$, denoted by $\delta\tau$, can be chosen arbitrarily small, so we will suppose that the volume is constant on each interval. This

allows the expression above to be simplified and we obtain the following approximate expression of $\mathbf{H_u}[i,j]$, which is used for the estimation of the growth rate in the figures of the main text, and in the `WellFARE` package:

$$\mathbf{H_u}[i,j] \simeq \tilde{V}(\tau_j)\max\left(0, \min(t_i - \tau_j, \delta\tau)\right).$$

B3.2 Efficient computation of the observation matrix for promoter activity estimation

In the main text we presented the following ODE model for the expression of the reporter gene:

$$\begin{cases} \dfrac{d}{dt}M(t) = k_M\, a(t)\, V(t) - d_M\, M(t) = k'_M\, a(t)\, \alpha\, V(t) - d_M\, M(t), \\ \dfrac{d}{dt}R_u(t) = k_U\, M(t) - (d_R + k_R)\, R_u(t), \\ \dfrac{d}{dt}R(t) = k_R\, R_u(t) - d_R\, R(t), \end{cases}$$

where $k'_M = k_M/\alpha$.

The observation matrix is of the form $\mathbf{H_w} = \begin{pmatrix}\mathbf{H_{x_0}} & \mathbf{H_u}\end{pmatrix}$. The element of $\mathbf{H_u}$ at position $[i,j]$ is computed by solving the ODE system, with $\begin{pmatrix}M(0) & R_u(0) & R(0)\end{pmatrix} = \mathbf{0}$ and

$$a(t) = \mathbf{1}_{[\tau_j,\tau_{j+1}[}(t),$$

and then evaluating $R(t)$ at time-point t_i. We can reformulate this as follows, using the input-output system notation from Section 2.1 of the main text (Chen, 1970):

$$\mathbf{H_u}[i,j] = R(t_i, \mathbf{1}_{[\tau_j,\tau_{j+1}[}, \mathbf{0}).$$

Computing $\mathbf{H_u}$ in this way, however, would require the solution of as many ODE systems as there are intervals $[\tau_j, \tau_{j+1}[$, typically on the order of 1000.

A more efficient procedure for computing $\mathbf{H_u}$ can be obtained by choosing a suitable approximation. Assume that the intervals $[\tau_j, \tau_{j+1}[$ are of equal

length $\delta\tau$ and small compared to the characteristic variation time of $V(t)$. As a consequence, $V(t)$ can be supposed constant over the interval $[\tau_j, \tau_{j+1}[$, and we can use the following approximated system to compute the $[i,j]$th element of $\mathbf{H_u}$:

$$\begin{cases} \dfrac{d}{dt}M(t) \simeq k'_M\,\alpha\,V(\tau_j)\,\mathbf{1}_{[\tau_j,\tau_{j+1}[}(t) - d_M\,M(t), \\ \dfrac{d}{dt}R_u(t) = k_U\,M(t) - (d_R + k_R)\,R_u(t), \\ \dfrac{d}{dt}R(t) = k_R\,R_u(t) - d_R\,R(t). \end{cases} \tag{B5}$$

It is easy to see that this system is linear in $\alpha\,V$, so that we can write

$$R(t_i, \mathbf{1}_{[\tau_j,\tau_{j+1}[}, \mathbf{0}) = \alpha\,V(\tau_j)\,R_1(t_i, \mathbf{1}_{[\tau_j,\tau_{j+1}[}, \mathbf{0}),$$

where R_1 is the output of the system of Equation B5 with $\alpha\,V(\tau_j)$ set to 1. Because the coefficients of this system are time-invariant, we also have that

$$R_1(t_i, \mathbf{1}_{[\tau_j,\tau_{j+1}[}, \mathbf{0}) = \begin{cases} 0 & \text{if } t_i < \tau_j, \\ R_1(t_i - \tau_j, \mathbf{1}_{[0,\delta\tau[}, \mathbf{0}) & \text{otherwise.} \end{cases}$$

This leads to the following approximation of $H_{\mathbf{u}}[i,j]$:

$$\mathbf{H_u}[i,j] = R(t_i, \mathbf{1}_{[\tau_j,\tau_{j+1}[}, \mathbf{0}) \simeq \begin{cases} 0 & \text{if } t_i < \tau_j, \\ \tilde{V}(\tau_j)\,R_1(t_i - \tau_j, \mathbf{1}_{[0,\delta\tau[}, \mathbf{0}) & \text{otherwise.} \end{cases}$$

The advantage of this approximate method for computing $\mathbf{H_u}$ is that it requires the ODE system of Equation B5 to be solved only once, replacing the term $\alpha\,V(\tau_j)\,\mathbf{1}_{[\tau_j,\tau_{j+1}[}(t)$ by $\mathbf{1}_{[0,\delta\tau[}(t)$ and evaluating the output at all time-points t_i, instead of solving N_u ODEs.

B3.3 Computation of the observation matrix for promoter activity estimation in a reduced gene expression model

In Section 3.3.3 of the main text, the production of mature GFP proteins was described as a three-step process (transcription, translation, and maturation).

In this section we consider a simplified version of this model, which allows us to explicitly formulate the observation matrix $\mathbf{H_w}$ as a function of the measured signal $\left(\tilde{V}(t_i)\right)_{1\leq i\leq N_y}$ and the degradation constant d_R.

When both transcription and maturation are fast as compared to the other processes involved in the expression of the reporter gene, it can be assumed that their effect on the dynamics of the folded reporter is negligible. In this case, the entire process of synthesizing mature GFP can be lumped into a single step and Equation 3.16 in the main text becomes:

$$\frac{d}{dt}R(t) = k'_R\,\alpha\,V(t)\,a(t) - d_R\,R(t), \tag{B6}$$

where k'_R denotes a lumped protein synthesis parameter.

The quantity of reporter protein $R(t)$ can be explicitly formulated as a function of $a(t)$ by solving Equation B6:

$$R(t) = R(0)\,e^{-d_R t} + e^{-d_R t}\int_0^t e^{d_R\sigma}k'_R\,\alpha\,V(\sigma)\,a(\sigma)\,d\sigma. \tag{B7}$$

In other words, the promoter activity is linearly related to the amount of reporter protein. In what follows, we set $k'_R = 1$, which allows the promoter activity to be estimated up to an unknown proportionality constant (Section B6).

The observation matrix $\mathbf{H_w}$ for the corresponding linear inversion problem is of the general form $\mathbf{H_w} = \begin{pmatrix}\mathbf{H_{x_0}} & \mathbf{H_u}\end{pmatrix}$, where $\mathbf{H_{x_0}}$ is given by:

$$\mathbf{H_{x_0}} = \begin{pmatrix} e^{-d_R t_0} \\ e^{-d_R t_1} \\ \vdots \\ e^{-d_R t_{N_y}} \end{pmatrix}.$$

The element of $\mathbf{H_u}$ at position $[i,j]$ is computed by evaluating Equation B7 at time t_i for $R(0) = 0$ and $a(t) = \mathbf{1}_{[\tau_j,\tau_{j+1}[}(t)$. This leads to

$$\mathbf{H_u}[i,j] = \begin{cases} 0 & \text{if } t_i < \tau_j, \\ e^{-d_R t_i}\int_{\tau_j}^{\min(\tau_j+\delta\tau,t_i)} e^{d_R\sigma}\alpha\,V(\sigma)\,d\sigma & \text{otherwise,} \end{cases} \tag{B8}$$

where $\delta\tau$ denotes the length of the time-interval $[\tau_j, \tau_{j+1}[$. We can exploit the fact that $\delta\tau$ can be chosen arbitrarily small to simplify the integral by assuming that the volume is approximately constant over the time-interval considered:

$$\begin{aligned}\int_{\tau_j}^{\min(\tau_j+\delta\tau,t_i)} e^{d_R\sigma}\alpha\, V(\sigma)\, d\sigma &\simeq \alpha\, V(\tau_j) \int_{\tau_j}^{\min(\tau_j+\delta\tau,t_i)} e^{d_R\sigma} d\sigma \\ &= \frac{1}{d_R}\,\alpha\, V(\tau_j)\left(e^{d_R \min(\tau_j+\delta\tau,t_i)} - e^{d_R\tau_j}\right) \\ &\simeq \frac{1}{d_R}\,\tilde{V}(\tau_j)\left(e^{d_R \min(\tau_j+\delta\tau,t_i)} - e^{d_R\tau_j}\right).\end{aligned}$$

As a consequence,

$$\mathbf{H_u}[i,j] \simeq \begin{cases} 0 & \text{if } t_i < \tau_j, \\ \frac{1}{d_R}\,\tilde{V}(\tau_j)\left(e^{d_R \min(\tau_j+\delta\tau,t_i)} - e^{d_R\tau_j}\right) & \text{otherwise.}\end{cases} \tag{B9}$$

The latter expression is used to compute promoter activities in the `WellFARE` package.

B3.4 Explicit formula for the observation matrix for protein concentration estimation in a reduced gene expression model

In the main text we have presented the production of a protein of interest and its reporter as multistep processes. The observation matrix allowing the estimation of the protein concentration $p(t)$ from the absorbance and fluorescence data can be computed by means of the procedure in Section 3.2.2. In this section, like in Section B3.3, we will simplify the problem by considering single-step gene expression models, enabling an explicit formulation of the observation matrix.

Using the same notation as in Section B3.3, $P'(t)$ (defined in Section 3.3.4) and $R(t)$ are driven by the following one-step gene expression models:

$$\frac{d}{dt}P'(t) = k'_P\,\alpha\,V(t)\,a(t) - d_P\,P'(t), \tag{B10}$$

$$\frac{d}{dt}R(t) = k'_R\,\alpha\,V(t)\,a(t) - d_R\,R(t), \tag{B11}$$

$$p(t) = P'(t)/\left(\alpha V(t)\right) \tag{B12}$$

where k'_P and k'_R denote lumped protein synthesis parameters. Notice that the degradation constants of the protein of interest (d_P) and the reporter protein (d_R) are different *a priori*.

This model enables $R(t)$ to be directly expressed as a function of $P'(t)$. We first derive the following expression of $\alpha\,V(t)\,a(t)$ from Equation B10:

$$\alpha\,V(t)\,a(t) = \frac{1}{k'_P}\left(d_P\,P'(t) + \frac{d}{dt}P'(t)\right),$$

and then inject this expression into Equation B11:

$$\frac{d}{dt}R(t) = K\left(d_P\,P'(t) + \frac{d}{dt}P'(t)\right) - d_R\,R(t),$$

where $K = k'_R/k'_P$. The above differential equation for $R(t)$, with input $P'(t)$, can be solved exactly, yielding

$$R(t) = \underbrace{K\,P'(t)}_{A(t)} + \underbrace{K\,(d_R - d_P)\,e^{-d_R t}\int_0^t e^{d_R\sigma}\,P'(\sigma)\,d\sigma}_{B(t)} + \underbrace{\left(R(0) - K\,P'(0)\right)e^{-d_R t}}_{C(t)}. \tag{B13}$$

The terms $A(t), B(t), C(t)$ in Equation B13 admit a simple interpretation. $A(t)$ shows that $R(t)$ will, at least partly, follow the variations of $P'(t)$. This is to be expected, as $P'(t)$ and $R(t)$ are driven by the same promoter activity $a(t)$. $B(t)$ is a correction term accounting for the difference in degradation constants of the reporter protein and the protein of interest. $C(t)$ accounts for the differences in initial conditions $P'(0)$ and $R(0)$. Equation B13 can be rewritten as

$$R(t) = R(0)\,e^{-d_R t} + K\left(\alpha\,V(t)\,p(t) - \alpha\,V(0)\,p(0)e^{-d_R t} \right. \\ \left. +(d_R - d_P)\,e^{-d_R t}\int_0^t e^{d_R\sigma}\,p(\sigma)\,\alpha\,V(\sigma)\,d\sigma\right). \tag{B14}$$

Note that this formulation shows the linear relationship between the output $R(t)$, the input $p(t)$, and initial condition $R(0)$, which correspond respectively to $y_2(t)$, $u(t)$, and $\mathbf{x}_{0,2}$ in Section 3.2.2 of the main text. Like in Section B3.3, we assume that $k'_R = k'_P = 1$ to simplify computations and obtain a proportional estimator for $p(t)$ (Section B6).

The observation matrix for the estimation problem is of the general form $\mathbf{H_w} = \begin{pmatrix}\mathbf{H_{x_0}} & \mathbf{H_u}\end{pmatrix}$, where

$$\mathbf{H_{x_0}} = \begin{pmatrix} e^{-d_R t_0} \\ e^{-d_R t_1} \\ \vdots \\ e^{-d_R t_{N_y}} \end{pmatrix},$$

obtained by setting $p(t) = 0$ for all t and $R(0) = 1$. The element of $\mathbf{H_u}$ at position $[i, j]$ is computed by evaluating Equation B14 at time t_i for $R(0) = 0$ and $p(t) = \mathbf{1}_{[\tau_j,\tau_{j+1}[}(t)$. This leads to

$$\mathbf{H_u}[i,j] = \begin{cases} 0 & \text{if } t_i < \tau_j, \\ \alpha\,V(t_i) + (d_R - d_P)\,e^{-d_R t_i}\int_{\tau_j}^{t_i} e^{d_R\sigma}\,\alpha\,V(\sigma)\,d\sigma & \text{if } \tau_j \le t_i < \tau_{j+1}, \\ (d_R - d_P)\,e^{-d_R t_i}\int_{\tau_j}^{\tau_{j+1}} e^{d_R\sigma}\,\alpha\,V(\sigma)\,d\sigma & \text{if } t_i \ge \tau_{j+1}. \end{cases} \tag{B15}$$

In this expression we can use $\tilde{V}$ instead of $\alpha\,V$, and approximate the integral, like in Section B3.3. This results in an approximate but practical formula

for $\mathbf{H_u}[i,j]$:

$$\mathbf{H_u}[i,j] = \begin{cases} 0 & \text{if } t_i < \tau_j, \\ \tilde{V}(\tau_j)\left(1 + \dfrac{d_R - d_P}{d_R}\left(1 - e^{d_R(\tau_j - t_i)}\right)\right) & \text{if } \tau_j \leq t_i < \tau_{j+1}, \\ \tilde{V}(\tau_j)\dfrac{d_R - d_P}{d_R}\left(e^{d_R(\tau_{j+1} - t_i)} - e^{d_R(\tau_j - t_i)}\right. & \text{if } t_i \geq \tau_{j+1}. \end{cases} \quad \text{(B16)}$$

The latter expression is used to compute protein concentrations in the `WellFARE` package.

B4 Reporter gene experiments: materials and methods

The *E. coli* wild-type strain used in this study is the strain BW25113 (Baba et al., 2006). The reporter strains were obtained by transforming the wild-type strain with a reporter plasmid, bearing a transcriptional fusion of the *crp*, *fis*, *gyrA* and *acs* promoter regions with the *gfp* reporter gene, and a promoter-less vector for background correction (Table A.1). The reporter gene codes either for a stable and fast-folding version of the GFP reporter (GFPmut2) or for a less stable allele (GFPmut3). More information on the half-live and maturation time can be found in (Berthoumieux et al., 2013).

Plasmid	Characteristics	Reference or source
pZE*gfp*	Amp^r, colE1 *ori*, *gfpmut3*	(de Jong et al., 2010)
pZE*fis-gfp*	Amp^r, colE1 *ori*, p*fis-gfpmut3*	(de Jong et al., 2010)
pZE*crp-gfp*	Amp^r, colE1 *ori*, p*crp-gfpmut3*	(Berthoumieux et al., 2013)
pZE*gyrA-gfp*	Amp^r, colE1 *ori*, p*gyrA-gfpmut3*	(Boyer et al., 2010)
pUA66*gfp*	Kan^r, pSC101 *ori*, *gfpmut2*	(Zaslaver et al., 2006)
pUA66*acs-gfp*	Amp^r, pSC101 *ori*, p*acs-gfpmut2*	(Baptist et al., 2013)

Table A.1: **Reporter plasmids used in this study.**

Glycerol stocks (-80°C) of the above-mentioned reporter strains were grown overnight (about 15 h) at 37°C, with shaking at 200 rpm, in M9 minimal

medium (Miller, 1972) supplemented with 0.3% glucose and mineral trace elements. For plasmid-carrying strains, the growth medium was supplemented with 100 μg ml^{-1} ampicilin or kanamycin. The overnight cultures were diluted into a 96-well microplate, so as to obtain an adjusted initial OD_{600} of 0.1. The wells of the microplate contain M9 minimal medium supplemented with 0.3% glucose, mineral trace elements, and 1.2% of the buffering agent HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) for maintaining physiological pH levels in the growth medium. No antibiotics were added at this stage. The wells were covered with 60 μl of mineral oil to avoid evaporation. The microplate cultures were then grown for about 24 h at 37°C, with agitation at regular intervals, in the Fusion microplate reader (Perkin Elmer).

During a typical experimental run, we acquire about 110 readings each of absorbance (600 nm) and fluorescence (485/520 nm). From the measured signal we remove the background signals of absorbance and fluorescence measured on wells containing growth medium only and strains carrying a promoterless reporter plasmid, respectively.

B5 Numerical evaluation of the linear inversion methods

In this section we test the ability of the proposed linear inversion methods to correctly estimate different shapes of growth rate, promoter activity, and protein concentration profiles. We generated 100 absorbance and fluorescence data sets for defined growth rate and promoter activity profiles, similar to those observed for the gene *acs* in the reporter gene experiments in Section 3.3.1 of the main text (Figure 3.3). In panel A the ability of the method to reconstruct different growth rate profiles is tested, whereas panels B and C consider different promoter activity and protein concentration profiles (with absorbance data from panel A1), respectively. In every case considered, the methods succeed in providing an almost unbiased estimate of the gene expression quantities.

We also compared the linear inversion methods with other methods, in particular indirect approaches that plug empirically smoothed versions of the data into the measurement models (Figure 3.4 in the main text). Below we

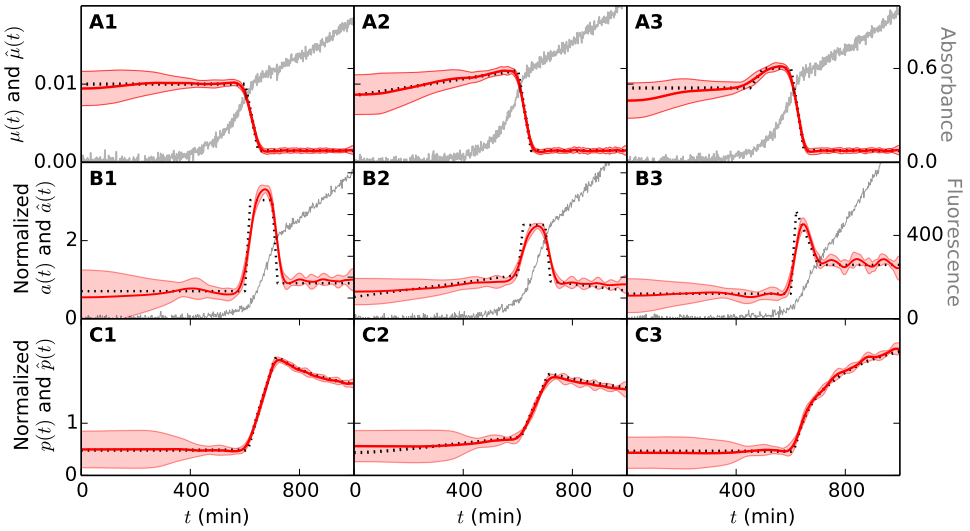


Figure B2: ***In-silico* experiments for testing the ability of the linear inversion methods to correctly estimate different growth rate, promoter activity, and protein concentration profiles.** The estimation results for growth rate, promoter activity, and protein concentration are shown in panels **A**-**C**, respectively. The dotted lines show the profiles used for generating the 100 data sets, the grey solid lines example absorbance and fluorescence time-series data, and the red solid line and the shaded area the mean ± one standard deviation of the 100 estimations, respectively.

extend this analysis, for the estimation of the growth rate from absorbance measurements, by showing that increasing the smoothing parameter to reduce the variance of the estimates introduces a strong bias (Figure B3). The growth rate is shown as the dotted curve and the absorbance data are the same as in Figure 3.4 in the main text.

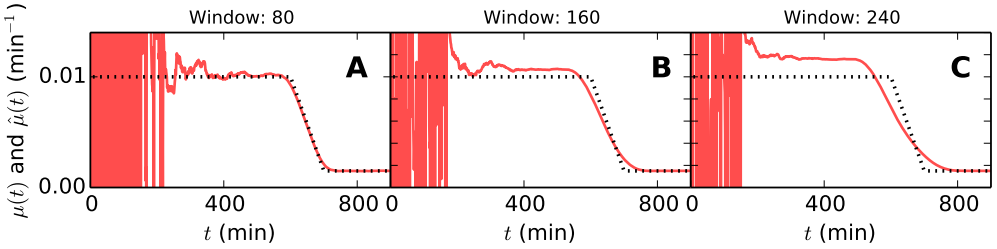


Figure B3: **Growth rate estimation by smoothing of the absorbance measurements.** The growth rate was computed from smoothed absorbance measurements of the volume by means of Equation 3.13 in the main text. The plots show the true value of the growth rate (dotted line), used to generate 100 absorbance data sets, and the mean (solid red line) ± one standard deviation (shaded area) of the growth rate estimations. Different levels of smoothing of the absorbance data were considered, using sliding windows of different length (80, 160, and 240 data points), corresponding to the panels **A**-**C**, respectively.

B6 Linear inversion when parameters in the gene expression model are unknown

Equation 3.16 in the main text describes the gene expression model on which the estimation methods are based:

$$\begin{cases} \dfrac{d}{dt}M(t) = k'_M a(t)\widehat{\alpha V}(t) - d_M M(t), \\ \dfrac{d}{dt}R_u(t) = k_U M(t) - (d_R + k_R)R_u(t), \\ \dfrac{d}{dt}R(t) = k_R R_u(t) - d_R R(t), \end{cases}$$

and we remind that $\tilde{R}(t_i) = \beta R(t) + \nu$ (Equation 3.11 in the main text). The constants k'_M, k_U, and β are generally unknown. In this section we show that the profile of the promoter activity can still be estimated, up to an (unknown) proportionality constant, using a linear inversion.

We consider the following transformed variables:

$$R^*(t) = \beta R(t), \quad R^*_u(t) = \beta R_u(t), \quad M^*(t) = k_U \beta M(t), \quad a^*(t) = k_U k'_M \beta a(t).$$

Replacing the variables M, R_u, R, a in the ODE system above by their starred counterpart we obtain the following system:

$$\begin{cases} \dfrac{d}{dt}M^*(t) = a^*(t)\widehat{\alpha V}(t) - d_M M^*(t), \\ \dfrac{d}{dt}R_u^*(t) = M^*(t) - (d_R + k_R)R_u^*(t), \\ \dfrac{d}{dt}R^*(t) = k_R R_u^*(t) - d_R R^*(t). \end{cases}$$

Moreover, $\tilde{R}(t_i) = R^*(t) + \nu$. This system is equivalent to Equation 3.16 in the main text when $\beta = k_U = k'_M = 1$ and can thus be used to estimate the profile $a^*(t)$, which is proportional to $a(t)$ (the proportionality constant being unknown).

The same approach can be applied to the gene expression system of Equation 3.17 in the main text to show that it is possible to obtain an estimator or the profile of the protein concentration $p(t)$ (up to an unknown proportionality constant) in the absence of reliable values for the parameters k_N and k_P.

B7 Software implementation of the linear inversion methods

The linear inversion methods discussed in this article have been implemented in the Python library `WellFARE` and are available online through the web application `WellInverter`. In this section we briefly describe `WellFARE` and `WellInverter`, and we refer to the dedicated web pages for more information.

B7.1 The `WellFARE` Python Package

`WellFARE` (well Fluorescence Analysis for Reporter Experiments) is a Python library released under an LGPL licence, implementing the methods for growth rate, promoter activity and protein concentration estimation developed in the main text. In addition, the library provides practical tools for the treatment of data from reporter experiments, such as automated outlier removal, data

synchronization, and parsing of Excel files generated by the TECAN Infinite Pro microplate reader. `WellFARE` uses extensively the Python core scientific library `SciPy` (Jones et al., 2001). The treated data can be exported in the form of JSON objects (Crockford, 2006). Source code, documentation and installation instructions for the `WellFARE` library and its command-line and JSON interfaces are available at the following address:

`https://github.com/ibis-inria/wellfare`

The code used to generate the figures of the main text is available in the `examples` folder of the library.

B7.2 The `WellInverter` web application

The `WellInverter` web application provides online access to the linear inversion methods without having to install the software locally. The server part of `WellInverter` is based on the Python library `WellFARE`, the computational core of the application. It also provides methods for managing experimental and user data as well as storing analysis parameters in JavaScript Object Notation (JSON) format (Crockford, 2006). The client part of `WellInverter` is the graphical user interface of the application, accessible through a web browser. It allows the user to upload, analyze, and visualize the results of a reporter gene experiment as well as downloading the results for further treatment. The client part is written in Javascript, and communicates with the server using Ajax (Asynchronous JavaScript and XML) calls (Garrett, 2006). More information on access to `WellInverter` and a tutorial are available at the following address:

`https://team.inria.fr/ibis/wellinverter`

A test account has been opened with username `guest` and password `guest2015`. The reporter gene data set on the server has been used in the main text.

Appendix B

Supplementary Information on Chapter 4

C1 Estimation of the CRP$^{(*)}$ degradation rate

In order to estimate the degradation rate d_{CRP}, bacteria were grown in conditions similar to Section 4.2.1 in the main text, where the only carbon source in the medium was acetate, which ensures a slow, steady growth of the population. In these conditions we found that activity of the pRM promoter was constant during the major part of the experiment (data not shown), which leads to a very simple model in which the observed modulations $a_{lac^*}(t)$ (dotted lines in Figure C1C) depend only on [CRP*] and can be predicted from the observed $a_{crp^*}(t)$ (Figure C1A) given the unknown parameters d_{CRP}, C_c, and f_a:

$$\begin{cases} a_{lac^*}(t) = f_a \dfrac{[CRP^*]}{C_c + [CRP^*]} \\ \dfrac{d}{dt}[CRP^*](t) = a_{crp^*}(t) - (d_{CRP} - \mu(t))\,[CRP^*](t) \end{cases} \tag{C1}$$

We call $a[f_a, d_{CRP}, C_c](t)$ the profile of $a_{lac^*}(t)$ corresponding to the set of parameters (f_a, d_{CRP}, C_c). As f_a is simply a multiplication coefficient, the optimal f_a given d_{CRP} and C_c, in the sense of the least of the least squares is given by the ratio between the mean of the observed activities $\hat{a}_{lac^*}(t)$ and the

activities predicted for the case $f_a = 1$:

$$f_{a,est.} = \frac{\bar{\hat{a}}_{lac^*}}{\bar{a}_{lac^*}[1, d_{CRP}, C_c]}$$

We fitted the parameters by evaluating the squared error between predictions and observation over a grid of values of d_{CRP} and C_c. The results are shown in Figure C1D. We see that taking $d_{CRP} = 0.001$ enables to predict well the activity of p*lac** at different levels of induction. More generally any value of d_{CRP} under 0.001 leads to a good fit, so it really is important to assume that CRP is a stable protein in *E. coli*. The decrease of [CRP*] concentration observed in Figure C1B is not due to CRP* degradation but to the dilution of CRP* through population growth.

C2 Plasmids used in this study

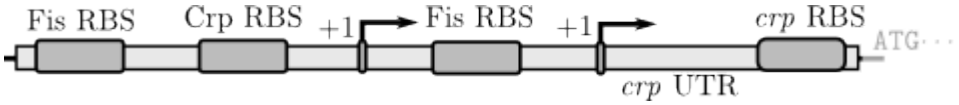
C2.1 Reporter plasmids

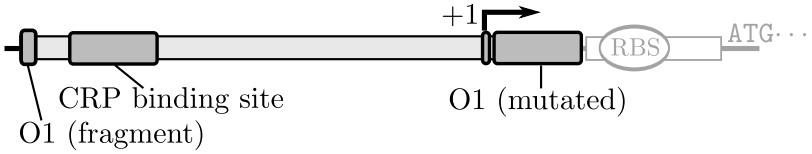
All reporter plasmids used in this study are derived from a pUA66 low-copy plasmid (Zaslaver et al., 2006). The pRM-*gfp* plasmid is as described in (Berthoumieux et al., 2013). For the p*lac***-gfp* and p*crp-gfp* plasmids, we replaced the kanamycin resistance gene of pUA66 by the *bla* gene, coding for resistance to ampicillin. The p*crp* sequence was amplified from the chromosome of *E. coli* BW25113, to ensure that the entire regulatory region as well as the 3'5'-UTR were exactly as on the natural chromosome. The 99-bp long p*lac** was synthetized using two primers with a 21-bp homology between them, and with each a 20-bp homology on the plasmid, as shown in Figure C2. All other constructions were made using Gibson Assembly. All the plasmids were sequenced. The sequences of the p*crp*, p*lac** and pRM promoters are shown in Table B.1.

C2.2 Induction plasmids

The plasmids for CRP* induction were constructed as indicated in Figure C3. The *crp** gene was cloned from the chromosome of a mutant strain carrying

p*crp*	CCACTGCGTCAATTTTCCTGACAGAGTACGCGTACTAACCAAATCGCGCA ACGGAAGGCGACCTGGGTCATGCTGAAGCGAGACACCAGGAGACACAAAG CGAAAGCTATGCTAAAACAGTCAG**G**ATGCTACAGTAATACATTGATGTAC TGCATGTATGCAAAGGACGTCACATTACCGTGCAGTACAGTTGATAGCCC CTTCCCAGGTAGCGGGAAGCATATTTCGGCAATCCAGAGACAGCGGCGTT ATCTGGCTCTGGAGAAAGCTTATAACAGAGGATAACCGCGC
p*lac**	ATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTAT GCTTCCGGCTCGTATGTTGTGTGC**A**ATTTACATCCTCCGCTAGGTTCACT TTAAGAAGGAGATATACAT
pRM	TCGAGCCTATCACCGCCAGAGGTAAAATAGTCAACACGCACGGTGTTAGAT ATTTATCCCTTGTGGTGATAGATTTAACGT**A**TCAGCACAAAAAAGAAACC





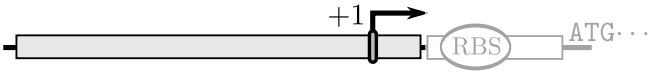


Table B.1: Sequences of the promoters used in the reporter constructions of this study. The schematic representations below each sequence show the regulatory elements contained in the promoter. These promoter fragments were cloned into pUA-Amp using Gibson Assembly or the method presented in Figure C2. The transcription start site is in bold.

this allele instead of the natural *crp* (Eppler and Boos, 1999), and inserted into the induction plasmid pZA which allows the gene to be controlled by the extracellular concentration of anhydro-tatracycline (ATc). The induction plasmid used as a template originally induced the gene *lux*, coding for luciferase, which made it easy to screen for valid colonies (an absence of luminescence meaning that the *lux* gene has been successfully replaced by *crp**). Finally, the *gfp* gene (along with its 5'3' UTR) from the p*lac**-*gfp* construction was inserted downstream of *crp** on the pZA plasmid.

We also constructed versions of the plasmids with the natural *crp* allele instead of *crp*, but those were not used in this study.

C3 M9 minimal medium

Table B.2 indicates the composition of the minimal growth medium used in this study for all the overnights and kinetic experiments. The glucose was replaced by another carbon depending on the experiment.

C4 Original Data

C5 Cyclic AMP measurements

During each of the diauxies presented in Figure 4.4a nd Figure C4, the microplates were regularly sampled over the course of the experiment. We took one sample per hour before stationary phase and one sample every 20 minutes during stationary phase. The samples were placed at -80°C directly after sampling. After analysis of the fluorescence data, twelve samples were selected for cAMP dosage in order to get an estimation of the cAMP concentration in the cells before and in the few hours following glucose exhaustion.

The estimation of cAMP in these samples is as in (Berthoumieux et al., 2013). The total amount of cAMP in the growth medium is measured using competitive ELISA Figure C5.

M9 + 0.03% glucose (50 mL)	
H_20	to 50 mL
CaCl2 1M	5 μL
$MgSO_4$ 1M	100 μL
20% glucose	750μL
1000x traces	45 μL
Fe solution	5μL
1% Thiamine	25μL
5x salts	10 mL

1000x traces (1 mL)	
H_2O	200μL
Na_2EDTA $2H_2O$	100μL
$ZnSO_4$ $7H_2O$	100μL
$CaCl_2$ $6H_2O$	100μL
$MnCl_2$ $4H_2O$	100μL
H_3BO_3	100μL
Na_2MoO_4 $2H_2O$	100μL
$CuSO_4$ $5H_2O$	100μL

5x Salts (10 mL)	
H_2O	10ml
$Na_2HPO_42H_20$	425 mg
KH_2HPO_4	150 mg
NaCl	25 mg
NH_4Cl	50 mg

Fe Solution	
H_20	1 mL
$FeSO_4$	30 mg
20% glucose (10mL)	
H_20	10 mL
D-glucose	2 g

Table B.2: Recipe for the minimal medium M9.

$$l([\text{cAMP}]) = d + \frac{a}{1 + (x/b)^c}$$

This enables to estimate cAMP concentration from luminescence in the samples using the inverse formula

$$[\text{cAMP}](l) = b\left(\frac{a}{l-d} - 1\right)^{1/c}$$

C5.1 Treatment of reporter gene data

This section presents the experimental data used in Chapter 4.

Removal of outliers

The microplate reader used in this study yields data of very good precision. However, the beads added to the wells (in order to enhance the stirring and oxygenation) can corrupt up to in 80% of the measurements, possibly because they deviate of the excitation light rays emitted by the plate reader, leading to less excitation of the GFP in the wells during the fluorescence measurements, and less light reaching the detectors in absorbance measurements. The resulting outliers are therefore almost all positively biased in the absorbance curves and negatively biased in the fluorescence curves (Figure C6A).

The high frequency of outliers makes classical filtering techniques useless, but the fact that they are mostly one-sided in each curve enables us to remove them using an ad-hoc procedure. The different steps are illustrated in Figure C6. In a first step we use the fact that, as they tend to produce *bumps* in the curve, outliers are point of very low or high second-order derivative. Therefore we first remove the points where this derivative is high in the fluorescence curve or low in the absorbance curve. The (discrete) second-order derivative of a discrete signal $(t_i, y_i)_{1 \leq i \leq N}$ is computed by differentiating the curve two times according to the formula

$$\frac{dy}{dt}(t_i) = \frac{y_{i+1} + y_i}{t_{i+1} - t_i}.$$

The points of the curves corresponding to the 50% higher (for the fluorescence) or lower (for the absorbance) values are removed. For better efficiency this filtering is carried over several times. The resulting data (Figure C6B) is generally very conservative, in the sense that many *sound* data points have been excluded as outliers. The next steps will aim at retrieving these points. We first smooth the data remaining after the first step using a moving window smoothing, in order to obtain a trend and an estimation of the standard deviation of the measurement noise (Figure C6C). The measurements conserved in the final curve (Figure C6D) are the ones whose value is less than three standard deviations away from the trend.

C5.2 Raw data

This section presents the original data behind Figure 4.2 and Figure 4.4 of the main text.

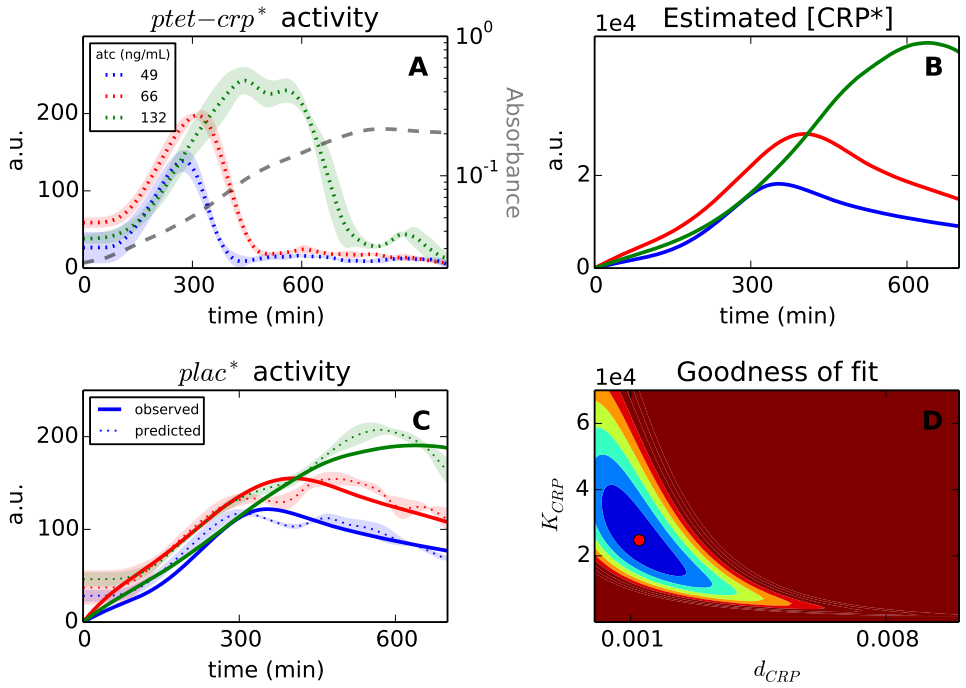


Figure C1: **Estimation of the degradation rate d_{CRP} through an induction experiment**. Line colors in panels A,B,C refer to different initial concentrations of ATc in the growth medium, given in panel A. In panels A and C, dotted lines and shaded regions indicate the mean ± one standard deviation of at least 4 replicates. **A.** Observed synthesis rate of CRP* proteins as as function of the ATc concentration in the medium. The population volume (OD) is indicated by the grey line. **B.** Intracellular concentration of CRP* estimated from the activity and volume profiles of panel A, using the coefficient d_CRP estimated in panel D. **C.** Oberved activity of the plac^* promoter (dotted lines) and profiles predicted from the concentrations in panel B and the parameter C_c estimated in panel D. **D.** Sum-of-squares error for different values of d_{CRP} and C_c. The red dot indicates the best fit, the dark red region indicates parameters yielding a squared error at least twice as large as the best fit.

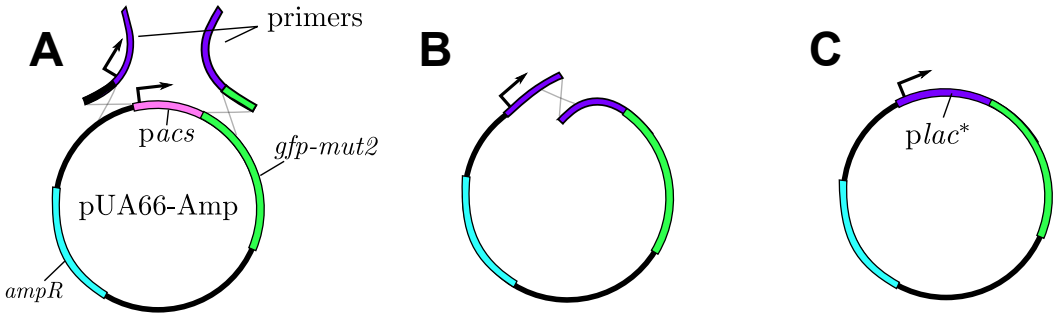


Figure C2: **Steps of the construction of the p*lac**-*gfp* reporter gene. A.** Two primers carrying the sequence of the p*lac** promoter are used to amplify the pUA66 template **B.** The linearized plasmid obtained in A is injected into *E. coli* NM100 (expressing naturally the λ-red recombinase) where the two homologous ends recombine to form a closed circle. **C.** Final plasmid, with a synthetic p*lac** reporter upstream of GFP.

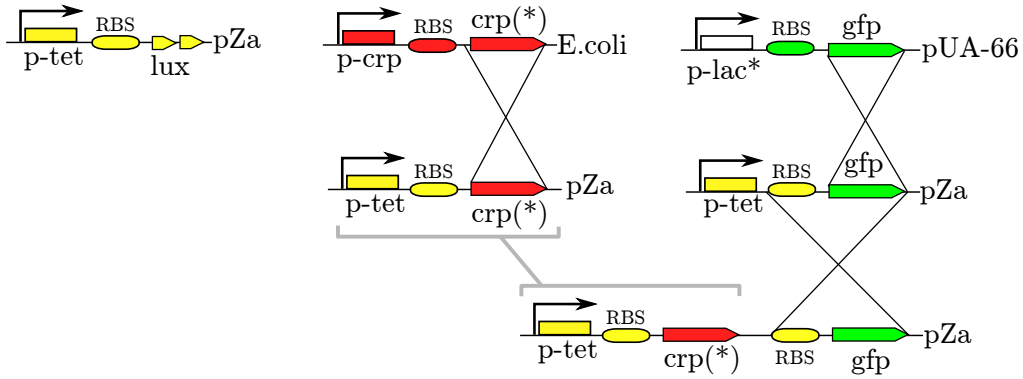


Figure C3: **Construction steps of the induction plasmids.** The plasmid pZA carries a constitutively expressed *tet* repressor gene and the luciferase operon downstream of the p*tet* promoter (left). We amplified the *crp** gene from the chromosome and inserted this fragment into pZA in place of the luciferase operon (center). In a similar manner, we inserted *gfp* into pZA downstream of the p*tet* promoter (right). In a final step, we placed *gfp*, including the strong RBS of pZA, downstream of the ATc-inducible pZA-*crp** plasmid (bottom).

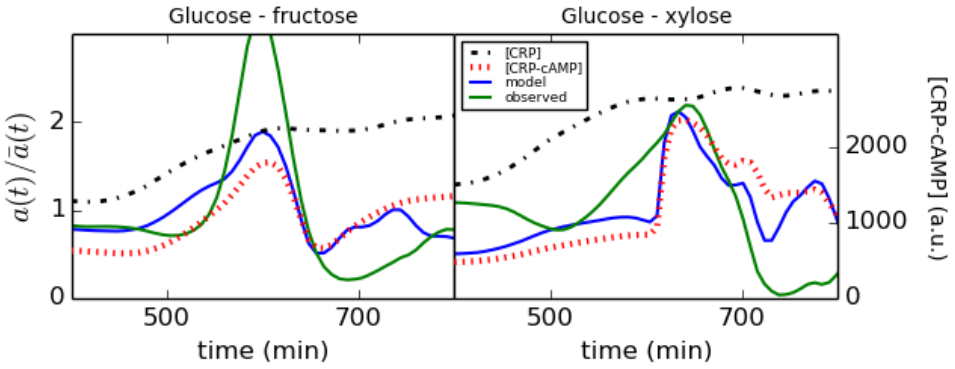


Figure C4: **Predictions of the model calibrated in Chapter 4 in the case of glucose-fructose and glucose-xylose diauxies.**

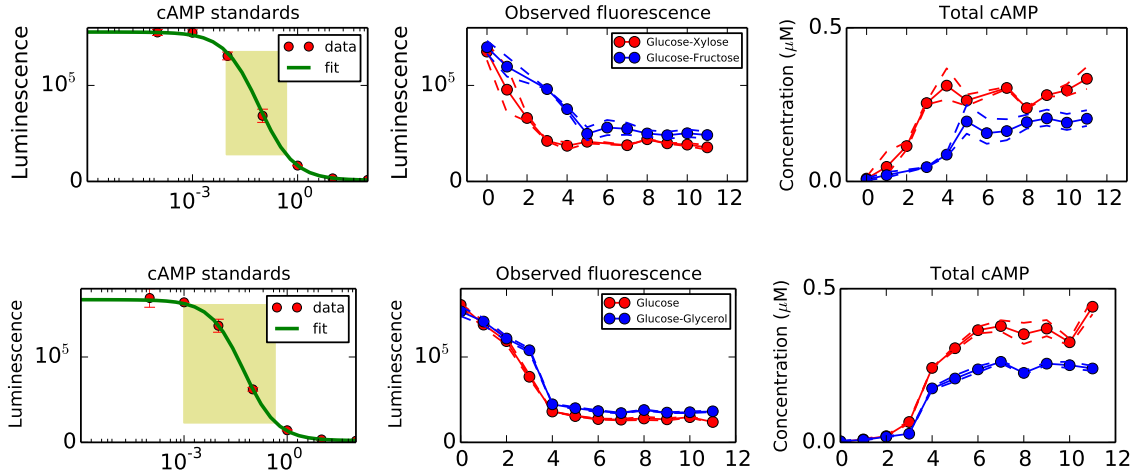


Figure C5: **Dosage of cAMP for the different diauxies discussed in this thesis**. Assays are performed in microplates by batches of ∼70. To be able to make triplicates measurements for each data point, we made separate assays for the diauxies on glucose(-acetate) and glucose-glycerol on one hand (first line), and glucose-fructose, glucose-xylose on the other hand (second line).

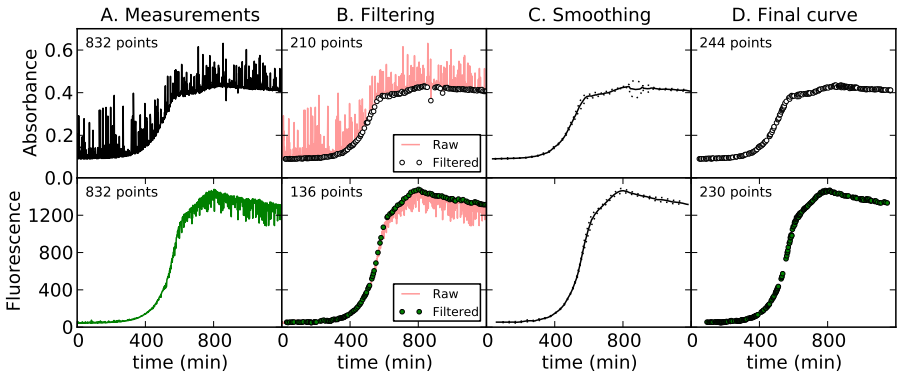


Figure C6: **Different steps of the outliers removal procedure.** Shown are the absorbance and fluorescence measure in a same microplate well containing bacteria growing on M9+Glucose. The measured data undergo different treatments detailed in Section C5.1 to remove all outliers while conserving a manimum of points. The number of points of the curves conserved after each treatment is indicated in the upper right corners.

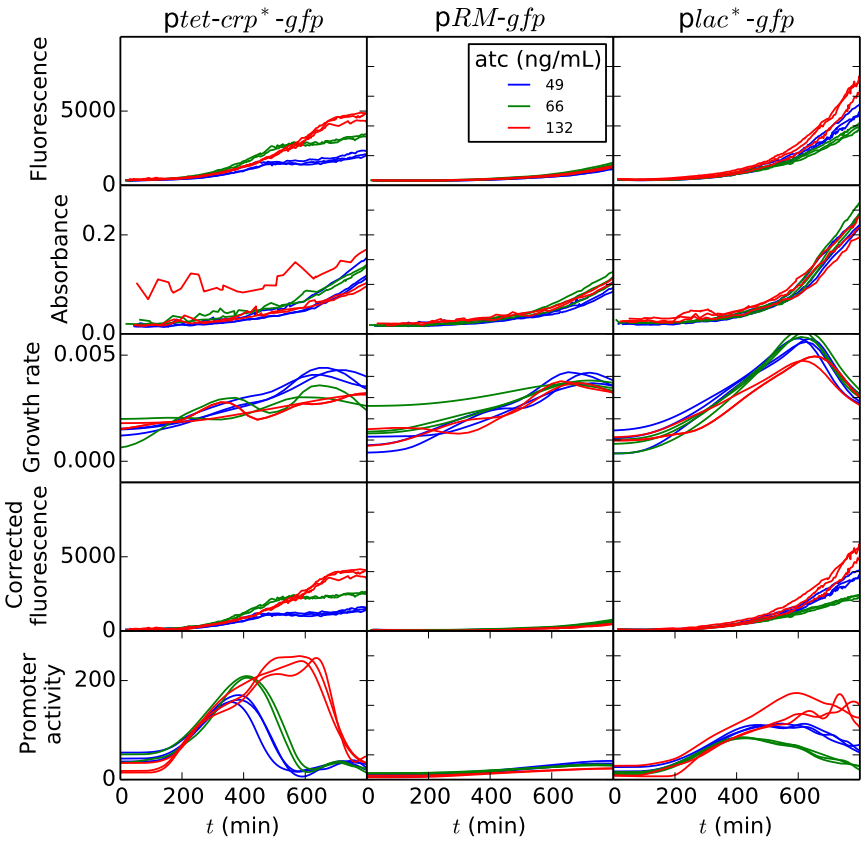


Figure C7: Analysis of the raw experimental data for the growth on glycerol shown in Figure 4.2.

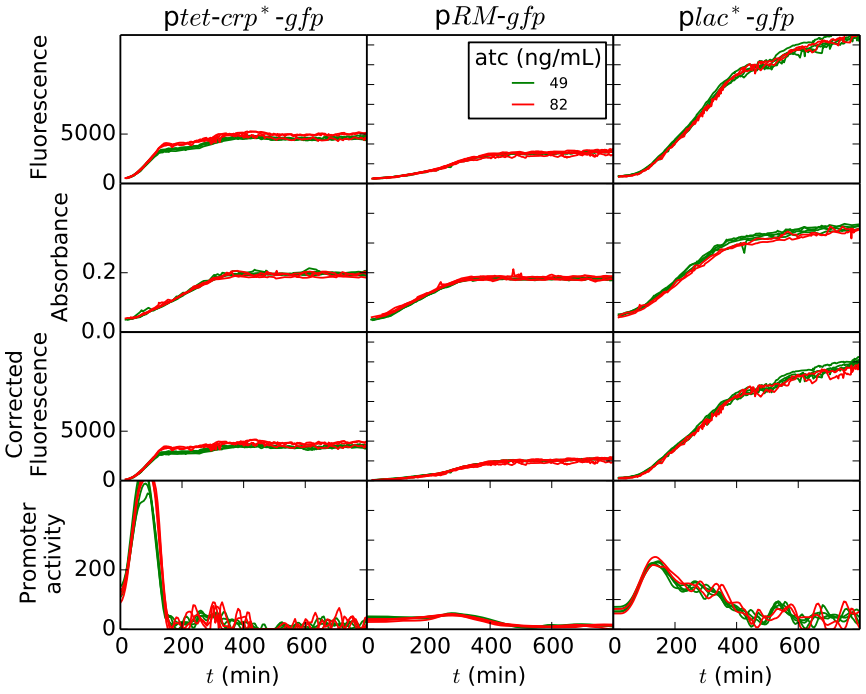


Figure C8: Analysis of the raw experimental data for the glucose-fructose diauxy shown in Figure 4.2.

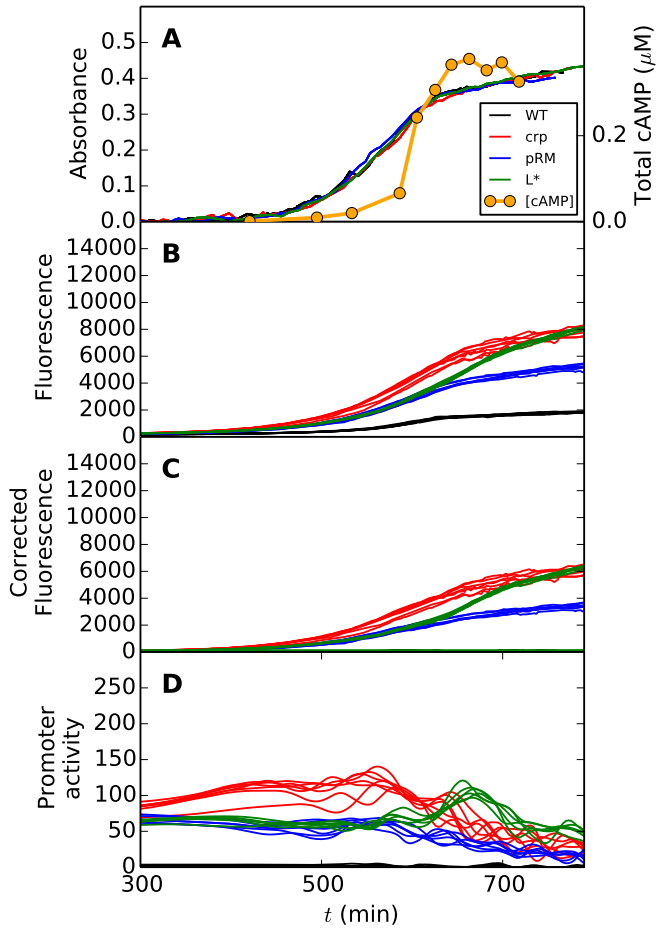


Figure C9: **Experimental data corresponding to growth on glucose in Figure 4.4 of the main text.** The corrected fluorescence profiles are obtained by subtracting the fluorescence of a wild-type strain carrying no reporter plasmid (black lines) to the observed fluorescence profiles. Promoter activities in the fourth panel are computed from absorbance and fluorescence data as explained in Chapter 3.

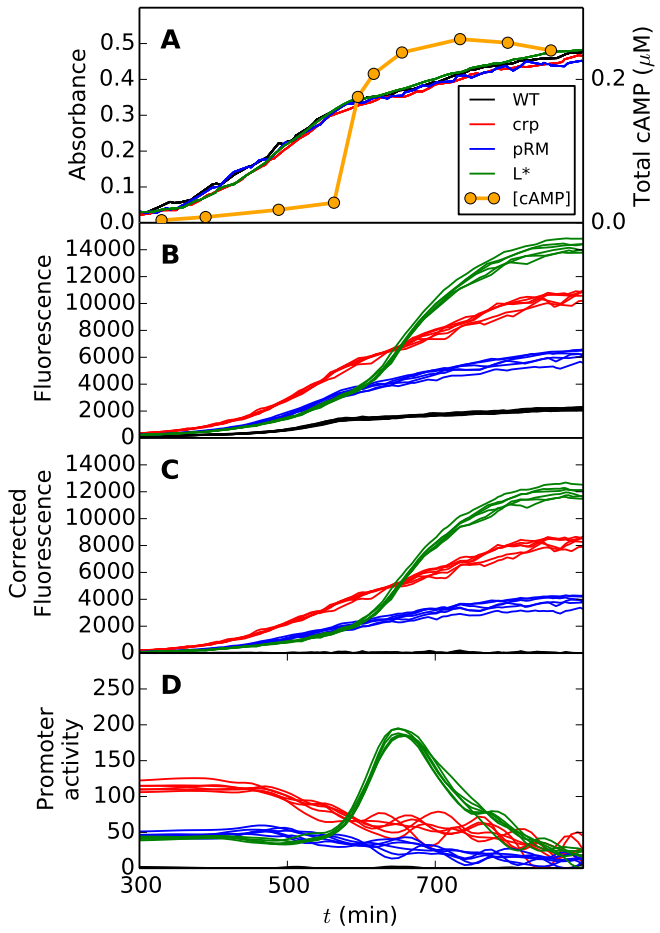


Figure C10: Experimental data used for the Glucose-glycerol diauxy shown in Figure 4.4. See Figure C9 for details.

Bibliography

Aïchaoui, L., Jules, M., Le Chat, L., Aymerich, S., Fromion, V., and Goelzer, A. (2012). BasyLiCA: a tool for automatic processing of a Bacterial Live Cell Array. *Bioinformatics*, 28(20):2705–6.

Alon, U. (2007). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Number 10 in CRC mathematical and computational biology series. Chapman&Hall/CRC.

Andersen, K. and von Meyenburg, K. (1980). Are growth rates of escherichia coli in batch cultures limited by respiration? *Journal of Bacteriology*, 144(1):114–123.

Andersen, M. S., Dahl, J., Liu, Z., and Vandenberghe, L. (2012). Interior-point methods for large-scale cone programming. In Sra, S., Nowozin, S., and Wright, S. J., editors, *Optimization for Machine Learning*, pages 55–83. MIT Press.

Azam, T. A., Iwata, A., Nishimura, A., Ueda, S., and Ishihama, A. (1999). Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *J. Bacteriol.*, 181(20):6361–70.

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

Bansal, L., Chu, Y., Laird, C., and Hahn, J. (2012). Determining transcription factor profiles from fluorescent reporter systems involving regularization of

inverse problems. In *Proc. 2012 American Control Conference (ACC 2012)*, pages 2725–30, Fairmont Queen Elizabeth, Montréal, Canada.

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

Baptist, G., Pinel, C., Ranquet, C., Izard, J., Ropers, D., de Jong, H., and Geiselmann, J. (2013). A genome-wide screen for identifying all regulators of a target gene. *Nucleic Acids Res.*, 41(17):e164.

Bauer, H., Kasper-Giebl, A., Löflund, M., Giebl, H., Hitzenberger, R., Zibuschka, F., and Puxbaum, H. (2002). The contribution of bacteria and fungal spores to the organic carbon content of cloud water, precipitation and aerosols. *Atmospheric Research*, 64(1-4):109–119.

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

Bertero, M. (1989). Linear inverse and ill-posed problems. *Adv. Electron. Electron Phys.*, 75:1–120.

Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D., and Geiselmann, J. (2013). Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.*, 9:634.

Bettenbrock, K. (2005). A Quantitative Approach to Catabolite Repression in Escherichia coli. *Journal of Biological Chemistry*, 281(5):2578–2584.

Bolouri, H. (2008). *Computational Modeling of Gene Regulatory Networks – A Primer*. Imperial College Press, London.

Borghans, J., de Boer, R., and Segel, L. (1996). Extending the quasi-steady state approximation by changing variables. *Bulletin of Mathematical Biology*, 58(1):43–63.

Boyer, F., Besson, B., Baptist, G., Izard, J., Pinel, C., Ropers, D., Geiselmann, J., and de Jong, H. (2010). WellReader: a MATLAB program for the analysis of fluorescence and luminescence reporter gene data. *Bioinformatics*, 26(9):1262–3.

Bradley, M., Beach, M., de Koning, A., Pratt, T., and Osuna, R. (2007). Effects of Fis on *Escherichia coli* gene expression during different growth stages. *Microbiology*, 153(9):2922–40.

Brock, T. D., Brock, K. M., Belly, R. T., and Weiss, R. L. (1972). Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low ph and high temperature. *Archiv für Mikrobiologie*, 84(1):54–68.

Chen, C. (1970). *Introduction to Linear System Theory*. Holt, Rinehart and Winston, New York.

Chen, W., Niepel, M., and Sorger, P. (2010). Classic and contemporary approaches to modeling biochemical reactions. *Genes and Development*, 24(17):1861–1876.

Chudakov, D., Matz, M., Lukyanov, S., and Lukyanov, K. (2010). Fluorescent proteins and their applications in imaging living cells and tissues. *Physiol. Rev.*, 90(3):1103–63.

Crasnier-mednansky, M., Officer, C. S., and Valley, P. (2008). Correspondance between Randy Schekman and Martine Crasnier-Mednansky.

Crockford, D. (2006). The application/json media type for JavaScript Object Notation (JSON). http://tools.ietf.org/html/rfc4627.

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

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

de Nicolao, G., Sparacino, G., and Cobelli, C. (1997). Nonparametric input estimation in physiological systems: Problems, methods, and case studies. *Automatica*, 33(5):851–70.

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

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

Dessein, a., Schwartz, M., and Ullmann, a. (1978). Catabolite repression in Escherichia coli mutants lacking cyclic AMP. *Molecular & general genetics : MGG*, 162(1):83–7.

Enjalbert, B., Letisse, F., and Portais, J.-C. (2013). Physiological and molecular timing of the glucose to acetate transition in *Escherichia Coli*. *Metabolites*, 3(3):820–37.

Eppler, T. and Boos, W. (1999). Glycerol-3-phosphate-mediated repression of malT in Escherichia coli does not require metabolism, depends on enzyme IIAGlc and is mediated by cAMP. *Molecular microbiology*, 33.

Epstein, W., Rothman-Denes, L. B., and Hesse, J. (1975). Adenosine 3':5'-cyclic monophosphate as mediator of catabolite repression in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 72(6):2300–4.

Escalante, A. and Calderón, R. (2010). Metabolic engineering for the production of shikimic acid in an evolved Escherichia coli strain lacking the phosphoenolpyruvate: carbohydrate. *Microbial Cell Factories*, 9(21).

Finkenstädt, B., Heron, E., Komorowski, M., Edwards, K., Tang, S., Harper, C., Davis, J., White, M., Millar, A., and Rand, D. (2008). Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24):2901–7.

Fraser, A. and Yamazaki, H. (1979). Effect of carbon sources on the rates of cyclic AMP synthesis, excretion, and degradation, and the ability to produce β-galactosidase in Escherichia coli. *Canadian journal of biochemistry*, 8(57):1073–1079.

Garrett, J. (2006). Ajax: A new approach to web applications. http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications/.

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

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

Glass, L. and Kauffman, S. (1973). The logical analysis of continuous non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129.

Goldbeter, A. (1996). *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour*. Cambridge Univ. Press.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good Ridge parameter. *Technometrics*, 21(2):215–23.

Görke, B. and Stülke, J. (2008a). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature reviews. Microbiology*, 6(8):613–24.

Görke, B. and Stülke, J. (2008b). Is there any role for cAMP-CRP in carbon catabolite repression of the Escherichia coli lac operon? Reply from Görke and Stülke. *Nature reviews. Microbiology*, 6(12):954; author reply 954.

Gosset, G., Zhang, Z., Nayyar, S., Cuevas, W., and Saier Jr, M. (2004a). Transcriptome analysis of Crp-dependent catabolite control of gene expression in *Escherichia coli*. *Journal of Bacteriology*, 186(11):3516–3524.

Gosset, G., Zhang, Z., Nayyar, S., Cuevas, W., and Saier Jr, M. (2004b). Transcriptome analysis of Crp-dependent catabolite control of gene expression in *Escherichia coli*. *J. Bacteriol.*, 186(11):3516–24.

Gottschalk, G. (1986). *Bacterial Metabolism*. Springer, New York, 2nd edition.

Gutierrez-Ríos, R., Freyre-Gonzalez, J., Resendis, O., Collado-Vides, J., Jr, M. S., and Gosset, G. (2007). Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. *BMC Microbiology*, 7:53.

Heinrich, R. and Schuster, S. (1996). *The Regulation of Cellular Systems*. Chapman & Hall, New York.

Hengge-Aronis, R. (2002). Signal transduction and regulatory mechanisms involved in control of the σ^S (RpoS) subunit of RNA polymerase. *Microbiol. Mol. Biol. Rev.*, 66(3):373–95.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

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

Inada, T., Takahashi, H., Mizuno, T., and Aiba, H. (1996). Down regulation of cAMP production by cAMP receptor protein in Escherichia coli: an assessment of the contributions of transcriptional and posttranscriptional control of adenylate cyclase. *Molecular general genetics MGG*, 253(1-2):198–204.

Ishizuka, H., Hanamura, A., Inada, T., and Aiba, H. (1994). Mechanism of the down-regulation of cAMP receptor protein by glucose in Escherichia coli: role of autoregulation of the crp gene. *the The European Molecular Biology Organization Journal*, 13(13):3077–3082.

Ishizuka, H., Hanamura, a., Kunimura, T., and Aiba, H. (1993). A lowered concentration of cAMP receptor protein caused by glucose is an important determinant for catabolite repression in Escherichia coli. *Molecular microbiology*, 10(2):341–50.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. http://www.scipy.org/.

Kao, K., Tran, L., and Liao, J. (2005). A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. *J. Biol. Chem.*, 280(43):36079–87.

Kaplan, S., Bren, A., Zaslaver, A., Dekel, E., and Alon, U. (2008). Diverse two-dimensional input functions control bacterial sugar genes. *Molecular cell*, 29(6):786–792.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401.

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

Khalil, H. (2001). *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd ed. edition.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2009). *Systems Biology: A Textbook*. Wiley-Blackwell.

Klumpp, S., Zhang, Z., and Hwa, T. (2009). Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1–21.

Kolb, A., Busby, S., Buc, H., Garges, S., and Adhya, S. (1993a). Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.*, 62:749–795.

Kolb, A., Busby, S., Buc, H., Garges, S., and Adhya, S. (1993b). Transcriptional regulation by cAMP and its receptor protein. *Annual Review of Biochemistry*, 62:749–795.

Kremling, A. (2007). Comment on mathematical models which describe transcription and calculate the relationship between mrna and protein expression ratio. *Biotechnology and Bioengineering*, 96(4):815–819.

Kuhlman, T., Zhang, Z., Saier Jr., M., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, 104(14):6043–8.

Kuo, J.-T., Chang, Y.-J., and Tseng, C.-P. (2003). Growth rate regulation of lac operon expression in Escherichia coli is cyclic AMP dependent. *FEBS Letters*, 553(3):397–402.

Larrabee, K., Phillips, J., Williams, G., and Larrabee, A. (1980a). The relative rates of protein synthesis and degradation in a growing culture of *Escherichia coli*. *Journal of Biological Chemistry*, 255(9):4125–4130.

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

Leveau, J. and Lindow, S. (2001). Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J. Bacteriol.*, 183(23):6752–62.

Lichten, C., White, R., Clark, I., and Swain, P. (2014). Unmixing of fluorescence spectra to resolve quantitative time-series measurements of gene expression in plate readers. *BMC Biotechnol.*, 14:11.

Locke, J., Southern, M., Kozma-Bognár, L., Hibberd, V., Brown, P., Turner, M., and Millar, A. (2005). Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular Systems Biology*. Article number: 2005.0013.

Mehra, A., Lee, K., and Hatzimanikatis, V. (2003). Insights into the relation between mRNA and protein expression patterns: I. Theoretical considerations. *Biotechnology and Bioengineering*, 84(7):822–833.

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

Monod, J. (1942). Recherches sur la croissance des cultures bacteriennes.

Morozova, N., Zinovyev, A., Nonne, N., Pritchard, L. L.and Gorban, A. N., and Harel-Bellan, A. (2012). Kinetic signatures of microRNA modes of action. *RNA*, 18(9):1635–1655.

Mosteller, R., Goldstein, R., and Nishimoto, K. (1980). Metabolism of individual proteins in exponentially growing *Escherichia coli*. *Journal of Biological Chemistry*, 255(6):2524–2532.

Müller, J., Oehler, S., and Müller-Hill, B. (1996). Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *Journal of molecular biology*, 257(1):21–9.

Narang, A. (2009a). cAMP does not have an important role in carbon catabolite repression of the Escherichia coli lac operon. *Nature reviews. Microbiology*, 7(3):250.

Narang, A. (2009b). Quantitative effect and regulatory function of cyclic adenosine 5'-phosphate in Escherichia coli. *Journal of biosciences*, 34(3):445–63.

Narang, A. and Pilyugin, S. S. (2007). Bacterial gene regulation in diauxic and non-diauxic growth. *Journal of theoretical biology*, 244(2):326–48.

Odaka, H., Arai, S., Inoue, T., and Kitaguchi, T. (2014). Genetically-encoded yellow fluorescent cAMP indicator with an expanded dynamic range for dual-color imaging. *PloS one*, 9(6):e100252.

Oehler, S. and Amouyal, M. (1994). Quality and position of the three lac operators of E. coli define efficiency of repression. *The EMBO ...*, 13(14):3348–3355.

Oh, M., Rohlin, L., Kao, K., and Liao, J. (2002). Global expression profiling of acetate-grown *Escherichia coli*. *Journal of Biological Chemistry*, 277(15):13175–13183.

Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and Van Oudenaarden, A. (2004). Multistability in the lactose utilization network of Escherichia coli. *Nature*, 427(6976):737–40.

Pastan, I. R. A. and Sankar, A. (1976). Cyclic Adenosine 5 ' -Monophosphate in Escherichia coli. *Bacteriological Reviews*, 40(3):527–551.

Pedraza, J. and van Oudenaarden, A. (2005). Noise propagation in gene networks. *Science*, 307(5717):1965–1968.

Politi, N., Pasotti, L., Zucca, S., Casanova, M., Micoli, G., Cusella De Angelis, M. G., and Magni, P. (2014). Half-life measurements of chemical inducers for recombinant gene expression. *Journal of biological engineering*, 8(1):5.

Polynikis, A., Hogan, S., and di Bernardo, M. (2009). Comparing different ODE modelling approaches for gene regulatory networks. *Journal of Theoretical Biology*, 261(4):511–530.

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

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

Roussel, M. and Fraser, S. (2001). Invariant manifold methods for metabolic model reduction. *Chaos*, 11(1):196–206.

Santillán, M. and Mackey, M. (2001). Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data. *Proceedings of the National Academy of Sciences of the USA*, 98(4):1364–1369.

Santillán, M. and Mackey, M. C. (2008). Quantitative approaches to the study of bistability in the lac operon of Escherichia coli. *Journal of the Royal Society, Interface / the Royal Society*, 5 Suppl 1(August):S29–39.

Schaechter, M., Maaløe, O., and Kjeldgaard, N. (1958). Dependency on medium and temperature of cell size and chemical composition during balanced grown of *Salmonella typhimurium*. *Journal of General Microbiology*, 19(3):592–606.

Segel, L. and Slemrod, M. (1989). The quasi-steady-state assumption: A case-study in perturbation. *SIAM Review*, 31(3):446–477.

Sekar, R. B. and Periasamy, A. (2003). Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations. *The Journal of cell biology*, 160(5):629–33.

Stefan, D., Pinel, C., Pinhal, S., Cinquemani, E., Geiselmann, J., and de Jong, H. (2015). Inference of quantitative models of bacterial promoters from time-series reporter gene data. *PLoS Comput. Biol.*, 11(1):e1004028.

Steinhoff, U. (2005). Who controls the crowd? New findings and old questions about the intestinal microflora. *Immunology letters*, 99(1):12–6.

Tagami, H., Inada, T., Kunimura, T., and Aiba, H. (1995). Glucose lowers CRP* levels resulting in repression of the lac operon in cells lacking cAMP. *Molecular Microbiology*, 17(2):251–258.

Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585.

Tournier, L. and Chaves, M. (2009). Uncovering operational interactions in genetic networks using asynchronous boolean dynamics. *Journal of theoretical biology*, 260(2):196–209.

Travers, A. and Muskhelishvili, G. (2005). DNA supercoiling: A global transcriptional regulator for enterobacterial growth? *Nat. Rev. Microbiol.*, 3(2):157–69.

Tsakraklides, V., Shaw, a. J., Miller, B. B., Hogsett, D. a., and Herring, C. D. (2012). Carbon catabolite repression in Thermoanaerobacterium saccharolyticum. *Biotechnology for biofuels*, 5(1):85.

Tsien, R. (1998). The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–544.

Turanyi, T., Tomlin, A. S., and Pilling, M. J. (1993). On the error of the quasi-steady-state approximation. *Journal of Physical Chemistry*, 97(1):163–172.

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

Volkmer, B. and Heinemann, M. (2011). Condition-dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling. *PLoS One*, 6(7):e23126.

von Dassow, G., Meir, E., Munro, E., and Odell, G. (2000). The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–192.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

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

Wanner, B., Kodaira, R., and Neidhardt, F. (1978). Regulation of lac operon expression: reappraisal of the theory of catabolite repression. *Journal of bacteriology*.

Wolfe, A. J. (2005). The acetate switch. *Microbiol. Mol. Biol. Rev.*, 69(1):12–50.

You, C., Okano, H., Hui, S., Zhang, Z., Kim, M., Gunderson, C. W., Wang, Y.-P., Lenz, P., Yan, D., and Hwa, T. (2013). Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature*, 500(7462):301–306.

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

Zhuang, K., Vemuri, G. N., and Mahadevan, R. (2011). Economics of membrane occupancy and respiro-fermentation. *Molecular systems biology*, 7(500):500.