UNIVERSITAT POLITÈCNICA DE VALÈNCIA

A systems engineering approach to model, tune and test synthetic gene circuits

PhD dissertation

Author: Yadira Boada Acosta

Supervisor: Jesús Picó i Marco

October 16, 2018

Instituto Universitario de Automática e Informática Industrial
Departamento de Ingeniería de Sistemas y Automática
Universitat Politècnica de València

This work has been partially supported by the Spanish Government (CICYT DPI2014-55276-C5-1) and the European Union (FEDER). The author was recipient of the grant *Formación de Personal Investigador* by the Universitat Politècnica de València, subprogram 1 (FPI/2013-3242).
She was also recipient of the competitive grants for pre-doctoral stays *Erasmus Student Placement*-European Programme 2015, and *FPI Mobility program* 2016 of the Universitat Politècnica de València.
She also received the competitive grant for a pre-doctoral stay *Becas de movilidad para Jóvenes Profesores e Investigadores* 2016, Programa de Becas Iberoamérica of the Santander Bank. Support from the Georgiev Lab and Department of Cybernetics from University of West Bohemia, Industrial Engineering and Systems Group at the Pontificia Universidade Católica do Paraná, and Mosaic Group from Center for Systems Biology Dresden at Technische Universität Dresden and Max Planck Institute of Molecular Cell Biology and Genetics. All of them are also acknowledged for accepting the author in their facilities as pre-doctoral stay and for their valuable collaboration sharing knowledge.

To my family Luis, Mariana, Alejandro, Anahí, and
my husband Alejandro.

Acknowledgements

A pesar de tratar que esta página no se extienda demasiado, cuando empecé a recordar a cada una de las personas que me han acompañado en la realización de esta tesis, el papel ha quedado algo corto.

En primer lugar quiero agradecer a Jesús Picó i Marco porque además de ser un director de tesis paciente y sabio, ha sido para mi un gran referente como profesional y ser humano. Siempre culto, amante de su trabajo y de la ciencia, pero siempre práctico y en constante aprendizaje que invita a seguir sin descanso como él lo hace. Muchas gracias Jesús porque desde hace 6 años he tomado nota de las valiosas lecciones de vida que me has enseñado en cada reunión, viaje, café o “cerveceta", y aspiro ser la mitad de buena que tú Jesús. Así mismo, le doy las gracias a José Luis Navarro por los consejos sobre modelos matématicos, biología, teoría de control y automática que no han sido pocos, y que personalmente me han sido de gran utilidad en la consecución de esta tesis.

En segundo lugar quiero agradecer a mis padres Mariana y Luis Fernando, así como a mi hermana Anahí y mi hermano Alejandro que aunque lejos desde Quito, siempre han sido y serán las personas que ilusionan cada paso que doy. Se han convertido en el soporte incondicional que viajó conmigo hace 7 años, pero que realmente me sostiene desde niña, ayudándome a superar momentos tristes, y regocijándome en los días más alegres. Perdón si me he ausentado este último tiempo. Pero mami, papi ustedes me enseñaron a volar con mis propias alas, y si cabe ese es el ejemplo que quiero dar a mis *ñaños* ahora que me he vuelto un poco más mayor. Gracias por su paciencia y su amor, este trabajo lo he culminado con su ayuda y ese mérito quiero dejarlo aquí escrito. De igual forma, muchas gracias a mis tías y tíos -de San Juan y San Carlos como los recuerdo cariñosamente-, a mis primos y primas por su *“que esté buenita niña, muchos saludos y cuídese*" que en cada conversación mientras comían y yo interrumpía, nunca dejaron de repetirme. Especialmente, gracias por el infinito cariño a la tía Tere.

Deseo agradecer a Ale, por sus incansables fuerzas, sus valiosos consejos de biología y automática, sus exquisitas recetas y por todo el amor que me ha brindado. Es justamente este amor el que nos dio el empujón del *si quiero* este año, y no pude encontrar mejor compañero de vida, gracias por todo mi vida!. Y muchas gracias también a toda mi nueva familia en Argentina.

In this paragraph, I would like to thank several smart people who help me to understand and enjoy synthetic biology. Thanks to Dr. Daniel Georgiev and his crew at the Georgiev Lab -particularly Pavel o Pablito- for introducing me the Wet-lab techniques and the exciting life in the lab. Thanks to Dr. Gilberto Reynoso-Meza at the PPGEPS group for mentoring me about Multi-objective optimization principles, algorithms and their applications. Most important, thank you for your friendship and advise. Finally, I'd like to thank Prof. Dr. Ivo Sbalzarini and his MOSAIC research group -especially Dr. Pietro Incadorna- for providing me the valuable help with the Parallel Particle Mesh library to perform very fast stochastic simulations.

Finalmente, a mis buenos amigos Vane y Juanpa que siempre me decían "*ánimo, ya falta poco*", muchas gracias por estar presentes y cuidar nuestra amistad, principalmente siempre sabré que podemos continuar con nuestras charlas de vida. Gracias por su amistad a Tati y Reyes, son una pareja maravillosa que han vuelto de València mi segundo hogar. A mis compañeros de *La sala* que ya no están: Jopipe, Diego, Jesusito, Iván Llopis y Manuel; y a quienes aún comparten conmigo: Vane -Vanessín o Vanessa con 2 s-, Henry, Alberto, JuanFer, Eslam, Clara, Iván y Fernando; les debo tantas risas, reuniones, viajes, congresos y hasta bodas. Innumerables experiencias que nunca pensé vivir, y que siempre recordaré. Finalmente, gracias a Roger porque aunque es el más joven, se ha convertido en gran amigo es estos dos últimos veranos.

Abstract

Synthetic biology is defined as the engineering of biology: the deliberate (re)design and construction of novel biological and biologically based parts, devices and systems to perform new functions for useful purposes, that draws on principles elucidated from biology and engineering. Methods and tools are needed to facilitate fast, reproducible and predictable construction of biological systems from sets of biological components.

This thesis raises multi-objective optimization as the proper framework to deal with common problems arising in rational design and optimal tuning of synthetic gene circuits. Using a classical systems engineering approach, the thesis mainly addresses: *i)* synthetic gene circuit modelling based on first-principles, *ii)* model parameters estimation from experimental data and *iii)* model-based tuning to achieve desired circuit performance.

Two gene synthetic circuits of different nature and with different goals and inherent problems have been used throughout the thesis: an Incoherent type 1 feedforward circuit (I1-FFL) that exhibits the important biological property of *adaptation*, and a Quorum sensing/Feedback circuit (QS/Fb) comprising two intertwined feedback loops –an intracellular one and a cell-to-cell communication-based one– designed to regulate the mean expression level of a protein of interest while minimizing its variance across the population of cells. Both circuits have been analyzed *in silico* and implemented *in vivo*.

In both cases, circuit modelling based on first-principles has been carried out. Then, special attention is paid to illustrate how to obtain reduced order models amenable for parameters estimation yet keeping biological significance.

Model parameters estimation from experimental data is considered in different scenarios, both using deterministic and stochastic models. For the I1-FFL circuit, deterministic models are considered. In this case, the thesis raises ensemble modelling using multi-objective optimization to perform model parameters estimation under scenarios with incomplete model structure (unmodeled dynamics). For the QS/Fb gene circuit, a feedback controlled structure, the lack of excitability of the signals is the problem addressed. The thesis proposes a two-stage estimation methodology using stochastic models. The methodology allows using population averaged time-course data and steady-state distribution measurements at the single-cell level.

Model-based circuit tuning to achieve desired circuit performance is also addressed using multi-objective optimization. First, for the QS/Fb feedback control circuit, a complete stochastic analysis is performed. Here, the thesis addresses how to correctly take into account both intrinsic and extrinsic noise, the two main sources of noise in gene synthetic circuits. The trade-off between both sources of noise, and the role played by in the intracellular single-cell feedback loop and the extracellular population-wide feedback is analyzed. The main conclusion being that the complex interplay between both feedback channels compel the use of multi-objective optimization for proper tuning of the circuit to achieve desired performance. Thus, the thesis wraps up all the previous results and uses them to address circuit tuning for desired performance. Here, besides the proper use of multi-objective optimization tools, the main concern is how to derive guidelines for circuit parameters tuning in silico that can realistically be applied in vivo in a standard laboratory. Thus, as an alternative to classical parameters sensitivity analysis, the thesis proposes the use of clustering techniques along the optimal Pareto fronts relating the performance trade-offs with regions in the circuits parameters space.

In summary, the thesis provides useful practical methods and tools for modelling, parameters estimation, analysis and practical tuning of synthetic gene circuits, both in the deterministic and stochastic domains, using multi-objective optimization as common framework.

Resumen

La biología sintética se define como la ingeniería de la biología: el (re)diseño y construcción de nuevas partes, dispositivos y sistemas biológicos para realizar nuevas funciones con fines útiles, que se basan en principios elucidados de la biología y la ingeniería. Para facilitar la construcción rápida, reproducible y predecible de estos sistemas biológicos a partir de conjuntos de componentes es necesario desarrollar nuevos métodos y herramientas.

Esta tesis plantea la optimización multiobjetivo como el marco adecuado para tratar los problemas comunes que surgen en el diseño racional y el ajuste óptimo de los circuitos genéticos sintéticos. Utilizando un enfoque clásico de ingeniería de sistemas, la tesis se centra principalmente en: *i)* el modelado de circuitos genéticos sintéticos basado en los primeros principios, *ii)* la estimación de parámetros de modelos a partir de datos experimentales y *iii)* el ajuste basado en modelos para lograr el desempeño deseado de los circuitos.

A lo largo de la tesis se han utilizado dos circuitos genéticos sintéticos de diferente naturaleza y con diferentes objetivos y problemas: un circuito de realimentación de tipo 1 incoherente (I1-FFL) que exhibe la importante propiedad biológica de adaptación, y un circuito de detección de quorum sensing y realimentación (QS/Fb) que comprende dos bucles de realimentación entrelazados -uno intracelular y uno basado en la comunicación de célula a célula- diseñado para regular el nivel medio de expresión de una proteína de interés mientras se minimiza su varianza a través de la población de células. Ambos circuitos han sido analizados *in silico* e implementados *in vivo.*

En ambos casos, se han desarrollado modelos de estos circuitos basado en primeros principios. Luego, se presta especial atención a ilustrar cómo obtener modelos de orden reducido susceptibles de estimación de parámetros, pero manteniendo el significado biológico.

La estimación de los parámetros del modelo a partir de los datos experimentales se considera en diferentes escenarios, tanto utilizando modelos determinísticos como estocásticos. Para el circuito I1-FFL se consideran modelos determinísticos. En este caso, la tesis plantea la utilización de modelos locales utilizando la optimización multiobjetivo

para realizar la estimación de parámetros del modelo bajo escenarios con estructura de modelo incompleta (dinámica no modelada). Para el circuito QS/Fb, una estructura controlada por realimentación, el problema tratado es la falta de excitabilidad de las señales. La tesis propone una metodología de estimación en dos etapas utilizando modelos estocásticos. La metodología permite utilizar datos de curso temporal promediados de la población y mediciones de distribución en estado estacionario a nivel de una sola célula.

El ajuste de circuitos basado en modelos para lograr el desempeño del circuito deseado también se aborda mediante la optimización multiobjetivo. En primer lugar, para el circuito de control de realimentación QS/Fb, se realiza un análisis estocástico completo. Aquí, la tesis aborda cómo tener en cuenta correctamente tanto el ruido intrínseco como el extrínseco, las dos principales fuentes de ruido en los circuitos genéticos sintéticos. Se analiza el equilibrio entre ambas fuentes de ruido y el papel que desempeñan en el bucle de realimentación intracelular, y en la realimentación extracelular de toda la población. La principal conclusión es que la compleja interacción entre ambos canales de realimentación obliga al uso de la optimización multiobjetivo para el adecuado ajuste del circuito. En esta tesis además del uso adecuado de herramientas de optimización multiobjetivo, la principal preocupación es cómo derivar directrices para el ajuste in silico de parámetros de circuitos que puedan aplicarse de forma realista in vivo en un laboratorio estándar. Así, como alternativa al análisis de sensibilidad de parámetros clásico, la tesis propone el uso de técnicas de clustering a lo largo de los frentes de Pareto, relacionando el compromiso de rendimiento con las regiones en el espacio de parámetros.

En resumen, la tesis proporciona métodos y herramientas prácticas útiles para el modelado, la estimación de parámetros, el análisis y el ajuste práctico de circuitos genéticos sintéticos, tanto en el dominio determinístico como en el estocástico, utilizando optimización multiobjetivo como marco común.

Resum

La biologia sintètica es defineix com l'enginyeria de la biologia: el (re) disseny i construcció de noves parts, dispositius i sistemes biològics per a realitzar noves funcions útils que es basen a principis elucidats de la biologia i l'enginyeria. Per facilitar la construcció ràpida, reproduïble i predictible de aquests sistemes biològics a partir de conjunts de components és necessari desenvolupar nous mètodes i eines.

Aquesta tesi planteja la optimització multiobjectiu com el marc adequat per a tractar els problemes comuns que apareixen en el disseny racional i l' ajust òptim dels circuits genètics sintètics. Utilitzant un enfocament clàssic d'enginyeria de sistemes, la tesi es centra principalment en: *i)* el modelatge de circuits genètics sintètics basat en primers principis, *ii)* l' estimació de paràmetres de models a partir de dades experimentals i *iii)* l' ajust basat en models per aconseguir el rendiment desitjat dels circuits.

Al llarg de la tesi s'han utilitzat dos circuits genètics sintètics de diferent naturalesa i amb diferents objectius i problemes: un circuit de prealimentació de tipus 1 incoherent (I1-FFL) que exhibeix la important propietat biològica d'adaptació, i un circuit de quorum sensing i realimentació (QS/Fb) que comprèn dos bucles de realimentació entrellaçats -un intracel·lular i un basat en la comunicació de cèl·lula a cèl·lula- dissenyat per regular el nivell mitjà d'expressió normal d'una proteïna d'interès mentre es minimitza la seua variació al llarg de la població de cèl·lules. Els dos circuits han estat analitzats *in silico* i implementats *in vivo*.

En tots dos casos, s'han desenvolupat models basats en primers principis d'aquests circuits. Després es presta especial atenció a delinear com obtenir models d'ordre reduït susceptibles de estimació de paràmetres, però mantenint el significat biològic. L' estimació dels paràmetres del model a partir de les dades experimentals es considera en diferents escenaris, tant utilitzant models determinístics com estocàstics. Per al circuit I1-FFL es consideren models determinístics. En aquest cas, la tesi planteja la utilització de models locals utilitzant la optimització multiobjectiu per realitzar l'estimació de parametres del model sota escenaris amb estructura de model incompleta (dinàmica no modelada). Per al circuit de QS/Fb, una estructura controlada per realimentació, el problema tractat és la manca d'excitabilitat dels senyals. La tesi proposa una metodologia de estimació en dues etapes utilitzant models estocàstics.

La metodologia permet utilitzar dades de curs temporal promediats de la població i mesures de distribució en estat estacionari a nivell d'una sola una cèl·lula.

L' ajust de circuits basat en models per aconseguir el rendiment desitjat dels circuits també s' aborda mitjançant la optimització multiobjectiu. Per al circuit de control de realimentació de QS/Fb, es fa un anàlisi estocàstic complet. Açí, la tesi aborda com tenir en compte correctament tant el soroll intrínsec com l' extrínsec, les dues principals fonts de soroll en els circuits genètics sintètics. S' analitza l'equilibri entre dues fonts de soroll i el paper que exerceixen en el bucle de realimentació intracel·lular, les i en la realimentació extracel·lular de tota la població. La principal conclusió es que la complexa interacció entre els dos canals de realimentació fa necessari l' ús de la optimització multiobjectiu per al adequat ajust del circuit. En aquesta tesi, a més de l'ús adequat d'eines d'optimització multiobjectiu, la principal preocupació és com derivar directives per al ajust *in silico* de paràmetres de circuits que puguin aplicar-se de forma realista en viu en un laboratori estàndard. Així, com a alternativa a l'anàlisi de sensibilitat de paràmetres clàssic, la tesi proposa l'ús de l' tècniques de l'agrupació al llarg dels fronts de Pareto, relacionant el compromís de dessempeny amb les regions en l'espai d'paràmetres. En resum, la tesi proporciona mètodes i eines pràctiques útils per a la modelització, l' estimació de paràmetres, l'anàlisi i l' ajust pràctic de circuits genètics sintètics, tant en el domini determinístic com en l'estocàstic, utilitzant optimització multiobjectiu com marc comú.

Contents

Chapter 1

Thesis outline and contributions

Thesis outline

In the remainder of this Thesis, Chapter 2 describes the state of the field of Synthetic biology, aiming to develop new and functional synthetic gene circuits. Its design-build-test cycle brings out some of the concepts further detailed in this work. Chapter 3 introduces two synthetic gene circuits used as case-studies. The well-known Incoherent type 1 feedforward circuit (I1-FFL) that presents Adaptation as an important biological feature, and the Quorum senging/Feedback circuit (QS/Fb) developed to reduce noise in protein production that incorporates feedback control to improve the system robustness. Chapter 4 describes the deterministic and stochastic models for these circuits using first-principles to capture both the single cell and the cells population dynamics. The models are systematically reduced to obtain more tractable models even for complex systems that do not suffer from over-parametrization as well as computational cost. Chapter 5 presents a multi-objective optimization framework for model parameter estimation, including experimental data of different nature for the same system identification process. Chapter 6 describes the stochastic analysis of a feedback control synthetic gene circuit. The chapter elucidates the benefits from the interplay between feedback and cell-to-cell communication in the QS/Fb gene circuit. Finally, Chapter 7 tunes the performance of a gene circuit via multi-objective optimization. This methodology allow us design efficient and optimal synthetic gene controllers.

Contributions

All the results of this work have been published in:

Refereed Journal Papers:

- Y. Boada, A. Vignoni, and J. Picó. Model reduction and multi-objective identification of a feedback synthetic gene circuit. *IEEE Transactions on Control Systems Technology*. *Accepted*.
- Y. Boada, A. Vignoni, and J. Picó. Engineered control of genetic variability reveals interplay among quorum sensing, feedback regulation, and biochemical noise. *ACS Synthetic Biology*, 6(10):1903–1912, 2017a. doi: 10.1021/acssynbio.7b00087.
- Y. Boada, G. Reynoso-Meza, J. Picó, and A. Vignoni. Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices: an adaptive network case. *BMC Syst Biol*, 10(1):27, 2016b.
- J. Picó, A. Vignoni, E. Picó-Marco, and Y. Boada. Modelling biochemical systems: from mass action kinetics to linear noise approximation. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, 12(3):241–252, 7 2015.

Conference and Presentation Papers, Posters:

- Y. Boada, A. Vignoni, D. Oyarzún, and J. Picó. Host-circuit interactions explain unexpected behaviours of a feedforward gene circuit. 2018. Foundations of Systems Biology in Engineering FOSBE. *Acepted*.
- Y. Boada, A. Vignoni, and J. Picó. Multi-objective optimization for gene expression noise reduction in a synthetic gene circuit. *IFAC-PapersOnLine*, 50(1): 4472 – 4477, 2017b. ISSN 2405-8963. 20th IFAC World Congress.
- Y. Boada, A. Vignoni, and J. Picó. Multi-objective identification of synthetic circuits stochastic models using flow cytometry data. *Proceedings 25th Mediterranean Conference on Control and Automation MED*, pages 1077–1082, 2017c.
- Y. Boada, J. Pitarch, A. Vignoni, G. Reynoso-Meza, and J. Picó. Optimization alternatives for robust model-based design of synthetic biological circuits. *IFAC-PapersOnLine*, 49(7):821 – 826, 2016a. ISSN 2405-8963. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB.

- Y. Boada, A. Vignoni, G. Reynoso-Meza, and J. Picó. Parameter identification in synthetic biological circuits using multi-objective optimization. volume 49, pages 77 – 82, 2016c. Foundations of Systems Biology in Engineering FOSBE.

- E. Picó-Marco, Y. Boada, J. Picó, and A. Vignoni. Contractivity of a genetic circuit with internal feedback and cell-to-cell communication. *IFAC-PapersOnLine*, 49(26):213 – 218, 2016. ISSN 2405-8963. Foundations of Systems Biology in Engineering FOSBE.

- Y. Boada, A. Vignoni, J. L. Navarro, and J. Picó. Improvement of a cle stochastic simulation of gene synthetic network with quorum sensing and feedback in a cell population. In *2015 European Control Conference (ECC)*, pages 2274–2279, 2015.

Conference and Presentation Papers in collaboration:

- G. Reynoso-Meza, J. Carrillo-Ahumada, Y. Boada, and J. Picó. Pid controller tuning for unstable processes using a multi-objective optimisation design procedure. *IFAC-PapersOnLine*, 49(7):284 – 289, 2016. ISSN 2405-8963. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB 2016.

Chapter 2

Foundations

2.1 Synthetic biology

Synthetic biology is a new field that has developed over the last 15 years thanks to the confluence of a number of factors. Advances in biology, genetics, genome sequencing, computational and automation technologies have enabled researchers to understand living systems in more detail.

As a multidisciplinary field, *synthetic biology* aims to design and engineer biologically based parts, novel biological devices and systems as well as redesigning existing, natural biological systems (The Royal Academy of Engineering, 2009). One of the key features of synthetic biology is the application of engineering principles to design, modeling, testing and validation of new biological devices that meet with defined specifications (see Fig.2.1). Thereby, Synthetic biology is a growing up area with tremendous impact on biotechnology industry, healthcare, chemistry, energy, environment and the general economy.

2.1.1 Gene expression

As this Thesis unfolds within the context of synthetic biology, some basic concepts about biological systems that will be used throughout the Thesis are briefly described in this Chapter. Living systems involve several main components (cells, proteins, genes) that allow biological microorganisms to grow and replicate. Understanding interactions among these components has led to understand that the rules governing how cells grow and replicate operate at various levels, from the individual cells to the populations of cells. At the molecular scale within an individual cell, the relationships between DNA, RNA and proteins are key to understand cell biology.

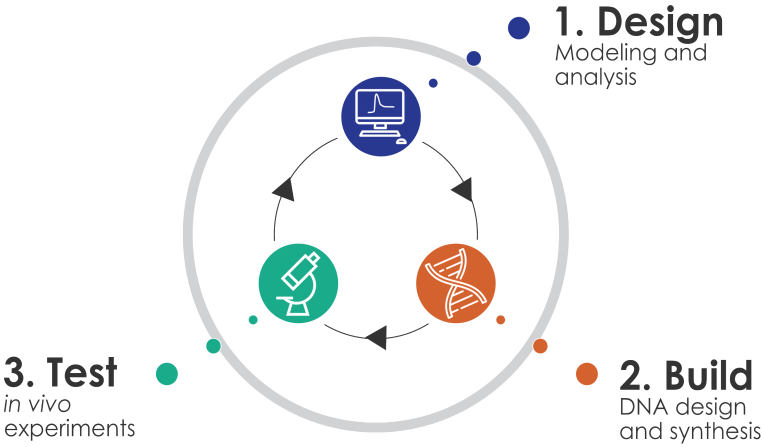


Figure 2.1. Synthetic biology cycle. Three fundamental pillars designing, building, and testing enable the development of synthetic biological systems in a more efficient and systematic way.

Gene expression was explained for the first time by Crick (**?**Crick, 1970) after the discovery of the structure for deoxyribose nucleic acid (DNA) by Watson and Crick (Watson et al., 1953). Crick established Gene expression as the so-called *Central Dogma of Molecular Biology*. The Central Dogma is a framework for understanding the transfer of sequence information between information-carrying molecular agents in living organisms. Transcription and translation are the two main stages of protein production, also known as **gene expression** (Alberts et al., 2009). Inside a single cell, genes store the genetic code (DNA) of a target protein. A DNA sequence is decoded into an intermediary messenger RNA (mRNA) in a process known as *transcription*. Then, the mRNA is translated into a protein in a process known as *translation*. Particularly, proteins are long chains of peptide molecules (also known as polypeptides) which fold in shapes that confer their active and structural properties.

Transcription starts when RNA polymerase (RNAP) binds to a specific DNA region that encodes a protein (see Fig.2.2). RNAP is an abundant molecular complex that catalyzes the production of an intermediary molecule of mRNA out of the coding region of a DNA molecule. If RNAP binds DNA and generates mRNA without other molecular compound intervening, transcription is said to proceed *constitutively*. Alternatively, transcription may be either helped or hampered by molecules binding to regulatory regions in DNA. These molecules are called transcription factors (TF). A TF may increase production of mRNA from DNA (**an activator**), or may reduce it (**a repressor**). Activation and repression by TFs is never completely on-off. It can be better approximated as a smooth switch-like characteristic function.

Translation starts when ribosomes attach the mRNA sequence to synthesize a protein (see Fig.2.2). Ribosomes are complexes of proteins that recognize a specific region

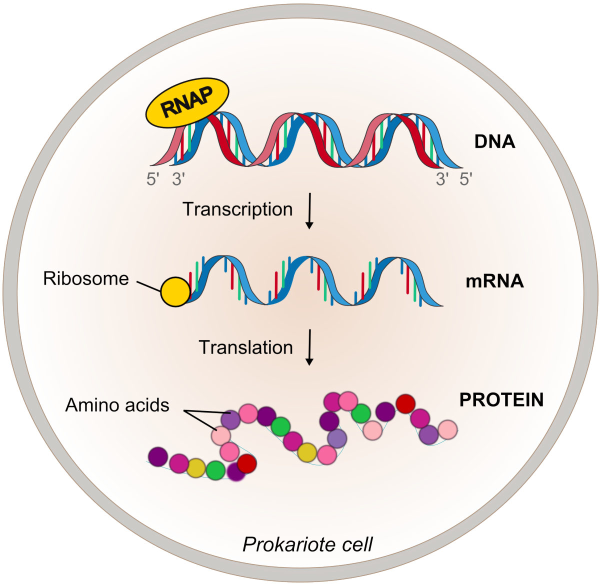


Figure 2.2. Gene expression process. DNA replication of a gene, then the Polymerase transcribes a single-strand copy of the DNA, finally the mRNA is translated into the corresponding protein by the ribosomes.

in the mRNA sequence called ribosomal binding site (RBS). The easiness with which ribosomes bind/undind the RBS (which in turn depends on the RBS sequence) will determine the rate of protein production. Once translation finishes producing the polypeptide chain, this folds onto itself due to electrochemical affinities among its constituents. This folded polypeptide is the functional protein.

In synthetic biology, the most commonly used model organisms include the procaryotic bacteria *Escherichia coli* (*E. coli*), and the eukaryotic yeast *Saccharomyces cerevisiae*. All the results obtained in this work are referred to *E. coli* bacteria as a host microorganism. Figure 2.3 depicts important gene expression timescales for *E. coli*. Transcription is a fast process that takes $\sim$ 1 minute since RNAP binds to DNA. But translation takes around 5-10 minutes since ribosomes release the protein completely translated and folded (Milo et al., 2016). Proteins are also naturally degraded by other proteins called proteases. Finally, cell division in bacteria ranges from 20 minutes (in a rich environment) to 1 hour.

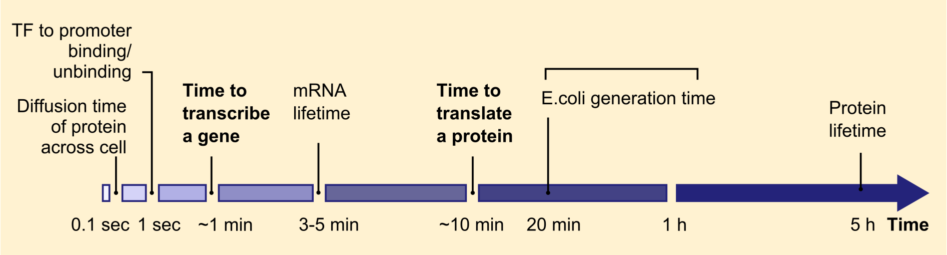

TF to promoter binding/ unbinding
Diffusion time of protein across cell
Time to transcribe a gene
mRNA lifetime
Time to translate a protein
E.coli generation time
Protein lifetime
0.1 sec
1 sec
~1 min
3-5 min
~10 min
20 min
1 h
5 h
Time

Figure 2.3. Relative timescales of biological processes. For *E. coli*bacteria, diffusion time of protein and binding of small molecule to protein are the fastest reactions (~ 1 msec). By contrast, proteins lifetime are on the order of hours that means the protein degradation process is slower than the other ones.

2.1.2 Synthetic gene circuit and parts

Advances in molecular biology and computer technologies have allowed researchers to manipulate DNA in bacteria, plants or animals. Synthetic biology aims to establish a rational framework for the DNA re-engineering, based on design and computational modelling. These principles have been used as the basis to build novel and artificial biological parts, devices and systems.

A synthetic gene circuit is a system where one or more genes interact between them and perform a specific function. Gene circuits are normally built from standard devices or transcriptional units, which in turn are built from standard bioparts. Each **biopart or part** is a modular DNA sequence that can be combined in the design of multiple transcriptional units. As shown in Fig.2.4b assembling a transcriptional unit needs four basic parts (each part has a glyph according the SBOL language (Cox et al., 2018)):

1. **Promoter.** A small part of DNA that recruits RNAP to transcribe mRNA from a DNA sequence. Promoters can be split into two types: constitutive or regulated. A *constitutive promoter* is always producing mRNA (as mentioned in section 2.1.1) at the same transcription rate. In contrast, a promoter can be an activator or a repressor. The *activator* attracts RNAP to begin transcription. The *repressor* obstructs RNAP thus inhibiting transcription.

2. **Ribosomal binding site (RBS).** A sequence found in mRNA where ribosomes bind and initiate translation. Depending of the RBS's affinity for the ribosomes, the effective translation rate will be higher or lower. The relative change in translation rate is known as the *RBS strength*.

3. **Coding sequence (CDS).** A long specific DNA sequence that is transcribed into mRNA and translated into its corresponding protein. That is, the coding sequence is the DNA sequence that corresponds to the sequence of amino acids (peptides) which constitute the protein.

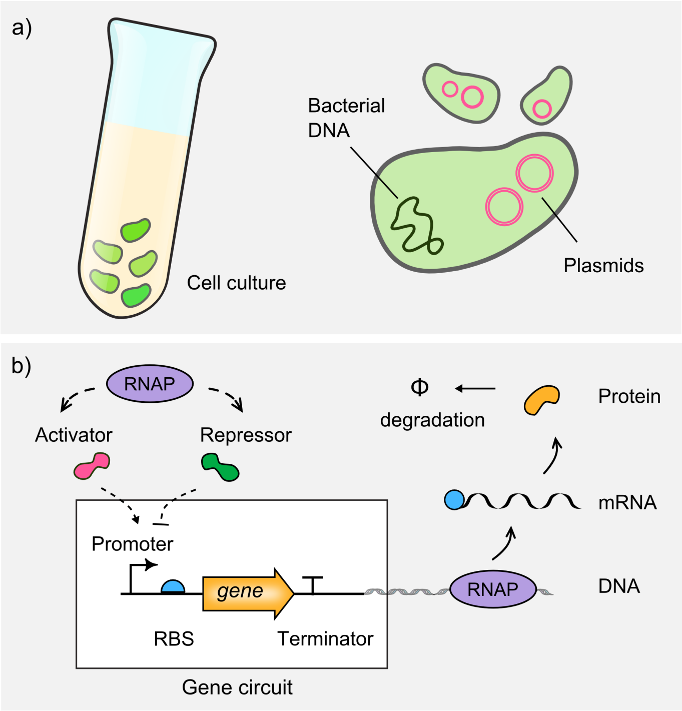


Figure 2.4. A gene circuit in a plasmid. a) Plasmids are DNA sequences containing the designed gene circuit. Several plasmids are inserted into *E. coli*bacteria to amplify the gene circuit. **b)** For a basic gene circuit, transcription of a desired gene is regulated by an activator or a repressor to either recruit or inhibit the binding of RNAP to the promoter. DNA is transcribed into mRNA, which contains a ribosome binding site (RBS). Ribosomes recognize the RBS and start translating the sequence of amino acids into a polypeptide. The resulting polypeptide is a folded protein with a specific 3D conformation.

4. **Terminator.** A genetic part that stops transcription by dissociating RNAP from DNA.

One of the most widely used library of biological parts in synthetic biology is the Registry of Standard Biological Parts (Biobrick Foundation, 2006). Here, a designed part is well characterized using a set of parameters so it can be easily re-used in other gene circuits. All the registered parts are compatible with the BioBrick$^{\text{TM}}$ (Knight, 2007) and the Gibson (Gibson et al., 2009) assembly methods. These two methods were used in this Thesis to create new longer parts and more complex transcriptional units, so they are described in annexes A.1 and A.2.

Figures 2.4a depicts separate circular DNA structures called **plasmids** (Zucca et al., 2013) inside a prokaryote cell (*E. coli*). Plasmids are akin to a piece of software loaded onto a computer that the computer runs to achieve the function. Hence plasmids are used to introduce foreign DNA in prokaryotes. Natural plasmids confer antibiotic resistance to bacteria. So, they arose as a protection mechanism. They replicate as the cell replicates, and a cell may carry many copies of the same plasmid (also known as *copy plasmid number*). Introducing foreign DNA in a plasmid is very simple. This way, one can achieve a cell that expresses the gene coded by the foreign DNA. The new cell is so-called a *recombinant* one. Assembling different foreign DNA parts of different origins (such us promoters, RBS, CDS or terminators) in a single plasmid is also possible. To do this, two sets of enzymes are used very much in the *cut & paste* spirit: the restriction enzymes to cut, and the ligase to paste.

The first step in the construction of a recombinant plasmid is the **restriction digest**. It is a process in which DNA is cut at specific sites by restriction enzymes, dictated by the surrounding DNA sequence (the protocol is described in annex A.1). There are hundreds of different restriction enzymes, allowing scientists to target a wide variety of recognition sequences. One or more digested parts are inserted into a compatibly digested *vector backbone*. This is another plasmid DNA sequence carrying both a bacterial origin of replication (particular sequence at which gene expression is initiated), and an antibiotic resistance gene for use as a selectable marker in bacteria. The final step is connecting the DNA inserts into the backbone using the reaction called **ligation** (refer to annex A.1) that is performed by the ligase enzyme. Now, the insert DNA is physically attached to the backbone and the complete plasmid can be introduced (transformed) into bacterial cells for propagation.

Transformation of bacteria with plasmids is important not only because is the process by which foreign DNA is introduced into a cell (Fig.2.4a), but also because bacteria are used as the means for both storing and replicating plasmids. Specific treatments have been discovered that increase the transformation efficiency and make bacteria more susceptible to either chemical or electrical based transformation, generating what are commonly referred to as “competent cells" (see annex A.3).

2.1.3 Cell density and Fluorescence

In this Thesis two main variables will be considered as measured ones: cell density and fluorescence. Cell density is a measure of the number of cell in a culture sample. It is measured as absorbance, also called **optical density (OD)** that is adimensional. Usually, the OD of a culture sample is quantified at a wave length of 600 nm. Hence, OD$_{600}$=1 contains 8×10^{11} cells per one litre. Fluorescent proteins are used as reporters informing about the expression level of a gene. The fluorescent protein content (sometimes linked to another target protein) in a bacterial culture is measured as **fluorescence intensity (F)** in relative light units (RLU).

In practice, these raw data are affected by the background signal from the culture medium absorbance $\mathrm{OD_b}$, and the auto-fluorescence $\mathrm{F_b}$ of the cells. These values should be eliminated by subtracting them from the corresponding data, obtaining the corrected absorbance ($\mathrm{OD} = \mathrm{OD}_{\mathrm{raw}} - \mathrm{OD}_{\mathrm{b}}$), and corrected fluorescence ($\mathrm{F} = \mathrm{F}_{\mathrm{raw}} - \mathrm{F}_{\mathrm{b}}$). In most cases, the ratio F/OD is taken as a measure of the fluorescent protein content per cell.

2.1.4 Measurements types

From the point of view of the way measurements are taken, there are three main ways measurements can be carried out *in vivo* (see Fig.2.5): (1) bulk data, (2) single-cell population snapshot data, and (3) time-series single-cell data.

1. **Bulk data** consist of measurements of a variable of interest for a culture of cells growing in a bioreactor. The measured variable is proportional to its sum for all cells in the population (Fig.2.5a). For instance, a plate reader machine can incubate cell cultures at a specific temperature and collect cell density and fluorescence information at regular time intervals. These two data will be described below.

2. **Population snapshot** data consist of measuring a variable for a large number of individual cells in a population (see Fig.2.5b). yet different from above, these same cells cannot be followed over time. Therefore at the next time point, a different set of cells are measured. Population snapshot measurements include techniques such as **flow cytometry**. Flow cytometers can process a cell culture and read both cell size and fluorescence for tens of thousands of cells within a few seconds.

3. **Time-series measurements** provide the track of some single cells variable of interest over time, using techniques like time-lapse fluorescence microscopy.

It is important to point out that the type of measurement collected can and should impact the choice of model for the system being studied (Hsiao et al., 2018). As we will see in Chapter 5, if the data collected are all bulk data, using a stochastic model that accounts for *noisy gene expression* in single cells could be unnecessarily complex, and a deterministic model based on ordinary differential equations might be a better choice.

2.2 Feedback control in synthetic biology

Control theory has arisen from the conceptualization and generalization of design strategies aimed at improving the stability, robustness and performance of physical systems, including the microscopic ones. Some of these techniques have been successfully implemented for the design of controllers for synthetic gene networks at either

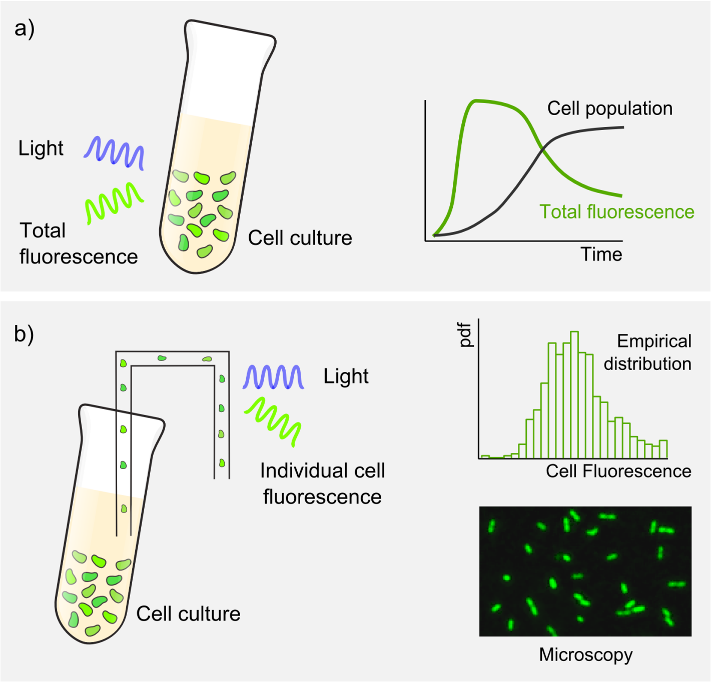


Figure 2.5. a) Bulk data are proportional the sum of the cells in the population. Cell density and total fluorescence of the culture can be quantified. **b)** Population snapshot data reveal measurements for individual cells. They usually include tools like flow cytometry (fluorescence and cell size), and microscopy.

single-cell or population level. However, the nature of biomolecular interactions has also supposed unavoidable challenges to the implementation of any feedback loop.

For example (Olson and Tabor, 2014; Olson et al., 2017) have achieved highly predictable gene expression programming by combining experimental characterization with simplified models of optogenetic systems in bacteria. Though the methods are open-loop, they enable a basic dynamic characterization approach of synthetic gene circuits and provide valuable information that can be used in the subsequent design of more complex or closed-loop gene circuits.

Inevitably, gene circuits are complex large-scale systems with intricate patterns, there are limited tools for measuring biomolecular interactions in real time to design an appropriated input/structure that follows a desired reference (Milias-Argeitis et al., 2011), and system identification entails great difficulty of isolating gene circuits from their cellular environment. Therefore, solving problems in synthetic biology using control theory requires much more than simply transplanting existing theories developed for engineering systems directly to a biomolecular setting. Engineers have to tackle these issues to design more stable, predictable and robust synthetic gene circuits.

2.2.1 Control strategies

There are two different ways to design controllers for synthetic gene networks. The first one is to design a gene circuit as *a controller in living single cells*, which in turn form multicellular systems (Hsiao et al., 2018; Menolascina et al., 2011). This controller must be implemented by the biological parts and their subsequent chemical species and reactions, hence it is highly constrained. Also, species and reactions are subject to stochastic fluctuations, so there are fundamental limits on the controller robustness in the process. The second way implements the *controller at the population level using a computer* in the loop (Menolascina et al., 2014). This can reduce the issues from the first strategy, but it does not allow for independent actuation on distinct single cells.

In this Thesis, the fist strategy –a gene circuit as a controller in individual cells– has been chosen to design new synthetic gene circuits that behave following the desired reference.

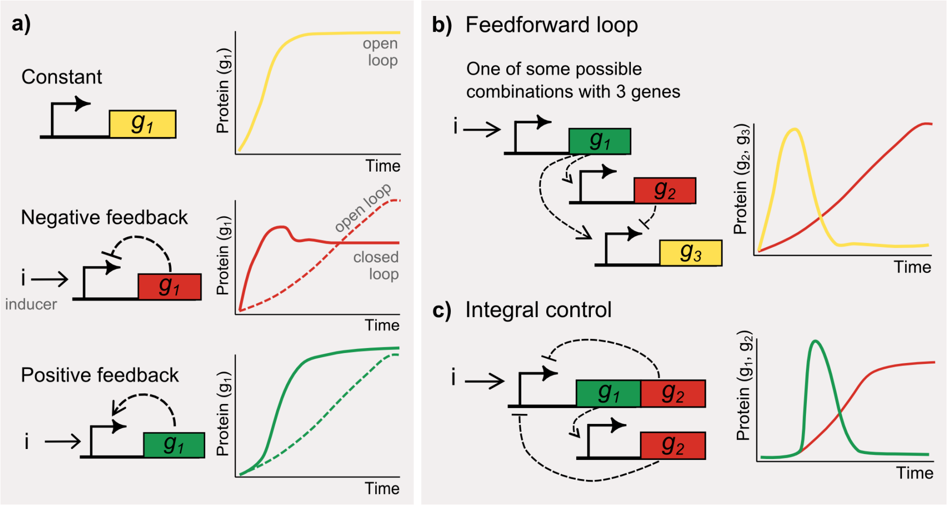


Figure 2.6. Implementation of feedback control in living cells. a) From top to bottom: synthetic feedback modules associated to the most common non-regulated and regulated actions *i)* constant protein production, *ii)* negative feedback and positive feedback protein expression. **b)** Some feedback controllers built to demonstrate the feasibility of feedback in engineered organisms. *Figure inspired in (Hsiao et al., 2018).*

2.2.2 Control design in single cells

There are several issues to deal with when one is interested in regulating the synthetic gene circuits behavior. Devising synthetic feedback control circuits to address the problems caused by uncertainty and burden-induced undesired dynamics and their scalability to industrial conditions in a systematic, conceptually unified, and modular way will require a set of enabling tools:

1. **Modular model building.** Models are key for both model-based design of (feedback control) gene circuits, and for computational analysis and design. Modular model building i.e. the systematic way of combining module components from a library of standard parts must address retroactivity and inter- and intra-circuit interactions. Retroactivity among modules can be addressed by considering fast binding/unbinding reactions to be at quasi-steady state (QSS), leading to cancellation of the interconnection terms in the dynamic balances. This approach will be fully described and used in section 2.3.1 and Chapter 4. This is equivalent to considering that fast reactions interconnecting modules have low output and high input impedances. Yet, this is not the case for slow reactions (dynamics). In such a case, a modular approximated matrix representation taking into account retroactivity has been devised in (Gyorgy and Del Vecchio, 2014) by assuming QSS and using sensitivity analysis on the resulting fast manifolds. This framework can be easily integrated within model structures accounting for genetic load like those in (Qian et al., 2017). Further integration of basic substrate uptake metabolism and cell growth using models like (Weiße et al., 2015; Beguerisse-Díaz et al., 2016) would be a key step towards comprehensive models capturing all relevant phenomena to deal with burden-induced dynamics and their interaction with the cell environment. Uncertainty could then be addressed using model consensus approaches (Villaverde et al., 2015). Alternatively, extension of these (generally deterministic) ODE models to the stochastic realm would be quite straightforward using Langevin-based approaches like the ones described in Chapter 4 and 6. Modular model building can be applied to the basic modules of gene circuits illustrated in Fig.2.6a. The models must reflect in a systematic way the positive and/or negative control actions that these basic modules can carry out.

2. **Model-based and automated design of feedback control gene circuits.** Feedback control mechanisms play a fundamental role in synthetic biology. Successful design and implementation of stress and genetic burden feedback control genetic biocircuits requires on the one hand availability of biological devices for building feedback mechanisms (see Fig.2.6b), and appropriate design methodologies on the other. The increasing number of biological devices available makes it possible to implement control structures within the cell (Del Vecchio et al., 2016; Folliard et al., 2017). Today there are transcriptional activators/repressors that act as gains with saturation; switches (Wittmann and Suess, 2012) that act as sensors that recognize metabolites or DNA sequences and activate/deactivate

genes; sensors (Dahl et al., 2013) that respond to different types of intracellular or extracellular metabolites, to metabolic states, or to external light signals (Olson et al., 2014). Feedback control methodologies for constrained systems and for systems with shared resources (e.g. information and electrical networks) already exist. In the first case, reference conditioning methods have been adapted to biological systems in (Picó et al., 2009) though not at the intracellular genetic scale. In the second one, initial attempts with basic control strategies have been tested in (Shopera et al., 2017). However, further work is clearly required. Considering the problem as the one of agents coordination (Vignoni et al., 2013a) under optimality conditions (Giordano et al., 2016) seem a promising approach. On the other hand, modular and systematic design of biocircuits, i.e. the systematic automated way of finding combinations of components from a library of standard parts allowing to optimally perform a pre-defined function can be formulated using an optimization framework (Otero-Muras and Banga, 2016, 2017). These approaches combine the efficiency of global mixed-integer nonlinear programming solvers with multi-objective optimization techniques (Sendin et al., 2010) using coarse-grain dynamic models of the biocircuit modules. They also can provide guidelines to tune the gains of few circuits parts (see section 2.1.2) to achieve a desired control metrics. Their combination with model-based approaches would certainly speed-up the design and analysis process, as we will see in Chapters 5 and 7.

Many important challenges remain to be addressed for synthetic biology to mature as an engineering discipline that take advantage of control theory (Church et al., 2014; Way et al., 2014b): (1) availability and characterization of biological parts, devices and systems, (2) systematic and modular design methods of synthetic circuits, (3) dealing with host interaction (metabolic burden and variability), and (4) scalability to industrial bioreactors. In the Thesis, we will focus on the first two challenges to develop systematic and modular design methods of synthetic gene circuits.

2.3 Deterministic modeling. Law of mass action

A model is a representation of a system in some form useful for a given purpose. Modeling allows the generation of new testable hypothesis and novel ways of intervention, as well as mechanistic explanations of experimental results. Models are key for both model-based design of biocircuits (feedback control), and for computational analysis and design.

Kinetic modeling of small biological circuits has a long fruitful tradition (Villaverde and Banga, 2014). Kinetic (i.e. dynamic) models are particularly important since they can explain and predict the functional behavior that emerges from the time-varying concentrations in cellular components (Villaverde and Banga, 2014; Chen et al., 2010a).

All reactions taking place inside the cell are stochastic by nature. That is, modeling the set of reactions should be formulated in terms of the probabilities of each of the reactions to occur. The resulting models are stochastic ones, and they will be considered in section 4.4. As a simpler alternative, deterministic models do not take into account the probabilistic nature of reactions. Instead, they assume the amount of species transformed by the reactions depend solely on the current amount of species, the rates at which the reactions proceed, and the stoichiometry of them. Furthermore, it is normally assumed that the amount of molecules are large enough, so one may consider that the amount of molecules is a continuous magnitude that varies continuously in time.

Consider for instance a simple model of gene transcription and mRNA degradation (see sections 2.1.1 and 2.1.2) given by the set of chemical reactions

$$\begin{aligned}\mathrm{DNA} &\xrightarrow{\mathrm{k}} \mathrm{DNA} + \mathrm{mRNA}\\ \mathrm{mRNA} &\xrightarrow{\mathrm{d}} \emptyset\end{aligned} \tag{2.1}$$

where DNA and mRNA are both reactants and products, k and d denote the *reaction rate constants*.

In the general case, given a set of chemical species X_i, $i = [1, \ldots, I]$ that interact through J reactions, each one may be expressed as

$$\sum_{i=1}^{I} \mathrm{s}_{ij}\mathrm{X}_i \xrightarrow{\mathrm{k}_j} \sum_{i=1}^{I} \mathrm{s}'_{ij}\mathrm{X}_i, \quad j = 1, \ldots, J \tag{2.2}$$

where s_{ij} and s'_{ij} are the *stoichiometric coefficients* denoting numbers of reactant and product molecules, respectively. Thus, reaction (2.1) can be rewritten as

$$\begin{aligned}\mathrm{X}_1 &\xrightarrow{\mathrm{k}} \mathrm{X}_1 + \mathrm{X}_2\\ \mathrm{X}_2 &\xrightarrow{\mathrm{d}} \emptyset\end{aligned} \tag{2.3}$$

where X_1 is the concentration of the DNA, and X_2 is the concentration of the mRNA.

The dynamics of a reaction network can be derived considering the dynamic balance for each chemical species and using (2.2). Hence, one can express

$$\dot{x} = \mathrm{S} \cdot \mathbf{v} \tag{2.4}$$

where $x \,\epsilon\, \mathbb{R}^I$ is the vector of chemical species, $\mathrm{S} = \mathrm{s}'_{ij} - \mathrm{s}_{ij}$, $\mathrm{S} \,\epsilon\, \mathbb{R}^{I \times J}$ is the stoichiometric matrix, and $\mathbf{v} \,\epsilon\, \mathbb{R}^J$ is the vector of reaction rates. Notice these balances can be set either in terms of mass (i.e. number of molecules as we will use later) or in terms of concentration (mass divided cell volume).

Thereby, for the example (2.3), the dynamic balance applying equation (2.4) will become system (2.5). Note that each row of S denotes the *i*-th species of the system, and each column shows the *j*-th reaction. We use $\mathrm{s}_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\mathrm{s}'_{ij} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{2.5}$$

The reaction rates can be expressed using the *Law of mass action kinetics (MAK)* (Chellaboina et al., 2009; Steinfeld et al., 1989; Horn and Jackson, 1972). The MAK states that for an elementary reaction (reaction where all of the stoichiometric coefficients of the reactants are one), the reaction rate is proportional to the product of the concentrations of the reactants raised to a given power defined by the stoichiometry of the reaction. The proportionality coefficient is the *specific reaction rate* (reaction rate constant). Thus, for system (2.1), the dynamic balance can be mathematically described by the ordinary differential equations or ODEs (2.6) that commonly referred to as kinetic equations (Alon, 2007; Heinrich and Schuster, 1996)

$$\begin{aligned} &\dot{x}_1 = 0, \quad x_1(0) = x_{10},\ t \geq 0, \\ &\dot{x}_2 = \mathrm{k}x_1(t) - \mathrm{d}x_2(t), \quad x_2(0) = x_{20} \end{aligned} \tag{2.6}$$

where the coefficients of the right-side of (2.6) are the stoichiometric coefficients from the matrix S. One single simulation of the ODE system (2.6) is shown in Fig.2.7. The mRNA is transcribed after around 16 minutes (typical transcription time in *E. coli*), and the DNA is an invariant state.

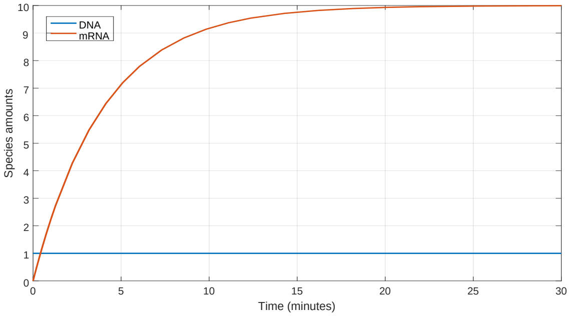


Figure 2.7. Deterministic simulation of constitutive gene transcription. Parameters used: the gene copy number $\mathrm{C_n}$=1 molecule is also the initial condition $x_1(0)$, transcription rate k=2.5 min^{-1}, degradation rate d=0.25 min^{-1}, and $x_2(0) = 0$.

For the MAK, if one of the required products is lacking, the reaction will not take place. The reaction proceeds faster as the concentration of the required substrates increase. The mass-action kinetics accounts for the probability of encounter (collision) among the reactants, doing it proportionally to the product of the reactant concentrations. Thus, the rationale behind mass-action kinetics is that the rate at which a reaction proceeds is proportional to the probability that the required reactants encounter. This probability, in turn, is proportional to the product of their concentrations. This is an idea that we will find again in section 4.4 when we will deal with stochastic models.

The MAK has numerous analytical properties that are of inherent interest from a dynamical systems perspective. For example, mass-action kinetics give rise to systems of ODEs having polynomial nonlinearities. Polynomial systems are notorious for their intricate analytical properties even in low-dimensional cases (Chellaboina et al., 2009; Jarrah et al., 2007). For example, consider the reversible reaction network of four species

$$\begin{aligned} \mathrm{X}_1 + \mathrm{X}_2 &\xleftrightarrow[\mathrm{k}_{-1}]{\mathrm{k}_1} \mathrm{X}_3 + 2\,\mathrm{X}_4 \\ 0 &\xrightarrow{\mathrm{k}_3} \mathrm{X}_3 \\ \mathrm{X}_4 &\xrightarrow{\mathrm{d}} \emptyset \end{aligned} \tag{2.7}$$

the MAK for system (2.7) implies that:

$$\begin{aligned} \dot{x}_1 &= -\mathrm{k}_1 x_1(t) x_2(t) + \mathrm{k}_{-1} x_3(t) x_4^2(t), \quad x_1(0) = x_{10},\ t \geq 0, \\ \dot{x}_2 &= -\mathrm{k}_1 x_1(t) x_2(t) + \mathrm{k}_{-1} x_3(t) x_4^2(t), \quad x_2(0) = x_{20} \\ \dot{x}_3 &= \mathrm{k}_1 x_1(t) x_2(t) - \mathrm{k}_{-1} x_3(t) x_4^2(t) + \mathrm{k}_3, \quad x_3(0) = x_{30} \\ \dot{x}_4 &= 2\mathrm{k}_1 x_1(t) x_2(t) - 2\mathrm{k}_{-1} x_3(t) x_4^2(t) - \mathrm{d} x_4(t), \quad x_4(0) = x_{40} \end{aligned} \tag{2.8}$$

where the second reaction represents mass addition of species X_3 at the k_3 production rate, and the last reaction represents mass removal of X_4 at the degradation rate d. The ODE system (2.8) is a second-order polynomial system with nonnegative solutions, because of physical considerations like nonnegative initial conditions. The MAK and the resulting kinetic equations are widely used formalism to perform a relatively straightforward analysis of a biological system's behavior. The resulting dynamic models have interesting properties that have been studied using the *Chemical Network Theory* (Angeli et al., 2007; Craciun and Feinberg, 2005; Feinberg, 1987).

However, the MAK fails to provide a valid description in cases where the effects of stochastic fluctuations become significant. This is typically the case when some reactions between the species involved occur at low number of molecules. Thereby stochastic modeling is needed as we will see in section 2.4. This is a common feature of biological systems.

2.3.1 Model reduction

Dynamic models obtained from the reaction networks by applying mass action kinetics are usually large order ones. *Model reduction techniques* can be put into practice, yielding models with less state variables, i.e. with less order. There are some advantages in reducing these dynamic models:

- Large order models have many parameters (i.e. specific reaction rates). The values of these parameters must be obtained using experimental data related to the corresponding reactions. This process is called parameters estimation. It turns out that the experimental difficulties and computational cost for the parameters estimation process increases a lot with the number of parameters. For instance, estimating binding rates is not an easy task.

- In practice, there are reactions in the network that proceed at much faster rates than others. For instance, the RNA polymerase binding/unbinding rates to the gene promoter are much faster than the translation or elongation rates. This means that there are very different time scales associated to each reaction. The large differences in the time scales among the different species in the reaction network (typically many orders of magnitude) originate huge difficulties for simulating the temporal evolution of the network and for understanding the basic principles of its operation.

In case we want a model that allow for some degree of mechanistic description of the system, the reduction process should yield a more amenable model for computational analysis, but avoiding excessive reduction that would lead to lack of biological relevance. In particular, the species in the reduced model must not be lumped ones, and the resulting lumped parameters in this reduced model must be easy to associate to experimental tuning knobs.

There are several methods that can be used for model reduction. The most widely used in the field of systems and synthetic biology is the so-called *Quasi Steady-State Approximation (QSSA)* (Segel and Slemrod, 1989; Kokotovic et al., 1986; Khalil, 1996), which is also know as Bodenstein-Semenov kinetics. Recently, the related *Layered Decomposition* was proposed in (Prescott and Papachristodoulou, 2014).

On the one hand, model reduction can be carried out by means of the QSSA. In essence, QSSA is a singular perturbation method that considers the time-scale separation among the different dynamics (Zagaris et al., 2004; Mélykúti et al., 2014). In particular, one can assume that some binding reactions occur very fast in comparison with those corresponding to transcription, translation and degradation. This results in considering that the time derivatives of fast state variables are zero. In other words, they are at quasi steady-state.

On the other hand, *layered decomposition* suggests that in many instances the parameters are such that it is the reaction rates which separate in time scale and not the state space variables. In standard singular perturbation analysis state space variables, chemical species or reactants in our case, are grouped into fast and slow subsets. But in biochemical networks, often, a reactant takes part in both fast and slow reactions so there is no way of classifying it this way. The problem is sometimes solved using state space transformations, introducing new variables whose behaviour is difficult to interpret in physical terms. For *layered decomposition*, the state space variables are expressed as the sum of a fast variables set and a slow one,

$$x = x_f + x_s \tag{2.9}$$

in such a way the original state equations can be rearranged as

$$\dot{x} = \frac{1}{\epsilon} S^f v^f(x) + S^s v^s(x) \tag{2.10}$$

where the v^f, v^s represent the corresponding fast and slow reaction rates, and the S^f, S^s are stoichiometric matrices. Using (2.9) to rewrite (2.10) in a form more amenable for singular perturbation analysis the following system is obtained

$$\begin{aligned} \epsilon \dot{x_f} &= S^f v^f(x_f, x_s) \\ \dot{x_s} &= S^s v^s(x_f, x_s) \end{aligned} \tag{2.11}$$

Either using QSSA or Layered Decomposition, additional algebraic relationships among variables can be obtained through *system invariants*. In the case of reaction networks, it can be observed that some reactions are a linear combination of other ones. Then, the linear combination of the concentrations of the species involved will keep constant in time. These linear combinations, so called *moieties*, can be understood as a kind of quasi-species that keep invariant, i.e. keep constant concentration.

In this Thesis QSSA and systems invariants will be used most often. next, to illustrate them, these concepts will be applied to reduce the following set of reactions taking place during transcription

$$\begin{aligned} \mathrm{RNAP} + \mathrm{DNA} &\underset{v_{-1}}{\overset{v_1}{\rightleftharpoons}} \mathrm{RNAP} \cdot \mathrm{DNA} \\ \mathrm{RNAP} \cdot \mathrm{DNA} &\xrightarrow{v_2} \mathrm{RNAP} + \mathrm{DNA} + \mathrm{mRNA} \\ \mathrm{mRNA} &\xrightarrow{v_d} \emptyset \end{aligned} \tag{2.12}$$

where DNA stands for the gene, and $\mathrm{RNAP} \cdot \mathrm{DNA}$ represents the compound resulting from binding of the RNA polymerase (RNAP) to the gene promoter. Notice the promoter can be taken to represent the whole gene. Indeed, from the reaction point of view, the important fact is that of binding of RNAP to the gene promoter. v_1, v_{-1}, v_2, and v_d denote the reaction rates. The binding of RNAP to the gene promoter has been assumed as a reversible reaction, while mRNA degradation is an irreversible reaction. We assume that after the transcription reaction finishes when RNAP encounters the terminator, so RNAP releases from the DNA. Thus, the gene promoter becomes free to bind to a new RNAP.

Setting the dynamic mass balances for the four chemical species of (2.12) implies the use of the MAK through the equation (2.4). This will give us the following model

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 & 0 \\ -1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} \triangleq \mathbf{S}v = \begin{bmatrix} -\mathrm{k}_1 x_1 x_2 + \mathrm{k}_{-1} x_3 + \mathrm{k}_\mathrm{m} x_3 \\ -\mathrm{k}_1 x_1 x_2 + \mathrm{k}_{-1} x_3 + \mathrm{k}_\mathrm{m} x_3 \\ \mathrm{k}_1 x_1 x_2 - \mathrm{k}_{-1} x_3 - \mathrm{k}_\mathrm{m} x_3 \\ \mathrm{k}_\mathrm{m} x_3 - \mathrm{d}_\mathrm{m} x_4 \end{bmatrix} \tag{2.13}$$

where $x_1 = [\mathrm{DNA}]$, $x_2 = [\mathrm{RNAP}]$, $x_3 = [\mathrm{RNAP} \cdot \mathrm{DNA}]$, and $x_4 = [\mathrm{mRNA}]$, $\mathbf{S}$ is the stoichiometry matrix, and v contents the proportionality constants $\mathrm{k}_1, \mathrm{k}_{-1}, \mathrm{k}_\mathrm{m}, \mathrm{d}_\mathrm{m}$ as the specific reaction rates. Notice the model takes the form of a system of four first order differential equations, each one corresponding to the balance at each species. Thus, the model is a fourth-order one.

Identifying moieties is a straight forward process. In this simple case, conservation laws can be inferred from simple inspection. Notice that $\dot{x}_1 + \dot{x}_3 = 0$, and $\dot{x}_2 + \dot{x}_3 = 0$. This implies that the sum of concentrations in both cases keep constant in time

$$\begin{aligned} x_1 + x_3 &= \mathrm{C_n} \\ x_2 + x_3 &= \mathrm{C_{RNAP}} \end{aligned} \tag{2.14}$$

where the constants $\mathrm{C_n}$, and $\mathrm{C_{RNAP}}$ correspond to the initial concentrations of $x_1 + x_3$, and $x_2 + x_3$ respectively.

In larger reaction networks where getting invariants by simple inspection may probe difficult, one can calculate the kernel of the stoichiometric transpose matrix $\mathbf{S}^T$. In our case, this kernel is spanned by the vectors $[1, 0, 1, 0]^T$, and $[0, 0, 1, 1]^T$. This implies the conservation laws above.

The first invariant in (2.14) simply means that the number of gene copies keep constant in time, being equal to the sum of the free and bound promoter. Therefore, the constant C_n equals the **gene copy number**. Recall that most often the gene will be introduced by means of plasmids, as was described in section 2.1.2. In the same cell, multiple copies can be introduced of a plasmid carrying the gene to be transcribed. Thus integer values of gene copy number greater than one are possible. The plasmid copy number is an important tuning knob for genetic circuits. In case the transcribed gene is in the cell chromosome, there is only one molecule of DNA in the case of prokaryotes[1].

[1] It is said that they are are haploid cells, i.e. they only have one copy of each chromosomal gene. Diploid cells, e.g. yeast cells, have two copies of each chromosomal gene.

The second invariant states the conservation of the RNAP. The polymerase is either free (x_2) or bound to the DNA under the form of x_3. This invariant may be misleading, and should be interpreted with care. The second differential equation of model (2.13) is not true. It would be if the only gene using RNAP is the one under study. But in the cell, other genes might be using RNAP in parallel transcriptions. Therefore, the dynamic balance for the RNAP should be written taking into account all reactions in the cell using it. This is unfeasible. A better approach to cope with this problem is to consider that the cell contains free RNA polymerase in excess enough to *serve* all the active genes transcribing at a given moment. In this case, one could consider that the free RNA polymerase concentration in the cell will not appreciably change in time. That is $\dot{x}_2 \approx 0$. But this means that the free RNA polymerase concentration will approximately keep constant with time $x_2 \approx \mathrm{C_{RNAPf}}$.

In practice, the free RNAP concentration may appreciably change with time if we consider long time intervals in which the cell goes through different situations. But if this appreciable change over long periods is very slow as compared to the time-scales at which the other species change. Therefore, we may consider that x_2 is kind of a slowly time-varying parameter. In the rest of the Thesis we will use this approach in all models.

In summary, by looking at reaction invariants, we have the algebraic constraints

$$\begin{aligned} x_1 + x_3 &= \mathrm{C_n} \\ x_2 &= \mathrm{C_{RNAP}} \end{aligned} \tag{2.15}$$

Now it is time to apply the QSSA approach to get rid of species whose dynamics are very fast as compared with the remaining ones. The underlying idea behind QSSA is that very fast reactions will quickly reach steady-state as compared with reactions proceeding slower. Therefore, one could neglect the dynamics of the fast reactions, and directly assume they are at steady-state. This will convert the corresponding differential equation into an algebraic one.

In our case, as we already mentioned above, one can safely assume that the RNAP binding/unbinding reactions to the gene promoter proceed much faster than elongation and mRNA degradation. This is reflected in the values of the reaction rates. How much difference there is between the RNAP binding/unbinding rates and e.g. the elongation rate will depend on the gene promoter affinity for RNAP, i.e. on the promoter strength. Recalling the model (2.13), we can define the so called perturbation parameter

$$\epsilon = \frac{1}{\mathrm{k_1}} \tag{2.16}$$

Pre-multiplying by ϵ both sides of the first equation of the model

$$\epsilon \dot{x}_1 = -x_1 x_2 + \frac{\mathrm{k}_{-1} + \mathrm{k}_\mathrm{m}}{\mathrm{k}_1} x_3 \tag{2.17}$$

If assuming that k_1 is large enough, ϵ will be a very small number. In the limit, ϵ will tend to zero as k_1 increases. The term $\frac{\mathrm{k}_{-1}+\mathrm{k}_\mathrm{m}}{\mathrm{k}_1}$ will not vanish if its numerator and denominator have the same or close orders of magnitude. That is, if RNAP binding and unbinding to the gene promoter have rates that are not different by several orders of magnitude. Under this condition, equation (2.18) can be approximated as

$$0 = -x_1 x_2 + \frac{\mathrm{k}_{-1} + \mathrm{k}_\mathrm{m}}{\mathrm{k}_1} x_3 \tag{2.18}$$

This is an algebraic equation that can be used, along with (2.15), to reduce the model (2.13). To this end, from (2.18) and (2.15) we can obtain

$$x_3 = \frac{\mathrm{C}_\mathrm{n}}{1 + \frac{1}{\mathrm{C}_\mathrm{RNAP}} \frac{\mathrm{k}_{-1}+\mathrm{k}_\mathrm{m}}{\mathrm{k}_1}} \tag{2.19}$$

Equation (2.19) gives the concentration of the complex formed by RNA polymerase bound to the gene promoter, $x_3 = [\mathrm{RNAP} \cdot \mathrm{DNA}]$. Notice the maximum possible concentration will equal the gene copy number concentration C_n. That is, all gene copies in the cell have their promoters occupied by RNA polymerase. This concentration can only be achieved in the limit as the available free RNA polymerase in the cell tends to infinity. See also that the ratio $\frac{\mathrm{k}_{-1}+\mathrm{k}_\mathrm{m}}{\mathrm{k}_1}$ is the one between the release rate of free RNAP from the promoter, and its capture rate. Therefore, the more affinity of RNAP for the gene promoter –the stronger the promoter– the closer x_3 will be to C_n.

Now, using equation (2.19) in (2.13) we reach the desired reduced model for constitutive gene transcription

$$\dot{x}_4 = \mathrm{k}_\mathrm{m} \frac{\mathrm{C}_\mathrm{n}}{1 + \frac{1}{\mathrm{C}_\mathrm{RNAP}} \frac{\mathrm{k}_{-1}+\mathrm{k}_\mathrm{m}}{\mathrm{k}_1}} - \mathrm{d}_\mathrm{m} x_4 \tag{2.20}$$

where recall $x_4 = [\mathrm{mRNA}]$.

This model can still be further simplified by lumping all parameters associated to promoter strength and transcription elongation rate under the umbrella of a new parameter reflecting effective transcription rate

$$\mathrm{k}_\mathrm{me} = \frac{\mathrm{k}_\mathrm{m}}{1 + \frac{1}{\mathrm{C}_\mathrm{RNAP}} \frac{\mathrm{k}_{-1}+\mathrm{k}_\mathrm{m}}{\mathrm{k}_1}} \tag{2.21}$$

Then we have

$$\dot{x}_4 = \mathrm{C_n k_{me}} - \mathrm{d_m} x_4 \tag{2.22}$$

Under this assumption, the two first reactions in the transcription reactions network (2.12) can be lumped together, resulting in the simplified reactions network

$$\begin{aligned} \mathrm{DNA} &\xrightarrow{\mathrm{k_{me}}} \mathrm{DNA} + \mathrm{mRNA} \\ \mathrm{mRNA} &\xrightarrow{\mathrm{d_m}} \emptyset \end{aligned} \tag{2.23}$$

If the gene copy number $\mathrm{C_n}$ is larger than one, we can either consider $\mathrm{C_n}$ parallel reactions DNA $\xrightarrow{\mathrm{k_{me}}}$ DNA + mRNA in the model (2.23), or consider that the reaction rate is the product $\mathrm{C_n k_{me}}$. The model structure given by equation (2.22) is the most often used for *constitutive* gene transcription, i.e. unregulated one.

2.4 Stochastic modeling and simulation

The continuous deterministic approach fails to capture many important details at molecular level (Samoilov et al., 2005; Eldar and Elowitz, 2010a). Noise is pervasive in the cellular mechanisms underlying gene expression (Raser and O'Shea, 2005). It propagates to downstream genes at the single-cell level, and eventually causes variation within an isogenic population (Raj and van Oudenaarden, 2008; Labhsetwar et al., 2013) that may determine the fate of individual cells and that of a whole population (Eldar and Elowitz, 2010b). As a consequence, a variation of protein expression levels appears in every cell across the population (Novick and Weiner, 1957). This stochasticity in protein expression levels is often referred to as *gene expression noise* (Elowitz et al.; Chalancon et al., 2012).

As said before, gene expression noise can not be avoided, and has relevant impact on cellular functions generating *phenotypic variability*. This variability can be it can be beneficial in some contexts and harmful in others. These situations include e.g. the stress response, metabolism, development, the cell cycle, circadian rhythms, and aging (Raj and van Oudenaarden, 2008; Acar et al., 2008).

2.4.1 Sources of gene expression noise

The sources of gene expression noise can be classified according its origin as *i)* intrinsic noise, and *ii)* extrinsic noise (Swain et al., 2002; Wilkinson, 2006; Gillespie, 1977). *Intrinsic noise* appears from the inherently randomness of chemical reactions at a single-cell level. It arises from the discrete nature of the molecular events of gene expression (see Fig.2.8a). *Extrinsic noise* becomes visible when other cellular processes interact with the system under study across a cell population, or when fluctuations come from extracellular environment (see Fig.2.8b). It is important to realize extrinsic noise can be theoretically isolated from the system. But intrinsic noise is the very essence (discrete nature) of the underlying molecular events and cannot be separated (even hypothetically) from the system. Thereby, both of them should be taken into account to perform stochastic modeling of gene circuits (Hilfinger and Paulsson, 2011; Wilkinson, 2009; Cai et al., 2006).

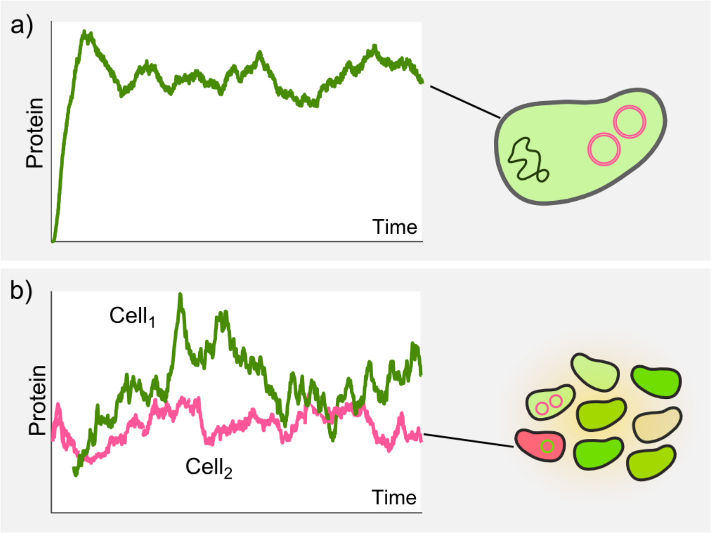


Figure 2.8. Gene expression noise. a) Intrinsic noise comes form the stochastic nature of biochemical reactions. It becomes harmful at low number of molecules. **b)** Extrinsic noise comes from other cellular processes and the environment. This noise .

As we will see in Chapter 4, intrinsic noise was modeled by using the stochastic Chemical Langevin Equation (Higham, 2008), and extrinsic noise was set by randomizing values of the model parameters (Joo et al., 2013; Toni and Tidor, 2013). Despite modeling extrinsic noise as an additive signal is a commonly used approach (Elowitz et al.; Swain et al., 2002), Chapter 6 will demonstrate that this assumption can disguise the magnitude and effects of extrinsic noise in a gene circuit.

2.4.2 Chemical Master Equation (CME)

The most accurate way to describe the stochasticity of a gene network (being a system of chemically reacting molecules) is by means of the *Chemical Master Equation* (CME) (Wilkinson, 2006; Van Kampen, 2011).

We seek a stochastic description of the chemical reaction network under *well-mixed* and *diluted conditions* in a closed compartment of volume Ω. Well-mixed means that the diffusion of particles in the compartment is the fastest time scale of the system. This implies that the spatial positions of molecules can be ignored and the dynamics of the system only depends on the total molecule numbers. Diluted means the combined volume of all the considered molecules is much smaller than the total volume, which in turn means that the molecules can be considered as point particles (Schnoerr et al., 2017).

Using these two conditions, it can be shown that the state of the system at any time is fully determined by the state vector $\mathbf{n}(t) = (n_1, \ldots, n_I)$, where n_i is the amount of molecules of species, I is the total number of species in the compartment at that time (Gillespie, 1992). The spatial locations and diffusion of molecules does not have to be modelled, and the system corresponds to a *continuous-time Markov jump process*. Hence, the probability for the *j*-th reaction to happen in an infinitesimal time step δt is given by $a_j(\mathbf{n})\,\delta t$, which is the **propensity function** of the *j*-th reaction and proportional to the number of combinations of reactant molecules in $\mathbf{n}(t)$. For example, a bimolecular reaction of the form $\mathrm{X}_1 + \mathrm{X}_2 \xrightarrow{\mathrm{k}_1} \mathrm{X}_3$ has $n_1 n_2$ as the number of pairs with one n_1 and one n_2 molecule. The corresponding propensity function is given by $\mathrm{k}_1 n_1 n_2/\Omega$. The scaling comes from the fact that the probability for two molecules to collide is proportional to the compartment size $1/\Omega$. The compartment size Ω is a scaling parameter that allows us to express the reactions in terms of species concentrations (Ullah and Wolkenhauer, 2011). Thus

$$n_i(t) = \Omega x_i(t) \tag{2.24}$$

where x_i is the concentration of the species i.

To mathematically describe the dynamics of this system, consider the probability distribution $P(\mathbf{n}, t) = P(\mathbf{n}, t|\mathbf{n}_0, t_0)$ for the system to be in state **n** at time t given that it was in state $\mathbf{n}_0$ at time t_0. The probability $P(\mathbf{n}, t + \delta t\ )$ after an infinitesimal time step δt is given by $P(\mathbf{n}, t)$ plus the probability to transition into state **n** from a different state $\mathbf{n}^*$ minus the probability to leave state $\mathbf{n}$:

$$P(\mathbf{n}, t + \delta t\ ) = P(\mathbf{n}, t) + \delta t\ \left(\sum_{j=1}^{J} a_j\left(\mathbf{n} - \mathbf{S}_j\right) P\left(\mathbf{n} - \mathbf{S}_j, t\right) - \sum_{j=1}^{J} a_j(\mathbf{n})\, P(\mathbf{n}, t) \right) \tag{2.25}$$

where $a_j\left(\mathbf{n}\right)\delta t$ is the probability of the *j*-th reaction occurs in an infinitesimal time interval δt , $\mathbf{S}_j$ is the *j*-th column of the stoichiometric matrix **S**. Subtracting $P\left(\mathbf{n}, t\right)$ and dividing by δt and taking the limit $\delta t \to 0$ gives the CME

$$\frac{\partial}{\partial t}P\left(\mathbf{n}, t\right) = \sum_{j=1}^{J}\left[a_j\left(\mathbf{n}-\mathbf{S}_j\right)P\left(\mathbf{n}-\mathbf{S}_j, t\right) - a_j\left(\mathbf{n}\right)P\left(\mathbf{n}, t\right)\right] \tag{2.26}$$

The CME in (2.26) is a very large system because **n** is an *unbounded* discrete-valued vector (typically is a coupled infinite-dimensional system of linear ODEs), where one solution for each possible state gives the probability of the system being in that particular state **n** at time t.

For example, recall the transcription network (2.1) rewritten below again

$$\begin{aligned}\mathrm{DNA} &\xrightarrow{\mathrm{k}} \mathrm{DNA} + \mathrm{mRNA}\\ \mathrm{mRNA} &\xrightarrow{\mathrm{d}} \emptyset\end{aligned} \tag{2.27}$$

Now, instead of considering the reaction rates as in the deterministic case, we will talk about the reaction probability rates. Later we will see the relationship between deterministic reaction rates and stochastic reaction probability ones. For the time being, consider the probability per time unit that one molecule of mRNA is transcribed is k, and the one molecule of mRNA degrades is d. Assume the gene copy number $\mathrm{C_n}$ is 1. If this is not the case, one could simply consider either the product $\mathrm{kC_n}$ as transcription probability rate or, more accurately, $\mathrm{C_n}$ parallel reactions like (2.27).

Let the probability of having n copies of mRNA at time t be denoted as $p(n, t)$. Let us set a balance to obtain the probability of having n copies at time $t+\delta t$. The rationale is as follows: the probability sought for equals the probability of having $n-1$ copies of mRNA at t, and the transcription reaction took place during the elapsed time δt, plus the probability of having $n+1$ copies of mRNA at t, and the mRNA degradation reaction took place during the elapsed time δt, plus the probability that there already were n copies of mRNA at t, and no reaction took place during δt. Notice addition of probabilities is used, because we assume that the first event, *or* the second, *or* the third may occur, and the elapsed time δt is taken small enough so that all three possible events are disjoint. If two events are mutually exclusive, then the probability of either occurring is the sum of the probabilities of each occurring.

The probability that a specific reaction takes place in a given time interval equals the product of the corresponding probability reaction rate, and the elapsed time. Thus, for instance, *each* molecule of mRNA has probability $\mathrm{d}\delta t$ of decaying in the time interval $[t, t+\delta t]$.

On the other hand, the probability that e.g. the degradation reaction took place in a time interval δt is proportional to the number of mRNA copies at t, the rate of degradation d, and the elapsed time δt. The product of the first two terms is often referred to as the *propensity* of the reaction (Higham, 2008).

With the elements above, the probability of having n copies of mRNA at time $t + \delta t$ is

$$p(n, t+\delta t) = p(n-1, t)\mathrm{k}\delta t + p(n+1, t)(n+1)\mathrm{d}\delta t + p(n, t)\left[1 - n\mathrm{d}\delta t - \mathrm{k}\delta t\right] \quad (2.28)$$

Rearranging terms, and taking the limit as δt goes to zero, the CME equation expressing the time evolution of the probability distribution, is reached

$$\frac{\partial p(n,t)}{\partial t} = \mathrm{d}\left[p(n+1,t)(n+1) - p(n,t)n\right] + \mathrm{k}\left[p(n-1,t) - p(n,t)\right] \quad (2.29)$$

Equation (2.29) is a linear infinite dimensional one. There is one ODE for each possible state of the system. That is, the CME has to be solved for all possible values of the mRNA copy number. Thus, we have

$$\begin{bmatrix} \frac{\partial p(0,t)}{\partial t} \\ \frac{\partial p(1,t)}{\partial t} \\ \frac{\partial p(2,t)}{\partial t} \\ \vdots \end{bmatrix} = \begin{bmatrix} -\mathrm{k} & \mathrm{d} & 0 & 0 & 0 & \cdots \\ \mathrm{k} & -(\mathrm{r}+\mathrm{d}) & 2\mathrm{d} & 0 & 0 & \cdots \\ 0 & \mathrm{k} & -(\mathrm{r}+2\mathrm{d}) & 3\mathrm{d} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} p(0,t) \\ p(1,t) \\ p(2,t) \\ \vdots \end{bmatrix} \quad (2.30)$$

But solving equation (2.30) is computationally a challenging problem. Despite its simple structure, solving the CME is an intractable problem from the analytical point of view, although there are some cases where the CME is possible to solve (Jahnke and Huisinga, 2007; Grima et al., 2012). Moreover, there are also numerical schemes like in (Munsky and Khammash, 2006; Kazeev et al., 2014) providing a solution in a truncated state-space. However, long-term predictions are not always possible, even for simple bimolecular reactions, which are the most common in biological networks. Instead of seeking an analytical solution of the CME, it is possible to approximate simulate its exact sample paths using stochastic simulation.

2.4.3 Approximation of the CME. Linear Noise Approximation (LNA)

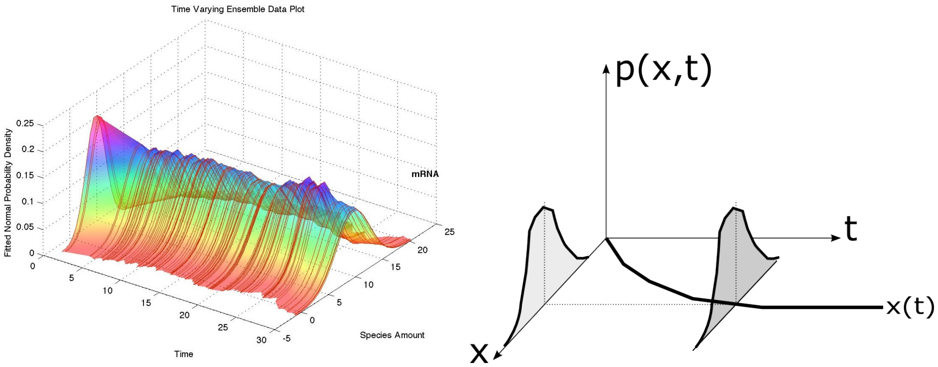


Figure 2.9. Constitutive gene transcription. **Left:** Stochastic simulation with SSA (see section 2.4.4). 3D representation of the time evolution of the mean and variance of mRNA copy number over 100 realizations. **Right:** A representation of the LNA approach. At each time instant, approximate the system response as that of the deterministic model, plus a fluctuation term with an associated gaussian variance-covariance characteristics.

The CME (2.26) can be written directly from the rate constants and stoichiometries of all the elementary reactions of a chemical system, but neither analytical nor numerical solutions are in general available. Fortunately, the CME can often be simplified by the so-called *Linear noise approximation* (LNA) (Van Kampen, 2011) using the Ω-expansion of the CME. The Ω-expansion means that the CME is Taylor-expanded near macroscopic system trajectories or stationary solutions in powers of $1/\Omega$, where Ω is the system volume (see Fig.2.9). This section highlights the key principles to obtain the mathematical expressions for the LNA. A complete development of it can be found in annex A.4.

The LNA tries to deal with noise in a deterministic setting, where analytical solutions are locally valid close to macroscopic trajectories (deterministic reaction rates) plus an additive noise called *fluctuation* term (see Fig.2.9 rigth). This section follows the arguments from (Ullah and Wolkenhauer, 2011) to derive the LNA from the CME. An alternative notation more suited for Taylor expansion uses the step operator $\mathrm{E}_j f(n) = f(n + \mathbf{S}_j)$ for the *j*-th reaction in the CME (2.26)

$$\frac{\partial}{\partial_t} P\left(\mathbf{n}, t\right) = \sum_{j=1}^{J} \left(\mathrm{E}_j^{-1} - 1\right) a_j\left(\mathbf{n}\right) P\left(\mathbf{n}, t\right) \tag{2.31}$$

The LNA can anticipate the way in which the solution of the CME $P\left(\mathbf{n}, t\right)$ will depend on the system size Ω (see Fig.2.9 left). Assuming that the continuous approximation

$\mathbf{n}(t)$ of the system fluctuates around a macroscopic trajectory (deterministic reaction rate) of order Ω with a fluctuation of order $\sqrt{\Omega}$

$$n_i = \Omega\phi_i + \sqrt{\Omega}\xi_i \tag{2.32}$$

where n_i is the molecules number of species i, ϕ is the macroscopic concentration defined in section 2.4.2, and ξ is a random variable from the random matrix $\Xi(t)$, which models the fluctuations around $\phi(t)$. The probability distribution $P(\mathbf{n}, t)$ transforms into the probability distribution $\Pi(\xi, t)$ of $\Xi(t)$

$$P(\mathbf{n}, t) = P\left(\Omega\phi_i + \sqrt{\Omega}\xi_i\right) = \Pi(\xi, t) \tag{2.33}$$

Now, it is necessary to define the propensity function $a_j(\mathbf{n})$ in terms of the fluctuation ξ, the deterministic rate $\nu_j(x)$, and the operator E_j for each *j*-th reaction through $a_j(\mathbf{n}) = \Omega\left[\nu_j\left(\phi + \Omega^{-1/2}\xi\right) + O\left(\Omega^{-1}\right)\right]$. Replacing $a_j(\mathbf{n})$ in the CME (2.31)

$$\frac{\partial}{\partial t}P(\mathbf{n}, t) = \Omega\sum_{j=1}^{J}\left(\mathrm{E}_j^{-\Omega^{-1/2}} - 1\right)\left[\nu_j\left(\phi + \Omega^{-1/2}\xi\right) + O\left(\Omega^{-1}\right)\right]\Pi(\xi, t) \tag{2.34}$$

where $O(x)$ is the first neglected order with respect to x in an expansion.

Equations (2.34) and (2.33) lead to the linear equation (further details in annex A.4)

$$\frac{\partial\Pi}{\partial t} = -\sum_{i,k}\mathbf{A}_{ik}\frac{\partial\left(\xi_k\Pi\right)}{\partial\xi_i} + \frac{1}{2}\sum_{i,k}\mathbf{BB}_{ik}^{T}\frac{\partial^2\Pi}{\partial\xi_i\partial\xi_k} \tag{2.35}$$

where $\mathbf{A} = \sum_{j=1}^{J}\mathbf{S}_{ij}\frac{\partial\nu_j}{\partial\phi_k}$ is the Jacobian matrix, and $\mathbf{B} = \sum_{j=1}^{J}\nu_j\mathbf{S}_{ij}\mathbf{S}_{kj}$ is the diffusion matrix. Both of these matrices depend on time thorough the deterministic rate concentration $\phi(t)$. The terms of order $\Omega^{-1/2}$ are proportional to $\frac{\partial\Pi}{\partial\xi_i}$, and ϕ corresponds to $\frac{d\phi_i}{dt} = \sum_{j=1}^{J}\mathbf{S}_{ij}\nu_j(\phi)$. The stationary solution of (2.35) is a multidimensional normal distribution $P(\xi) = \left((2\pi)^{I/2}\sqrt{\det\mathbf{\Xi}}\right)^{-1}\exp\left(-\xi^T\mathbf{\Xi}\xi/2\right)$, which has a covariance matrix $\mathbf{\Xi} = \langle\xi\xi^T\rangle$, and follows a Lyapunov equation

$$\mathbf{A\Xi} + \mathbf{\Xi A}^T + \mathbf{BB}^T = 0. \tag{2.36}$$

The correlation matrix of the stationary process is $\langle\xi(t)\xi^T(s)\rangle = \mathbf{\Xi}\exp\left(\mathbf{A} \mid t - s \mid\right)$. Thereby, we can determine the symmetric **covariance matrix** as $\mathbf{C} = \Omega\mathbf{\Xi}$. The LNA solutions together with the matrix $\mathbf{C}$ often give very accurate descriptions of the size of molecule number fluctuations and how they are correlated.

As an example, consider the reaction network (2.27) for the transcription of a gene. Applying the LNA, we can obtain the mean and the variance (μ, σ^2 respectively) of the species involved in the system. The temporal dynamics of $n_1 = [\mathrm{DNA}]$ and $n_2 = [\mathrm{mRNA}]$ is described by the ODE model (see section 2.3)

$$\begin{aligned} \dot{n}_1 &= 0 \\ \dot{n}_2 &= \mathrm{k}n_1(t) - \mathrm{d}n_2(t), \quad n_2(0) = n_{20},\ t \geq 0 \end{aligned} \tag{2.37}$$

Consider the steady-state $n = \overline{n}$ solution of the system (2.37), where $\overline{n}_1 = n_1$ is a constant number of DNA molecules, and $\overline{n}_2 = \frac{\mathrm{k}}{\mathrm{d}}n_1$ is the molecules number of mRNA. The LNA implies solving $\mathbf{A\Xi} + \mathbf{\Xi A}^T + \mathbf{BB}^T = 0$ with

$$\mathbf{A} = \left[\begin{array}{cc} \mathrm{k} & -\mathrm{d} \end{array}\right]_{n=\overline{n}}, \qquad \mathbf{B} = \left[\begin{array}{cc} 0 & \mathrm{k}n_1 + \mathrm{d}n_2 \end{array}\right]_{n=\overline{n}}, \qquad \mathbf{\Xi} = \left[\begin{array}{cc} \sigma^2_{n_1} & \sigma_{n_1 n_2} \\ \sigma_{n_2 n_1} & \sigma^2_{n_2} \end{array}\right]$$

The number of molecules of n_1 is fixed. Hence its variance and covariances between n_1 and n_2 are null ($\sigma^2_{n_1}$, $\sigma_{n_1 n2}$, and $\sigma_{n_2 n1}$, respectively). For the variance of mRNA ($\sigma^2_{n_2}$) we have

$$\sigma^2_{n_2} = \left.\frac{(\mathrm{k}n_1 + \mathrm{d}n_2)^2}{2d}\right|_{n=\overline{n}} \tag{2.38}$$

Replacing the steady-states values $\overline{n}_1, \overline{n}_2$ in (2.38), we obtain the variance for the stationary distribution of mRNA as $\sigma^2_{n_2} = \frac{2\mathrm{k}^2}{\mathrm{d}}n_1^2$. Figure 2.10 illustrates the simulated solution of the number of molecules evolution using the LNA.

The LNA is very accurate for systems where the fluctuations around the stationary state do not drive the system to regions far away from the equilibrium point at which the linearization was carried out (Elf and Ehrenberg, 2003). For large systems is difficult to deal with the Jacobian matrix $\mathbf{A}$ or the diffusion matrix. Having several species or having a large population of cells also increase the computational load of the probability distributions simulations. As an alternative to solve the CME, it is possible to simulate the exact sample paths of the CME by using stochastic simulation.

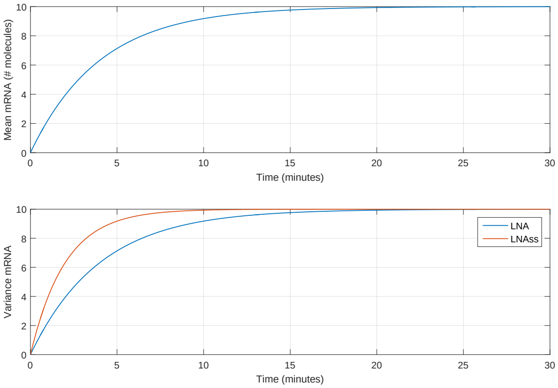


Figure 2.10. (Top) Deterministic output for the mRNA concentration. Units are number of molecules. **(Bottom)** Comparison between the variance of $\xi \triangleq \sqrt{V}\xi(t)$ obtained using equations (2.4.3) and (2.38). Units are number of molecules squared per cell. Parameters used: the gene copy number $\mathrm{C_n}$=1 molecule is also the initial condition $n_1(0)$ (not showed), transcription rate k=2.5 min^{-1}, degradation rate d=0.25 min^{-1}, variable step time in seconds, and $x_2(0) = 0$.

2.4.4 Discrete stochastic simulation. The Gillespie algorithm

Typically, the CME is too high-dimensional to deal with computationally due to the high dimensionality of the domain of the probability distribution $P(\mathbf{n}, t)$, which leads to an exponential increase of computational and memory cost with the network size. The *stochastic simulation algorithm* (SSA) or Gillespie algorithm gets around this issue by computing single realizations of the state vector rather than an entire probability distribution (Gillespie, 1977, 2007).

The Gillespie algorithm allows us to draw exact samples (also called realizations or runs) of the Markov jump process to obtain a numerical solution from the underlying stochastic process. It generates time course trajectories of the system state over a given time window, starting from a given initial system state $n_0(t)$. The SSA algorithm is *exact* in the sense that each run is an independent realization from the true underlying process. The properties deduced about the probabilistic nature of the process from multiple runs can be made arbitrarily accurate by averaging over a sufficient number of runs to reduce the *Monte Carlo error* associated with the estimates (Wilkinson, 2006).

The Gillespie algorithm introduce $P_0(\tau|\mathbf{n}, t)$ as the probability that no reaction takes place in the time interval $[t, t+\tau)$ (Higham, 2008). Considering the time interval $[t+\tau, t+\tau+\delta\tau)$ and assuming that what happens over this interval is independent of what happens in the first interval $[t, t+\tau)$, we have

$$\begin{aligned} P_0\left(\tau+\delta\tau|\mathbf{n},t\right) &= P_0\left(\tau|\mathbf{n},t\right)\left(1-\sum_{j=1}^{J}a_j\left(\mathbf{n}\right)\delta\tau\right) \\ \frac{P_0\left(\tau+\delta\tau|\mathbf{n},t\right)-P_0\left(\tau|\mathbf{n},t\right)}{\delta\tau} &= -P_0\left(\tau|\mathbf{n},t\right)a_{sum}\left(\mathbf{n}\right) \end{aligned} \tag{2.39}$$

where $a_{sum}\left(\mathbf{n}\right)=\sum_{j=1}^{J}a_j\left(\mathbf{n}\right)$ is the sum of all propensity functions. Solving equation (2.39) when $\delta\tau\to 0$ and the initial condition is $P_0\left(0|\mathbf{n},t\right)=1$

$$P_0\left(\tau|\mathbf{n},t\right)=e^{-a_{sum}(\mathbf{n})\tau} \tag{2.40}$$

The key quantity for the SSA is $P\left[\tau,j|\mathbf{n},t\right)$, which is defined by given $\mathbf{n}(t)=n,\ P\left(\tau,j|\mathbf{n},t\right)\delta\tau$ as the probability that the next reaction *i)* will be the *j*-th reaction, and *ii)* will occur in the time interval $[t+\tau,t+\tau+\delta\tau)$. Using the definitions of P_0 and $a_j\left(\mathbf{n}\right)$, we obtained an expression for $P\left(\tau,j|\mathbf{n},t\right)$ during the mentioned interval. Replacing equation (2.40) in (2.39) one obtain

$$\begin{aligned} P\left(\tau,j|\mathbf{n},t\right)\delta\tau &= P_0\left(\tau|\mathbf{n},t\right)a_j\left(\mathbf{n}\right)\delta\tau \\ P\left(\tau,j|\mathbf{n},t\right) &= \frac{a_j\left(\mathbf{n}\right)}{a_{sum}\left(\mathbf{n}\right)}a_{sum}\left(\mathbf{n}\right)e^{-a_{sum}(\mathbf{n})\tau} \end{aligned} \tag{2.41}$$

where equation (2.41) was conveniently rewritten to obtain: *i)* the next reaction index $\frac{a_j(\mathbf{n})}{a_{sum}(\mathbf{n})}$ as a discrete random variable where the chance of picking the *j*-th reaction is proportional to $a_j\left(\mathbf{n}\right)$, and *ii)* the time until next reaction $a_{sum}\left(\mathbf{n}\right)e^{-a_{sum}(\mathbf{n})\tau}$ is the density function for a continuous random variable with an *exponential distribution*. These two random variables formally describe $P\left(\tau,j|\mathbf{n},t\right)$ as their joint density function. The essential structure of this discrete event simulation algorithm is outlined below as in (Wilkinson, 2006):

- Step 1: set the initial number of molecules of each biochemical species in the reaction network and set the simulation time to zero.

- Step 2: on the basis of the current molecular abundances, calculate the propensity for each possible reaction event.

- Step 3: using the current propensities, simulate the time to the next reaction event, and update the simulation time accordingly (the larger the reaction propensities, the shorter the time to the next event).

- Step 4: pick a reaction event at random, with probabilities determined by the reaction propensities (higher propensities lead to higher probability of selection), and update the number of molecules accordingly.

- Step 5: record the new simulation time and state.

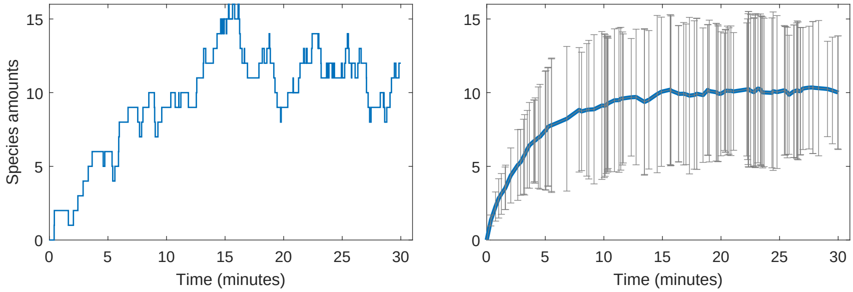


Figure 2.11. Left:. A single run of the SSA for constitutive gene transcription. **Right:** Mean of mRNA over 100 runs together with mean ± SD (error bars). The parameters used are the same ones for the LNA. Step time $\delta\tau = 0.3$ seconds

- Step 6: check the simulation time. If the simulation is not yet finished, return to step 2.

To give an explicit example, the Gillespie algorithm was implemented to generate one realization from the stochastic discrete model (2.42) of the transcription network (2.27). Figure 2.11 shows the stochastic simulation results for the species DNA and mRNA in equation (2.42)

$$P\left(\tau, j|\mathbf{n}, t\right)\delta\tau = P\left(\tau, 1|n+1, t\right)\mathrm{k}\, x_1\delta\tau + P\left(\tau, 2|n-1, t\right)\mathrm{d}\, x_2\delta\tau \tag{2.42}$$

The SSA is exact, in the sense that the statistics from the CME are reproduced precisely. But it comes at a high computational cost even for few species. In particular, if the molecule numbers have large fluctuations or if many reactions happen per unit time. In the first case a large number of samples have to be simulated to obtain statistically accurate results, whereas in the second case single simulations become expensive since the time between reaction events becomes small (Schnoerr et al., 2017). We can try to speed up the SSA by *lumping together* reactions and only updating the state vector after many reactions have fired. This is the so-called *tau-leaping approximation* that introduces errors that will be small as long as the state vector updates are relatively small (Higham, 2008). Therefore, pushing the approximation further leads to a continuous stochastic simulation.

However, we will see in Chapter 3 that having an interconnected population of cells in a synthetic gene network jeopardizes the possibility of employing SSA for several reasons. First there are different volumes involved, extracellular and intracellular. The diffusion of an intracellular species through the membrane depends on the concentration gradient of the extracellular one, which in turn is function of the extracellular volume. It makes the account for the probability of this diffusion process more complicated. Second, when using SSA, several realizations or trajectories of the system are needed in order to obtain an accurate estimation of the moments, making the use of SSA in a population of interconnected cells a computationally very demanding task. These kind of systems have been modeled and simulated before as ODE perturbed with white noise (Koseska et al., 2009). However, this representation does not capture the intrinsic noise phenomena as desired.

2.4.5 Continuous stochastic simulation. Chemical Langevin Equation (CLE)

The *Chemical Langevin Equation* (CLE) is a practical way to model gene expression noise. The CLE is a stochastic differential equation driven by zero-mean Gaussian noise that describes the system when the molecules of reactants into a cell population are sufficiently large (Gillespie, 2007, 2000). From the CLE, one can obtain statistical parameters to analyze the noise generated during gene expression. Moreover, to analyze a stochastic system, statistical moments such as the mean of gene expression (μ), variance (σ^2) or standard deviation (σ) are used. An alternative measure of gene expression noise is the noise strength ($\eta = \sigma^2/\mu^2$), that shows of the dispersion of a probability distribution for the number of molecules of certain species (Paulsson, 2004). Thus, the SSA and the CLE methods allow us to obtain certain statistical parameters for analyzing the fluctuations of a synthetic gene circuit.

For a system in state $\mathbf{x}$ at time t suppose a $\delta t > 0$ that satisfies the condition that the number of times each reaction fires in the next infinitesimal time interval $[t, t + \delta t)$ follows a Poisson distribution[2] with mean (and variance) $a_j(\mathbf{x})\delta t$. The term $a_j(\mathbf{x})\delta t \gg 1$ is known as the reaction propensity, and we have

$$\mathbf{X}(t + \delta t) = \mathbf{x} + \sum_{j=1}^{M} \nu_j \mathcal{P}(a_j(\mathbf{x})\delta t, a_j(\mathbf{x})\delta t) \tag{2.43}$$

where $\mathbf{x} = \mathbf{X}(t)$ is the number of molecules, M is the number of reactions, and ν_j is the stoichiometry (change in the molecular population) caused by reaction j. Denoting the normal (Gaussian) random variable with mean μ and variance σ^2 by $\mathcal{N}(\mu, \sigma^2)$, and using the fact that a Poisson random variable with a mean and variance much larger than one (μ, $\sigma^2 \gg 1$ respectively) can be approximated as a continuous normal random variable with that same mean and variance. The equation (2.43) can be approximates as

[2] The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known average rate and independently of the time since the last event.

$$\begin{aligned}\mathbf{X}(t+\delta t) &= \mathbf{x}(t) + \sum_{j=1}^{M} \nu_j \mathcal{N}_j(a_j(\mathbf{x})\delta t, a_j(\mathbf{x})\delta t) \\ &= \mathbf{x}(t) + \sum_{j=1}^{M} \nu_j \left[a_j(\mathbf{x})\delta t + \sqrt{a_j(\mathbf{x})\delta t}\mathcal{N}_j(0,1)\right]\end{aligned} \tag{2.44}$$

where we used the well-known property of the normal random variable that $\mathcal{N}(\mu, \sigma^2) = \mu + \sigma(0,1)$. Collecting terms from (2.44) gives the Chemical Langevin Equation (CLE)

$$\mathbf{X}(t+\delta t) = \mathbf{x} + \sum_{j=1}^{M} \nu_j a_j(\mathbf{x})\delta t + \sum_{j=1}^{M} \nu_j \sqrt{a_j(\mathbf{x})}\mathcal{N}_j(0,1)\sqrt{\delta t} \tag{2.45}$$

where $\mathcal{N}_j(0,1)$ is a statistically independent normal random variable. Note that equation (2.45) has sense only if the populations of all reaction substrates are sufficiently large, which ensures the reaction will be fired during δt.

The simulation results the reaction network (2.27) using the CLE equation (2.45) is depicted in Fig.2.12.

Since a CLE is a special form of the general SDE, the *Euler-Maruyama method* (Higham, 2001) can be used for generating sample paths of the stochastic process driven by a CLE. The method involves generation of random numbers from the Gaussian distribution to represent the *Wiener processes* and then using the update rule (2.45) at each time step (Ullah and Wolkenhauer, 2011). As we sill see in Chapter 4, the stochastic dynamics of all reactions involved in a gene circuit was modeled and simulated by using CLE together with the Euler-Maruyama method.

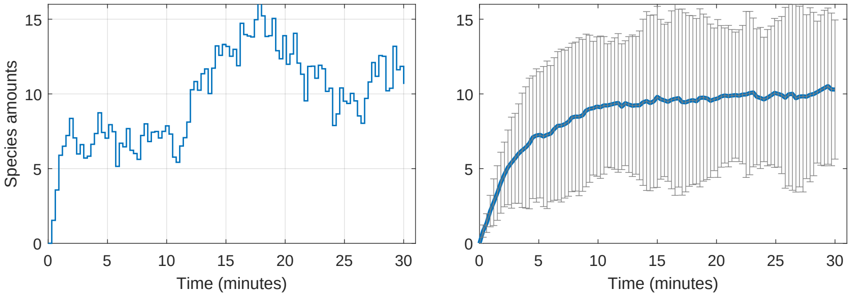


Figure 2.12. Left:. A single run of the CLE for constitutive gene transcription. **Right:** Mean of mRNA over 100 runs together with mean ± SD (error bars). The parameters used are the same for the SSA algorithm.

2.5 Multi-objective optimization

Optimization can be used as a computational tool to search the best solution for a given problem in a systematic and efficient way. In the context of synthetic biology, coupling mathematical modeling and optimization with suitable simulation of synthetic gene circuits contributes to discern some of their principles and behavior under defined conditions.

This thesis raises **Multi-Objective Optimization Design (MOOD)** as the proper framework to deal with common problems arising in rational design and tuning of synthetic gene circuits. Using a classical systems engineering approach, the thesis mainly addresses: *i)* synthetic gene circuit modelling based on first-principles, *ii)* model parameters estimation from experimental data, and *iii)* model-based tuning to achieve desired circuit performance. The optimization problem will be a multi-objective one in the general case. Typically, some of the objectives will be in conflict, so a trade-off among solutions is required. *Ad hoc* weighting of the different objectives may be used to transform the problem into a single-objective one (Mattson and Messac, 2005). Alternatively, thresholds on each of the objectives may be set in order to run multiple times a single-objective optimization. Instead, we can address the problem as a truly multi-objective optimization design one.

In MOOD all objectives are important, so all of them are optimized simultaneously. Thus, the solution rarely is unique, but a set of solutions is called the **Pareto Front**. In this sense, all Pareto-optimal solutions differ from each other in a trade-off between the objectives that each one represents. Then, the design carefully reduces the desired dynamics into the objectives as an optimization problem in the MOOD framework (Meza, 2014). As a result, the designer obtains qualitative regions/intervals of parameters along the Pareto Front giving rise to the predefined behavior of the circuit. Contrarily to the passive search for solutions of Monte Carlo-based approaches, the MOOD may actively searches for all the optimal solutions as a first step. The MOOD framework also naturally provides a classification of the parameters along the Pareto front, by taking into account their effect on each of the goals. Moreover, this framework makes easy to analyze the impact of context on the synthetic devices to be designed. This can be done by just incorporating information about the relationship between the device and the context. In general, this means we only need to know where do we connect the device which is being designed and how we are connecting it. Including this information in the optimization problem, we obtain a qualitative region of parameters taking into account the effect of the context on the device.

To successfully implement the multi-objective optimization approach, at least three fundamental steps are required (Miettinen et al., 2008), as depicted in the figure 2.13:

1. the *multi-objective problem definition (MOP)*: defining the circuit behavior specifications in a proper way,

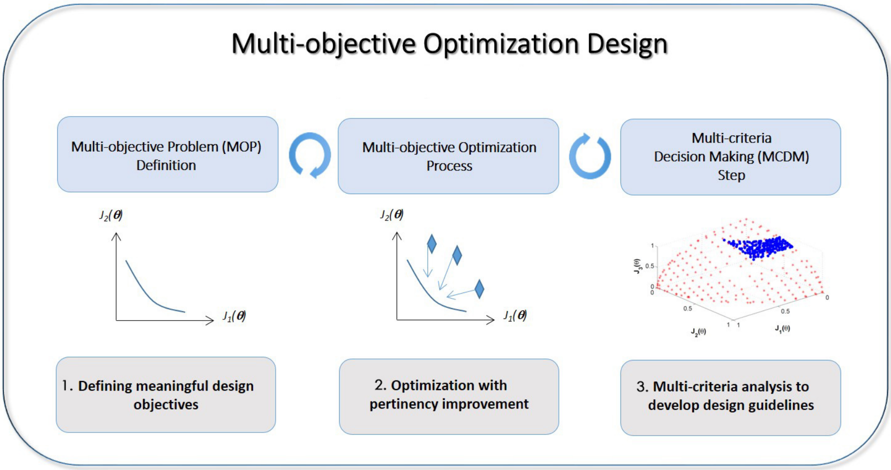

Multi-objective Optimization Design
Multi-objective Problem (MOP) Definition
Multi-objective Optimization Process
Multi-criteria Decision Making (MCDM) Step
$J_2(\theta)$
$J_1(\theta)$
$J_2(\theta)$
$J_1(\theta)$
$J_3(\theta)$
$J_2(\theta)$
$J_1(\theta)$
1. Defining meaningful design objectives
2. Optimization with pertinency improvement
3. Multi-criteria analysis to develop design guidelines

Figure 2.13. Steps for the multi-objective optimization design (MOOD). *Figure inspired in* (Meza, 2014).

2. the *optimization process*: tuning the parameters using multi-objective global optimization (MOO).

3. and the *multi-criteria decision making stage (MCDM)*: obtaining tuning guidelines useful for the wet-lab implementation.

2.5.1 Multi-objective problem definition (MOP)

The main goal of this first stage (see Fig.2.13-1) is to articulate a multi-objective statement that precisely describes the problem to address and ensures that the obtained solutions match the needs of the decision-maker (Paula et al., 2017). First, the context must be considered: which is the aim of the analysis?, which are the variables under study?, and what rules connect them?. These questions answer how many objectives of the procedure we should define, and which are their corresponding decision variables. For this purpose, generally, a parametric model is used, relating the variables among them and describing the system's behavior. As we saw in sections 2.3 and 2.4, a model will set some limitations to the values that the variables can take, and will establish the outputs that result when those variables take particular values. Taking into account the model, the objectives and the constraints strongly influence the solutions that can be found through the process. Thus, this is a fundamental step to ensure the quality of the results (Mattson and Messac, 2005).

As said above, in this Thesis the multi-objective optimization problem definition is applied to two problems: *i)* model parameters estimation from experimental data, and *ii)* model-based tuning to achieve desired circuit performance. In both cases, this stage will be properly defined and solved in Chapters 5 and 7.

2.5.2 Multi-objective optimization (MOO)

In this second stage (see Fig.2.13-2), the multi-objective optimization process itself seeks to approximate the best values of the decision variables θ_P^* that give the best Pareto front approximation J_P^* (Miettinen, 1999). Such search could be done through a random Monte-Carlo sampling in the decision variables space θ –the set of parameters determining our biological model–, followed by filtering of the solutions in order to obtain the decision variables θ_P^* (hereinafter Pareto set) that defines the Pareto front approximation J_P^*. This could be a good option for problems with few decision variables. For problems with a large number of decision variables, as synthetic gene circuits, it is more efficient to use an appropriate multi-objective optimization algorithm to approximate this solution.

Convergence and spreading properties of the solutions are considered *a must* in multi-objective optimization (see Fig.2.14). An additional required characteristic regards to pertinency of solutions, i.e. getting interesting solutions from the designer point of view. It might happen that some obtained solutions are not useful, due to the a strong degradation in some objectives. This is a characteristic to take into account.

In the Thesis, the Pareto front of solutions was obtained via **spMODE**, a multi-objective optimization algorithm based on differential evolution (Reynoso-Meza et al., 2010, 2013b) implemented in Matlab, available at Matlab Central[3]. The algorithm spMODE actively searches for all the solutions in the parameter space along the Pareto front. Thereby, it:

- improves convergence by using an external file to store high-quality solutions and include them in the evolutionary strategy itself. The main idea is instead of using the classical relation for dominance, one can follow concepts such as: *i)* a solution dominates the solutions that are less fit for all the objectives, or *ii)* a solution dominates the solutions inside a distance that is less than a parameter ϵ as a dominance measure (Herrero et al., 2005).

- improves spreading by using the spherical pruning mechanism based on spherical relations in the objective space (Reynoso-Meza et al., 2010). This technique shows good flexibility dealing with diverse geometries in m-dimensional Pareto fronts, so it achieves a well-spread set of solutions (see Fig.2.14),

- improves the pertinency of the solutions, i.e. getting interesting solutions from the designer's point of view, by means of a basic bound mechanism in the objective space, as described in (Reynoso-Meza et al., 2012).

[3] http://es.mathworks.com/matlabcentral/fileexchange/39215

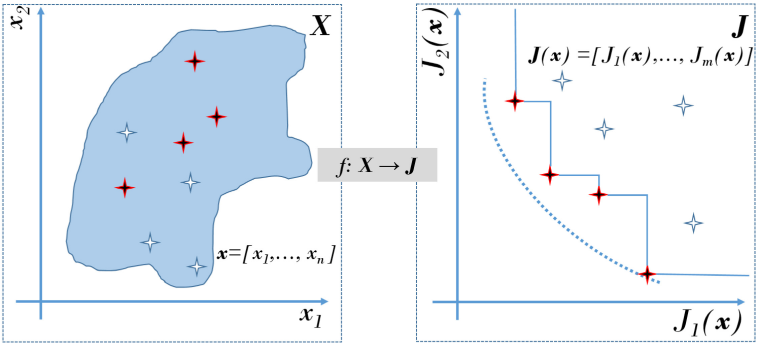


Figure 2.14. Decision variable's space (left) showing the optimal values (in color black) for the variables x_1 and x_2 that corresponds to the Pareto front (right) computed by the spMODE algorithm.

2.5.3 Multi-criteria decision making (MCDM)

The selection of the preferable solution according to designer's criteria is the final stage of the MOOD framework (Fig.2.13-3). It takes place in an *a-posteriori* multi-criteria analysis of the Pareto front approximation. Using tools that simplify the visualization and the analysis of the trade-off among competing objectives. Such visualization and analysis is not a trivial task when the number of objectives is larger than three and/or the number of decision variables in the Pareto set is large. Several tools are available for designers, but in any case, characteristics to analyze and visualize the results are desirable:

- The tool must enable to compare design alternatives (analyze different solutions).
- It must enable to compare design concepts, that is, analyze different Pareto front approximations.
- Completeness: all relevant information should be contained in the visualization.
- Persistence: all the relevant information should be retained in the designer's mind.
- Simplicity: the visualization should be easily understandable.

The last three characteristics are related to the degree of training and/or familiarity of the designer with a given tool. The first two depend on the required multi-criteria analysis.

In this Thesis, the visualization tool known as **Level Diagrams** (LD) (Blasco et al., 2008; Reynoso-Meza et al., 2013a) was used. It has a freely available implementation for designers[4]. LD-Tool allows the designer to correlate design objectives with decision variables. It classifies the calculated optimal parameters θ_P^* with respect to each objective $J_q(\theta)$ normalized with respect to its minimum and maximum value. A graph for each objective is displayed as in Fig.2.15 top. The Y-axis is the p-norm $\|\hat{J}(\theta)\|_p$ of the objectives vector, and the X-axis corresponds to the objective value or decision variable depending on the case. A second graph displays $\|\hat{J}(\theta)\|_p$ with respect to each decision variable (see Fig.2.15 bottom). This characteristics make it helpful in order to propagate the information between the design objectives space and the decision variables space. Thus, a given solution will have the same value -y in all graphs. LD-Tool enables the simple comparison of alternative design solutions.

[4] Tool available at http://www.mathworks.com/matlabcentral/fileexchange/24042

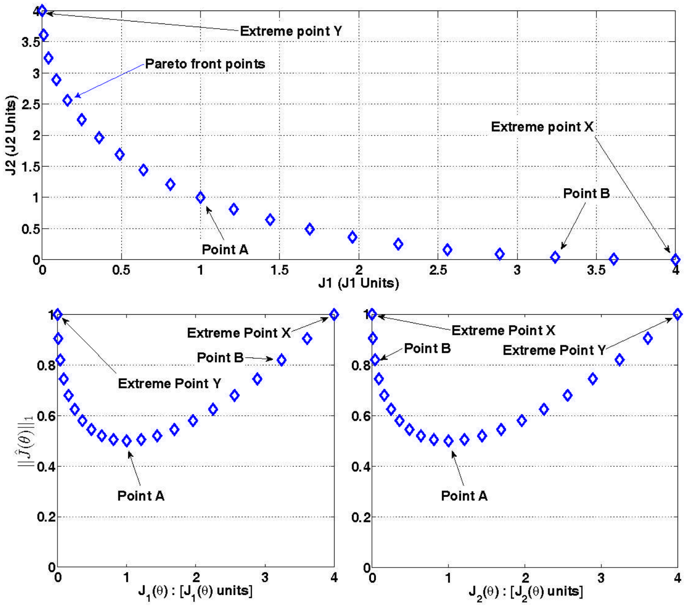


Figure 2.15. Example of Level Diagrams tool for a bi-objective Pareto front (top). The Y-axis is the p-norm $\|\hat{J}(\theta)\|_p$ (or distance to the ideal point) of the objectives vector $\mathbf{J}(\theta)$. *Figure inspired in* (Blasco et al., 2008)

Chapter 3

Two case studies: incoherent feedforward and quorum sensing/feedback gene circuits.

3.1 Introduction

Synthetic biology broadly encompasses the genetic engineering of microorganisms to implement and test new biological functions. To this end, synthetic gene circuits are built following engineering principles of mathematical design and modeling, feedback control and optimization. This Thesis has considered and contributed to solve different problems arising in all four aspects. Thus, drawing from a practical problem-based engineering approach, the Thesis has considered problems arising in two specific applications, and the corresponding synthetic gene circuits used in both cases:

1. The **Incoherent type 1 feedforward circuit (I1-FFL)**. This is a well-known gene circuit which presents an interesting behavior for many applications: its output temporarily responds to a change in its input and then returns to the value it had prior to the application of the stimulus. This behavior, often referred to as *adaptative*, is relevant in many biological processes. This explains why this circuit motif is so widespread in natural biological networks.

2. The **Quorum sensing/Feedback circuit (QS/Fb)**. This is a gene circuit entirely developed and implemented during this work. Its goal is to reduce noise in

protein production while maintaining a desired mean value. The circuit combines two subsystems: *i)* an intracellular negative feedback loop, and *ii)* and extracellular feedback loop based on cell-to-cell communication via *quorum sensing*.

Both the I1-FFL and the QS/Fb circuits will be used as two case studies in this Thesis. In this chapter, the structure, construction in the laboratory, the biochemical reactions, and the dynamics of both circuits are described.

3.2 Incoherent type 1 feedforward circuit

Adaptation is an important property of biological systems, linked to homeostasis (Alon, 2007) and to the generation of responses that depend on the fold-change in the input signal and not on its absolute level (Goentoro et al., 2009). It is defined as the particular ability of biological circuits to respond to a change in its input and return to the value it had prior to the stimulus, as depicted in Fig.3.1. Due to its relevance, synthetic gene circuits showing adaptation have received much attention for a long time (Alon, 2007; Ma et al., 2009; Rodrigo and Elena, 2011; Rahi et al., 2017).

Although currently there are no *catalogues* of functional modules, there is a vast literature in the systems biology area on network motifs producing a variety of dynamic behaviors like adaptation. Circuit topologies giving rise to adaptive behavior have been extensively studied (Alon, 2007). Feedforward circuits with such adaptive behavior are an important case. In (Ma et al., 2009) all three-node possible network topologies that present adaptive dynamical behavior are analyzed using function-topology maps based on Monte Carlo sampling in the parameters space. Using a simple enzymatic model, the authors draw design principles of adaptation circuits. They show that there are only two core solutions that achieve robust adaptation: negative feedback loops and incoherent feed-forward ones.

In particular, the **Incoherent type 1 feedforward loop (I1-FFL)** is one of these feedforward network motifs. Its three-node structure is the second most common feedforward type in *E. coli*, yeast, as well as higher microorganisms. Figure 3.1 illustrates the I1-FFL. Feedforward motifs consist of two paths from the input to the output: a direct path, and an indirect one. The sign of the indirect path is opposite to the one of the direct one. Thus, the effect of a change in the input affects the output via both paths, either increasing or decreasing the effect, and at different time instants due to the different length of both paths. Specifically in the case of the I1-FFL circuit, the direct path (C block in Fig.3.1) is positive and the indirect path (A and B blocks) is negative.

Theoretical studies in (Alon, 2007; Behar et al., 2007) suggest the I1-FFL can act as pulse generator and even as a sign-sensitive accelerator. Recently (Rahi et al.,

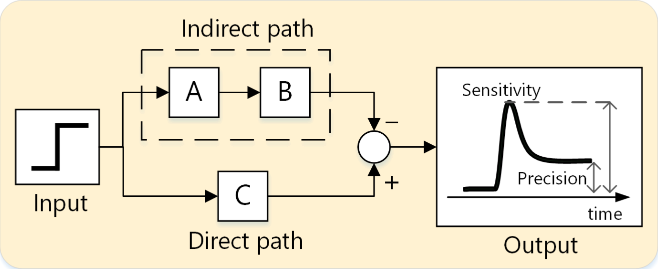


Figure 3.1. Input-output adaptive behavior. I1-FFL is made of two parallel but antagonistic regulation paths. The direct path activates the Actuator and the subsequent output, but it also activates the feedforward path that represses the actuator as well as the output.

2017) has enunciated an orthogonal approach to probe *response signatures* (i.e. , characteristic input-output features) in response to oscillatory stimulation for adapting systems like the I1-FFL circuit. Moreover, the ability to sense the environment is a fundamental trait of biological systems. in this context, the I1-FFL circuit has been presented as one of the sensory systems in cells and organisms, which shares a recurring property called fold-change detection (Adler and Alon, 2017). The circuit can be included as sensor in signaling pathways and the bacterial chemotaxis system that guides motion toward attractants. Therefore, researchers have begun to map the space of feedforward circuits and the functions they can provide.

Though the I1-FFL circuit structure can achieve adaptive behavior, its parameters must be tuned in order to actually achieve it. In (Chiang et al., 2014), the incoherent feedforward adaptive enzyme network structure derived in (Ma et al., 2009), is used as case-study. A method is proposed to make inferences on the contribution of individual parameters to specific components of the system. Classes of kinetic parameters are obtained that may correspond to varying strengths of enzymatic reactions that can be measured and classified experimentally. The authors show that, for a given network structure, certain types of values, or motifs, also exist for kinetic parameters in order to achieve specific system dynamics. Clustering in the parameters space to detect kinetic motifs, i.e. sets of parameters yielding desired circuit dynamics, is used in (Chiang and Hwang, 2013).

3.2.1 I1-FFL circuit structure

Different implementations of the I1-FFL circuit are possible, including enzyme reaction networks (Ma et al., 2009; Chiang et al., 2014), gene circuits (Basu et al., 2004; Rodrigo and Elena, 2011) and *in vitro* transcriptional networks (Kim et al., 2014).

In this Thesis an implementation based on a circuit with three genes has been used, as shown in figure 3.2. The gene *gfp* represents the direct path of the I1-FFL circuit, while the genes *luxR* and *cI* are part of the indirect path as it was depicted in Fig.3.1.

Thus, the main biochemical species expressed by these genes are: proteins LuxR, cI, GFP, and the external inducer $\mathrm{AHL}_{\mathrm{ext}}$. GFP is a fluorescent protein considered as the output of the circuit. To introduce a step-like input signal to the circuit, the addition of small inducer molecules N-acyl-L-homoserine lactone (AHL) (Kaplan and Greenberg, 1985; Fuqua et al., 2001) is considered. $\mathrm{AHL}_{\mathrm{ext}}$ molecules diffuse from the extracellular culture inside the cell. Most of these inducers undergo an heterodimerization, i.e. the inducer binds to one of the circuit species thus effectively providing an input to the circuit. Most of them subsequently dimerize. These phenomena are present in both the I1-FFL and the QS/Fb circuits, and they will be modeled in Chapter 4.

Particulary, protein LuxR binds to the inducer AHL and forms a monomer (LuxR.AHL), which in turn dimerizes. The dimer (LuxR.AHL)$_2$ is the transcription factor activating expression of both downstream genes *cI* and *gfp*. It directly activates expression of GFP, and indirectly represses it via activation of the repressor protein cI. In turn, protein cI becomes a transcription factor of the *gfp* inhibiting expression of GFP. Thereby, (LuxR.AHL)$_2$ acts as an activator of the hybrid promoter $\mathrm{P}_{\mathrm{lux/cI}}$ and the promoter $\mathrm{P}_{\mathrm{lux}}$, whilst protein cI is the repressor of the promoter $\mathrm{P}_{\mathrm{lux/cI}}$. As a result, when a signal causes node *luxR* to assume its active conformation (dimer (LuxR.AHL)$_2$), GFP is produced, but after some time cI accumulates, eventually attaining the repression threshold for the gene *gfp*.

3.2.2 I1-FFL circuit construction

The I1-FFL synthetic gene circuit was implemented in *E. coli* with two different plasmids (see Fig.3.3). The genes *luxR* and *gfp* with their corresponding transcriptional units are in one plasmid, and the gene *cI* is in another different one (see further details in annex B.1).

On the one hand, in the plasmid **pCB14mut**, the gene coding for the protein LuxR (BBa_C0062) is constitutively expressed under the control of a medium strength promoter (BBa_J23106) and a strong RBS (BBa_B0034). Also in the same plasmid, a hybrid promoter $\mathrm{P}_{\mathrm{lux/cI}}$ (BBa_K415032) drives the expression of protein GFP (BBa_K082003) with a strong RBS (BBa_B0034). This two cassettes are placed in a pBR322 plasmid backbone.

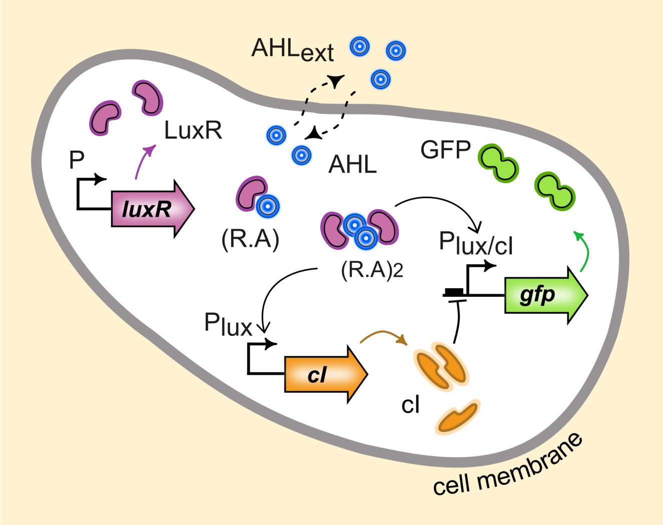


Figure 3.2. I1-FFL circuit. It shows an important biological property so-called *Adaptation*. The gene *luxR* produces LuxR protein, which in turn binds to the inducer AHL forming the transcription factor $(\text{LuxR.AHL})_2$. It activates both *gfp* and *cI*, so GFP expression begins. Then, cI protein generated by *cI* represses GFP expression until eventually it achieves the same level as before activation.

On the other hand, the plasmid **pCB11a** contains the gene *cI* (BBa_K327018) controlled by the $\mathrm{P}_{\mathrm{lux}}$ repressible promoter (BBa_R0062), and a mild ribosome binding site (RBS part BBa_B0033) in the pACYC184 plasmid backbone.

The *cI* and *gfp* coding sequences are followed by the terminator BBa_B0015 in both plasmids. All parts were taken from the Registry of Standard Biological Parts described in section 2.1.2, and cloned using the 3 Antibiotic Assembly method from Biobrick's foundation (see annex A.1).

Additionally, GFP proteins is tagged for faster degradation than the remaining proteins LuxR and cI. This was done to effectively captures the peak of I1-FFL dynamics. Therefore only for these last two proteins the main degradation component is due to the growth related dilution. Thus, their dynamics can be considered as equivalent. Finally, both plasmids pCB14mut and pCB11a were co-transformed in competent cells (Top10, Invitrogen).

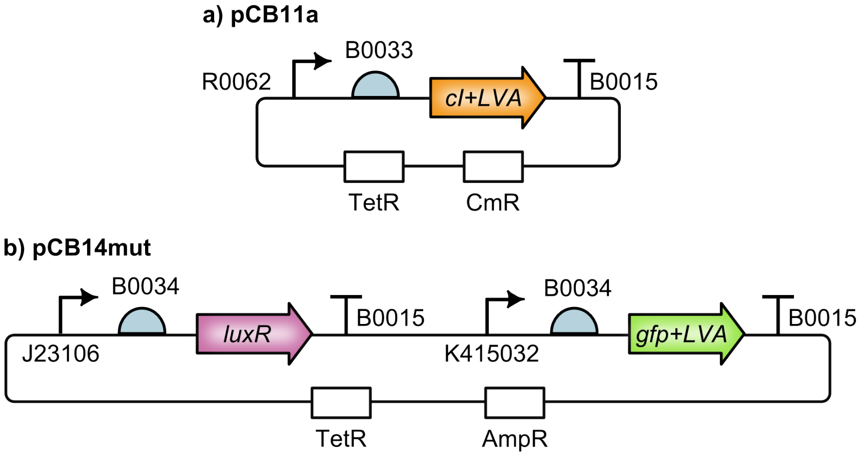


Figure 3.3. Glyph of the I1-FFL gene circuit corresponding to the lab-construction using both plasmids pCB11tc and pCb14mut inside the cell. Further details in annex B.1.

3.2.3 Biochemical reactions

The biochemical reactions considered can be split in two main classes: the *gene expression* reactions, and the *induction* ones.

In the *gene expression* block, the main processes and assumptions considered for each of the three proteins LuxR, cI, and GFP are:

- the binding of the RNA polymerase (RNAP) to each promoter,
- binding of the transcription factors to the genes promoters,
- degradation of mRNA and proteins.

In the *induction* part, the main processes considered are:

- hetero- and homodimerization reactions involving the inducer, like binding between the protein LuxR and AHL to form the monomer, and dimerization of this monomer to form the dimer,
- diffusion of the inducer through the cell membrane,
- binding of the dimer (LuxR.AHL)$_2$ to both *cI* and *gfp* promoters ($\mathrm{P}_{\mathrm{lux}}$ and $\mathrm{P}_{\mathrm{lux/cI}}$, respectively),
- binding between the activator and/or repressor to the *gfp* hybrid promoter ($\mathrm{P}_{\mathrm{lux/cI}}$), and

Table 3.1. Species of the I1-FFL circuit.

	Species	Description	Unit
1	gR	unbound $\mathrm{P_C}$ promoter	nM
2	RNAP	RNA polymerase	nM
3	gR.RNAP	RNAP bound $luxR$	nM
4	mR	$luxR$ messenger RNA	nM
5	R	LuxR protein	nM
6	A	AHL intracellular inducer	nM
7	(R.A)	LuxR and AHL monomer	nM
8	$(R.A)_2$	dimer of (R.A)	nM
9	gI	unbound $\mathrm{P_{lux}}$ promoter	nM
10	gI.$(R.A)_2$	dimer-bound $\mathrm{P_{lux}}$ promoter	nM
11	gG	unbound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
12	gG.$(R.A)_2$	dimer-bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
13	gG.I	cI-bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
14	gG.$(R.A)_2$.I	cI-dimer-bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
15	gI.$(R.A)_2$.RNAP	RNAP-dimer-bound P_{lux} promoter	nM
16	mI	cI messenger RNA	nM
17	I	cI protein	nM
18	mG	gfp messenger RNA	nM
19	G	GFP protein	nM
20	$\mathrm{A_e}$	AHL extracellular inducer	nM
21	gI.RNAP	RNAP bound cI	nM
22	gG.$(R.A)_2$.I.RNAP	bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
23	gG.$(R.A)_2$.RNAP	dimer-RNAP bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
24	gG.RNAP	RNAP bound gfp	nM

- degradation of monomer, dimer and inducer.

All the biochemical species involved in the I1-FFL circuit are listed in Table 3.1. Notice the genes *luxR*, *cI* and *gfp* are denoted as their corresponding unbound promoters.

The resulting set of biochemical reactions has three different subsets (3.1-3.3) for each gene or node *luxR*, *cI* and *gfp*. Species degradation is denoted as $\emptyset$.

gene *luxR*:

$$
\begin{gathered}
\mathsf{r}_1: \quad \mathrm{gR} + \mathrm{RNAP} \xrightleftharpoons[\mathrm{k}_{-1}]{\mathrm{k}_1} \mathrm{gR}\cdot\mathrm{RNAP} \\
\mathrm{gR}\cdot\mathrm{RNAP} \xrightarrow{\mathrm{k_{mR}}} \mathrm{gR} + \mathrm{RNAP} + \mathrm{mR} \\
\mathrm{mR} \xrightarrow{\mathrm{k_{pR}}} \mathrm{mR} + \mathrm{R} \\
\mathrm{mR} \xrightarrow{\mathrm{d_{mR}}} \emptyset \\
\mathrm{R} \xrightarrow{\mathrm{d_R}} \emptyset \\
\mathsf{r}_2: \quad \mathrm{R} + \mathrm{A} \xrightleftharpoons[\mathrm{k{-}2}]{\mathrm{k}_2} (\mathrm{R}\cdot\mathrm{A}) \\
\emptyset \xrightarrow{\mathrm{k_e}} \mathrm{A_e} \\
\mathrm{A_e} \xrightleftharpoons[\mathrm{k_d}]{\mathrm{k_d}} \mathrm{A} \\
\mathrm{A} \xrightarrow{\mathrm{d_A}} \emptyset \\
\mathsf{r}_3: \quad 2\,(\mathrm{R}\cdot\mathrm{A}) \xrightleftharpoons[\mathrm{k}_{-3}]{\mathrm{k}_3} (\mathrm{R}\cdot\mathrm{A})_2 \\
(\mathrm{R}\cdot\mathrm{A}) \xrightarrow{\mathrm{d_{RA}}} \emptyset \\
(\mathrm{R}\cdot\mathrm{A})_2 \xrightarrow{\mathrm{d_{RA2}}} \emptyset
\end{gathered}
\tag{3.1}
$$

gene *cI*:

$$
\begin{gathered}
\mathsf{r}_4: \quad \mathrm{gI} + (\mathrm{R}\cdot\mathrm{A})_2 \xrightleftharpoons[\mathrm{k}_{-4}]{\mathrm{k}_4} \mathrm{gI}\cdot(\mathrm{R}\cdot\mathrm{A})_2 \\
\mathsf{r}_7: \quad \mathrm{gI}\cdot(\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{RNAP} \xrightleftharpoons[\mathrm{k}_{-7}]{\mathrm{k}_7} \mathrm{gI}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{RNAP} \\
\mathrm{gI}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{RNAP} \xrightarrow{\mathrm{k_{mI}}} \mathrm{gI}\cdot(\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{RNAP} + \mathrm{mI} \\
\mathsf{r}_{11}: \quad \mathrm{gI} + \mathrm{RNAP} \xrightleftharpoons[\mathrm{k}_{-11}]{\mathrm{k}_{11}} \mathrm{gI}\cdot\mathrm{RNAP} \\
\mathrm{gI}\cdot\mathrm{RNAP} \xrightarrow{\mathrm{k_{mIb}}} \mathrm{gI} + \mathrm{RNAP} + \mathrm{mI} \\
\mathrm{mI} \xrightarrow{\mathrm{k_{pI}}} \mathrm{mI} + \mathrm{I} \\
\mathrm{mI} \xrightarrow{\mathrm{d_{mI}}} \emptyset \\
\mathrm{I} \xrightarrow{\mathrm{d_I}} \emptyset
\end{gathered}
\tag{3.2}
$$

gene *gfp*:

$$\begin{aligned}
\mathsf{r}_5:\quad & \mathrm{gG} + (\mathrm{R}\cdot\mathrm{A})_2 \xrightleftharpoons[\mathrm{k}_{-5}]{\mathrm{k}_5} \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\\
\mathsf{r}_6:\quad & \mathrm{gG}\cdot\mathrm{I} + (\mathrm{R}\cdot\mathrm{A})_2 \xrightleftharpoons[\mathrm{k}_{-6}]{\mathrm{k}_6} \mathrm{gG}\cdot\mathrm{I}\cdot(\mathrm{R}\cdot\mathrm{A})_2\\
\mathsf{r}_8:\quad & \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{I} \xrightleftharpoons[\mathrm{k}_{-8}]{\mathrm{k}_8} \mathrm{gG}\cdot\mathrm{I}\cdot(\mathrm{R}\cdot\mathrm{A})_2\\
& \mathrm{gG} + \mathrm{I} \xrightleftharpoons[\mathrm{k}_{-9}]{\mathrm{k}_9} \mathrm{g}_\mathrm{G}\cdot\mathrm{I}\\
\mathsf{r}_{10}:\quad & \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{RNAP} \xrightleftharpoons[\mathrm{k}_{-10}]{\mathrm{k}_{10}} \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{RNAP}\\
& \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{RNAP} \xrightarrow{\mathrm{k}_{\mathrm{mG}}} \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{RNAP} + \mathrm{mG}\\
\mathsf{r}_{12}:\quad & \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{I} + \mathrm{RNAP} \xrightleftharpoons[\mathrm{k}_{-12}]{\mathrm{k}_{12}} \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{I}\cdot\mathrm{RNAP}\\
& \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{I}\cdot\mathrm{RNAP} \xrightarrow{\mathrm{kb}_{\mathrm{G1}}} \mathrm{gG}\cdot(\mathrm{R}\cdot\mathrm{A})_2\cdot\mathrm{I} + \mathrm{RNAP} + \mathrm{mG}\\
\mathsf{r}_{13}:\quad & \mathrm{gG} + \mathrm{RNAP} \xrightleftharpoons[\mathrm{k}_{-13}]{\mathrm{k}_{13}} \mathrm{gG}\cdot\mathrm{RNAP}\\
& \mathrm{gG}\cdot\mathrm{RNAP} \xrightarrow{\mathrm{kb}_{\mathrm{G2}}} \mathrm{gG} + \mathrm{RNAP} + \mathrm{mG}\\
& \mathrm{mG} \xrightarrow{\mathrm{k}_{\mathrm{pG}}} \mathrm{mG} + \mathrm{G}\\
& \mathrm{mG} \xrightarrow{\mathrm{d}_{\mathrm{mG}}} \emptyset\\
& \mathrm{G} \xrightarrow{\mathrm{d}_{\mathrm{G}}} \emptyset
\end{aligned}\tag{3.3}$$

Finally, reactions (3.1-3.3) were obtained under the following assumptions:

1. during transcription, the cell contains enough free RNAP to serve all the active genes transcribing at a given moment. In this way, it is assumed that the free RNAP concentration in the cell will not appreciably change in time,

2. each basal expression (i.e. transcription even in saturating presence of the repressor) of the promoters $\mathrm{P}_{\mathrm{lux}}$ and $\mathrm{P}_{\mathrm{lux/cI}}$ is a nonzero minimal expression level,

3. Translation is not a simple process (Alberts et al., 2009). It was modeled as an irreversible reaction with an average transcription rate accounting for the fact that binding of ribosomes to the ribosome binding site (RBS) is indeed reversible, and several ribosomes may translate a single messenger RNA copy (mRNA) simultaneously,

4. Transcription of genes *luxR*, *cI*, and *gfp* is irreversible, so that $\mathrm{k}_{\mathrm{m_R}}$, $\mathrm{k}_{\mathrm{m_I}}$ and $\mathrm{k}_{\mathrm{m_G}}$ are the effective transcription rates of *luxR*, *cI*, and *gfp* respectively,

5. gR, gI, and gG are the plasmid copy numbers of genes *luxR*, *cI*, and *gfp* respectively,

6. LuxR and AHL binding is a fast and reversible reaction,

7. Dimerization of $(\text{LuxR.AHL})_2$ or $(\text{R.A})_2$ is a reversible reaction,

8. Interactions between AHL and $\mathrm{AHL}_{\mathrm{ext}}$ represent the physical passive diffusion process for cell-to-cell communication via quorum sensing, so that was modeled as a reversible pseudo-reaction,

9. $\mathrm{V_c} = \mathrm{V_{cell}}/\mathrm{V_{ext}}$ is the ratio between the cellular and the environment volumes to quantify the AHL/$\mathrm{AHL}_{\mathrm{ext}}$ effect,

10. The monomer (LuxR.AHL) or (R.A) only degrades at each cell cycle[1], and

11. Messengers RNA, proteins and transcription factor degradation are irreversible reactions (denoted as $\emptyset$).

3.3 Quorum sensing/Feedback circuit

In section 2.4 it was commented that noise is pervasive in the cellular mechanisms underlying gene expression (Raser and O'Shea, 2005). It propagates to downstream genes at the single-cell level, and eventually causes variation within an isogenic population (Raj and van Oudenaarden, 2008; Labhsetwar et al., 2013) that may determine the fate of individual cells and that of a whole population (Eldar and Elowitz, 2010b; Labhsetwar et al., 2013).

At the gene level, noise can be traced back to intrinsic sources due to stochastic fluctuations in transcription and translation mechanisms, and extrinsic ones corresponding to gene independent fluctuations in protein expression due to external factors (Eldar and Elowitz, 2010b; Chalancon et al., 2012; Jones et al., 2014). To minimize the deleterious effects of noise, cells have evolved different strategies at the single-cell level: from different transcription and translation efficiency so as to reduce translation burst rates in key genes (Kærn et al., 2005) to more elaborated strategies, such as negative feedback regulation to reduce noise by shifting the noise spectrum to a higher frequency region (Raser and O'Shea, 2005). Yet, cells live in communities, forming a population. At this level, extracellular signaling propagates intracellular stochastic fluctuations across the population (Tabbaa et al., 2013). Thus, cells have adapted their communication mechanisms in order to improve the signal-to-noise ratio (Weber and Buceta, 2013). One of such communication mechanisms is quorum sensing.

Quorum sensing (QS), initially discovered in *V. fisheri* and *P. putida*, is a cell-to-cell communication mechanism whereby bacteria exchange chemical signaling molecules, called autoinducers, whose external concentration depends on the cell population den-

[1] Cell doubling time or cell cycle is the period of time required for a quantity to double in size or value (see section 2.1.1)

sity. It is known that synchronization and consensus protect from noise (Tabareau et al., 2010). Cells detect a threshold concentration of QS autoinducers and alter gene expression accordingly (Fuqua et al., 2001), driving the population as a whole to achieve a desired consensus gene expression level despite the individual noise of each member of the population. Cells consensus induced by QS is thought to reduce extrinsic noise by reducing the transmission of fluctuating signals in the low-frequency domain (Tanouchi et al., 2008), enhances intrinsic stochastic fluctuations (Tabbaa et al., 2013), and allows entrainment of a noisy population when faced to environmental changing signals (Nelson et al., 2013). Therefore QS seems an effective tool to control the phenotypic variability in a population of cells (Weber and Buceta, 2013).

Phenotypic variability has important practical relevance in many applications in the areas of biomedicine, biotechnology and other branches of biological science (Geiler-Samerotte et al., 2013) as the presence of heterogeneous subpopulations may have significant impact on the yield and productivity of industrial cultures (Müller et al., 2010; Fernandes et al., 2011; Carlquist et al., 2012). Thus, improving homogeneity of protein expression in industrial cultures is a goal of economic relevance for microbial cell factory processes that has traditionally been attempted either by optimizing environmental conditions in the culture or by careful selection of the strain. Open loop strategies based on sensitivity analysis have been used to provide guides as to how properly tune transcriptional and translational parameters so that the noise levels can be controlled while the mean values can be simultaneously adjusted to desired values (Kim and Sauro, 2012). While sensitivity analysis gives very valuable insights, open loop control is not robust against system uncertainty and/or variations. There is an ever-growing appreciation that biological complexity requires new bioprocess design principles.

Synthetic biology, sometimes defined as the engineering of biology, has the potential to engineer genetic circuits to perform new functions for useful purposes in a systematic, predictable, robust, and efficient way (Way et al., 2014a). In the last years, several synthetic circuits have been proposed with the ultimate goal of dealing with gene expression noise (Zechner et al., 2016; Zhang et al., 2016). Though circuits using **Negative feedback (Fb)** have been proved to decrease gene expression noise (Dublanche et al., 2006), single-cell intracellular feedback loops do not take into account that in practice one is interested in controlling gene expression mean value and noise across a population of cells. In other words, having a feedback loop inside one cell helps to reduce noise in that cell. However for a cell population, the effect of these individual feedback loops can be improved by incorporating information from the remaining cells of the population.

Feedback across a population of cells can be implemented by means of quorum sensing-based strategies, and has been shown to reduce noise effects (Tanouchi et al., 2008; Weber and Buceta, 2011, 2013). Indeed, cell-to-cell communication by means of quorum sensing induces consensus among cells (Russo and Slotine, 2010), that is, contributes to reduce the difference of internal state among cells in a population.

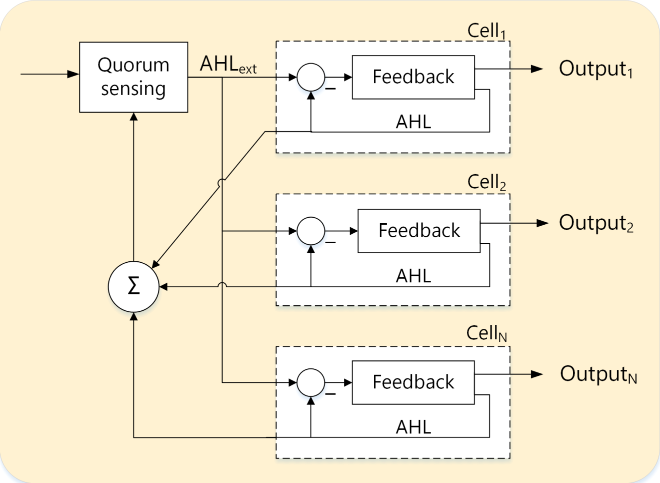


Figure 3.4. QS/Fb circuit. It combines the quorum sensing (QS) subsystem, and the negative feedback (Fb) subsystem to control noise in a protein of interest. The QS part achieves consensus across N cells during protein expression. The Fb part regulates the same protein production inside every cell.

This, in turn, may contribute to protect from noise (Tabareau et al., 2010). Thus, the idea of joining both *intracellular negative feedback and extracellular feedback via quorum sensing* is a natural one, that has been suggested in (Vignoni et al., 2013b; Zargar et al., 2016). Together quorum sensing and the intracellular feedback loop make the Quorum senging/Feedback (QS/Fb) circuit, which is illustrated in the diagram of Fig.3.4.

3.3.1 QS/Fb circuit structure

Figure 3.5a depicts the QS/Fb circuit to reduce gene expression noise while achieving a desired mean expression level in a protein of interest (PoI). It couples two functional subsystems already implemented in *E. coli*(Vignoni et al., 2013b): *i)* quorum sensing interconnecting cell population with the signal inducer AHL, and *ii)* a feedback loop regulating expression of the PoI protein with another two proteins LuxR and LuxI.

The first subsystem implements a cell-to-cell communication mechanism via quorum sensing (QS). It is based on the exchange of the small signaling autoinducer molecule N-acyl-L-homoserine lactone (AHL) (Kaplan and Greenberg, 1985; Fuqua et al., 2001) to induce population cell consensus. AHL molecules passively diffuses across the cellular membrane from inside the cell to the external environment, and viceversa (is was also described for the previous circuit). Intracellular AHL is synthesized by the LuxI

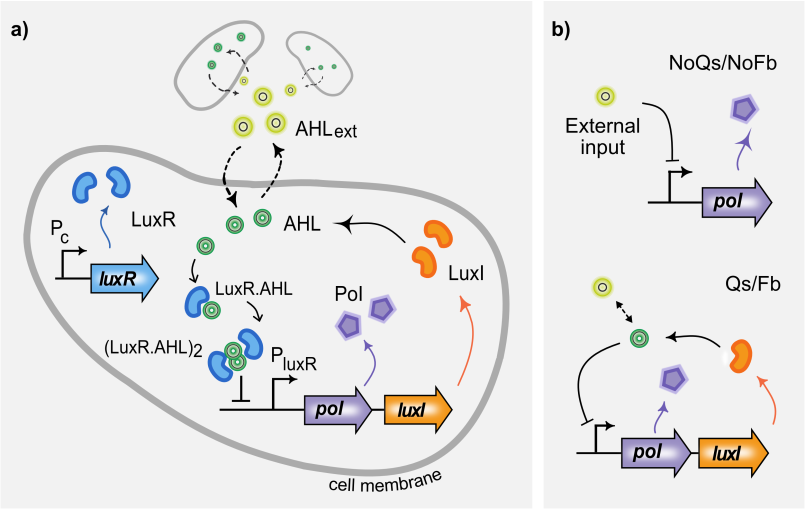


Figure 3.5. QS/Fb circuit. **a)** Representation of a *E. coli* cell incorporating the engineered QS/Fb synthetic circuit including cell-to-cell communication system based on quorum sensing and an intracellular negative feedback loop. **b)** NoQS/NoFb gene circuit to asses the roles of both quorum sensing and feedback.

protein previously expressed by an homolog of the *luxI* gene from *V. fisheri* (Schaefer et al., 1996).

The second subsystem introduces an intracellular negative feedback loop (Fb) to control expression of the protein of interest Pol using two genes *luxR* and *pol/luxI*. First, LuxR protein is expressed by *luxR* under the constitutive promoter $\mathrm{P_c}$. Then, LuxR and AHL bind forming the heterodimer (LuxR.AHL), which subsequently dimerizes as the heterotetramer transcription factor (LuxR.AHL)$_2$. The dimer represses co-expression (simultaneous gene expression) of Pol and LuxI, when it is attached to the synthetic repressible promoter $\mathrm{P_{luxR}}$ designed in (Egland and Greenberg, 2000). In other words, the production excess of protein Pol is regulated by (LuxR.AHL)$_2$, when it accumulates and eventually attaining the repression threshold for the gene *pol/luxI* promoter.

To assess the role played by feedback and QS, the QS/Fb circuit illustrated in Fig.3.5a was compared with another one that has only constitutive expression (**NoQS/NoFb** in Fig.3.5b), and with a circuit with feedback but no QS (**NoQS/Fb**).

3.3.2 QS/Fb circuit construction

Following the same implementation presented in 3.2.2, the QS/Fb and NoQS/NoFb circuits were implemented using components taken from the Lux operon in the *V. fisheri* (Kaplan and Greenberg, 1985) quorum sensing system, and a green fluorescent protein (GFP) as a reporter. As in section 2.1.2, all the bioparts from both circuits were inserted into different plasmids, and taken from the iGEM Registry of Standard Biological Parts. Finally, they were cloned using the Biobrick's foundation 3 Antibiotic Assembly method, and confirmed by sequencing. Both the QS/Fb and the NoQS/NoFb gene circuits were implemented in *E. coli*containing two different plasmids for each construction, as was detailed in annexes B.2 and B.3.

The QS/Fb circuit couples two subsystems: *i)* QS-based cell-to-cell communication mechanism, and *ii)* a negative feedback loop. So, this is an auto-regulated gene circuit, whose output is the protein of interest PoI co-expressed together with protein LuxI. The QS/Fb system was split in two subunits integrated in different plasmids.

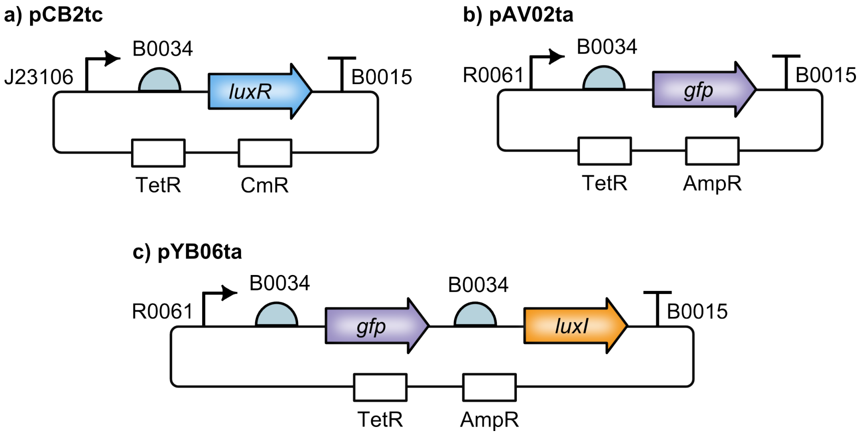


Figure 3.6. Glyph of the QS/Fb and NoQS/NoFb gene circuits built in the lab. Plasmids a) and c) correspond to the QS/Fb gene circuit transformed in every cell. Similarly, plasmids a) and b) represent the NoQS/NoFb circuit inside the cell. Further details in B.2.

On the one hand, plasmid **pCB2tc** (Fig.3.6) contains the gene *luxR* (part BBa_C0062) coding for protein LuxR that is constitutively expressed under the control of a medium strength promoter $\mathrm{P_c}$ (part BBa_J23106), and a strong RBS (part BBa_B0034). This insert was cloned into the pACYC184 plasmid cloning vector (p15A origin, chloramphenicol/tetracycline).

Then, plasmid **pYB06ta** (Fig.3.6) comprises of gene *luxI* (part BBa_C0161) under control of the repressible promoter $\mathrm{P_{luxR}}$ (part BBa_R0062), and a strong RBS (part BBa_B0034). Another strong RBS (BBa_B0034) and the green fluorescent protein GFP (part BBa_E0040) were inserted using GIBSON assembly (NEB Catalog Number

E2611S) upstream of *luxI*, right after the $\mathrm{P}_{\mathrm{luxR}}$ promoter. In this way, GFP represents the protein of interest PoI (Fig.3.5), and it is co-expressed with LuxI to also reports its dynamics during experiments. These constructions were inserted into the pBR322 plasmid cloning vector (pMB1 origin, ampicillin/tetracycline).

On the other hand, the control circuit NoQS/NoFb was also implemented for removing both QS and the feedback loop subsystems. To this end, the plasmid pCB2tc above was also co-transformed with the plasmid **pAV02ta** (see Fig.3.6) that contains only GFP downstream of the $\mathrm{P}_{\mathrm{luxR}}$ repressible promoter (part BBa_R0062), and the same previous strong RBS (part BBa_B0034). pAV02ta was cloned in the pBR322 plasmid cloning vector (pMB1 origin, ampicillin/tetracycline). As a consequence, the only structural difference between both circuits is the gene *luxI* downstream in the QS/Fb system.

None of the proteins are tagged for degradation (an additional coding sequence for fast protein turnover), therefore the main degradation component is due to the growth related dilution for both proteins. Thus, their dynamics can be considered as equivalent. Finally, both plasmids pCB2tc and pYB06ta were co-transformed in competent cells (DH-5α, Invitrogen).

3.3.3 Biochemical reactions

For the QS/Fb circuit, the main biochemical reactions can be split in two main classes: the *gene expression* reactions, and the *induction* ones (similar to the I1-FFL circuit).

For the *gene expression* block, the key processes and assumptions considered for the genes *luxR* and *poI/luxI* are:

- the binding of RNA polymerase (RNAP) to each promoter is a fast reaction, therefore it has not been evaluated, even though it was considered in 3.2.3 of the I1-FFL circuit,
- binding of the transcription factors to the genes promoters,
- degradation of mRNA, transcription factors and proteins.

In the *induction* part using the autoinducer AHL, the main processes considered are:

- hetero- and homodimerization reactions involving the inducer, like binding between the protein LuxR and AHL to form the monomer, and dimerization of this monomer to form the dimer,
- diffusion of the inducer through the cell membrane,
- binding of the dimer to the $\mathrm{P}_{\mathrm{luxR}}$ promoter,
- degradation of monomer, dimer and inducer.

The biochemical species involved in the *gene expression* and the *induction* parts are shown in the Table 3.2.

genes *luxR* and *pol/luxI*:

$$
\begin{aligned}
&\xrightarrow{C_R} mR \\
gPI &\xrightarrow{k_{e_I}} gPI + mPI \\
mR &\xrightarrow{p_R} mR + R \\
mPI &\xrightarrow{p_I} mPI + PI \\
I &\xrightarrow{k_A} A + I \\
R + A &\xrightleftharpoons[k_{-1}]{k_{-1}/k_{d1}} (R \cdot A) \\
{}_2(R \cdot A) &\xrightleftharpoons[k_{-2}]{k_{-2}/k_{d2}} (R \cdot A)_2 \\
gPI + (R \cdot A)_2 &\xrightleftharpoons[k_{lux}]{k_{lux}/k_{dlux}} gPI \cdot (R \cdot A)_2 \\
gPI \cdot (R \cdot A)_2 &\xrightarrow{\alpha k_{e_I}} gPI \cdot (R \cdot A)_2 + mPI \\
A &\xrightleftharpoons[DV_c]{D} A_{ext} \\
mPI &\xrightarrow{dm_I} \emptyset \\
mR &\xrightarrow{dm_R} \emptyset \\
PI &\xrightarrow{d_I} \emptyset \\
R &\xrightarrow{d_R} \emptyset \\
A &\xrightarrow{d_A} \emptyset \\
A_{ext} &\xrightarrow{d_{A_e}} \emptyset \\
(R \cdot A) &\xrightarrow{d_{RA}} \emptyset \\
(R \cdot A)_2 &\xrightarrow{d_{RA_2}} \emptyset
\end{aligned}
\tag{3.4}
$$

The species react following the set of biochemical reactions (3.4) under the following assumptions:

1. During transcription, the cell contains enough free RNAP to serve all the active genes transcribing at a given moment. In this way, it is assumed that free RNAP concentration in the cell will not appreciably change in time,

Table 3.2. Species of the QS/Fb circuit.

	Species	Description	Unit
1	mPI	*poI/luxI* messenger RNA	molecules
2	mR	*luxR* messenger RNA	molecules
3	PI	co-expression of proteins PoI/LuxI	molecules
4	R	LuxR protein	molecules
5	(R.A)	LuxR and AHL monomer	molecules
6	(R.A)$_2$	dimer of (R.A)	molecules
7	gPI	unbound $\mathrm{P_{luxR}}$ promoter gene $poI/luxI$	molecules
8	gPI.(R.A)$_2$	bound $\mathrm{P_{luxR}}$ promoter of gene $poI/luxI$	molecules
9	A	AHL intracellular inducer	molecules
10	$\mathrm{A_{ext}}$	AHL extracellular inducer	molecules

2. Basal expression or *leakiness* (i.e. transcription even in saturating presence of the repressor) of the repressible promoter $\mathrm{P_{luxR}}$ is a nonzero minimal expression level,

3. Transcription of genes *poI/luxI* and *luxR* is irreversible, so that $\mathrm{k_{e_I}}$ and $\mathrm{C_R}$ are the effective transcription rates of *poI/luxI* and *luxR* respectively,

4. α is the basal expression (leakage) of *poI/luxI*,

5. $\mathrm{C_R}$ is the plasmid copy number times the effective constitutive transcription rate of *luxR*,

6. Translation is not a simple process (Alberts et al., 2009). It was modeled an irreversible reaction with an average transcription rate accounting for the fact that binding of ribosomes to the ribosome binding site (RBS) is indeed reversible, and several ribosomes may translate a single messenger RNA copy (mRNA) simultaneously,

7. LuxR and AHL binding is a fast and reversible reaction,

8. Dimerization of (LuxR.AHL)$_2$ is a reversible reaction,

9. Interactions between AHL and $\mathrm{AHL_{ext}}$ represent the physical passive diffusion process for cell-to-cell communication via quorum sensing, so that it is a reversible pseudo-reaction

10. $\mathrm{V_c = V_{cell}/V_{ext}}$ is the ratio between the cellular and the environment volumes to quantify the AHL/$\mathrm{AHL_{ext}}$ effect,

11. Monomer (LuxR.AHL) only degrades at each cell cycle, and

12. Messengers RNA, proteins and transcription factor degradation are irreversible reactions (denoted as $\emptyset$).

3.4 Summary

Two synthetic gene circuits of different nature and with different goals and inherent problems will be used in this Thesis: (1) an Incoherent type 1 feedforward circuit (I1-FFL) that exhibits the important biological property of *adaptation*, and (2) a Quorum sensing/Feedback circuit (QS/Fb) comprising two intertwined feedback loops –an intracellular one and a cell-to-cell communication-based one– designed to regulate the mean expression level of a protein of interest, while minimizing its variance across the population of cells. Both circuits will be analyzed *in silico* as we will see in Chapter 4.

Chapter 4

Modeling and simulation

4.1 Introduction

This Chapter deals with the deterministic and stochastic modeling and simulation of two synthetic gene circuits used throughout the Thesis: the I1-FFL and the QS/Fb circuits already described in Chapter 3.

The choice of a modeling framework (i.e. deterministic or stochastic) is determined by the complexity of the system being modeled, the level of investigation, and, consequently, the question being asked (R Dougherty and L Bittner, 2010). The answer is different for each of the two circuits analyzed in the Thesis.

In the first case, we are concerned about the I1-FFL circuit average behavior. Thus, a **deterministic modeling** framework will be used. In the second case, the Thesis focuses in the QS/Fb circuit capacity to deal with variability due to intrinsic and extrinsic noise. Therefore, a **stochastic modeling** approach is used.

One of goals of this Chapter is to obtain reduced-order models more amenable for computational analysis, but avoiding excessive reduction that would lead to lack of biological relevance. Model reduction comes at a cost, for identification of the parameters in the reduced order model may only valid around some operating region. A multi-objective identification methodology is propose to tackle with this problem (see Chapter 5). The results of this Chapter have been published in

- Y. Boada, A. Vignoni, D. Oyarzún, and J. Picó. Host-circuit interactions explain unexpected behaviours of a feedforward gene circuit. 2018. Foundations of Systems Biology in Engineering FOSBE,

- Y. Boada, A. Vignoni, and J. Picó. Engineered control of genetic variability reveals interplay among quorum sensing, feedback regulation, and biochemical noise. *ACS Synthetic Biology*, 6(10):1903–1912, 2017a. doi: 10.1021/acssynbio.7b00087,

- Y. Boada, A. Vignoni, and J. Picó. Model reduction and multi-objective identification of a feedback synthetic gene circuit. *IEEE Transactions on Control Systems Technology*, which is accepted.

Additionally, since stochastic simulations demand high speed computers with hundreds of megabytes for storing data, an efficient computational framework for stochastic simulation was described in

- Y. Boada, A. Vignoni, J. L. Navarro, and J. Picó. Improvement of a cle stochastic simulation of gene synthetic network with quorum sensing and feedback in a cell population. In *2015 European Control Conference (ECC)*, pages 2274–2279, 2015.

The Chapter is organized as follows. Sections 4.2 and 4.3 describe the deterministic modeling and the model reduction approach for both gene circuits: the I1-FFL and the QS/Fb. Section 4.4 describes stochastic CLE-based approach to model gene circuit stochasticity for a cell population. Aspects related to stochastic modeling like efficient simulation, data storage, and modeling functions are introduced in this section. Finally in section 4.5, the main conclusions are drawn.

4.2 The I1-FFL gene circuit

As I said before, an average behaviour is considered for the I1-FFL circuit. therefore, a deterministic model will be sought after. The deterministic approach starts from the corresponding sets of biochemical reactions for the I1-FFL gene circuit. Then, dynamic balances for the species of the I1-FFL circuit were obtained using the mass-action kinetics formalism described in section 2.3. The resulting ODE model is a high dimensional one and depends on several parameters that need to be estimated before the model can be used. It will be reduced in subsection 4.2.2.

4.2.1 Circuit model

For the I1-FFL ODE model, the key regulatory interactions between the concentrations of the main biochemical species presented in section 3.2 were taking into account. Its main species are proteins LuxR, cI, and GFP, and inducer AHL. The inducer AHL is the input of the system, and the protein GFP is the output signal.

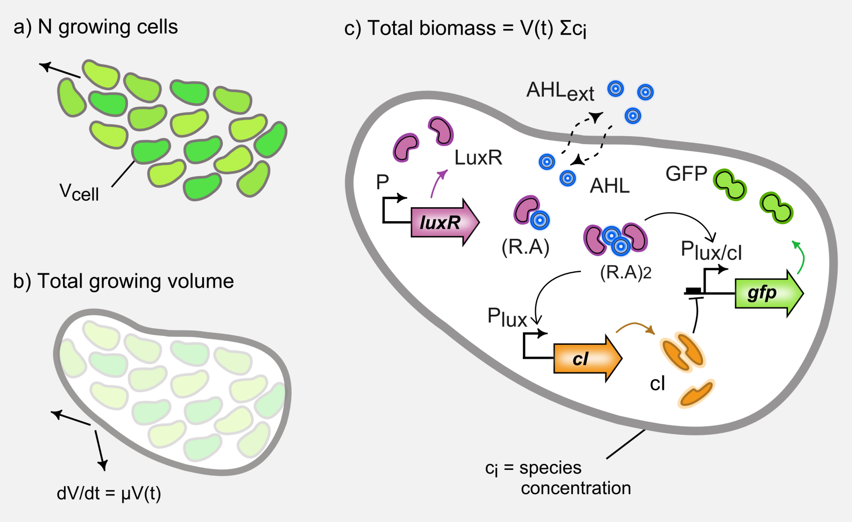


Figure 4.1. I1-FFL model. a) Population of N growing cells with different sizes. **b)** Total volume V(t) of a growing population of cells. **c)** Total biomass and concentration c_i of species i inside V(t).

Recall the reaction sets (3.1) - (3.3) summarize the regulatory interactions for the AHL with genes *luxR*, *cI* and *gfp* respectively. In turn, these genes produce their corresponding proteins LuxR, cI, and GFP. Again, the I1-FFL circuit (Fig.4.1c) comprises a gene *gfp* under the control of the hybrid promoter $P_{\mathrm{lux/cI}}$. Expression of GFP is activated by the dimer (LuxR.AHL)$_2$ or (R.A)$_2$ that acts as transcription factor for the hybrid promoter, and repressed by protein cI. The dimer (LuxR.AHL)$_2$ also acts as transcription factor activating the promoter $\mathrm{P_{lux}}$. Protein LuxR is constitutively expressed, and bounds to the inducer AHL. The inducer can passively diffuse across the cell membrane. Though the input signal to the circuit is the intracellular inducer concentration AHL, the experimental input signal is the external application of the inducer in the broth $\mathrm{AHL_{ext}}$.

Using the law of mass-action kinetics (described in section 2.3), reactions (3.1) - (3.3) can be used to formulate the corresponding ODE equations for the all species concentrations in the I1-FFL circuit. These equations can be derived either by inspection, or using specific software to automate the process. For example, software packages like Facile (Siso-Nadal et al., 2007), BioNetGen (Blinov et al., 2004) or COPASI (Mendes et al., 2009) allow us to obtain the dynamic kinetic model from either the set of reactions or from SBML files encoding them. Here, the ODE model (4.1) was obtained by inspection.

$$
\begin{aligned}
\dot{x}_1 &= -\mathrm{k}_1\mathrm{c}_{\mathrm{RNAP}}x_1 + (\mathrm{k}_{-1} + \mathrm{k}_{\mathrm{mR}})x_3 - \mu x_1 + \mu(x_1 + x_3)\\
\dot{x}_2 &= 0\\
\dot{x}_3 &= \mathrm{k}_1\mathrm{c}_{\mathrm{RNAP}}x_1 - (\mathrm{k}_{-1} + \mathrm{k}_{\mathrm{mR}})x_3 - \mu x_3\\
\dot{x}_4 &= \mathrm{k}_{\mathrm{mR}}x_3 - (\mathrm{d}_{\mathrm{mR}} + \mu)x_4\\
\dot{x}_5 &= \mathrm{k}_{\mathrm{pR}}x_4 - (\mathrm{d}_{\mathrm{R}} + \mu)x_5 - \mathrm{k}_2x_5x_6 + \mathrm{k}_{-2}x_7\\
\dot{x}_6 &= -\mathrm{k}_2x_5x_6 + \mathrm{k}_{-2}x_7 + \mathrm{k}_{\mathrm{d}}(x_{20} - x_6) - (\mathrm{d}_{\mathrm{A}} + \mu)x_6\\
\dot{x}_7 &= \mathrm{k}_2x_5x_6 - \mathrm{k}_{-2}x_7 + 2\mathrm{k}_{-3}x_8 - 2\mathrm{k}_3x_7^2 - (\mathrm{d}_{\mathrm{RA}} + \mu)x_7\\
\dot{x}_8 &= -\mathrm{k}_{-3}x_8 + \mathrm{k}_3x_7^2 - \mathrm{k}_4x_8x_9 + \mathrm{k}_{-4}x_{10} - \mathrm{k}_5x_8x_{11} + \mathrm{k}_{-5}x_{12} - \mathrm{k}_6x_8x_{13}\\
&\quad + \mathrm{k}_{-6}x_{14} - (\mathrm{d}_{\mathrm{RA2}} + \mu)x_8\\
\dot{x}_9 &= -\mathrm{k}_4x_8x_9 + \mathrm{k}_{-4}x_{10} - \mathrm{k}_{11}\mathrm{c}_{\mathrm{RNAP}}x_9 + (\mathrm{k}_{-11} + \mathrm{k}_{\mathrm{mIb}})x_{21} - \mu x_9\\
&\quad + \mu(x_9 + x_{10} + x_{15} + x_{21})\\
\dot{x}_{10} &= \mathrm{k}_4x_8x_9 - \mathrm{k}_{-4}x_{10} - \mathrm{k}_7\mathrm{c}_{\mathrm{RNAP}}x_{10} + (\mathrm{k}_{-7} + \mathrm{k}_{\mathrm{mI}})x_{15} - \mu x_{10}\\
\dot{x}_{11} &= -\mathrm{k}_9x_{11}x_{17} + \mathrm{k}_{-9}x_{13} - \mathrm{k}_5x_{11}x_8 + \mathrm{k}_{-5}x_{12} - \mathrm{k}_{13}\mathrm{c}_{\mathrm{RNAP}}x_{11}\\
&\quad + (\mathrm{k}_{-13} + \mathrm{k}_{\mathrm{mGb}_2})x_{24} - \mu x_{11} + \mu(x_{11} + x_{12} + x_{13} + x_{14} + x_{22} + x_{23} + x_{24})\\
\dot{x}_{12} &= \mathrm{k}_5\mathrm{x}_{11}x_8 - \mathrm{k}_{-5}x_{12} - \mathrm{k}_8x_{12}x_{17} + \mathrm{k}_{-8}x_{14} - \mathrm{k}_{10}\mathrm{c}_{\mathrm{RNAP}}x_{12}\\
&\quad + (\mathrm{k}_{-10} + \mathrm{k}_{\mathrm{mG}})x_{23} - \mu x_{12}\\
\dot{x}_{13} &= \mathrm{k}_9x_{11}x_{17} - \mathrm{k}_{-9}x_{13} - \mathrm{k}_6x_{13}x_8 + \mathrm{k}_{-6}x_{14} - \mu x_{13}\\
\dot{x}_{14} &= \mathrm{k}_6x_{13}x_8 - \mathrm{k}_{-6}x_{14} + \mathrm{k}_8x_{12}x_{17} - \mathrm{k}_{-8}x_{14} - \mathrm{k}_{12}\mathrm{c}_{\mathrm{RNAP}}x_{14}\\
&\quad + (\mathrm{k}_{-12} + \mathrm{k}_{\mathrm{mGb}_1})x_{22} - \mu x_{14}\\
\dot{x}_{15} &= \mathrm{k}_7\mathrm{c}_{\mathrm{RNAP}}x_{10} - (\mathrm{k}_{-7} + \mathrm{k}_{\mathrm{mI}})x_{15} - \mu x_{15}\\
\dot{x}_{16} &= \mathrm{k}_{\mathrm{mI_b}}x_{21} + \mathrm{k}_{\mathrm{mI}}x_{15} - (\mathrm{d}_{\mathrm{mI}} + \mu)x_{16}\\
\dot{x}_{17} &= \mathrm{k}_{\mathrm{pI}}x_{16} - \mathrm{k}_9x_{11}x_{17} + \mathrm{k}_{-9}x_{13} - \mathrm{k}_8x_{12}x_{17} + \mathrm{k}_{-8}x_{14} - (\mathrm{d}_{\mathrm{I}} + \mu)x_{17}\\
\dot{x}_{18} &= \mathrm{k}_{\mathrm{mG}}x_{23} + \mathrm{k}_{\mathrm{mGb}_1}x_{22} + \mathrm{k}_{\mathrm{mGb}_2}x_{24} - (\mathrm{d}_{\mathrm{mG}} + \mu)x_{18}\\
\dot{x}_{19} &= \mathrm{k}_{\mathrm{pG}}x_{18} - (\mathrm{d}_{\mathrm{G}} + \mu)x_{19}\\
\dot{x}_{20} &= -\mathrm{k}_{\mathrm{d}}\frac{\mathrm{NV}_{\mathrm{cell}}}{\mathrm{V}_{\mathrm{ext}}}(x_{20} - x_6) - \mathrm{d}_{\mathrm{Ae}}x_{20}\\
\dot{x}_{21} &= \mathrm{k}_{11}\mathrm{c}_{\mathrm{RNAP}}x_9 - (\mathrm{k}_{-11} + \mathrm{k}_{\mathrm{mIb}} + \mu)x_{21}\\
\dot{x}_{22} &= \mathrm{k}_{12}\mathrm{c}_{\mathrm{RNAP}}x_{14} - (\mathrm{k}_{-12} + \mathrm{k}_{\mathrm{mGb}_1} + \mu)x_{22}\\
\dot{x}_{23} &= \mathrm{k}_{10}\mathrm{c}_{\mathrm{RNAP}}x_{12} - (\mathrm{k}_{-10} + \mathrm{k}_{\mathrm{mG}} + \mu)x_{23}\\
\dot{x}_{24} &= \mathrm{k}_{13}\mathrm{c}_{\mathrm{RNAP}}x_{11} - (\mathrm{k}_{-13} + \mathrm{k}_{\mathrm{mGb}_2} + \mu)x_{24}
\end{aligned}
\tag{4.1}
$$

where $x_{20}(0) = \mathrm{k}_{\mathrm{e}}$ is the initial concentration of extracellular inducer, and $x_2 = \mathrm{c}_{\mathrm{RNAP}}$ is the free RNA polymerase (RNAP). This one is assumed to be large enough so its fast time-varying fluctuations due to its use and release in the cell reactions can be neglected, and only its slow time-varying average amount is taken into account. Thus,

x_2 can be considered as a fixed parameter. The remaining variables are listed in Table 4.1. The parameters correspond to the reaction rates in (3.1), (3.2), and (3.3).

Table 4.1. Variables of the I1-FFL circuit.

Variable	Species	Description	Unit
x_1	gR	unbound $\mathrm{P_C}$ promoter	nM
x_2	RNAP	RNA polymerase	nM
x_3	gR.RNAP	RNAP bound $luxR$	nM
x_4	mR	$luxR$ messenger RNA	nM
x_5	R	LuxR protein	nM
x_6	A	AHL intracellular inducer	nM
x_7	(R.A)	LuxR and AHL monomer	nM
x_8	$(\mathrm{R.A})_2$	dimer of (R.A)	nM
x_9	gI	unbound $\mathrm{P_{lux}}$ promoter	nM
x_{10}	$\mathrm{gI.(R.A)_2}$	dimer-bound $\mathrm{P_{lux}}$ promoter	nM
x_{11}	gG	unbound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
x_{12}	$\mathrm{gG.(R.A)_2}$	dimer-bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
x_{13}	gG.I	cI-bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
x_{14}	$\mathrm{gG.(R.A)_2.I}$	cI-dimer-bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
x_{15}	$\mathrm{gI.(R.A)_2.RNAP}$	RNAP-dimer-bound P_{lux} promoter	nM
x_{16}	mI	cI messenger RNA	nM
x_{17}	I	cI protein	nM
x_{18}	mG	gfp messenger RNA	nM
x_{19}	G	GFP protein	nM
x_{20}	$\mathrm{A_e}$	AHL extracellular inducer	nM
x_{21}	gI.RNAP	RNAP bound cI	nM
x_{22}	$\mathrm{gG.(R.A)_2.I.RNAP}$	bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
x_{23}	$\mathrm{gG.(R.A)_2.RNAP}$	dimer-RNAP bound $\mathrm{P_{lux/cI}}$ hybrid promoter	nM
x_{24}	gG.RNAP	RNAP bound gfp	nM

The full model (4.1) takes into account the dilution effect caused by cells growth, where a specific growth rate μ (min^{-1}) has been assumed. The number of copies of the genes *luxR*, *cI*, and *gfp* is considered to keep constant through time. This is indeed the case if the genes are chromosomal ones. In case the genes are located in *plasmids* (as it was saw in section 2.1.1), one can assume that at each cell division, *plasmids* are first duplicated, and then half of them will be inherited by each of the offspring cells. This is a valid approximation for the model (4.1) that is a model for an **average cell** (see Fig.4.1a). These kind of models do not distinguish individual cells but lump them into an aggregate volume (as in Fig.4.1b), which expands along time (De Jong et al., 2017). To account for this, the dilution terms due to cell growth were compensated in the balance expressions corresponding to the three genes (x_1, x_9, and x_{11}). Notice also that extracellular species, like the external inducer $\mathrm{AHL_{ext}}$ are not subject to dilution by growth rate. On the other hand, the external inducer is introduced as a *bolus* injection. That is, at time $t = 0$ an amount of $\mathrm{AHL_{ext}}$ is injected in the culture, so that its concentration in the culture equals $\mathrm{k_e}$. This value of external inducer concentration is taken as the initial condition for its corresponding dynamic balance (with $x_{20}(0) = \mathrm{k_e}$).

For the output protein GFP and its messenger mRNA (x_{19} and x_{18}, respectively), binding between the activating dimer (LuxR.AHL)$_2$ and the *gfp* hybrid promoter ($\mathrm{P}_{\mathrm{lux/cI}}$) is always possible, even if the repressor cI is already bound to the promoter, and vice-versa. Yet, it was considered that whenever the repressor cI is bound to the promoter $\mathrm{P}_{\mathrm{lux/cI}}$, the RNA polymerase cannot bind the promoter. Thus, the cI repressor x_{17} and the dimer concentration x_8 include the corresponding balance terms to model this dynamics, that is the corresponding terms for the species from x_{10} to x_{14}.

For the $\mathrm{AHL}_{\mathrm{ext}}$ extracellular input signal x_{20}, though its diffusion has been expressed as a biochemical reaction, it is not at all, but a physical process modeled using a lumped approximation of the Fick's law (Alberts et al., 2009; Weiss, 1996). The $\mathrm{AHL}_{\mathrm{ext}}$ concentration is measured with respect to the total volume occupied by the cells $\mathrm{NV}_{\mathrm{cell}}$ and the liquid medium $\mathrm{V}_{\mathrm{ext}}$, where there is a population of N cells. In practice $\mathrm{NV}_{\mathrm{cell}} \ll \mathrm{V}_{\mathrm{ext}}$, so that the external inducer concentration can be measured with respect to the medium volume. The diffusion coefficient is k_{d}. In addition, the AHL intracellular inducer concentration x_6 is the same for all cells. With this simplifying assumption, the dynamics of the external inducer concentration x_{20} depends on x_6 instead of on the average of x_{20} across the population of cells. The same diffusion process is assumed for the second circuit, the QS/Fb, as it will be seen in section 4.3.

4.2.2 Model reduction

As already said in section 2.3.1, this Thesis aims at obtaining a reduced model more amenable for computational analysis, but avoiding excessive reduction that would lead to lack of biological relevance. In particular, the species in the reduced model must not be lumped ones and, and the resulting lumped parameters in the reduced model must be easy to associate to experimental tuning knobs (Hancock et al., 2015).

The model (4.1) is a large order one, with 24 state variables and around 96 parameters. This makes difficult the parameters estimation process that will be carried out later on. Moreover, the large differences in the time scales among the different species in the synthetic gene network (typically many orders of magnitude) originate huge difficulties for simulating the temporal evolution of the network and for understanding the basic principles of its operation. Therefore, this model will be reduced using time-scale separation and detection of system invariants, as it was described in section 2.3.1.

To obtain a reduced order model, we will look for *system invariants* resulting from conservation laws. Then, the Quasi Steady-State Approximation (QSSA) and the layered decomposition techniques will be applied to the fast chemical species. In particular, it is assumed that binding reactions occur very fast as compared to transcription, translation and degradation. Special attention will be paid to the reduced expressions obtained for the promoters. Finally, standard values of the parameters will be taken from the literature. These values will also be useful to assess the reaction time-scales during the model reduction process.

System invariants

The number of copies of the genes *luxR*, *cI*, and *gfp* were kept as constant values through time. The model (4.1) considers this by implicitly incorporating the invariants

$$\begin{aligned} \mathrm{C_R} &= x_1 + x_3 \\ \mathrm{C_I} &= x_9 + x_{10} + x_{15} + x_{21} \\ \mathrm{C_G} &= x_{11} + x_{12} + x_{13} + x_{14} + x_{22} + x_{23} + x_{24} \end{aligned} \tag{4.2}$$

where $\mathrm{C_R}$, $\mathrm{C_I}$, and $\mathrm{C_G}$ are the gene copy numbers of the corresponding genes.

Reduction of the reactions associated to gene luxR

For gene *luxR*, we assume that the binding/unbinding reaction r_1 (see reactions 3.1) of RNAP is much faster than transcription and dilution by cells growth. Indeed, the rate constant of binding of RNAP to the promoter will depend on its affinity for the promoter and, thus, will depend on the promoter sequence. Typical values for the binding and unbinding rates k_1 and k_{-1} of RNAP for a constitutive strong promoter are in the order of magnitude of $\mathrm{k}_1 = 600\,\mathrm{nM}^{-1}\,\mathrm{min}^{-1}$ (Berg et al., 2002), and $\mathrm{k}_{-1} = 180\,\mathrm{min}^{-1}$ (Skinner et al., 2004). On the other hand, if we approximate $\mathrm{k_{mR}}$ by the transcription elongation rate, a typical value for *E. coli* is $\mathrm{k_{mR}} = 50\,\mathrm{nt\,sec}^{-1}$, where nt represents one nucleotide.

Considering the mean transcript length of 1000 base pairs for *E. coli*, and the possibility of several transcripts occurring simultaneously, the mRNA transcription rate is in the order of magnitude of $\mathrm{k_{mR}} = 6\,\mathrm{min}^{-1}$. On the other hand, the growth rate for a doubling time of 20 minutes corresponds to $\mu = 0.035\,\mathrm{min}^{-1}$. Therefore, applying the layered decomposition approach to the x_3 dynamics in equation (4.1), one can set $\mathrm{k_1 c_{RNAP}} x_1 - \mathrm{k}_{-1} x_3 = 0$. This, along with the first invariant in (4.2) gives the relationship

$$x_3 = \frac{\mathrm{C_R}}{1 + \frac{1}{\mathrm{c_{RNAP}}} \frac{\mathrm{k}_{-1}}{\mathrm{k}_1}} \triangleq \frac{\mathrm{C_R}}{1 + \frac{\mathrm{k_{d1}}}{\mathrm{c_{RNAP}}}} \tag{4.3}$$

where $\mathrm{k_{d1}}$ is the dissociation constant. The values given above correspond to a strong promoter, where $\mathrm{k_{d1}} = 0.3\,\mathrm{nM}$. This value may increase several orders of magnitude for weak promoters (Brewster et al., 2012). Notice that using the QSSA approximation $(\dot{x}_3 = 0)$ would approximately give the same result, as $\frac{\mathrm{k}_{-1} + \mathrm{k_{mA}} + \mu}{\mathrm{k}_1} \approx \mathrm{k}_{-1}/\mathrm{k}_1$.

The concentration of free RNAP ($\mathrm{c_{RNAP}}$) depends on the growth rate. For a doubling time of 20 minutes estimates of $\mathrm{c_{RNAP}} = 1\mu M$ were obtained in (Bremer et al., 2003; Klumpp and Hwa, 2008). This value drops to $0.5\,\mu\mathrm{M}$ for a doubling time of 120 minutes.

In summary, for a strong enough promoter it is assumed $\frac{\mathrm{k_{d1}}}{\mathrm{c_{RNAP}}} \ll 1$ in (4.3). Thus, the reduced set of equations for gene *luxR* is considered

$$\begin{aligned} \dot{x}_4 &= \mathrm{k_{mR}}\mathrm{C_R} - (\mathrm{d_{mR}} + \mu)x_4 \\ \dot{x}_5 &= \mathrm{k_R}x_4 - (\mathrm{d_R} + \mu)x_5 - \mathrm{k_2}x_5x_6 + \mathrm{k_{-2}}x_7 \end{aligned} \tag{4.4}$$

For weaker promoters, one may consider that the *effective* transcription rate $\mathrm{k_{mR}}$ takes lower values with respect to the nominal maximum one. Alternatively one might use $\mathrm{k'_{mR}} = \alpha\mathrm{k_{mR}}$, with $\alpha \in [0, 1]$, in (4.4).

The term $\mathrm{r_5}(\mathbf{x}) \triangleq -\mathrm{k_2}x_5x_6 + \mathrm{k_{-2}}x_7$ in the dynamics of x_5, corresponds to the formation and dissociation of the monomer complex (R.A) in the reaction r_2. Thus, it represents the loading effect of the monomer formation of the amount of protein LuxR. The unbinding rate $\mathrm{k_{-2}} = 10\,\mathrm{min}^{-1}$ (Weber and Buceta, 2013), while a dissociation constant $\mathrm{k_{d2}} = \mathrm{k_{-2}}/\mathrm{k_2} = 100\,\mathrm{nM}$ Schwarz-Schilling et al. (2016); Urbanowski et al. (2004) results in the binding rate $\mathrm{k_2} = 0.1\,\mathrm{nM}^{-1}\mathrm{min}^{-1}$. Thus, reaction $\mathrm{r_2}$ is not a fast one.

Reduction of the reactions associated to gene cI

For gene *cI*, it is again assumed that the binding/unbinding reactions of RNA polymerase to the promoter of *cI* (i.e. reactions $\mathrm{r_4}$ and $\mathrm{r_7}$ in 3.2) are fast ones. Thus, it is considered

$$\begin{aligned} x_{15} &= \frac{\mathrm{k_7}}{\mathrm{k_{-7}}}\mathrm{c_{RNAP}}x_{10} \triangleq \frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d7}}}x_{10} \\ x_{21} &= \frac{\mathrm{k_{11}}}{\mathrm{k_{-11}}}\mathrm{c_{RNAP}}x_9 \triangleq \frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d11}}}x_9 \end{aligned} \tag{4.5}$$

where $\mathrm{k_{d7}}$ and $\mathrm{k_{d11}}$ are the dissociation constants accounting for the inverse of the affinity of RNA polymerase for the *cI* promoter when the activator is bound and when it is not, respectively. Notice that at low basal transcription, the RNA polymerase will have low affinity for the promoter unless the activator is bound to it. This accounts to consider that $\mathrm{k_{d7}} \ll \mathrm{k_{d11}}$.

On the other hand, it was also assumed that binding of the transcription factor (R.A)$_2$ to the *cI* promoter $\mathrm{P_{lux}}$ (reaction $\mathrm{r_4}$ in 3.2) is also a fast reaction, as compared e.g. with transcription and translation ones. Thus, taking into account (4.5)

$$x_{10} = \frac{x_8x_9}{\frac{\mathrm{k_{-4}}}{\mathrm{k_4}}} \triangleq \frac{x_8x_9}{\mathrm{k_{d4}}} \tag{4.6}$$

Now, using (4.5), (4.6) and the second invariant in (4.2) for x_9

$$x_9 = \frac{\mathrm{C_I}}{1 + \frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d11}}} + \frac{1}{\mathrm{k_{d4}}}\left(1 + \frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d7}}}\right) x_8} \tag{4.7}$$

On the other hand, it makes sense to consider that the transcription rates $\mathrm{k_{mI}}$ and $\mathrm{k_{mIb}}$ both correspond to the elongation rate once RNA polymerase has achieved to bind the promoter. Thus, $\mathrm{k_{mIb}} = \mathrm{k_{mI}}$. Therefore, the dynamics for the messenger RNA of *cI* (mI: x_{16}) using these parameters becomes

$$\begin{aligned} \dot{x}_{16} &= \frac{\frac{1}{\mathrm{k_{d11}}} + \frac{x_8}{\mathrm{k_{d4}k_{d7}}}}{\frac{1}{\mathrm{c_{RNAP}}} + \frac{1}{\mathrm{k_{d11}}} + \frac{x_8}{\mathrm{k_{d4}}}\left(\frac{1}{\mathrm{c_{RNAP}}} + \frac{1}{\mathrm{k_{d7}}}\right)} \mathrm{k_{mI}C_I} - (\mathrm{d_{mI}} + \mu)x_{16} \\ &\approx \frac{\frac{1}{\mathrm{k_{d11}}} + \frac{x_8}{\mathrm{k_{d7}k_{d4}}}}{\frac{1}{\mathrm{c_{RNAP}}} + \frac{1}{\mathrm{k_{d11}}} + \frac{x_8}{\mathrm{k_{d4}k_{d7}}}} \mathrm{k_{mI}C_I} - (\mathrm{d_{mI}} + \mu)x_{16} \end{aligned} \tag{4.8}$$

where in the last approximation $\mathrm{k_{d7}} \ll \mathrm{c_{RNAP}}$. Recall, it was already assumed as $\mathrm{k_{d7}} \ll \mathrm{k_{d11}}$.

Notice the dynamics for x_{16} contains a **Hill-like function** (Hill, 1910; Weiss, 1997) for the promoter kinetics. The Hill-like function estimates the number of ligand molecules that are required to bind to a receptor to produce a functional effect. In other words, it describes how many transcription factors (dimer (R.A)$_2$: x_8 in this case) are required to start transcription of the gene *cI* following the expression

$$\dot{x}_{16} = \frac{\alpha_0 + \alpha_1 x_8}{1 + \alpha_1 x_8} \mathrm{k_{mI}C_I} - (\mathrm{d_{mI}} + \mu)x_{16} \tag{4.9}$$

where

$$\begin{aligned} \alpha_0 &= \frac{1}{1 + \frac{\mathrm{k_{d11}}}{\mathrm{c_{RNAP}}}} \\ \alpha_1 &= \frac{\mathrm{k_{d11}}}{1 + \frac{\mathrm{k_{d11}}}{\mathrm{c_{RNAP}}}} \frac{1}{\mathrm{k_{d7}k_{d4}}} \end{aligned} \tag{4.10}$$

Notice that as the unbinding of RNA polymerase from the unactivated promoter is much more favorable than the binding reaction, $\mathrm{k_{d11}}$ will increase, while the basal term α_0 will decrease. Simultaneously, the first term in α_1 will tend to one. Furthermore, the affinity of the transcription factor (R.A)$_2$ for the *cI* promoter ($1/\mathrm{k_{d4}}$), and the one of the RNA polymerase for the induced promoter ($1/\mathrm{k_{d7}}$) will be the main factors determining the half-concentration constant (Ang et al., 2013).

***Reduction of the reactions associated to gene* gfp**

For gene *gfp*, the same procedure as for genes *luxR* and *cI* was followed. Thus, the fast dynamics associated to the bound species formed by the *gfp* complexes and RNA polymerase from the reactions set 3.3 were considered

$$\begin{aligned} x_{22} &= \frac{\mathrm{k}_{12}}{\mathrm{k}_{-12}} \mathrm{c}_{\mathrm{RNAP}} x_{14} \triangleq \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}12}} x_{14} \\ x_{23} &= \frac{\mathrm{k}_{10}}{\mathrm{k}_{-10}} \mathrm{c}_{\mathrm{RNAP}} x_{12} \triangleq \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}10}} x_{12} \\ x_{24} &= \frac{\mathrm{k}_{13}}{\mathrm{k}_{-13}} \mathrm{c}_{\mathrm{RNAP}} x_{11} \triangleq \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}13}} x_{11} \end{aligned} \tag{4.11}$$

where the dissociation constants $\mathrm{k}_{\mathrm{d}10}$, $\mathrm{k}_{\mathrm{d}12}$ and $\mathrm{k}_{\mathrm{d}13}$ correspond to the inverse of the affinity of RNA polymerase for the *gfp* promoter when the activator is bound, when both activator and repressor are bound, and when none of them are bound, respectively.

Also, it was assumed that fast binding of the activator and repressor to gene *gfp*. Using (4.11) leads to

$$\begin{aligned} \mathrm{r}_5(\mathbf{x}) + \mathrm{r}_8(\mathbf{x}) + \mathrm{r}_{10}(\mathbf{x}) = 0 &\rightsquigarrow x_8 x_{11} - \left(\frac{\mathrm{k}_{-5}}{\mathrm{k}_5} + \frac{\mathrm{k}_8}{\mathrm{k}_5} x_{17}\right) x_{12} + \frac{\mathrm{k}_{-8}}{\mathrm{k}_5} x_{14} = 0 \\ \mathrm{r}_6(\mathbf{x}) + \mathrm{r}_9(\mathbf{x}) = 0 &\rightsquigarrow x_{11} x_{17} - \left(\frac{\mathrm{k}_{-9}}{\mathrm{k}_9} + \frac{\mathrm{k}_6}{\mathrm{k}_9} x_8\right) x_{13} + \frac{\mathrm{k}_{-6}}{\mathrm{k}_9} x_{14} = 0 \\ \mathrm{r}_6(\mathbf{x}) + \mathrm{r}_8(\mathbf{x}) + \mathrm{r}_{12}(\mathbf{x}) = 0 &\rightsquigarrow x_8 x_{13} - \left(\frac{\mathrm{k}_{-6}}{\mathrm{k}_6} + \frac{\mathrm{k}_{-8}}{\mathrm{k}_6}\right) x_{14} + \frac{\mathrm{k}_8}{\mathrm{k}_6} x_{17} x_{12} = 0 \end{aligned} \tag{4.12}$$

Using (4.11) along with the third invariant in (4.2)

$$\mathrm{C}_{\mathrm{G}} = \left(1 + \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}13}}\right) x_{11} + \left(1 + \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}10}}\right) x_{12} + x_{13} + \left(1 + \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}12}}\right) x_{14} \tag{4.13}$$

Now, lets make the following assumptions:

- The affinity of the inducer (R.A)$_2$ for the hybrid promoter $\mathrm{P}_{\mathrm{lux/cI}}$ is high, so that $\mathrm{k}_{\mathrm{d}5} = \frac{\mathrm{k}_{-5}}{\mathrm{k}_5} \ll 1$, i.e. it is sufficiently small,
- Once the repressor cI is bound to the hybrid promoter, the inducer has low affinity for it, so that $\mathrm{k}_{\mathrm{d}6} = \frac{\mathrm{k}_{-6}}{\mathrm{k}_6} \gg 1$,
- The repressor strength is enough so that $\mathrm{k}_{\mathrm{d}9} = \frac{\mathrm{k}_{-9}}{\mathrm{k}_9}$ takes a small value, but it does not necessarily allow for a strong repressor.

From (4.12) and (4.13), the system of following equations is derived

$$\begin{bmatrix} x_8 & -\left(\frac{\mathrm{k}_{-5}}{\mathrm{k}_5}+\frac{\mathrm{k}_8}{\mathrm{k}_5}x_{17}\right) & 0 & \frac{\mathrm{k}_{-8}}{\mathrm{k}_5} \\ x_{17} & 0 & -\left(\frac{\mathrm{k}_{-9}}{\mathrm{k}_9}+\frac{\mathrm{k}_6}{\mathrm{k}_9}x_8\right) & \frac{\mathrm{k}_{-6}}{\mathrm{k}_9} \\ 0 & \frac{\mathrm{k}_8}{\mathrm{k}_6}x_{17} & x_8 & -\frac{\mathrm{k}_{-6}+\mathrm{k}_{-8}}{\mathrm{k}_6} \\ \left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d13}}}\right) & \left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d10}}}\right) & 1 & \left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d12}}}\right) \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathrm{C_G} \end{bmatrix} \tag{4.14}$$

Solving (4.14) gives rational polynomial expressions for x_{11} to x_{14}. The denominator is a second order polynomial of the form $p(x_8, x_{17}, x_8x_{17}, x_8^2, x_{17}^2, x_8^2x_{17}, x_8x_{17}^2)$. In order to simplify the resulting expressions, the limit cases where x_8, and x_{17} are large enough, and when they are both in small amounts, and when one of the is large and the other one small were taken into account.

First, lets consider the case when the affinity of the inducer (R.A)$_2$ and the repressor protein cI for the promoter of gene *gfp* are large enough, and/or their amounts, so that the approximations

$$\begin{aligned} \frac{\mathrm{k}_{-5}}{\mathrm{k}_5} &\ll \frac{\mathrm{k}_8}{\mathrm{k}_5}x_{17} \\ \frac{\mathrm{k}_{-9}}{\mathrm{k}_9} &\ll \frac{\mathrm{k}_6}{\mathrm{k}_9}x_8 \end{aligned} \tag{4.15}$$

hold for values of inducer and repressor over some small minimum concentration. Under these assumptions, solving (4.14) gives

$$\begin{aligned} x_{11} &= 0 \\ x_{12} &= \frac{\mathrm{k}_6\mathrm{k}_{-8}x_8}{\left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d10}}}\right)\mathrm{k}_6\mathrm{k}_{-8}x_8 + \mathrm{k}_8\mathrm{k}_{-6}x_{17} + \left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d12}}}\right)\mathrm{k}_6\mathrm{k}_8x_8x_{17}}\mathrm{C_G} \\ x_{13} &= \frac{\mathrm{k}_8\mathrm{k}_{-6}x_{17}}{\left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d10}}}\right)\mathrm{k}_6\mathrm{k}_{-8}x_8 + \mathrm{k}_8\mathrm{k}_{-6}x_{17} + \left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d12}}}\right)\mathrm{k}_6\mathrm{k}_8x_8x_{17}}\mathrm{C_G} \\ x_{14} &= \frac{k_6k_8x_8x_{17}}{\left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d10}}}\right)\mathrm{k}_6\mathrm{k}_{-8}x_8 + \mathrm{k}_8\mathrm{k}_{-6}x_{17} + \left(1+\frac{\mathrm{c_{RNAP}}}{\mathrm{k_{d12}}}\right)\mathrm{k}_6\mathrm{k}_8x_8x_{17}}\mathrm{C_G} \end{aligned} \tag{4.16}$$

The expressions (4.16) are not valid when any of the approximations (4.15) do not hold. This is the case when either the inducer $x_8 : (\mathrm{R.A})_2$ or the repressor $x_{17} : \mathrm{cI}$ or both have very low concentrations. Accordingly, the limit cases considered are:

Case $x_8 = 0,\ \frac{\mathrm{k}_{-5}}{\mathrm{k}_5} \ll \frac{\mathrm{k}_8}{\mathrm{k}_5} x_{17}$:

$$\begin{aligned} x_{11} &= \frac{\mathrm{k}_{-9}}{\left(1+\frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}13}}\right)\mathrm{k}_{-9} + \mathrm{k}_9 x_{17}} \mathrm{C}_{\mathrm{G}} \\ x_{12} &= x_{14} = 0 \\ x_{13} &= \frac{k_9 x_{17}}{\left(1+\frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}13}}\right)\mathrm{k}_{-9} + \mathrm{k}_9 x_{17}} \mathrm{C}_{\mathrm{G}} \end{aligned} \tag{4.17}$$

Case $x_{17} = 0,\ \frac{\mathrm{k}_{-9}}{\mathrm{k}_9} \ll \frac{\mathrm{k}_6}{\mathrm{k}_9} x_8$:

$$\begin{aligned} x_{11} &= \frac{\mathrm{k}_{-5}}{\left(1+\frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}13}})\mathrm{k}_{-5} + (1+\frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}10}}\right)\mathrm{k}_5 x_8} \mathrm{C}_{\mathrm{G}} \\ x_{12} &= \frac{\mathrm{k}_5 x_8}{\left(1+\frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}13}})\mathrm{k}_{-5} + (1+\frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{k}_{\mathrm{d}10}}\right)\mathrm{k}_5 x_8} \mathrm{C}_{\mathrm{G}} \\ x_{13} &= x_{14} = 0 \end{aligned} \tag{4.18}$$

Case $x_8 = 0,\ x_{17} = 0$:

$$\begin{aligned} x_{11} &= \frac{\theta_{13}}{\mathrm{c}_{\mathrm{RNAP}} + \theta_{13}} \mathrm{C}_{\mathrm{G}} \\ x_{12} &= x_{13} = x_{14} = 0 \end{aligned} \tag{4.19}$$

One can assume $\mathrm{k}_{\mathrm{mGb}_1} = \mathrm{k}_{\mathrm{mGb}_2} = \mathrm{k}_{\mathrm{mGb}}$ because these basically correspond to the transcription elongation rates once the RNA polymerase has achieved to bind the $\mathrm{P}_{\mathrm{lux/cI}}$ promoter, that will depend on the gene CDS sequence. Therefore, for the normal case (4.16), the dynamics for the messenger RNA of gene *gfp* can be approximated as

$$\begin{aligned} \dot{x}_{18} &= \mathrm{k}_{\mathrm{mG}} \mathrm{c}_{\mathrm{RNAP}} \left(\frac{1}{\mathrm{k}_{\mathrm{d}13}} x_{11} + \frac{1}{\mathrm{k}_{\mathrm{d}10}} x_{12} + \frac{1}{\mathrm{k}_{\mathrm{d}12}} x_{14} \right) - (\mathrm{d}_{\mathrm{mG}} + \mu) x_{18} \\ &= \frac{\mathrm{k}_{\mathrm{mG}} \mathrm{C}_{\mathrm{G}} \left(\frac{\mathrm{k}_6 \mathrm{k}_{-8}}{\mathrm{k}_{\mathrm{d}10}} x_8 + \frac{\mathrm{k}_6 \mathrm{k}_8}{\mathrm{k}_{\mathrm{d}12}} x_8 x_{17} \right)}{\left(\frac{1}{\mathrm{c}_{\mathrm{RNAP}}} + \frac{1}{\mathrm{k}_{\mathrm{d}10}} \right) \mathrm{k}_6 \mathrm{k}_{-8} x_8 + \frac{\mathrm{k}_{-6} \mathrm{k}_8}{\mathrm{c}_{\mathrm{RNAP}}} x_{17} + \left(\frac{1}{\mathrm{c}_{\mathrm{RNAP}}} + \frac{1}{\mathrm{k}_{\mathrm{d}12}} \right) \mathrm{k}_6 \mathrm{k}_8 x_8 x_{17}} \\ &\quad - (\mathrm{d}_{\mathrm{mG}} + \mu) x_{18} \end{aligned} \tag{4.20}$$

Notice that $\mathrm{k}_{\mathrm{d}12}$ should be large ($10^3 - 10^4$ nM), corresponding to low affinity of RNA polymerase for the repressed promoter. The dissociation constant $\mathrm{k}_{\mathrm{d}13}$ will be large ($\sim 10^4$ nM), corresponding to low basal transcription, while $\mathrm{k}_{\mathrm{d}10}$ will be small ($1 - 10^2$ nM), corresponding to medium to strong activation of the hybrid promoter by the inducer (R.A)$_2$.

For the other limit cases:

Case $x_8 = 0$, $\frac{\mathrm{k}_{-5}}{\mathrm{k}_5} \ll \frac{\mathrm{k}_8}{\mathrm{k}_5} x_{17}$:

$$\dot{x}_{18} = \frac{\mathrm{k}_{-9}}{(1 + \frac{\mathrm{k}_{\mathrm{d13}}}{\mathrm{c}_{\mathrm{RNAP}}})\mathrm{k}_{-9} + \mathrm{k}_9 \frac{\mathrm{k}_{\mathrm{d13}}}{\mathrm{c}_{\mathrm{RNAP}}} x_{17}} \mathrm{k}_{\mathrm{mG}} \mathrm{C}_{\mathrm{G}} - (\mathrm{d}_{\mathrm{mG}} + \mu) x_{18} \tag{4.21}$$

Case $x_{17} = 0$, $\frac{k_{-9}}{k_9} \ll \frac{k_6}{k_9} x_8$:

$$\dot{x}_{18} = \frac{\frac{\mathrm{k}_{-5}}{\mathrm{k}_{\mathrm{d13}}} + \frac{\mathrm{k}_5}{\mathrm{k}_{\mathrm{d10}}} x_8}{\left(\frac{1}{c_{RNAP}} + \frac{1}{\mathrm{k}_{\mathrm{d13}}}\right) \mathrm{k}_{-5} + \left(\frac{1}{\mathrm{c}_{\mathrm{RNAP}}} + \frac{1}{\mathrm{k}_{\mathrm{d10}}}\right) k_5 x_8} \mathrm{k}_{\mathrm{mG}} \mathrm{C}_{\mathrm{G}} - (\mathrm{d}_{\mathrm{mG}} + \mu) x_{18} \tag{4.22}$$

Case $x_8 = 0$, $x_{17} = 0$:

$$\dot{x}_{18} = \mathrm{k}_{\mathrm{mG}} \frac{\mathrm{c}_{\mathrm{RNAP}}}{\mathrm{c}_{\mathrm{RNAP}} + \mathrm{k}_{\mathrm{d13}}} \mathrm{C}_{\mathrm{G}} - (\mathrm{d}_{\mathrm{mG}} + \mu) x_{18} \tag{4.23}$$

Reduction of the monomer

One last assumption concerns the large production of monomer (LuxR.AHL) as compared to the dimer one. Therefore, one can apply the QSSA approach already mentioned in section 2.3.1 to the monomer dynamics x_7 in the model (4.1). Thus, the monomer is assumed $\dot{x}_7 = 0$, and the resulting algebraic expression

$$x_7 = -\frac{\mathrm{d}_{\mathrm{RA}} + \mathrm{k}_{-2}}{4\mathrm{k}_3} + \frac{1}{4\mathrm{k}_3} \sqrt{(\mathrm{d}_{\mathrm{RA}} + \mathrm{k}_{-2})^2 + 8\mathrm{k}_3(\mathrm{k}_2 x_5 x_6 + 2\mathrm{k}_{-3} x_8)} \tag{4.24}$$

which can be replaced in the involved dynamics for the species $x_5 =$[LuxR], $x_6 =$[AHL], and $x_8 =$[(LuxR.AHL)$_2$] in equation (4.1).

Besides, a phenomenological model for the hybrid promoter, that includes all the previous cases is

$$\dot{x}_{18} = \mathrm{k}_{\mathrm{mG}} \mathrm{C}_{\mathrm{G}} \frac{\gamma_0 + \gamma_1 x_8 + \gamma_2 x_8 x_{17}}{\gamma_3 + \gamma_4 x_8 + \gamma_5 x_{17} + \gamma_6 x_8 x_{17}} - (\mathrm{d}_{\mathrm{mG}} + \mu) x_{18} \tag{4.25}$$

Interestingly, this phenomenological model is very similar to the one used in (Rodrigo and Elena, 2011). Notice from (4.20), and (4.22) that in the limit case when there is no repressor ($x_{17} = 0$), the promoter activity increases as the concentration of the inducer x_8 does, and it reaches a maximum value of 1 as $x_8 \to \infty$ when $\mathrm{k}_{\mathrm{d10}}$ is low, as described above. This behavior will be captured in the phenomenological model (4.25) making $\gamma_1 = \gamma_4$. Again, from (4.20) and (4.22), this is the case when $\mathrm{k}_{\mathrm{d10}}$ is low so that $\frac{1}{\mathrm{c}_{\mathrm{RNAP}}} + \frac{1}{\mathrm{k}_{\mathrm{d10}}} \approx \frac{1}{\mathrm{k}_{\mathrm{d10}}}$. On the other hand, the relationship $\gamma_6 > \gamma_2$ can be easily inferred.

Besides, the basal level corresponds in all cases to

$$\frac{\gamma_0}{\gamma_3} = \frac{\mathrm{c_{RNAP}}}{\mathrm{c_{RNAP}} + \mathrm{k_{d13}}} \tag{4.26}$$

Therefore, the model sensibly predicts that for a low basal level, the affinity of RNA polymerase for the empty hybrid promoter must be very low, i.e. $\mathrm{k_{d13}}$ large enough. Moreover, the more free RNA polymerase in the cell, the less affinity is required. Additionally, leakiness of the hybrid promoter in absence of either x_{17} or the product $x_8 x_{17}$ are denoted as β_1 and β_2.

To sum up, after the model reduction process, we obtain the ODE model (4.1) with 9 state variables (coming from 24 initially), and one additional algebraic equation (4.24). The reduced order model contains 26 parameters.

$$\begin{aligned}
\dot{x}_1 &= \mathrm{k_{mR}C_R} - \mathrm{d_{mR}}x_1\\
\dot{x}_2 &= \mathrm{k_{pR}}x_1 - \mathrm{k_2}x_2x_3 + \mathrm{k_{-2}}M - \mathrm{d_R}x_2\\
\dot{x}_3 &= -\mathrm{k_2}x_2x_3 + k_{-2}M + \mathrm{k_d}(x_9 - x_3) - \mathrm{d_A}x_3\\
\dot{x}_4 &= \mathrm{k_3}M^2 - \mathrm{k_{-3}}x_4 - \mathrm{d_{RA2}}x_4\\
\dot{x}_5 &= \mathrm{k_{mI}C_I}\frac{x_4}{\gamma_1 + x_4} - \mathrm{d_{mI}}x_5\\
\dot{x}_6 &= \mathrm{k_{pI}}x_5 - \mathrm{d_I}x_6\\
\dot{x}_7 &= \mathrm{k_{mG}C_G}\frac{\gamma_0 + x_4 + \beta_1\gamma_4x_6 + \beta_2\gamma_5x_4x_6}{\gamma_2 + \gamma_3x_4 + \gamma_4x_6 + \gamma_5x_4x_6} - \mathrm{d_{mG}}x_7\\
\dot{x}_8 &= \mathrm{k_{pG}}x_7 - \mathrm{d_G}x_8\\
\dot{x}_9 &= \mathrm{K_{cells}k_d}\left(x_3 - x_9\right) - \mathrm{d_{Ae}}x_9\\
M &= -\frac{\mathrm{d_{RA}} + \mathrm{k_{-2}}}{4\mathrm{k_3}} + \frac{1}{4\mathrm{k_3}}\sqrt{(\mathrm{d_{RA}} + \mathrm{k_{-2}})^2 + 8\mathrm{k_3}(\mathrm{k_2}x_2x_3 + 2\mathrm{k_{-3}}x_4)}
\end{aligned} \tag{4.27}$$

where M is the monomer concentration, and $\mathrm{K_{cells}} = \frac{\mathrm{V_{cell}*N_{cells}}}{\mathrm{V_{medium}}}$ the volumes relationship required to take into account the concentration outside the $\mathrm{N_{cells}}$ cells. Table 4.3 lists the values of the parameters used in the I1-FFL reduced model (4.27), and Table 4.2 describes the species involved.

Notice the transport term $(x_3 - x_9)$, depends only on the difference of the concentrations inside and outside the cells. The $\mathrm{K_{cells}}$ constant reflects the amount that goes out (or in, depending on the sing) from all the cells into the extracellular volume. Finally, the deterministic simulations use $\mathrm{V_{cell}} = 1 \times 10^{-15}\,\mathrm{L}$, which is the typical volume of an *E. coli* cell, and the population is $\mathrm{N_{cells}} = 4.32 \times 10^7$ that is the number of cells of a culture sample with OD$= 0.3$ (see section 2.1.3) placed in a well containing $\mathrm{V_{medium}} = 180\,\mu\mathrm{L}$.

Table 4.2. Variables of the I1-FFL reduced model.

Variable	Species	Description	Units
x_1	mR	*luxR* messenger RNA	nM
x_2	R	LuxR protein	nM
x_3	A	AHL intracellular inducer	nM
M	(R.A)	LuxR and AHL monomer	nM
x_4	(R.A)$_2$	dimer of (R.A)	nM
x_5	mI	*cI* messenger RNA	nM
x_6	I	cI protein	nM
x_7	mG	*gfp* messenger RNA	nM
x_8	G	GFP protein	nM
x_9	$\mathrm{A_e}$	AHL extracellular inducer	nM

4.3 The QS/Fb gene circuit

As said at the beginning of this Chapter, the QS/Fb circuit deals with variability caused by both intrinsic and extrinsic noise sources. Before studying this variability, an amenable model has to be proposed. This section presents the QS/Fb deterministic model and its systematic model reduction as a first step before considering stochasticity in section 4.4.

4.3.1 Circuit model

To obtain the QS/Fb model, a different point of view is adopted with respect to that in the I1-FFL gene circuit. The I1-FFL model quantified a molecular species as a function of the growing population in terms of its expanding volume. This model is appropriate when one is interested in the **average evolution of the population** of cells along time. In contrast, the QS/Fb circuit needs to distinguish **individual cells** for understanding and dealing with the population heterogeneity (see figure 4.2a). We need models suitable for describing molecular species at the single-cell level rather than at the cells population level (Kiviet et al., 2014; Westermayer et al., 2016). Thereby, all species involved in the QS/Fb circuit are considered for every cell in the population, as depicted in Fig.4.2b).

Recall the set of chemical reactions (3.4) and the 10 species involved in the QS/Fb circuit. The dynamical balance equations describe the evolution of the all molecule species as a function of time t, using the mass-action kinetics formalism (section 2.3). The resulting ODE model is given by the set of equations (4.28) that describes each species dynamics inside the *i*-th cell in a population of N cells.

Table 4.3. Parameters of the I1-FFL circuit.

Parameter	Description	Value	Unit	Reference
$\mathrm{d_{mR}}, \mathrm{d_{mI}}, \mathrm{d_{mG}}$	mR, mI, mG degradation rates	0.3624	min^{-1}	(Milo and Phillips, 2015)
$\mathrm{k_{p_R}}, \mathrm{k_{p_I}}, \mathrm{k_{p_G}}$	mR, mI, mG translation rates	80, 40, 15	min^{-1}	(Alon, 2007; Milo and Phillips, 2015)
$\mathrm{d_R}$	LuxR degradation rate	0.035	min^{-1}	(Boada et al., 2016b), and refs. therein
$\mathrm{k_d}$	Inducer diffusion rate	0.06	min^{-1}	(Weiss, 1996; Nilsson et al., 2001)
$\mathrm{k_2}, \mathrm{k_3}$	(R.A), (R.A)$_2$ association rates	0.1	min^{-1}	estimated
$\mathrm{k_{-2}}$	(R.A) dissociation rate	20	min^{-1}	(Weber and Buceta, 2013)
$\mathrm{k_{-3}}$	(R.A)$_2$ dissociation rate	1	min^{-1}	(Boada et al., 2017a)
β_1	$\mathrm{P_{lux/cI}}$ promoter leakage	0.05	adim, nM^{-1}	estimated
β_2	$\mathrm{P_{lux/cI}}$ promoter basal expression	0.05	adim, nM^{-1}	estimated
$\mathrm{d_A}, \mathrm{d_{Ae}}$	Intra/extra cellular inducer degradation rates	0.0164	min^{-1}	(Kaufmann et al., 2005; Schaefer et al., 1996; Kaplan and Greenberg, 1985)
$\mathrm{d_{RA}}, \mathrm{d_{RA2}}$	(R.A), (R.A)$_2$ degradation rates	0.035	min^{-1}	(Buchler et al., 2005) and refs. therein
$\mathrm{k_{mR}C_R}$	Transcription rate times plasmid copy number of $luxR$	30	min^{-1}	(Boada et al., 2015) and refs. therein
$\mathrm{k_{mI}C_I}$	Transcription rate times plasmid copy number of cI	10	min^{-1}	(Lewis et al., 2011; Milo and Phillips, 2015)
$\mathrm{k_{mG}C_G}$	Transcription rate times plasmid copy number of gfp	30	min^{-1}	(Boada et al., 2015)
$\mathrm{d_I}, \mathrm{d_G}$	cI and GFP degradation rates	0.1733	min^{-1}	(Dodd et al., 2001; Lippincott-Schwartz et al., 2001)
γ_0	$\mathrm{P_{lux/cI}}$ promoter coefficient times basal expression	0.0001	nM	estimated
γ_1	$\mathrm{P_{lux}}$ promoter Hill constant	50	nM	estimated
γ_2	$\mathrm{P_{lux/cI}}$ promoter Hill constant	0.02	nM	(Lewis et al., 2011)
γ_3	$\mathrm{P_{lux/cI}}$ promoter coefficient	0.1	adim	estimated
γ_4	$\mathrm{P_{lux/cI}}$ promoter coefficient	1.42	adim	(Milo and Phillips, 2015)
γ_5	$\mathrm{P_{lux/cI}}$ promoter coefficient	70	nM^{-1}	(Milo and Phillips, 2015)

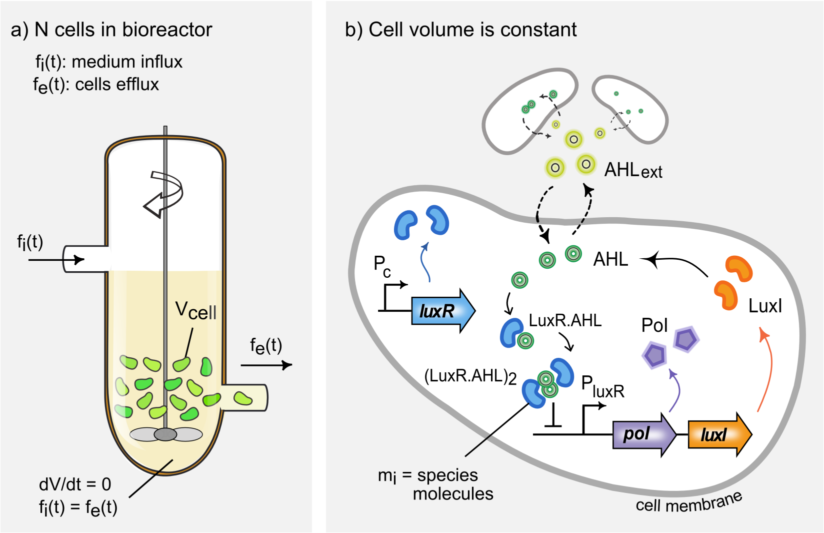


Figure 4.2. QS/Fb circuit model. a) Population of N cells growing in a bioreactor. Dilution is proportional to the efflux $\mathrm{f}_e(t)$. The influx $\mathrm{f}_i(t)$ is the same as the efflux $\mathrm{f}_e(t)$, so the cells are washed away and the total volume $\mathrm{V}(t)$ is kept constant. **b)** The QS/Fb circuit inside of the *i*-th cell in the population.

$$
\begin{aligned}
\dot{n}_1^i &= \mathrm{k}_{\mathrm{e_I}} n_7^i + \alpha \mathrm{k}_{\mathrm{e_I}} n_8^i - \mathrm{d}_{\mathrm{m_I}} n_1^i \\
\dot{n}_2^i &= \mathrm{C_R} - \mathrm{d}_{\mathrm{m_R}} n_2^i \\
\dot{n}_3^i &= \mathrm{p_I} n_1^i - \mathrm{d_I} n_3^i \\
\dot{n}_4^i &= \mathrm{p_R} n_2^i + \mathrm{k}_{-1} n_5^i - \mathrm{d_R} n_4^i - \frac{\mathrm{k}_{-1}}{\mathrm{k_{d1}}} n_9^i n_4^i \\
\dot{n}_5^i &= 2\mathrm{k}_{-2} n_6^i + \frac{\mathrm{k}_{-1}}{\mathrm{k_{d1}}} n_9^i n_4^i + \left(-\mathrm{k}_{-1} - \mathrm{d_{RA}} - 2\frac{\mathrm{k}_{-2}}{\mathrm{k_{d2}}} n_5^i \right) n_5^i \\
\dot{n}_6^i &= \mathrm{k_{lux}} n_8^i + \frac{\mathrm{k}_{-2}}{\mathrm{k_{d2}}} {n_5^i}^2 + \left(-\mathrm{k}_{-2} - \mathrm{d}_{\mathrm{RA_2}} - \frac{\mathrm{k_{lux}}}{\mathrm{k_{dlux}}} n_7^i \right) n_6^i \\
\dot{n}_7^i &= \mathrm{k_{lux}} n_8^i - \frac{\mathrm{k_{lux}}}{\mathrm{k_{dlux}}} n_6^i n_7^i \\
\dot{n}_8^i &= -\mathrm{k_{lux}} n_8^i + \frac{\mathrm{k_{lux}}}{\mathrm{k_{dlux}}} n_6^i n_7^i \\
\dot{n}_9^i &= \mathrm{D} \left(\mathrm{V_c} n_{10} - n_9^i \right) - \left(\frac{\mathrm{k}_{-1}}{\mathrm{k_{d1}}} n_4^i + \mathrm{d_A} \right) n_9^i + \mathrm{k}_{-1} n_5^i + \mathrm{k_A} n_3^i \\
\dot{n}_{10} &= \mathrm{D} \left(-N \mathrm{V_c} n_{10} + \sum_{i=1}^{N} n_9^i \right) - \mathrm{d}_{\mathrm{A_e}} n_{10}
\end{aligned}
\tag{4.28}
$$

Table 4.4. Species of the QS/Fb circuit.

Variable	Species	Description	Unit
n_1	mPI	*pol/luxI* messenger RNA	molecules
n_2	mR	*luxR* messenger RNA	molecules
n_3	PI	co-expression of proteins PoI/LuxI	molecules
n_4	R	LuxR protein	molecules
n_5	(R.A)	LuxR and AHL monomer	molecules
n_6	$(R.A)_2$	dimer of (R.A)	molecules
n_7	gPI	unbound $\mathrm{P_{luxR}}$ promoter gene $poI/luxI$	molecules
n_8	$gPI.(R.A)_2$	bound $\mathrm{P_{luxR}}$ promoter of gene $poI/luxI$	molecules
n_9	A	AHL intracellular inducer	molecules
n_{10}	$\mathrm{A_{ext}}$	AHL extracellular inducer	molecules

The deterministic model (4.28) has 10 state variables listed in Table 4.4. Each equation in (4.28) corresponds to one *i*-th cell, and the last equation $\dot{n}_{10}$ accounts for the $\mathrm{AHL_{ext}}$ dynamics for the whole population.

For the quorum sensing subsystem, passive diffusion of $\mathrm{AHL_{ext}}$ outside the cell and AHL inside the cell across the cell membrane are modeled using an approximation of the Fick's law (Alberts et al., 2009; Weiss, 1996). Thus, in the dynamic balances of AHL and $\mathrm{AHL_{ext}}$ (n_9^i and n_{10}, respectively), a flow of molecules proportional to the gradient of concentrations is considered. The diffusion coefficient $\mathrm{D = SP_n/V_{cell}\,(min^{-1})}$ depends on the cell surface area S, of its membrane permeability $\mathrm{P_n}$ and the cell volume.

For the $\mathrm{AHL_{ext}}$ dynamics, the fact that there is a population of N cells is used to obtain the net inflow. The effect of this inflow over one cell in the population corresponds to the term $\mathrm{V_c = \frac{V_{cell}}{V_{ext}}}$, where $\mathrm{V_{ext}}$ is the volume of the culture medium (refer section 3.3.3). It is assumed all N cells contribute equally. Therefore no effect of the spatial distribution is taken into account. this is the main simplification made. In practice, spatial distribution could be considered while keeping the simple structure of the model, by weighting the summation term of the accumulative cells contributions.

The degradation rate for the 9 intracellular species in the model (4.28) is the sum of the growth dilution and their specific degradation rates. Extracellular species like the external inducer $\mathrm{AHL_{ext}}$ are not subject to dilution by growth rate. On the other hand, the number of copies of the genes *poI/luxI* and *luxR* (n_1 and n_2, respectively), keeps constant through time t. This is indeed the case if the genes are chromosomal ones. In case the genes are located in plasmids, one can assume that at each cell cycle, plasmids are first duplicated, and then half of them will be inherited by each of the offspring cells. This is a valid approximation if we assume that we model the average cell.

Finally, the model parameters summarized in Table 4.5 were calculated taking into account the remarks above and considering:

1. The transcription rate $\mathrm{k_{e_I}}$ is the minimum PoI/LuxI transcription rate. The typical transcription rate in *E. coli.* is ≈ 600-6000 bp/min [1] (Alberts et al., 2009). The LuxI length is 582 bp (part BBa_C0161) (Biobrick Foundation, 2006). Therefore, $\mathrm{k_{e_I}} =$ (600 bp/min)/582 bp = 1.03 min^{-1},

2. The rate $\mathrm{C_R}$ was obtained as the transcription rate obtained as before times the LuxR plasmid copy number. We use the vector pACYC184 with 10 copies/cell, the minimum transcription rate 600 bp/min, and the LuxR length 756 bp (part BBa_C0062) (Biobrick Foundation, 2006). Hence, the plasmid copy number times LuxR transcription rate is $\mathrm{C_R} = (10 * 600$ bp/min)/756 bp = 7.9 molecules·min^{-1},

3. The translation rate can be tuned using a ribosome-binding site (RBS) of different strengths. In bacteria, the translation rate is ≈ 30-60 bp/sec (Alberts et al., 2009). Accordingly, the minimum PoI/LuxI translation rate is $\mathrm{p_I}=$ (1800 bp/min)/582 bp = 3.09 min^{-1}, and the minimum LuxR translation rate is $\mathrm{p_R}=$ (1800 bp/min)/756 bp = 2.38 min^{-1},

4. The degradation rates $\mathrm{dm_I}, \mathrm{dm_R}, \mathrm{d_I}, \mathrm{d_R}, \mathrm{d_A}, \mathrm{d_{RA}}$ include the dilution effect due to the cell growth. The specific growth rate $\mu_{spe} = 0.017$ min^{-1} corresponds to a cell doubling time of 40 min,

5. The degradation rate $\mathrm{d_{RA_2}} = 0.017$ min^{-1} of the dimer (LuxR.AHL)$_2$ only depends of the specific growth rate μ_{spe}, since (R.A)$_2$ is much more stable than the other species in the system (Basu et al., 2005; Buchler et al., 2005),

6. The diffusion coefficient was calculated as $\mathrm{D} = \frac{\mathrm{SP_n}}{\mathrm{V_{cell}}}$ min^{-1}. It depends on the cell surface area $\mathrm{S} = 4\pi\mathrm{r}^2$ (spherical area with r=10 μm), the membrane permeability $\mathrm{P_n} = 3 \times 10^{-3} \mu\mathrm{m} \cdot \mathrm{min}^{-1}$ and the typical *E. coli.* volume $\mathrm{V_{cell}} = 1.1 \times 10^{-9}\ \mu\mathrm{L}$,

7. The dissociation rate of (LuxR.AHL)$_2$ to the $\mathrm{P_{luxR}}$ promoter $\mathrm{k_{lux}}$ was not required by the mathematical model (4.28).

4.3.2 Model reduction

As in the I1-FFL case, the QS/Fb model (4.28) can be further simplified by applying the QSSA to the fast species, and taking *system invariants* as the result of conservation laws. We aimed at obtaining a reduced model more amenable for computational analysis, but avoiding excessive reduction that would lead to lack of biological relevance. In particular, the species we obtained in the reduced model are not lumped ones. Reduced models accounting for total mRNA and total transcription factor have been proposed to match modeled species with measurable ones (Hancock et al., 2015). In

[1] bp/min is one unit of DNA base pair coded per minute.

Table 4.5. Parameters of the QS/Fb model.

Parameter	Description	Value	Unit	Reference
$\mathrm{C_R}$	Plasmid copy number times gR transcription rate	7.9	molecules·min^{-1}	(Boada et al., 2015)
$\mathrm{k_{e_I}}$	gPI transcription rate	17.5	molecules·min^{-1}	(Boada et al., 2015)
α	$\mathrm{P_{luxR}}$ promoter basal expression	0.01	-	estimated
$\mathrm{p_R}$	mR translation rate	10	min^{-1}	(Alon, 2007; Milo and Phillips, 2015)
$\mathrm{p_I}$	mPI translation rate	3.09	min^{-1}	(Alon, 2007; Milo and Phillips, 2015)
$\mathrm{k_A}$	Synthesis rate of AHL by LuxI	0.04	min^{-1}	(Vignoni et al., 2013b)
$\mathrm{k_{-1}}$	Dissociation rate of (R.A)	10	min^{-1}	(Weber and Buceta, 2013)
$\mathrm{k_{-2}}$	Dissociation rate of dimer (R.A)$_2$	1	min^{-1}	(Boada et al., 2017a)
$\mathrm{k_{d1}}$	Dissociation constant of (R.A)	100	molecules	(Urbanowski et al., 2004)
$\mathrm{k_{d2}}$	Dissociation constant of (R.A)$_2$	20	molecules	(Harman, 2001)
$\mathrm{k_{dlux}}$	Dissociation constant of (R.A)$_2$ to the $\mathrm{P_{luxR}}$ promoter	100	molecules	(Buchler et al., 2005) and refs. therein
$\mathrm{d_I}$	PI degradation rate	0.027	min^{-1}	(Goryachev et al., 2006; Milo et al., 2016)
$\mathrm{d_R}$	R degradation rate	0.2	min^{-1}	(Boada et al., 2016b), and refs. therein
$\mathrm{d_A}$	A degradation rate	0.057	min^{-1}	(Kaufmann et al., 2005; Schaefer et al., 1996)
$\mathrm{d_{A_e}}$	A degradation rate in culture medium	0.04	min^{-1}	(Kaufmann et al., 2005; Kaplan and Greenberg, 1985)
$\mathrm{d_{RA}}$	(R.A) degradation rate	0.156	min^{-1}	(Buchler et al., 2005) and refs. therein
$\mathrm{d_{RA_2}}$	(R.A)$_2$ degradation rate	0.017	min^{-1}	(Boada et al., 2017a)
$\mathrm{dm_I}$	mPI degradation rate	0.247	min^{-1}	(Roberts et al., 2006; Santillán and Mackey, 2001)
$\mathrm{dm_R}$	mR degradation rate	0.247	min^{-1}	(Milo et al., 2016; Santillán and Mackey, 2001)
D	Diffusion rate of AHL through the cell membrane	2	min^{-1}	(Weiss, 1996; Nilsson et al., 2001)
$\mathrm{V_{cell}}$	Typical volume of *E. coli.*	1.1×10^{-9}	μL/cell	(Milo and Phillips, 2015)
$\mathrm{V_{ext}}$	Typical volume of microfluidic device	1×10^{-3}	μL	(Olson et al., 2014)

the QS/Fb case, we explicitly modeled bound and unbound forms of the transcription factor, but the model accounts for the total LuxI protein. For this gene circuit this is a good proxy for the amount of protein of interest if both are co-expressed, and transcriptional noise dominates. In the best case, when the protein of interest is in self-cleavable tandem fusion with LuxI, both will express in 1:1 stoichiometric ratio (Chen et al., 2010b). Moreover, the resulting lumped parameters in the reduced model are easy to associate to tuning knobs available in the wet-lab implementation in the relevant cases (Arpino et al., 2013), and their values are amenable to be obtained experimentally.

The first assumption concerns *system invariants* (as in section 4.2.2) resulting from conservation laws in the model (4.28). The amount of DNA from the gene *pol/luxI* keeps as a constant along time t. As a result, the sum of free DNA plus the bound DNA.(R.A)$_2$ (n_7^i and n_8^i, respectively) leads to

$$\dot{n}_7^i + \dot{n}_8^i = 0 \rightsquigarrow n_7^i + n_8^i = \mathrm{P_N} \tag{4.29}$$

This implies that the sum of free and bound promoter $\mathrm{P_{luxR}}$ is constant and equal to the *plasmid copy number* $\mathrm{P_N}$.

The second consideration assumes that the transcription factor binding/unbinding reactions to the $\mathrm{P_{luxR}}$ promoter proceed much faster than translation and mRNA degradation, so they can be assumed to be at quasi-steady state. Applying the *Quasi Steady-State Approximation* (QSSA), this is also equivalent to consider that $\mathrm{k_{lux}}$ is large enough so, using (4.29), this can be approximated

$$\begin{aligned} \frac{1}{\mathrm{k_{lux}}}\dot{n}_7^i = 0 \rightsquigarrow n_7^i = \mathrm{P_N}\left(\frac{\mathrm{k_{dlux}}}{\mathrm{k_{dlux}} + n_6^i}\right) \\ \frac{1}{\mathrm{k_{lux}}}\dot{n}_8^i = 0 \rightsquigarrow n_8^i = \mathrm{P_N}\left(\frac{n_6^i}{\mathrm{k_{dlux}} + n_6^i}\right) \end{aligned} \tag{4.30}$$

The third QSSA assumption is related to the transcription reactions of genes *pol/luxI* and *luxR* (n_1^i and n_2^i, respectively). It is assumed that messenger RNA for both *pol/luxI* and *luxR* are produced and degraded much faster than the proteins. Hence, applying QSSA

$$\dot{n}_1^i = 0 \rightsquigarrow n_1^i = \frac{\mathrm{k_{e_I}}}{\mathrm{d_{m_I}}}\left(n_7^i + \alpha n_8^i\right) \tag{4.31}$$

and

$$\dot{n}_2^i = 0 \rightsquigarrow n_2^i = \frac{\mathrm{C_R}}{\mathrm{dm_R}} \tag{4.32}$$

From (4.31) and (4.30), replacing in (4.28) we have

$$\dot{n}_3^i = \frac{\mathrm{P_N k_{e_I} p_I}}{\mathrm{dm_I}} \left(\frac{\mathrm{k_{dlux}} + \alpha n_6^i}{\mathrm{k_{dlux}} + n_6^i} \right) - \mathrm{d_I} n_3^i \tag{4.33}$$

and using (4.32) and (4.30) in (4.28) provides

$$\dot{n}_4^i = \frac{\mathrm{C_R p_R}}{\mathrm{dm_R}} + \mathrm{k_{-1}} n_5^i - \left(\frac{\mathrm{k_{-1}}}{\mathrm{k_{d1}}} n_9^i + \mathrm{d_R} \right) n_4^i \tag{4.34}$$

Also replacing (4.30) in (4.28) approximates

$$\dot{n}_6^i = \frac{\mathrm{k_{-2}}}{\mathrm{k_{d2}}} (n_5^i)^2 - (\mathrm{k_{-2}} + \mathrm{d_{RA_2}}) n_6^i \tag{4.35}$$

Finally, the fourth QSSA assumption concerns the large production of monomer (R.A) or (LuxR.AHL) as compared to the dimer one. It follows the same procedure described in subsection 4.2.2 for the I1-FFL gene circuit. Thus, the monomer (R.A) is assumed $\dot{n}_5^i = 0$, and the resulting algebraic expression for n_5^i can be replaced in the species n_4^i =[LuxR], n_6^i =[(LuxR.AHL)$_2$], and the intracellular inducer n_9^i =[AHL].

All these approximations (with a renumbering of the variable names, so as to have continuous numbering) lead to the reduced-order model (4.36) for the *i*-th cell in a population of N cells. It consists of four differential equations and one algebraic equation per cell. Additionally, there is one differential equation describing the external $\mathrm{AHL_{ext}}$ dynamics for the whole population.

$$\begin{aligned}
\dot{n}_1^i &= \frac{\mathrm{C_I p_I}}{\mathrm{dm_I}} \left(\frac{\mathrm{k_{dlux}} + \alpha n_3^i}{\mathrm{k_{dlux}} + n_3^i} \right) - \mathrm{d_I} n_1^i \\
\dot{n}_2^i &= \frac{\mathrm{C_R p_R}}{\mathrm{dm_R}} + \mathrm{k_{-1}} n_6^i - \left(\frac{\mathrm{k_{-1}}}{\mathrm{k_{d1}}} n_4^i + \mathrm{d_R} \right) n_2^i \\
\dot{n}_3^i &= \frac{\mathrm{k_{-2}}}{\mathrm{k_{d2}}} (n_6^i)^2 - (\mathrm{k_{-2}} + \mathrm{d_{RA_2}}) n_3^i \\
\dot{n}_4^i &= \mathrm{k_{-1}} n_6^i + \mathrm{k_A} n_1^i + \mathrm{D} \left(\mathrm{V_c} n_5 - n_4^i \right) - \left(\frac{\mathrm{k_{-1}}}{\mathrm{k_{d1}}} n_2^i + \mathrm{d_A} \right) n_4^i \\
\dot{n}_5 &= \mathrm{D} \left(-\mathrm{N V_c} n_5 + \sum_{i=1}^{\mathrm{N}} n_4^i \right) - \mathrm{d_{A_e}} n_5 \\
n_6^i &= \frac{\mathrm{k_{d2}}(\mathrm{d_{RA}} + \mathrm{k_{-1}})}{4\mathrm{k_{-2}}} \left[\sqrt{\frac{8\mathrm{k_{-2}}(2\mathrm{k_{-2}k_{d1}} n_3^i + \mathrm{k_{-1}} n_2^i n_4^i)}{\mathrm{k_{d1}k_{d2}}(\mathrm{d_{RA}} + \mathrm{k_{-1}})^2} + 1} - 1 \right]
\end{aligned} \tag{4.36}$$

Table 4.6. Species for the QS/Fb reduced model.

Variable	Species	Unit
n_1	LuxI protein	molecules
n_2	LuxR protein	molecules
n_3	Dimer of (R.A)	molecules
n_4	AHL intracellular inducer	molecules
n_5	$\mathrm{AHL_{ext}}$ extracellular inducer	molecules
n_6	Monomer (R.A)	molecules

The species involved are listed in Table 4.6. The parameter $\mathrm{C_I}$ is the plasmid copy number times the *luxI* transcription rate. $\mathrm{C_I} = \mathrm{P_N}\, \mathrm{k_{e_I}} =$ 17.5 molecules·min^{-1}, where $\mathrm{P_N}$ is the LuxI plasmid copy number (vector pBR322 with $\approx$ 17 copies described in 3.3.2), and $\mathrm{k_{e_I}} =$ 1.03 min^{-1} from Table 4.5. The remaining parameters are the same as those of the full model listed in Table 4.5.

It is important to note the first term on the right hand side of the dynamics of n_1^i in the model (4.36). This is a Hill-like function (Hill, 1910; Alon, 2007) with a hill coefficient $\mathrm{n} = 1$, which together with the monomer algebraic equation n_6^i in (4.36) describe the transcription factor regulatory effect, as in the I1-FFL reduced model (4.27).

$$h(x) = k_{max} \frac{K^{\mathrm{n}}}{K^{\mathrm{n}} + x^{\mathrm{n}}} \tag{4.37}$$

where x is the transcription factor, k_{max} is the maximum transcription rate, K is the repression threshold, and n is the Hill coefficient. The Hill coefficient estimates the number of molecules of transcription factor required to bind the promoter and to inhibit gene expression of a desired protein (recall section 4.2.2 for the I1-FFL circuit).

For the n_1^i dynamics, the Hill-like function shows the inhibiting effect on PoI/LuxI expression caused by the negative feedback through the dimer $n_3^i =$ [(LuxR. AHL)]$_2$, i.e. with

$$h(n_3^i) = \frac{\mathrm{C_I p_I}}{\mathrm{dm_I}} \left(\frac{\mathrm{k_{dlux}} + \alpha n_3^i}{\mathrm{k_{dlux}} + n_3^i} \right) \tag{4.38}$$

where one molecule ($n = 1$) of the transcription factor $(\mathrm{LuxR.AHL})_2$ is the required Hill coefficient to start the repression of PoI/LuxI expression in the *i*-th cell. The dissociation constant $\mathrm{k_{dlux}}$ and the transcription-translation rate $\frac{\mathrm{C_I p_I}}{\mathrm{dm_I}}$ are relative to the $\mathrm{P_{luxR}}$ promoter strength. Additionally, αn_3^i represents the basal expression (leakage) of the promoter $\mathrm{P_{luxR}}$, as in the previous full model. These expressions are equivalent to the ones that one can obtain using (Hancock et al., 2015) for the multimer-dominant case.

To compare the full model with the reduced order one, a series of *in silico* experiments were performed. Fig.4.3 shows some of the results demonstrating the good agreement

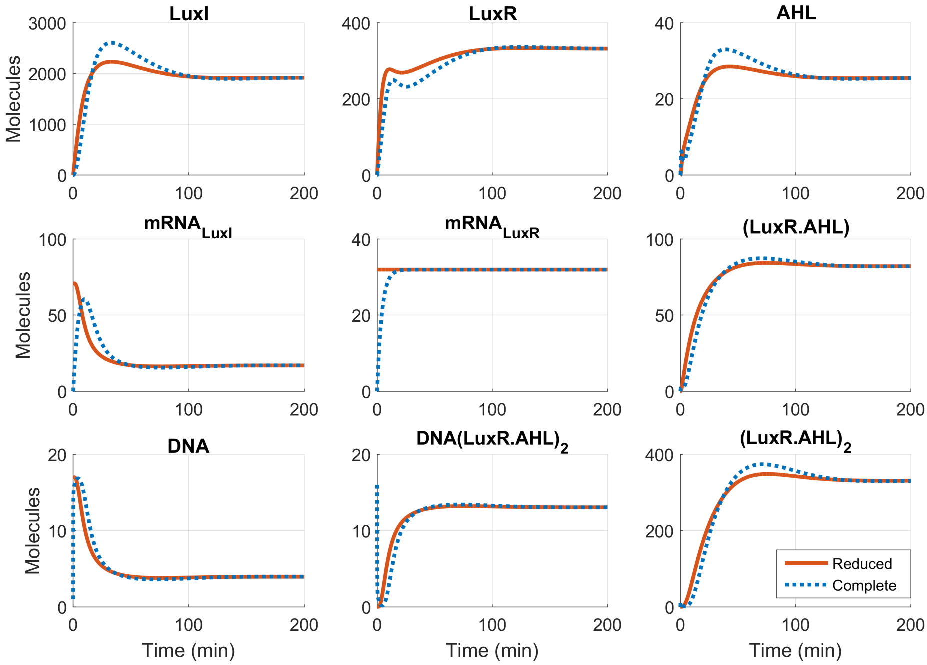


Figure 4.3. Comparison of the QS/Fb full and reduced models. Simulation during 250 minutes for a single cell of both the reduced (solid line) and the complete model (dashed line). In both cases the simulations were performed with the same initial conditions, same parameter values and same step size: $\delta t = 1 \times 10^{-3}$ seconds.

between the results provided by both the complete and the reduced models. The principal biochemical species LuxI, LuxR and AHL are plotted on the top of Fig.4.3 for the reduced model (solid line), and for the full one (dashed line).

The plots of the five species eliminated by the model reduction: *luxR* and *pol/luxI* messengers RNA ($\mathrm{mRNA_{luxI}}$ and $\mathrm{mRNA_{luxR}}$, respectively), unbound and bound DNA of gene *pol/luxI* (DNA and DNA.(LuxR.AHL)$_2$), and the monomer (LuxR.AHL) were calculated algebraically from the remaining species. This simulation was carried over a single cell (N=1). Therefore, the amount of molecules of AHL and $\mathrm{AHL_{ext}}$ is similar, hence, the $\mathrm{AHL_{ext}}$ plot was omitted in this figure. The agreement between the results of both models was good enough for our purposes, without requiring any *ad hoc* adjustment. From a qualitative point of view, the transient regime of the complete model is similar to the reduced one for all species. The length of the transients and the steady-state values coincide in both models.

In Section 3.3, another synthetic gene circuit with no QS nor feedback was proposed. The NoQS/NoFb circuit (see Fig.3.5b) assesses the role played by both feedback and

QS subsystems. As we saw, synthesis of AHL molecules by LuxI protein is the only one chemical reaction missed in the NoQS/NoFb circuit from the original QS/Fb system. As a consequence, two main aspects should be considered:

1. the NoQS/NoFb ODE model is similar to the QS/Fb one, except for there is no term of AHL synthesis $\mathrm{k_A} n_1$ in the equation $\dot{n}_4^i$ from (4.36), due to the synthesis rate $\mathrm{k_A}$ is null, and
2. n_5 represents the extracellular $\mathrm{AHL_{ext}}$ molecules. The initial condition $n_5(0)$ represents a bolus (pulse-like) amount of extracellular $\mathrm{AHL_{ext}}$ added to the medium.

4.4 Stochastic CLE-based model for cell population

The principal idea in this section is to use the deterministic model, as a basis to formulate a simplified stochastic model for analyzing noise in gene expression. Additionally, this section will address the practical problems arising when besides the individual states in each cell, one has a global cell population state. In the QS/Fb circuit, this global state as seen in section 4.3 arises from quorum sensing-based cell-to-cell communication.

Section 2.4 reviewed how biochemical reactions are subject to fluctuations producing gene expression noise, and how this stochasticity can be modeled using the Chemical Master Equation (CME), the Gillespie's Algorithm (SSA), or the Chemical Langevin Equation (CLE). Section 2.4.2 showed that the most accurate way to depict the time evolution of a biological system is using the (CME): a set of linear ODES, one for each possible state of the system at time t. Yet, the CME becomes intractable for large systems like the QS/Fb circuit. One option to sample the CME is to use Gillespie's Algorithm or Stochastic Simulation Algorithm (SSA), described in section 2.4.4: time trajectories numerically generated based on a Monte Carlo process, which are in exact accordance with the CME. However, having an interconnected population of cells like in the QS/Fb system jeopardizes the possibility of employing SSA for several reasons.

First there are different volumes involved, extracellular and intracellular. The diffusion through the membrane depends on the concentration gradient of the small molecule in both of them, making the account for the probability of reaction more complicated. Second, when using SSA, several realizations or trajectories of the system are needed in order to obtain an accurate estimation of the moments, making the use of SSA in a population of interconnected cells a computationally very demanding task. Although, these kind of systems have been modeled and simulated before as ODE perturbed with white noise (Koseska et al., 2009), this does not capture the intrinsic noise phenomena as desired.

If the expected number of firings of each reaction event during a time interval δt is greater than one, it is possible to use a Langevin-type equation called the Chemical Langevin Equation (CLE): a set of nonlinear SDEs, where the solution of each equation at time t is a real-valued random variable representing the amount of every species in the system (refer to section 2.4.5).

Interestingly, generating sample paths is orders of magnitude faster than doing the same for the CME, because it essentially needs generation of normal random numbers instead of probability functions. In fact, the CLE (and also SSA) allows us to obtain certain statistical parameters for analyzing fluctuations in a gene circuit, no matter if there is cell-to-cell communication.

4.4.1 CLE-based model

To model gene expression noise produced by extrinsic and intrinsic sources, a stochastic CLE-based model of the QS/Fb system is derived, whose mean corresponds to that of the deterministic reduced model (4.36) for each species.

To consider intrinsic noise and since a CLE is a special form of the general SDE, the *Euler-Maruyama discretization method* (Higham, 2001) was used for generating sample paths of the stochastic process driven by a CLE for each species

$$n(t+\delta t) = n(t) + \mathbf{S}\cdot\mathbf{a}(n)\delta t + \mathbf{S}\cdot\mathcal{N}\cdot\sqrt{\mathbf{a}(n)}\sqrt{\delta t} \tag{4.39}$$

where $n(t)$ is the number of molecules of each species in the population of the QS/Fb system, **S** is the stoichiometry matrix, $\mathbf{a}(n)$ is the reaction propensity, $\mathcal{N}(0,1)$ is a statistically independent normal random variable, representing a Brownian motion, and δt is the step time.

Now, in order to define the variables **S**, $\mathbf{a}(n)$, and $\mathcal{N}(0,1)$, a set of equivalent pseudo-reactions from the model (4.36) was deduced. There are four intracellular species, and one extracellular species interacting through thirteen biochemical reactions (J=13) in every *i*-th cell from the population. The biochemical reactions are listed in (4.40), and the species involved in the gene circuit (see Table 4.6) are denoted as: proteins PoI/LuxI (PI) and LuxR (R), monomer (LuxR.AHL) or (R.A), dimer (LuxR.AHL)$_2$ or (R.A)$_2$, and intra/extracellular inducer AHL (A and $\mathrm{A}_{\mathrm{ext}}$, respectively).

$$
\begin{aligned}
(\mathrm{R}\cdot\mathrm{A})_2 &\xrightarrow{f(n_3,t)} \mathrm{PI} + (\mathrm{R}\cdot\mathrm{A})_2 \\
\mathrm{I} &\xrightarrow{\mathrm{k_A}} \mathrm{I} + \mathrm{A} \\
&\xrightarrow{\mathrm{tt_{LuxR}}} \mathrm{R} \\
\mathrm{R} + \mathrm{A} &\xrightleftharpoons[\mathrm{k_{-1}}]{\frac{\mathrm{k_{-1}}}{\mathrm{k_{d1}}}} (\mathrm{R}\cdot\mathrm{A}) \\
{}_2(\mathrm{R}\cdot\mathrm{A}) &\xrightleftharpoons[g(n_6,t)]{g(n_6,t)} (\mathrm{R}\cdot\mathrm{A})_2 \\
\mathrm{A} &\xrightleftharpoons[\mathrm{DV_c}]{\mathrm{D}} \mathrm{A_{ext}} \\
\mathrm{PI} &\xrightarrow{\mathrm{d_I}} \emptyset \\
\mathrm{R} &\xrightarrow{\mathrm{d_R}} \emptyset \\
(\mathrm{R}\cdot\mathrm{A})_2 &\xrightarrow{\mathrm{d_{RA}}} \emptyset \\
\mathrm{A} &\xrightarrow{\mathrm{d_A}} \emptyset \\
\mathrm{A_{ext}} &\xrightarrow{\mathrm{d_{Ae}}} \emptyset
\end{aligned}
\tag{4.40}
$$

Note that function $f(n_3,t) \triangleq \frac{\mathrm{C_I p_I}}{\mathrm{dm_I}}\left(\frac{\mathrm{k_{dlux}}+\alpha_I n_3^i}{\mathrm{k_{dlux}}+n_3^i}\right)$ denotes the Hill-like function associated to PI expression, $g(n_6,t)$ corresponds to the dimerization reflected in the last equation of (4.36), $\mathrm{tt_{LuxR}} = \frac{\mathrm{C_R p_R}}{\mathrm{dm_R}}$ represents the transcription-translation activity of *luxR*, $\mathrm{V_c} = \frac{\mathrm{V_{cell}}}{\mathrm{V_{ext}}}$ is the ratio between the cell volume and the culture medium volume, and $\emptyset$ denotes species degradation. All the parameters are the same from Table 4.5.

Thus for example for the QS/Fb circuit, the amount of molecules of protein PoI/LuxI in the *i*-th cell can be expressed and numerically solved applying equation (4.39)

$$
\begin{aligned}
n_1(t+\delta t)^i = & n_1^i + \left[f(n_3^i,t) - \mathrm{d_I} n_1^i\right]\delta t + \\
& \left[\sqrt{f(n_3^i,t)}\mathcal{N}_1(0,1)^i - \sqrt{\mathrm{d_I} n_1^i}\mathcal{N}_2(0,1)^i\right]\sqrt{\delta t}
\end{aligned}
\tag{4.41}
$$

Summing up, the CLE-based model for cell population describes the QS/Fb dynamics as having five states formally written

$$
\mathbf{n}(t+\delta t) = \mathbf{n}(t) + \mathbf{S}\cdot\mathbf{a}(\mathbf{n})\delta t + \mathbf{S}\cdot\mathcal{N}\cdot\sqrt{\mathbf{a}(\mathbf{n})}\sqrt{\delta t} \tag{4.42}
$$

where $\mathbf{n}(t) = [\mathbf{n}^i(t), \ldots \mathbf{n}^{\mathrm{N}}, n_5]^T$ contains all vectors $\mathbf{n}^i(t)$ for the *i*-th cell in a population of N cells, whose elements represent the molecules number of the intracellular species $n_1, \ldots, n_4$ listed in Table 4.6, and n_5 is the molecules number of the extracellular inducer $\mathrm{AHL_{ext}}$. Remember the monomer n_6 : (LuxR.AHL) is obtained by means of an algebraic equation that depends on LuxR and AHL (n_2 and n_4, respectively) and also interacts with the dimer n_3 : (LuxR.AHL)$_2$. Recall the output of this system is the co-expression of proteins PoI and LuxI denoted as n_1.

For equation 4.42, the stoichiometry matrix **S**, whose elements for the *i*-th cell are the stoichiometry sub-matrices $\mathbf{S_{cell}}$ and the extracellular stoichiometry $\mathbf{S_{ext}}$, has the following structure

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{S_{cell}} \otimes \mathbf{I_N} & \mathbf{0_{4N\times 1}} \\ \hline \mathbf{S_{ext}} \otimes \mathbf{1_{1\times N}} & -1 \end{array}\right] \tag{4.43}$$

where $\otimes$ is the Kronecker product, $\mathbf{I_N}$ the identity matrix of dimension N× N, $\mathbf{0_{4N\times 1}}$ and $\mathbf{1_{1\times N}}$ are vectors of zeroes and ones respectively, J=13 is the total number of reactions, and the coefficients in the stoichiometry matrices $\mathbf{S_{cell}}_{\ [\mathbf{4\times J}]}$ and $\mathbf{S_{ext}}_{\ [\mathbf{1\times J}]}$ were obtained from the set of pseudo-reactions (4.40) as follows

$$\begin{aligned} \mathbf{S_{cell}} &= \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{bmatrix}, \\ \mathbf{S_{ext}} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \end{aligned} \tag{4.44}$$

Also in (4.42), the term $\mathbf{a}(\mathbf{n})$ (4.45) is the associated vector of reaction propensities, whose elements are in turn the $a(n)^i$ propensities vector for every *i*-th cell in the whole population.

$$\mathbf{a}(\mathbf{n}) = \left[\begin{array}{c} \mathbf{a}(\mathbf{n})^1 \\ \mathbf{a}(\mathbf{n})^2 \\ \vdots \\ \mathbf{a}(\mathbf{n})^N \\ \hline \mathrm{d_{A_e}} n_5 \end{array}\right], \quad \mathbf{a}(\mathbf{n})^i = \begin{bmatrix} f(n_3^i, t) \\ \mathrm{d_I} n_1^i \\ \frac{\mathrm{C_R p_R}}{\mathrm{dm_R}} \\ \mathrm{k_{-1}} n_6^i \\ \frac{\mathrm{k_{-1}}}{\mathrm{k_{d1}}} n_2^i n_4^i \\ \mathrm{d_R} n_2^i \\ \frac{\mathrm{k_{-2}}}{\mathrm{k_{d2}}} (n_6^i)^2 \\ \mathrm{k_{-2}} n_3^i \\ \mathrm{d_{RA_2}} n_3^i \\ \mathrm{k_A} n_1^i \\ \mathrm{d_A} n_4^i \\ \mathrm{D} n_4^i \\ \mathrm{DV_c} n_5 \end{bmatrix}. \tag{4.45}$$

In addition, the intrinsic noise term $\mathcal{N}_{\mathbf{(JN+1)\times(JN+1)}}$ is a diagonal matrix of continuous normal random variables with zero mean and unit variance ($\mathcal{N}(\mu, \sigma^{\mathbf{2}}) = \mathcal{N}(\mathbf{0}, \mathbf{1})$), where again there are $J = 13$ biochemical reactions for every cell.

Now, we can extend the CLE-based model (4.42) to add extrinsic noise. Extrinsic noise was modeled by randomizing the values of the model parameters as in (Joo et al., 2013; Toni and Tidor, 2013). The approach can easily be integrated within the CLE framework. Here, extrinsic noise assumes a normal distribution with zero mean and fifteen percent variance $\mathcal{N}(0, 0.15)$ to generate the model parameters of every *i*-th cell in the population of N cells.

Figure 4.4 illustrates the results obtained with the resulting model (4.42). The species abundance in a temporal window of 800 min for N=240 cells were simulated. Each species has a normalized endpoint distributions computed over one realization of the model (4.42), including intrinsic and extrinsic noise. Each species mean and its standard deviation ($\mu \pm \sigma$) was computed using the last one third data of each species in the whole population, in order to avoid the effect of the transient. The same parameters and initial conditions as the ODE model (4.36) were used.

Note that the average molecules of species Pol/LuxI, LuxR, (LuxR.AHL)$_2$, and AHL show slight differences between the deterministic (Fig.4.3) and the mean trajectories from the CLE simulation. The comparison between both the deterministic and stochastic models of the QS/Fb circuit is illustrated in Fig.4.5. All parameters and initial conditions in both cases are the same. This difference comes from both intrinsic and extrinsic noise, and it arises from the nonlinearity of the propensity of n_1^i together with the variance of extrinsic noise added.

As we saw in section 2.4, noise level in biological systems can be measured by the **noise strength** ($\eta^2 = \sigma^2/\mu^2$) that expresses how close to Poisson a given process is. From the CLE simulations, we see that the noise level at low number of molecules is higher than the one at large amount of molecules. This is clearly appreciated in the AHL case (see Fig.4.4 and Fig.4.5 bottom), where the noise strength of AHL is the highest ($\eta_A^2 = 0.2$) compared to the other ones ($\eta_R^2 = 0.18, \quad \eta_{RA_2}^2 = 0.17$, and $\eta_{PI}^2 = 0.1$). Additionally, all endpoint histograms of the population show a well-shaped normal distribution, and they differ only in their means and noise strengths.

Finally, to validate this model, it is necessary to consider three important aspects: *i)* could the NoQS/NoFb stochastic model can be deduced from the QS/Fb one?, *ii)* can we justify the use of non-linear functions as propensities in the stochastic model?, and *iii)* how do we assess the population density effect in the QS/Fb circuit to obtain representative data?.

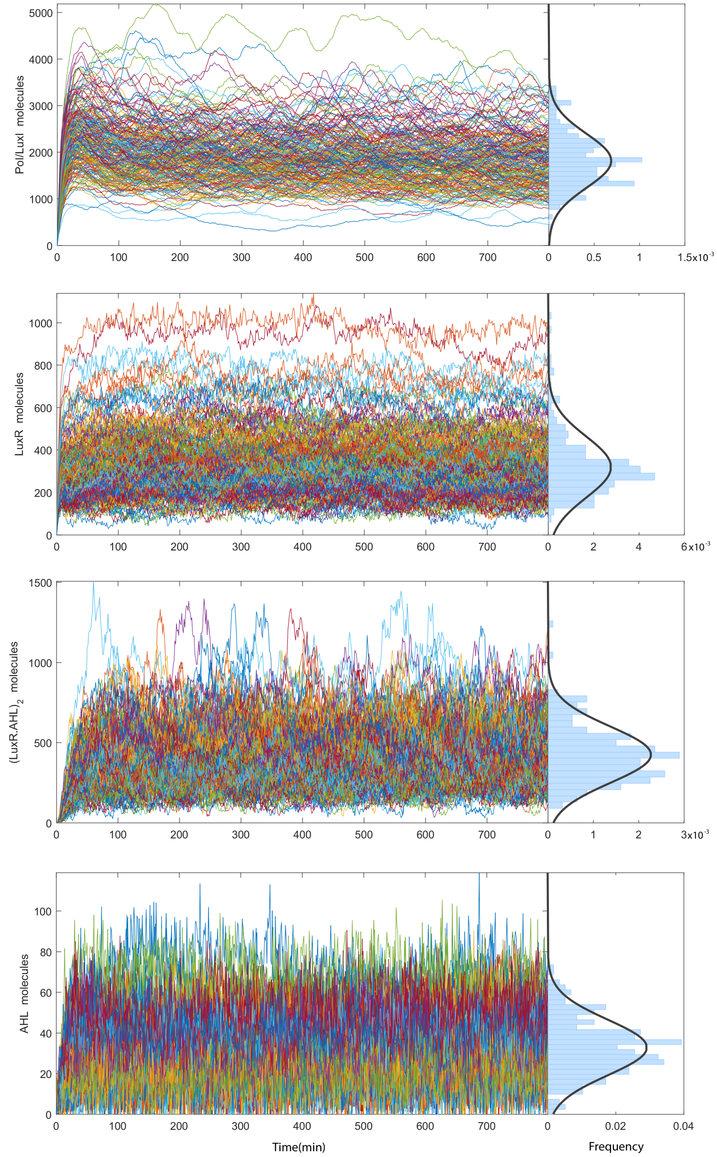


Figure 4.4. Stochastic simulation of the QS/Fb CLE-based model for a cell population. (Left) A single run computed over 800 minutes for cells for the four intracellular species, considering a population of 240 cells. A step time $\delta = 2.5 \times 10^{-3}$ seconds and the same parameters of Table 4.5 were used. **(Right)** Normalized endpoint histograms for the whole population with their corresponding probability density functions (solid line).

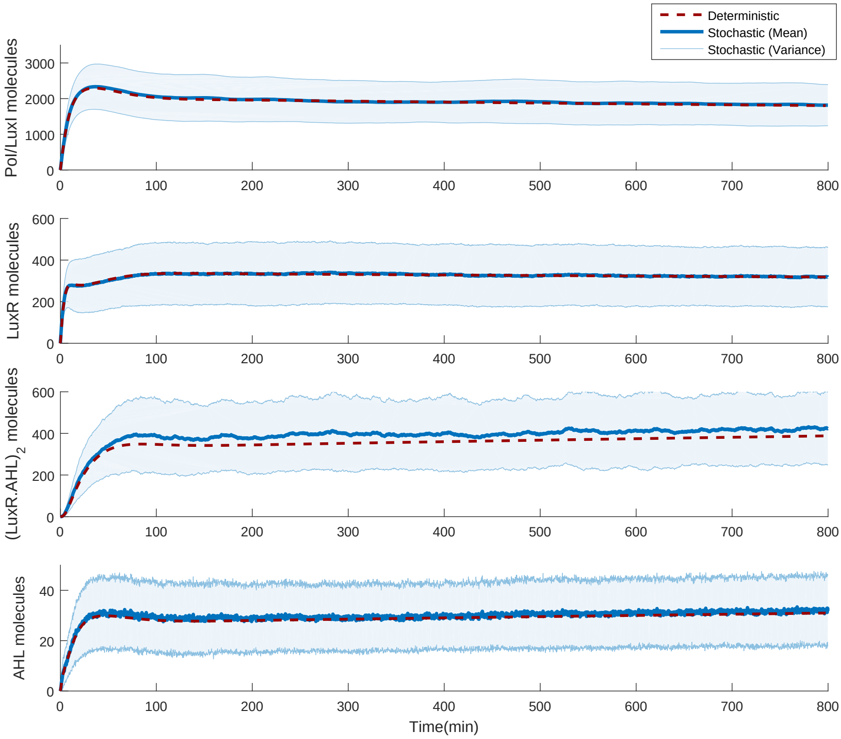


Figure 4.5. Comparison of stochastic and deterministic results. A single realization computed over 800 minutes for the four intracellular species of the QS/Fb circuit, considering a population of 240 cells. The stochastic (solid line) and deterministic (dashed line) are two independent simulations but under the same conditions. In both simulations, the average molecules number of each species closely match.

NoQS/NoFb CLE-based model

The term $\mathrm{k_A}$ in the propensity vector (4.45) corresponds to the synthesis rate of AHL by LuxI protein (n_1^i). LuxI protein is not present in the NoQS/NoFb circuit as was described in sections 3.3 and 4.3.2. Therefore, the value of the rate $\mathrm{k_A}$ in the NoQS/NoFb CLE-based model is set to zero $(\mathrm{k_A} = 0)$, and the initial condition of $\mathrm{AHL_{ext}}$ $n_5(0)$ will be changed to model pulse-like (bolus) addition of external AHL in the medium. This slight difference between both QS/Fb and NoQS/NoFb models will be used in the next Chapter 5 for system identification.

Validation of the non-linear propensities

In many cases the use of higher-order terms in stochastic simulation is indeed justified for the stochastic modelling and computational implementation (Cao et al., 2005; Rao and Arkin, 2003). Usually, stochastic algorithms treat all the reaction events alike. They will spend the great majority of their time simulating the many relatively uninteresting fast reaction events than explicitly simulate only the slow reactions.

In biological systems, it is common that fast and slow reactions share some species. For instance, slow reactions dependency on the fast ones is mathematically approximated using different approaches e.g. QSSA, species invariance, or deterministic reaction-rate equation. The first two methods have been used in this Thesis. Then, these approaches try to treat the new rational slow reaction stochastically, and consequently their approximations lead to the appearance of higher-order terms known as propensity functions.

In the QS/Fb case, there is one lumped propensity function derived from the CLE-based model (4.42) and its set of reactions (4.40): the Hill-like function $f(n_3, t)$ for the PoI/LuxI repression effect. A similar framework has already been used in (Woods et al., 2016). This high-order propensity function was validated by simulating the pseudo-reaction associated to $f(n_3, t)$ using CLE, and then comparing this result with the one obtained by simulating the set of corresponding original reactions using Gillespie's direct method SSA.

Particularly denoting n_3^i : [(R.A)$_2$] for the *i*-th cell, the Hill-like function $f(n_3^i, t) \triangleq \frac{\mathrm{C_I p_I}}{\mathrm{dm_I}} \left(\frac{\mathrm{k_{dlux}} + \alpha_I n_3^i}{\mathrm{k_{dlux}} + n_3^i} \right)$ represents how the dimer (LuxR.AHL)$_2$ inhibits transcription of DNA from the corresponding gene *pol/luxI* into messenger RNA of *pol/luxI* (mPI), which in turn is translated into protein PoI/LuxI (PI). Recalling the fast reactions (3.4) involving messenger mPI

$$\begin{aligned} \mathrm{gPI} &\xrightarrow{\mathrm{C_I}} \mathrm{gPI} + \mathrm{mPI} \\ \mathrm{gPI} + (\mathrm{R}\cdot\mathrm{A})_2 &\xrightleftharpoons[\mathrm{k_{lux}}]{\frac{\mathrm{k_{lux}}}{\mathrm{k_{dlux}}}} \mathrm{gPI}\cdot(\mathrm{R}\cdot\mathrm{A})_2 \\ \mathrm{gPI}\cdot(\mathrm{R}\cdot\mathrm{A})_2 &\xrightarrow{\alpha\mathrm{C_I}} \mathrm{gPI}\cdot(\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{mPI} \\ \mathrm{mPI} &\xrightarrow{\mathrm{dm_I}} \emptyset \end{aligned} \tag{4.46}$$

They were approximated into two equivalent reactions as (4.40) describes

$$\begin{aligned} (\mathrm{R}\cdot\mathrm{A})_2 &\xrightarrow{f(n_3,t)} (\mathrm{R}\cdot\mathrm{A})_2 + \mathrm{mPI} \\ \mathrm{mPI} &\xrightarrow{\mathrm{dm_I}} \emptyset \end{aligned} \tag{4.47}$$

where $f(n_3, t)$ describes the time evolution of mPI in the same way than in (4.46).

To validate the propensity function $f(n_3, t)$, both set of reactions were simulated. For one single-cell $(i = 1)$ and with the same conditions, reactions (4.47) were ran using the CLE, and reactions (4.46) were simulated using the Gillespie direct method (SSA).

Figure 4.6 depicts, for one realization, how the SSA trajectory (left-top) matches very well with the CLE trajectory (right-top) during the whole simulation time. Both SSA and CLE trajectories have similar distributions with small differences between their first statistical moments ($\mu_{SSA} \approx \mu_{CLE}$, and $\sigma_{SSA} \approx \sigma_{CLE}$) (see Fig.4.6B). It can be seen that the noise strength of $\mathrm{mRNA_{poI/luxI}}$ for the SSA distribution ($\eta^2_{SSA} = 0.008$) matches closely with the same for the CLE ($\eta^2_{CLE} = 0.0072$).

Finally, figure 4.6C shows the Box-and-Whisker plots of messenger RNA of *poI/luxI* SSA and CLE realizations. Their medians (red line) are practically the same, and the Kruskal-Wallis test (Kruskal and Wallis, 1952) reveals that there is no statistically significant difference between their medians at the 95.0 % confidence level ($[test\ statistic,\, p - value] = [-2.09067 \times 10^6,\, 1.0]$).

Population density effect

The CLE-based modeling approach for cell populations allow us to analyze synthetic gene circuits with cell-to-cell communication. For the QS/Fb circuit, quorum sensing is the signaling language among cells to send information and achieve its specific function. As a consequence, defining the amount of interconnected microorganisms in a culture is a key parameter to be discuss.

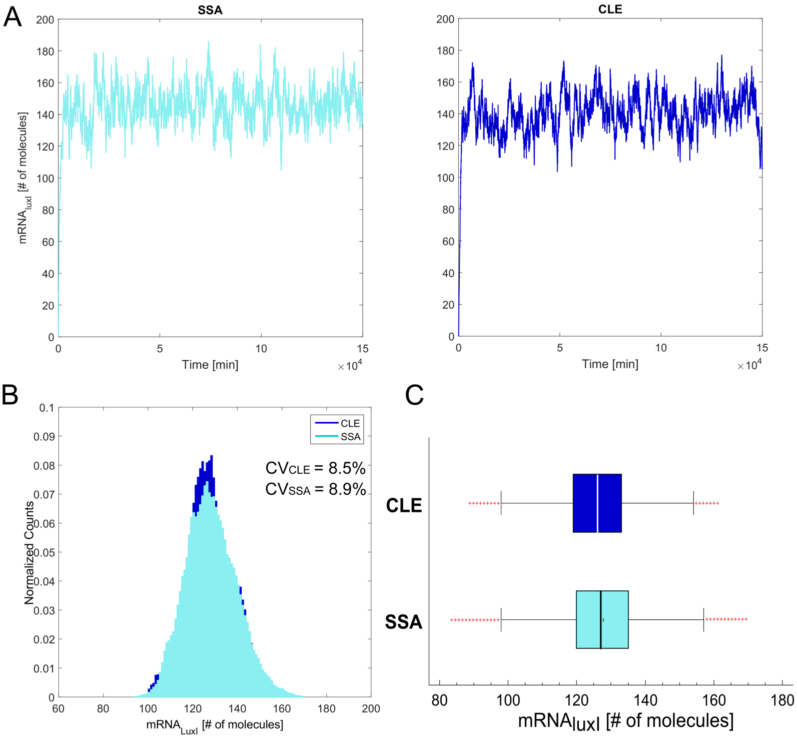


Figure 4.6. SSA and CLE comparison for a complex propensity. (A) One realization of $\mathrm{mRNA}_{\mathrm{luxI}}$ made using the SSA (cyan color) and the CLE (blue color) respectively. Both trajectories match during a large temporal window (15×10^4 min). (B) Histograms with similar means and covariances. (C) Two medians $\widetilde{\mathrm{n}1}_{\mathrm{SSA}} = 127.7$ and $\widetilde{\mathrm{n}1}_{\mathrm{CLE}} = 126.1$ molecules are statistically the same in the Box-and-Whisker plots.

The optical density (OD) is a common method for estimating the concentration of bacterial or other cells in a liquid culture (Sutton, 2006). Typically, the OD of a cell sample is measured at a wavelength of 600 nm, and it is denoted as OD$_{600}$. The OD (adim) of a cell culture depends on the number of cells, and the volume of that culture. In the CLE-based model simulations, the number of N cells were changed in order to obtain different OD values following the equation

$$\mathrm{OD} = \mathrm{N}\frac{1}{\mathrm{V_{ext}}} * \frac{1}{\mathrm{N_{OD=1}}} \tag{4.48}$$

where N is the number of cells (N = 240 bringing the OD to 0.3), $\mathrm{V_{ext}} = 1 \times 10^{-3}$ μL, N$= 8 \times 10^5$ is the quantity of cells contained in 1 μL of bacterial culture when the OD is 1 (*Source:* **Agilent**, *E. coli* Cell Culture Concentration from OD$_{600}$ Calculator).

In order to see whether quorum sensing effect on our circuit depends on the cell density, the OD was changed as a function of the number of cells and the volume. Figure 4.7A shows the PoI/LuxI noise strength obtained at different values of OD ranging from 0.005 to 5. First, the number of cells was kept constant (N=240 cells), and the culture volume $\mathrm{V_{ext}}$ was changed from 0.06 to 0.0003 μL (blue squares). The OD ratio is tabulated in Table 4.7. Next, we changed the cell number N and the external volume $\mathrm{V_{ext}}$ simultaneously, so as to have volumes in more realistic range for microfluidic settings (green squares). Their values (see Table 4.7) were chosen trying to keep the same cell densities as in the first case.

Table 4.7. OD changing the cell number and volume.

Cell number fixed						
N (cells)	240	240	240	240	240	240
$\mathrm{V_{ext}}$ (μL)	0.06	0.03	0.006	0.003	0.0006	0.0003
OD$_{600}$	0.005	0.01	0.05	0.1	0.5	1
Cell number and external volume are variable						
N (cells)	240	240	1200	2400	4800	12000
$\mathrm{V_{ext}}$ (μL)	0.03	0.006	0.015	0.006	0.006	0.003
OD$_{600}$	0.01	0.05	0.1	0.5	1	5

Table 4.8. OD fixed.

N (cells)	240	1200	2400	4800	12000
$\mathrm{V_{ext}}$ (μL)	0.001	0.005	0.01	0.02	0.05
OD$_{600}$	0.3	0.3	0.3	0.3	0.3

Moreover, to evaluate how representative of a cell population is a simulation with N = 240 cells, the number of cells and the volume were changed to achieve a constant cell density at different cell numbers. The cell numbers and volumes used in this case are in the intervals: N = $[240,\ 12000]$ cells and $\mathrm{V_{ext}} = [0.001,\ 0.05]\,\mu$L (see Table 4.8). Figure 4.7B shows the PoI/LuxI noise strength for different values of N ranging from 240 cells to 12000 cells. In all cases PoI/LuxI noise strength did not appreciably change.

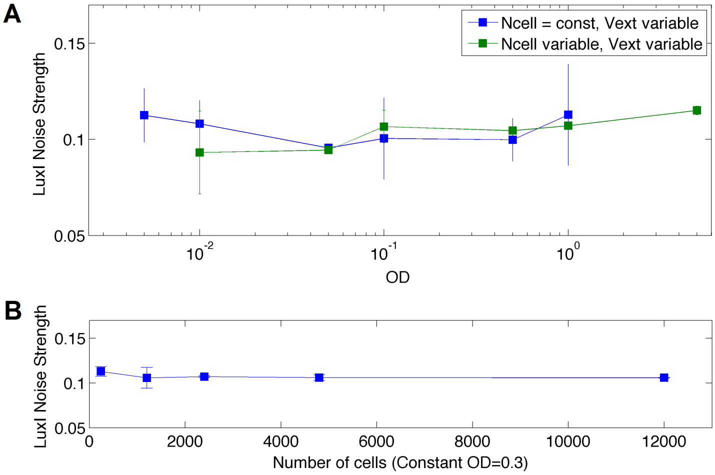


Figure 4.7. LuxI noise strength comparison at different OD$_{600}$ values. (A) LuxI noise strength does not appreciably change for different OD= [0.005, 0.01, 0.05, 0.1, 0.5, 1, 5], obtained either changing only the volume and keeping the cell number constant in N=240 (blue squares) or when changing both the cells number together with the volume (green squares). (B) LuxI noise strength for different number of cells and volume, but keeping constant OD$_{600}$ $= 0.3$.

4.4.2 Efficient stochastic simulation

This subsection revolves around four different improvements of the stochastic simulation for the CLE-based model for a cell population in synthetic gene circuits. Using the QS/Fb circuit, the improvements include: *i)* selection of the simulation step time, *ii)* defining the number of realizations for each simulation, *iii)* decimation, and *iv)* setting simulations using a particles simulation method.

Time step selection

Probability theory says that a Poisson random variable with large mean is well approximated by a normal random variable with the same mean and variance. For instance, if *every* reaction is expected to fire many times over $[t, t+\delta t)$, its corresponding propensity would change from Poisson to normal. Section 2.4.5 pointed out that a reaction firing has a Poisson distribution probability with mean $X_j \delta t$ if we consider a short time interval.

The requirement of δt being small enough assumes a *constant propensity* during the interval $[0, T]$. It is the first condition required to use CLE approximation. The second condition demands δt to be large enough so that the expected number of occurrences of each reaction in $[t, t+\delta t)$ be much larger than 1. Even though both conditions obviously present a trade-off, they can be simultaneously satisfied by having large molecular population numbers (Gillespie, 2000).

In this Thesis, the discretized Euler-Maruyama paths were specifically computed to generate the increments δt needed in the CLE-based model (4.42). The resulting step time is $\delta t = 2.5$ ms. This step size was selected as the largest one that ensures the stability and convergence of the simulation (Higham, 2001). In turn, initial conditions for proteins PoI/LuxI, LuxR, (LuxR.AHL)$_2$, and the inducers AHL and $\mathrm{AHL_{ext}}$ were defined as at least double than the maximum molecules number involved in one reaction.

Number of realizations

For stochastic simulations, it is necessary quantify the number of CLE realizations or CLE runs. Thus, it was analyzed if one realization of the CLE-based model for the whole population is *enough* to characterize the long-term statistics such as mean, variance or the noise strength of each modeled species.

Three experiments where performed with the same set of parameters and conditions as in section 4.4.1. First, only the steady-state portion of CLE simulations for every cell in each experiment were selected. Then, these portions were performed on time average, resulting in an *averaged* number of molecules for each species in the *i*-th cell of the population. In Fig.4.8 is possible to see the matrix scatter plot of the three realizations together. It is evident that each one of the four species long-term distributions is unimodal and well shaped.

A Multivariable Analysis of Variance (MANOVA) (Hair and Suárez, 1999) analysis on the three realizations of the four species together was performed. The MANOVA results reflect no statistically significance to reject the hypothesis of the three realizations having the same mean and variance, with $p-value = [0.1374, 0.7403]$, Wilk's lambda $\lambda = [0.9829, 0.9983]$, and the Mahalanobis distance between the means resulted in $[0.0106, 0.0748, 0.0700]$. These results justify that one realization of the population of N interconnected cells provided with enough time to perform the time average, is useful to obtain representatives of the long-term moments of the population.

Analyzing the QS/Fb output protein PoI/LuxI when quorum sensing (QS) occurs, another three experiments were ran. There are three CLE realizations under the same parameters and conditions for two different circuits: QS/Fb and NoQS/NoFb. As above, the steady-state portion (120 last samples) of the CLE simulation of every *i*-th cell in each experiment was chosen and performed the time average, resulting in an *averaged* number of molecules for protein PoI/LuxI in each cell. The obtained distributions of protein PoI/LuxI (three for each circuit) are shown in Fig.4.9. Again, these long-term distributions for each NoQS/NoFb and QS/Fb circuit (Fig.4.9 top and bottom, respectively) are unimodal, well shaped, and match closely between them.

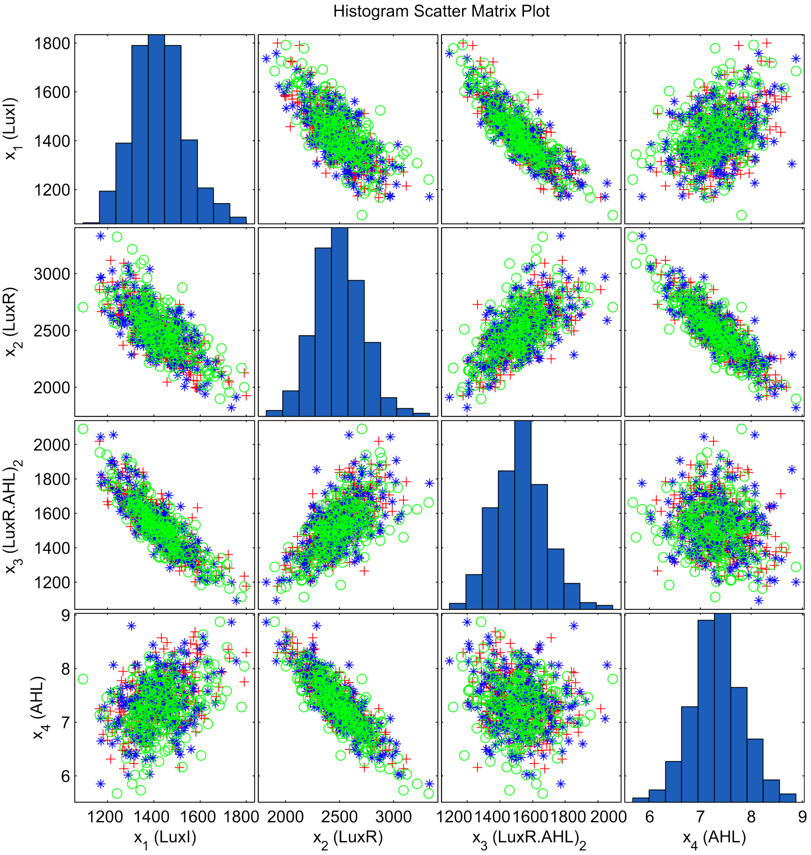


Figure 4.8. Several realizations. Scatter matrix plot for the four intracellular species depicts the agreement of three realizations for the CLE-based model (first realization one in blue *, second in red + , and third in green o).

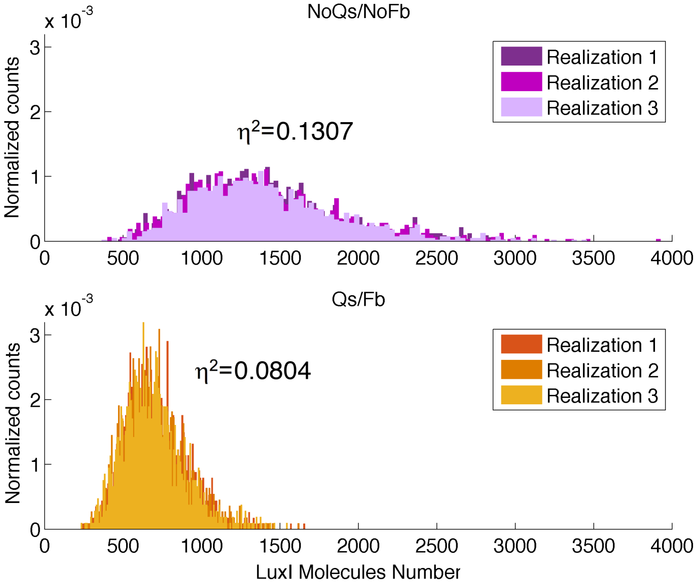


Figure 4.9. Three different realizations for the QS/Fb and NoQS/NoFb circuits. The Kruskal-Wallis test shows there is not a statistically significant difference amongst the medians of the 3 different realizations (under the same conditions) for both NoQS/NoFb (top) and QS/Fb (bottom) circuits.

For the NoQs/NoFb circuit in Fig.4.9 (top), the results of the Kruskal-Wallis analysis on the three realizations demonstrate no statistically significance to reject the hypothesis of the three realizations having the same Pol/LuxI median and average noise strength $\eta^2 = 0.1307$, with $[test\ statistic, p - value] = [0.0018, 0.9991]$. The same conclusion is shown in Fig.4.9 (bottom) for the QS/Fb network with $[test\ statistic, p - value] = [0.0006, 0.980714]$. Since the p-value is greater than or equal to 0.05, there is not a statistically significant difference amongst the medians at the 95.0 % confidence level.

Decimation

One realization for a population of 240 cells as in section 4.4.1, and the first moments were evaluated for each of the species under decimation of the signals. In this case there is only one realization. We are taking into advantage the ergodicity of the gene expression process for only one cell. So, the moments were calculated using the temporal average and quadratic error to estimate the mean and variance of the process (Gupta et al., 2014). The sampling rate of the signals was decreased following an iterative decimation process. In every iteration we reduced the sampling rate to the half of its previous value, obtaining in this way a new signal with half number of samples and double time between samples.

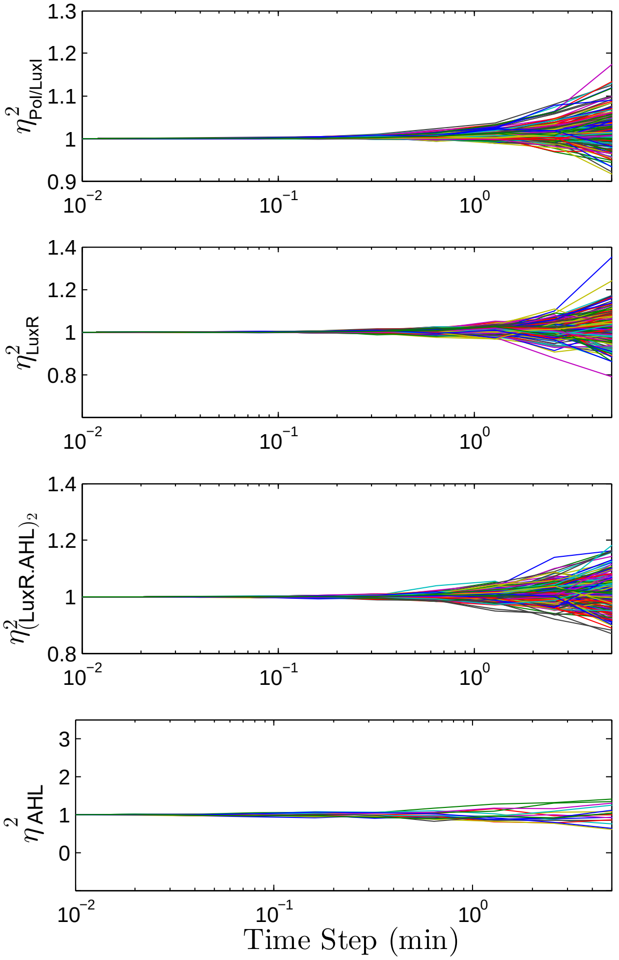


Figure 4.10. Relationship between the noise strength of each species after decimation of the signals. The noise strength starts damaging after the time step $\delta t = 0.2$ min, specially at low number molecules as in the AHL case.

Figure 4.10 plots how the decimation affects the noise strength measurement in all the species of the QS/Fb circuit. Degradation of the moments after decimation starts to occur when increasing the step time δ (or decreasing the sample rate $1/\delta$) by two orders of magnitude in the stochastic simulations. Notice the species where degradation starts early is the one with smaller mean (AHL).

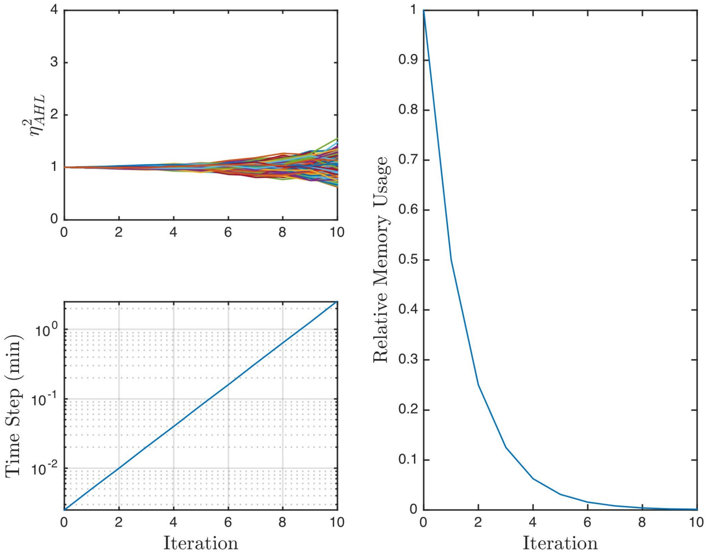


Figure 4.11. Decimation analysis for one species of the QS/Fb circuit. (Left top) Decimation effect in one species noise (η^2_{AHL}) *vs.* iteration number. **(Left bottom)** Time step increases its value as function of the iteration number. **(Rigth)** The use of memory for saving simulation data decreases when the iteration number increases.

Figure 4.11 shows how the use of memory decreases when the number of samples is reduced by the decimation process in every iteration. From this result is possible to see that, e.g. decimation to a time step of 10^{-1} implies a reduction of the required memory space to the 5% of the original. These improvements allows us to manage the simulation results in a more easier and flexible way, when we deal with *in silico* experiments including cells population like the QS/Fb circuit.

Particles simulation with OPEN FPM

In this Thesis, the CLE-based model (4.42) was numerically solved using the Euler-Maruyama discretization generating stochastic sample paths corresponding to every CLE, where each fluctuation is a scalar independent Brownian motion (Higham, 2008). The implementation of the Euler-Maruyama algorithm, of the QS/Fb circuit with a population of N cells, was done in C++ using OpenFPM[2]: a scalable open framework for particle and particle-mesh codes on parallel computers (Incardona et al., 2018). OpenFPM is a new improved version of the framework PPM (Sbalzarini et al., 2006) allowing efficient parallel computation. The framework provides a transparent and scalable infrastructure for shared-memory and distributed-memory implementations

[2] Software available at `http://openfpm.mpi-cbg.de/`

of particle and mesh-particles algorithms. Among other features, it has a number of abstractions including parametric data structures (particles and meshes), domain decomposition strategies of the physical domain into several sub-domains to distribute the work load between the different calculation nodes (processors), mappings, dynamic load-balancing and iterators together with file I/O.

The implementation of any algorithm is transparent, meaning that it is the same work to code an algorithm for a single core, than for a multi-processors workstation or even a cluster using always the same sets of data structures and iterations. All this allowed us to use this implementation within an optimization design, where the objective function includes the simulation of the circuit and has to be evaluated thousands of times. For N$= 240$ cells and $\delta t = 2.5 \times 10^{-3}$ sec, and during 400 min, the load of computational cost is $\approx$ 7 sec per simulation of one set of model parameters (one evaluation of the cost function). This, compared with the previous Matlab implementation, is 500-fold faster (Matlab implementation took around 2 hours).

4.5 Summary

As we have seen to analyze how our genetic circuit affects intrinsic and extrinsic noise, we needed an appropriate model and a computationally efficient simulation. For either deterministic or stochastic models, both aspects are intertwined. This chapter aimed at obtaining a reduced model more amenable for computational analysis, but avoiding excessive reduction that would lead to lack of biological relevance. For both the I1-FFL and the QS/Fb circuits, the model reduction process started from semi-mechanistic biological models based on first-principles with high dimension and a large number of parameters. The resulting reduced nonlinear models present incomplete parameter identifiability, so that many parameter combinations could fit the data equally well. We will deal with this in Chapter 5, where the type of measurements collected from a gene circuit can and should impact the choice of model for the system being studied.

Chapter 5

Model parameter estimation

5.1 Introduction

Identification of model parameters is an established problem in control systems technology. The new uprising of synthetic biology complexity, together with mathematical models and several reduction techniques associated to this kind of systems have revived the problem. Now, synthetic biology is reaching the situation where traditional approaches from systems theory and identification of model parameters no longer fulfill the needs of the field, due to the complexity and nonlinear character of the gene circuits being designed.

Chapter 4 exposed how to obtain a deterministic or stochastic model of a synthetic gene circuit at a single-cell or a population levels. However, parameter estimation of these models remains as a challenging issue because system identification is highly dependent on the collected data, and they can typically be measured for only a few outputs at limited time resolution. Furthermore, the interest in finding parameter values is that a well-characterized mathematical model of a gene circuit can be used to accurately explain the system behaviour, and design e.g. *novo* feedback controllers as we will see in the next Chapter 6.

The problem of parameter identification, that is, the indirect determination of the unknown parameters from measurements of other quantities, is a key issue in computational and systems and synthetic biology (Lillacci and Khammash, 2010). Some model parameters like binding and unbinding rates, or production and degradation coefficients have a physical meaning. Other parameters are lumped arising from model reductions and/or approximations. This makes more difficult the biological interpretation of their values. Nevertheless, generally both kind of **parameters have unknown values** for a particular model. Accurate parameter identification is crucial whenever one wants to obtain quantitative, or even qualitative information from the models. Recently, much attention has been given to this problem in the systems biology community, either building experimental platforms to obtain better data (Fiore et al., 2013), or using optimization techniques such as linear and nonlinear least squares (Mendes and Kell, 1998), genetic algorithms (Srinivas and Patnaik, 1994), and evolutionary computation (Ashyraliyev et al., 2008; Moles et al., 2003). Evolutionary computation is one of the suggested optimization techniques for the large parameter estimation problems present in systems and synthetic biology.

Structural identification approaches of synthetic gene circuits have been proposed as a first step of parameter estimation (Cinquemani, 2017; Porreca et al., 2008). Authors in (Chis et al., 2011) suggest the lack of identifiability is related to the structure of the model, i.e. the system dynamics plus the observation function. So, methods including local analyses are based on the computation of local sensitivities, the Fisher Information Matrix, the covariance matrix, or the Hessian of the least-squares function (Rodriguez-Fernandez et al., 2006; Srinath and Gunawan, 2010) have been proposed. Nevertheless, the problem becomes especially hard in models where the ratio between the number of observables and the number of parameters is low, or when complicated nonlinear terms, such as Michaelis-Menten or Hill kinetics, are present (Balsa-Canto et al., 2010). This means that for both the I1-FFL and the QS/Fb models already developed in Chapter 4, the structural identifiability shows limited operation modes. That is, because both circuits have nonlinearities such as the monomer evolution, and Hill functions to model the promoters dynamics. Hence, it will be impossible to compute a unique value for the parameters independently of the available experimental data.

Estimating parameters in nonlinear dynamic models remains a very challenging inverse problem due to its nonconvexity, and ill-conditioning caused by over-parametrization, experimental measurement errors, data scarcity and uncertainty (Gábor and Banga, 2015; Kaltenbach et al., 2009). Moreover, for nonlinear models, the amount of information collected from an experiment may strongly depend on the true value of the parameters (Pronzato and Pázman, 2013). Thus, parameter identification has been mostly addressed by optimizing the weighted combination of different prediction errors to obtain a single solution. Typically, this single-objective approach has demonstrated be inadequate to address often problems found in gene networks: *i)* the lack of identifiability of some of them, *ii)* multimodal circuits with separated parts of parameter space providing adequate fits to the experimental data, and *iii)* when the parameter selection must somehow trade-off model fit against model complexity, or against extra desired objectives. To address these situations, in this Thesis parameter identification is approached using a **Multi-objective optimization design (MOOD)** (see Fig.5.1) including a *global* multi-objective evolutionary algorithm, and a multi-criteria decision making strategy to select the most suitable solutions. One of the more important applications of optimization is to use it for parameter estimation, so-called inverse problem. That is, given a set of experimental data, calibrate the model so as to reproduce the experimental results in the best possible way (Moles et al., 2003). As said in section 2.5, in the Multi-objective optimization (MO) all objectives are important. Therefore all of them are optimized simultaneously to obtain a set of the best solutions called the *Pareto Front* (not a unique solution).

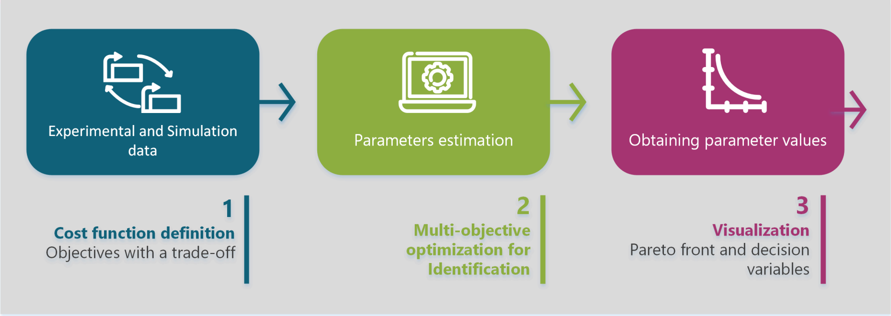

Experimental and Simulation data
Parameters estimation
Obtaining parameter values
1
Cost function definition
Objectives with a trade-off
2
Multi-objective optimization for Identification
3
Visualization
Pareto front and decision variables

Figure 5.1. MOOD framework for model parameter estimation of synthetic gene circuits.

The first case to show the applicability of this novel methodology is the **MO-based identification** of the I1-FFL model parameters that present a trade-off between different experimental scenarios. As a result, ensembles of local models with/without different parameters appropriate for all experimental scenarios describe the I1-FFL circuit's dynamics when the input is changing. The MOOD framework uses a global multi-objective evolutionary algorithm, and a multi-criteria decision making strategy to select the most suitable solutions. It finds an approximation to the Pareto optimal set of model parameters that correspond to each experimental scenario. Thus, the MO-based identification can fully harness parameter estimation to ensemble local models for gene circuits. These results have been published in

- Y. Boada, A. Vignoni, G. Reynoso-Meza, and J. Picó. Parameter identification in synthetic biological circuits using multi-objective optimization. volume 49, pages 77 – 82, 2016c. Foundations of Systems Biology in Engineering FOSBE.

In the second case, the QS/Fb model parameters were estimated using the same methodology. To enhance the information content of the measurable variables, we first identify an open-loop version of the circuit (NoQS/NoFb circuit) using averaged time-course experimental data obtained from plate readers. Then, we use steady-state stochastic distributions provided by flow cytometry to identify the remaining feedback gain in the QS/Fb circuit. The MO-based identification gives good identification results for ensemble models and it is particularly useful for easily combining both experimental flow cytometry with experimental plate reader data. Part of these results have been published in the articles

- Y. Boada, A. Vignoni, and J. Picó. Multi-objective identification of synthetic circuits stochastic models using flow cytometry data. *Proceedings 25th Me-*

diterranean Conference on Control and Automation MED, pages 1077–1082, 2017c.

- Y. Boada, A. Vignoni, and J. Picó. Model reduction and multi-objective identification of a feedback synthetic gene circuit. *IEEE Transactions on Control Systems Technology*.

The outline of this Chapter goes as follows: The first section 5.2 describes the use of *local models* as a result of the MO-based identification framework of the I1-FFL model parameters. In the next section 5.3, the MO-based stochastic parameter estimation for the QS/Fb model using diverse nature of experimental data is presented. At the end of this Chapter, section 5.4 present some important remarks.

5.2 Using local models to identify the I1-FFL circuit

One of the main problems associated with standard optimization methods for performing parameter estimation, is that they may not perform well in the case of significant difference in the system response to different inputs. One of the reason is related to the large-order of mechanistic models involving several variables, and used as a start point to explain the systems's behavior. As we saw in sections 4.2.2 and 4.3.2, if the order model was reduced, it would not entirely capture the system dynamics previously modeled. However, reduction process is especially needed to perform model parameter estimation using experimental data of a limited number of model variables. Additionally, most identification methods rely on single-objective optimization, and try to find only one solution (i.e. only one value for each parameter)–that is, the *best fit*. This best solution can be good for one set of experiments and bad for others, or it can be acceptable for all the experiments but not really good for any one.

Several approaches have been proposed to tackle these problems. Among them, **ensembles of local models** have received much attention in the last years, when a single set of parameters is not appropriate for all experimental scenarios. In (Steuer et al., 2006), local linear models at each point in parameter space where used to circumvent lack of knowledge about the structure of kinetics by a parametric representation of the Jacobian matrix. Then, the authors used the ensemble of models to elucidate the parameter regions associated with experimentally observed specific dynamical behaviors. A similar approach was used by (Samee et al., 2015). Ensembles of models, i.e. sets of models with different structures and/or parameter values have also been used in (Villaverde et al., 2015), where the final prediction is obtained from a consensus one among the models.

Here, the multi-objective optimization design (MOOD) to perform parameter identification leading to nonlinear local models of gene circuits is proposed to carry out parameter estimation using local models for the **I1-FFL circuit**. Figure 5.2 recalls the main species involved in the circuit and their roles. But, active green fluorescent protein (aGFP) is the output taking into account in this section.

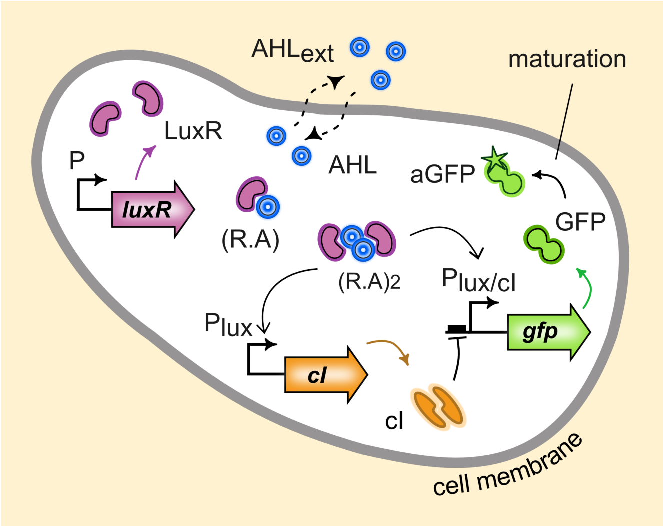


Figure 5.2. Representation of a I1-FFL synthetic circuit incorporating the maturation process of GFP. Active GFP (aGFP) is the system output, whose fluorescence is measured in the lab.

5.2.1 Obtaining experimental data

The first step of parameter estimation is collected enough experimental data from the system behavior. The I1-FFL gene circuit has two plasdmis Therefore, all plasmids mentioned in this Chapter and the next ones were entirely built in this Thesis, and they can be found in section B.

Following the implementation in (Basu et al., 2004), the I1-FFL circuit was engineered and implemented in the laboratory using two plasmids already described in section 3.3.2. Their components were taken from the Lux operon in the *V. fisheri* (Kaplan and Greenberg, 1985) quorum sensing system, the lambda cI repressor promoter and a green fluorescent protein (GFP) as a reporter.

The next temporal experiments were performed. *E. coli*cells (Top 10, NEB) carrying the pCB11a and pCB16mut plasmids were grown overnight in Luria-Bertani (LB)[1] medium with the appropriate antibiotics. Then, 96 well-plates were inoculated at an optical density $OD_{600} \approx 0.025$ and incubated to reach an $OD_{600} \approx 0.2$. At this point, selected wells were induced with appropriate concentrations of AHL (N-3-Oxohexanoyl-

[1] LB broth is the most widely used medium for the growth of bacteria (Bertani, et al., 1951).

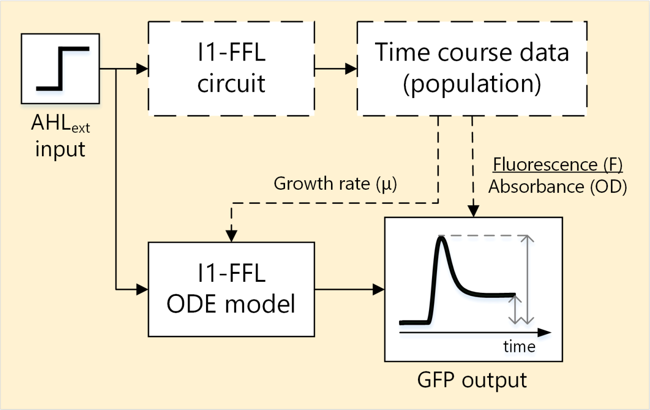


Figure 5.3. Procedure for I1-FFL identification using experimental (dashed line) and simulation (solid line) data.

L-homoserine lactone, Santa Cruz Biotecnology Catalog Number SC205396) and incubated for 200 minutes.

From each well, we obtained bulk temporal data (see section 2.1.4) for fluorescence (F) and absorbance (OD). These data were also post-processed following the procedure detailed in section 2.1.4. Assuming that F is proportional to the amount of GFP per unit of volume, and OD to the number of cells per unit of volume (Leveau and Lindow, 2001), then the quotient F/OD (FOD, in RLU per OD unit) is the cell density-normalized fluorescence, and it is proportional to the amount of GFP per cell. For example, the FOD ratio for the I1-FFL gene circuit was directly compared with the GFP output from the I1-FFL ODE model during the MO-based identification as in Fig.5.3. Additionally, the growth rate (μ) min^{-1} deduced from the OD data will be used for the I1-FFL model to compute the expanding number of cells in the culture, as it was described in section 4.2.

For instance, both OD and F measurements were taken with a POLARstar Omega plate reader (BMG Labtech GmBh) with the following protocol: 2 min shaking, OD measurement, then 15 sec pause, and Fluorescent measurement. Each condition was performed in 4 replicates (samples under the same conditions) for 2 different days, making a total of 8 data sets for each condition. In summary, each experiment has 5 conditions (ranging from 0 nM to 55 nM AHL), 8 data sets, absorbance and fluorescence measurements every $\triangle t \approx 5$ min during 200 min of incubation after AHL induction. These experimental

Table 5.1. Variables of the I1-FFL model including maturation.

Variable	Description	Units
x_1	*luxR* messenger RNA	nM
x_2	LuxR protein	nM
x_3	AHL intracellular inducer	nM
M	LuxR and AHL monomer (LuxR.AHL)	nM
x_4	Dimer of (LuxR.AHL)	nM
x_5	*cI* messenger RNA	nM
x_6	cI protein	nM
x_7	Dimer of cI	nM
x_8	*gfp* messenger RNA	nM
x_9	GFP protein	nM
x_{10}	Active GFP protein	nM
x_{11}	AHL extracellular inducer	nM
x_{12}	Number of cells in the culture	cells

5.2.2 I1-FFL extended model

As a second step of parameter estimation, the ODE model (4.27) was adapted to closely relate simulation results with the obtained experimental fluorescence and absorbance data. For the I1-FFL circuit, the new model will quantify the output GFP fluorescence emitted from individual bacteria in the growing population as a function of the input signal $\mathrm{AHL_{ext}}$. This leads to extend the model (4.27) with two additional states: the *mature GFP* concentration, and the *number of cells* of the culture.

Once a fluorescent protein is expressed, it must be fold into its fluorescent state in a process known as **maturation** (Miyawaki et al., 2003; Lippincott-Schwartz et al., 2001). Maturation can take different times depending on the fluorescent protein, its chemical structure and its efficiency. Maturation times range from 20 (lab-enhanced proteins) to 120 min (wild-type proteins). Maturation can be modeled as a first-order reaction (see section 2.3), where its rate can be calculated as ln(2) divided by the time constant of the fluorescent protein maturation, $\mathrm{k_{mat}}$.

On the other hand, the cells population growth among time in a realistic environment is subject to constraints like lack of nutrients or oxygen that eventually prevent exponential growth. Hence, the **number of cells** (the population size x) considering a maximum carrying capacity is given by the logistic equation $\dot{x} = \mu x \left(1 - x/\mathrm{K_{max}}\right)$. μ is the specific growth rate, and $\mathrm{K_{max}}$ is the maximum growth capacity for that particular population (Nowak, 2006).

Including GFP maturation and population size leads to the ODE extended model (5.1) that describes the I1-FFL dynamics using 12 states: messengers RNA with their corresponding proteins mLuxR and LuxR, mcI and cI, mGFP and GFP with its mature form aGFP, the input $\mathrm{AHL_{ext}}$, the intracellular AHL, the transcription factors (LuxR.AHL)$_2$ and (cI)$_2$, and finally the number of cells in the culture (see Table 5.1).

$$
\begin{aligned}
\dot{x}_1 &= \mathrm{k_{mR}}C_R - \mathrm{d_{mR}}x_1 - \mu x_1 \\
\dot{x}_2 &= \mathrm{k_{pR}}x_1 - \mathrm{k_2}x_2x_3 + \mathrm{k_{-2}}M - \mathrm{d_R}x_2 - \mu x_2 \\
\dot{x}_3 &= -\mathrm{k_2}x_2x_3 + k_{-2}M + \mathrm{k_d}(x_9 - x_3) - \mathrm{d_A}x_3 - \mu x_3 \\
\dot{x}_4 &= \mathrm{k_3}M^2 - \mathrm{k_{-3}}x_4 - \mu x_4 \\
\dot{x}_5 &= \mathrm{k_{mI}C_I}\frac{x_4}{\gamma_1 + x_4} - \mathrm{d_{mI}}x_5 - \mu x_5 \\
\dot{x}_6 &= \mathrm{k_{pI}}x_5 - 2\mathrm{k_4}{x_6}^2 + 2\mathrm{k_{-4}}x_7 - \mathrm{d_I}x_6 - \mu x_6 \\
\dot{x}_7 &= \mathrm{k_4}{x_6}^2 - \mathrm{k_{-4}}x_7 - \mu x_7 \\
\dot{x}_8 &= \mathrm{k_{mG}C_G}\frac{x_4 + \beta_1\gamma_2 + \beta_2\gamma_5x_4x_7}{\gamma_2 + \gamma_3x_4 + \gamma_4x_7 + \gamma_5x_4x_7} - \mathrm{d_{mG}}x_8 - \mu x_8 \\
\dot{x}_9 &= \mathrm{k_{pG}}x_8 - \frac{\ln 2}{\mathrm{k_{mat}}}x_9 - \mathrm{d_G}x_9 - \mu x_9 \\
\dot{x}_{10} &= \frac{\ln 2}{\mathrm{k_{mat}}}x_9 - \mathrm{d_G}x_{10} - \mu x_{10} \\
\dot{x}_{11} &= \mathrm{K_{cells}k_d}(x_3 - x_{11}) - \mathrm{d_{Ae}}x_{11} \\
\dot{x}_{12} &= \mu x_{12}\left(1 - \frac{x_{12}}{\mathrm{K_{max}}}\right) \\
M &= \frac{\mathrm{d_M} + \mathrm{k_{-2}} + \mu}{4\mathrm{k_3}} + \frac{\sqrt{(\mathrm{d_M} + \mathrm{k_{-2}} + \mu)^2 + 8\mathrm{k_3}(\mathrm{k_2}x_2x_3 + 2\mathrm{k_{-3}}x_4)}}{4\mathrm{k_3}}
\end{aligned}
\tag{5.1}
$$

where $\mathrm{K_{cells}} = \frac{\mathrm{V_{cell}} \cdot x_{12}}{\mathrm{V_{medium}}}$ is the ratio between cell population and medium. It is used to transform between extra and intracellular concentrations. Notice that the term for the leakiness of $\mathrm{P_{lux/cI}}$ ($\beta_1\gamma_4x_6$ in the model (4.27)) was eliminated from the mGFP dynamics ($\dot{x}_8$), because the experimental value β_1 is very low. Additionally, $\mathrm{V_{cell}} = 1 \times 10^{-15}\,L$ is the typical volume of an *E.coli* cell, and $\mathrm{V_{medium}} = 180\ \mu\mathrm{L}$ is the cell culture used in each 96-well plate reader in the experimental set up. The values for both the specific growth rate μ and the maximum growth capacity $\mathrm{K_{max}}$ were extracted from the corresponding OD bulk data.

The model (5.1) has 35 parameters enumerated in Table 5.2. Out of them, 18 are known from the literature and were kept fixed (top of Table 5.2). Parameter estimation was carried out to find values for the remaining model parameters (bottom of Table 5.2). The model (5.1) allows easier theoretical and computational analysis than the model (4.27). However, it will be difficult if not impossible that a single set of parameters will be appropriate for all experimental scenarios anymore. This is due to several reasons. One is practical identifiability from just one available measured variable (active fluorescent protein x_{10}). Most important, approximations done during the model reduction process and un-modeled dynamics may show up. The strongly nonlinear character of the model may amplify the effect of model structure mismatch for certain regions of the state space, and/or magnitudes of the input signal.

Table 5.2. Parameters of the I1-FFL model.

Fixed Parameter	**Description**	**Value**
$k_{m_{\text{luxR}}}$	luxR transcription rate	1 min^{-1}
$k_{p_{\text{luxR}}}$	LuxR translation rate	50 min^{-1}
k_d, k_{-d}	AHL diffusion rate	2 min^{-1}
k_4, k_{-4}	(cI)$_2$ association, dissociation rate	0.0009, 0.6 min^{-1}
C_{p1}	Plasmid pBR322 copy number	17
C_{p2}	Plasmid pACYC184 copy number	15
γ_2	Hybrid pLuxR/cI promoter coefficient	0.02 nM
$d_{m_{\text{luxR}}}, d_{m_{\text{cI}}}, d_{m_{\text{gfp}}}$	mRNAs degradation rates	0.23 min^{-1}
d_{LuxR}	LuxR degradation rate	0.0174 min^{-1}
$d_{\text{AHL}}, d_{\text{AHL}_\text{e}}$	AHL degradation rates	0.01 min^{-1}
d_M	Monomer degradation rate	0.0174 min^{-1}
$K_{\max}$	maximum growth capacity	1.62×10^8 cells
μ	Specific growth rate	0.028 min^{-1}
Unknown Parameter	**Description**	**Range of values**
$d_{\text{cI}}, d_{\text{GFP}}$	cI, GFP degradation rate	[0.01 0.3] min^{-1}
γ_1	pLux Promoter Hill constant	[50 100] nM
γ_3	Hybrid pLuxR/cI promoter coefficient	[0.0001 0.5]
γ_4	Hybrid pLuxR/cI promoter coefficient	[0.0005 5]
γ_5	Hybrid pLuxR/cI promoter coefficient	[1 100]
$k_{p_{\text{cI}}}, k_{p_{\text{gfp}}}$	cI, GFP translation rate	[1 60], [1 100] min^{-1}
β_1	Hybrid promoter basal expression	[0 0.01]
β_2	Hybrid promoter leakiness	[0 0.01]
$k_{m_{\text{cI}}}, k_{m_{\text{gfp}}}$	cI, gfp transcription rate	[0.1 75], [0.1 25] min^{-1}
k_{-2}, k_{-3}	Monomer and dimer dissociation rate	[0.05 0.3], [0.1 1] min^{-1}
k_2, k_3	Monomer and dimer association rate	[0.0006 0.06] min^{-1}
k_{mat}	GFP maturation time	[20 120] min

5.2.3 MO-based identification of the I1-FFL extended model

In order to successfully implement this multi-objective optimization for identification approach (see Fig.5.1), at least three fundamental steps are required as it was exhibited in section 2.5: *i)* the multi-objective problem definition (MOP), *ii)* the optimization process, and *iii)* the multi-criteria decision making stage (MCDM). This overall procedure enables us to analyze trade-offs between the objectives, and accordingly select a preferable solution.

In the spirit of an ensemble modeling approach, the error measurements between the experimental data and the model predictions for each inducer concentration or condition were formulated as independent objectives to be optimized. Thus, the mean squared errors (MSE) of the active GFP fluorescence for each input concentration $\mathrm{AHL_{ext}} = \{5, 15, 25, 35, 55\}$ nM are the **5 objectives to be optimized**. Several samples for every inducer concentration were taken at every $t = 5.65\,\mathrm{min}$ during approximately 200 min. Again, the bottom part of Table 5.2 enumerates the model parameters of (5.1) to be identified. These will be the **17 decision variables** of the MO-based identification.

The design objectives $J(\theta)$ as a function of the decision variables θ can be expressed by using the mean squared error (MSE) indexes

$$J_{[i=1,...,5]}(\theta) = \quad \frac{1}{n}\sum_{q=1}^{n}\frac{1}{m}\sum_{k=1}^{m}\Big(x^m_{10_{iq}}(k) - x_{10_{iq}}(kT)\Big)^2 \tag{5.2}$$

where i is the design objective for each input value, n is the number of observation copies measured at the instant time k for the same objective, m is the total number of experimental observations, x^m_{10} and x_{10} are the experimental and predicted observations of active GFP at the instant k respectively. The external input signal is applied by using the $\mathrm{AHL_{ext}}$ inducer stimulus at $t_0 = 0$.

We look for a set of values for the 17 decision variables θ that minimize all objectives $J(\theta)$. These five objectives are in conflict. If one tries to identify a single ensemble of parameters, as it will be described later in Fig.5.4, the best parameters achieving minimum MSE for one concentration of the external inducer worsen the model prediction performance at other concentrations of the external input signal. So, a trade-off must be reached, treating this problem as a multi-objective case

$$\begin{aligned} \min_{\theta\in\Re^{17}} J(\theta) = &\quad [J_1(\theta), ..., J_5(\theta)] \in \Re^5 \\ \text{subject to:} &\quad\quad \text{I1-FFL model} \ \ (5.1) \end{aligned} \tag{5.3}$$

A first optimization of (5.3) to obtain an initial estimation of the Pareto front and the unknown parameters in the model was carried out. This preliminary optimization

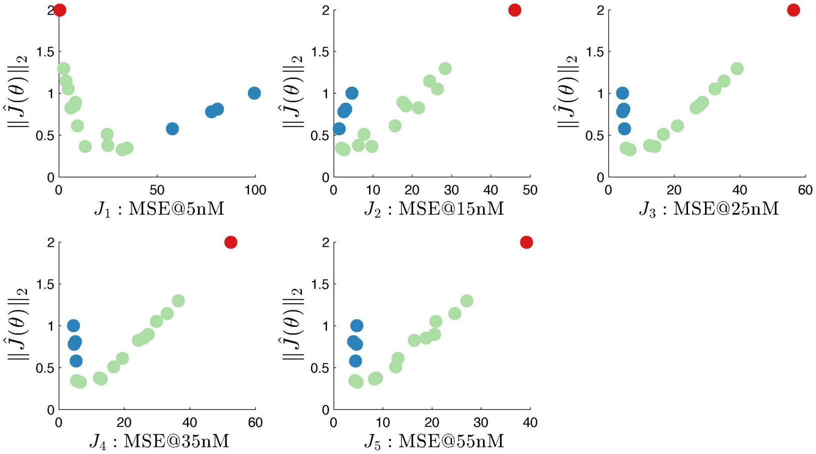


Figure 5.4. MO-based identification results. LD-modified representation of the Pareto Front for each objective $J_i(\theta)$ colored with the three resulting clusters. Each point represent a solution of the MOP.

can find minimum and maximum limits for the MSE of each objective $J_i(\theta)$, where $i = [1, \ldots, 5]$. In this way, the so-called *pertinency* of each $J_i(\theta)$ was assessed. Pertinency enclosures and enhances the search of the Pareto front, making the search region of the decision variables space narrower. Thereby, the subsequent MO-based identification process was done using the spMODE algorithm, with an initial population of candidate solutions chosen randomly from a uniform distribution in the narrowed parameters space.

In the next step, an approximation of the Pareto front with 17 solutions was obtained (Figure 5.4), together with the Pareto set (Fig.5.5) containing their corresponding parameters. These solutions were classified using the *kmeans* algorithm into three clusters showing a trade-off between the different objectives corresponding to the different $\mathrm{AHL_{ext}}$ induction levels. This *hierarchical clustering* helps to choose best parameters for different cases. It is seen that parameters in cluster 1 (in red dots) present small errors for the 5 nM induction, but they present larger errors for medium and/or higher induction values. Meanwhile, solutions in clusters 2 and 3 (green and blue respectively) perform better for medium (15 nM) and high (from 25 to 55 nM) inductions than solutions in cluster 1. The Pareto front shows the classical trade-offs. Thus, the red point, which the best solution for J_1 (5 nM) is the worst solution for all the other objectives. Green solutions in cluster 2, are better for J_2 (15 nM) but not that good for J_1. Finally, blue solutions in cluster 3 are good for J_3, J_4 and J_5 (from 25 to 55 nM) , but not so good for J_2, and bad for J_1.

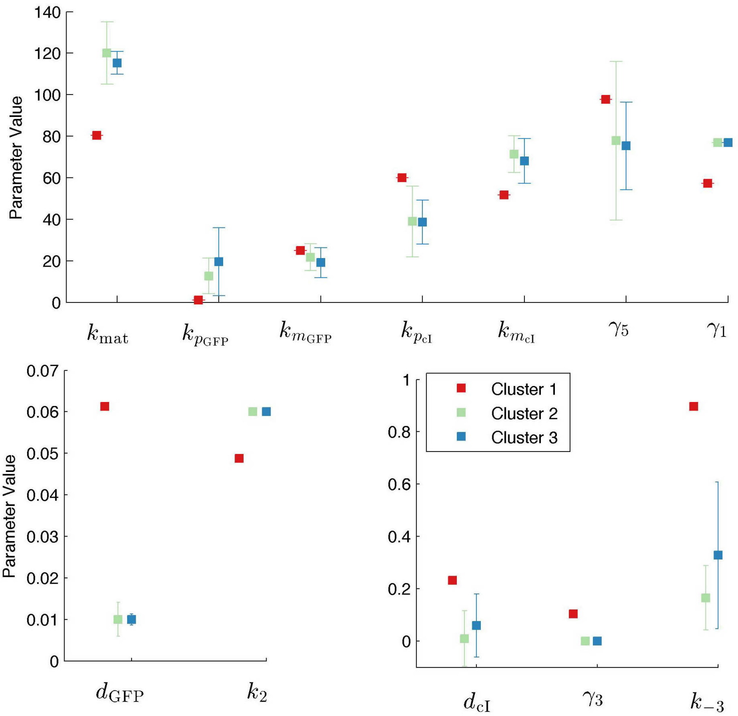


Figure 5.5. MO-based identification of parameters. The decision variables represented in the Pareto set have different values depending on the cluster (blue, red, and green).

The Pareto front looks quite similar for objectives J_3, J_4 and J_5, corresponding to inputs ranging from 25 nM to 55 nM respectively, as seen in Figure 5.4. Moreover the minimum values for these objectives are slightly larger than those obtained for objectives J_1 and J_2. Recall that these all three objectives fell within the same cluster. This similarity may be related to the high dependence of promoter activity on the concentration of the $\mathrm{AHL_{ext}}$ induction. Although it has been shown this dependence can saturate and reduce the $\mathrm{P_{lux}}$ and $\mathrm{P_{lux/cI}}$ promoters activity at levels of $\mathrm{AHL_{ext}} >$ 40 nM, there is no much difference in promoters activity for inductions larger than 20 nM (Egland and Greenberg, 2000). This saturation is observed in the experimental data (see Fig.5.6), and captured by the model (5.1). Yet, also a delayed peak is observed in the experimental results for large concentrations of the $\mathrm{AHL_{ext}}$ inducer that is not completely captured by the model.

Out of the 17 estimated parameters, 5 parameters had the same value in all clusters: $\gamma_4 = 1.42$, $\beta_1 =$ 0.008, $\beta_2 = 0.0014$, $\mathrm{k}_3 = 0.0006$ min^{-1} and $\mathrm{k}_{-2} = 0.2$ min^{-1}. Figure 5.5 shows the range of values obtained for the 12 parameters with different values in each cluster. Notice that parameter values in clusters 2 (15 nM) and 3 (from 25 to 55 nM) are quite alike, and distinctively different from parameter values in cluster 1 (5 nM). As said before, this is a consequence of the saturation when the $\mathrm{AHL}_{\mathrm{ext}}$ concentration is increased.

A few parameters account for the difference between the model for low inducer concentration (5nM), and for medium-high one (clusters 2 and 3). Note, in particular, the difference in the monomer (LuxR.AHL) association rate k_2 and the dimer dissociation rate k_{-3}. The LuxR-family of transcription factors are believed to be largely disordered (i.e., unfolded) in the monomeric form, becoming folded only upon dimerization in the presence of the external inducer (Buchler et al., 2005). Thus, for low values of inducer one may expect larger formation of monomer (larger k_2), and dissociation of dimer (lower k_{-3}), as is depicted in Fig.5.5.

In this context, the results presented also enable us to see that there is a big difference between values of the degradation rate d_{G} in the different clusters. The turnover of active GFP clearly decreases in presence of inducer. This may be related to the maximal capacity of the proteases that are present in a bacterial cell (Leveau and Lindow, 2001). Beyond a certain concentration of active GFP, their combined proteolytic activity is not enough to reduce the increment of the GFP content.

In addition, the parameters associated with the formation of the monomer and dimer, and the degradation rate of GFP vary, as already suggested by other studies. A difference in the peak time between the experimental and estimated circuit output is observed for large concentrations of the inducer. The delayed experimental response (dots in Fig.5.6) may be due to saturation effects not taken into account in the model, and deserve further work. Although the estimated maturation time of GFP was large, it cannot fully account for the peak delay. Some extra dynamics seem to be at play. The overall agreement between experimental and predicted data is remarkable.

Finally, from the resulting clusters the median value of every parameter in each cluster was selected. Figure 5.6) illustrates the comparison between model predictions using the selected values of the parameters (solid lines), and experimental data (points with bars) for different AHL input concentrations. This validation was performed with data sets not previously used for identification. Note that responses to induction levels from 25nM to 55nM are very similar among them, in agreement with the Pareto front analysis done before.

5.3 Using stochastic data to estimate the QS/Fb model

Parameter estimation in nonlinear stochastic dynamic models remains a very challenging inverse problem due to its nonconvexity, and ill-conditioning caused by over-parametrization, experimental measurement errors, data scarcity and uncertainty (Gábor and Banga, 2015; Kaltenbach et al., 2009). As we have seen, the multi-objective optimization for identification (MO-based identification) technique discussed

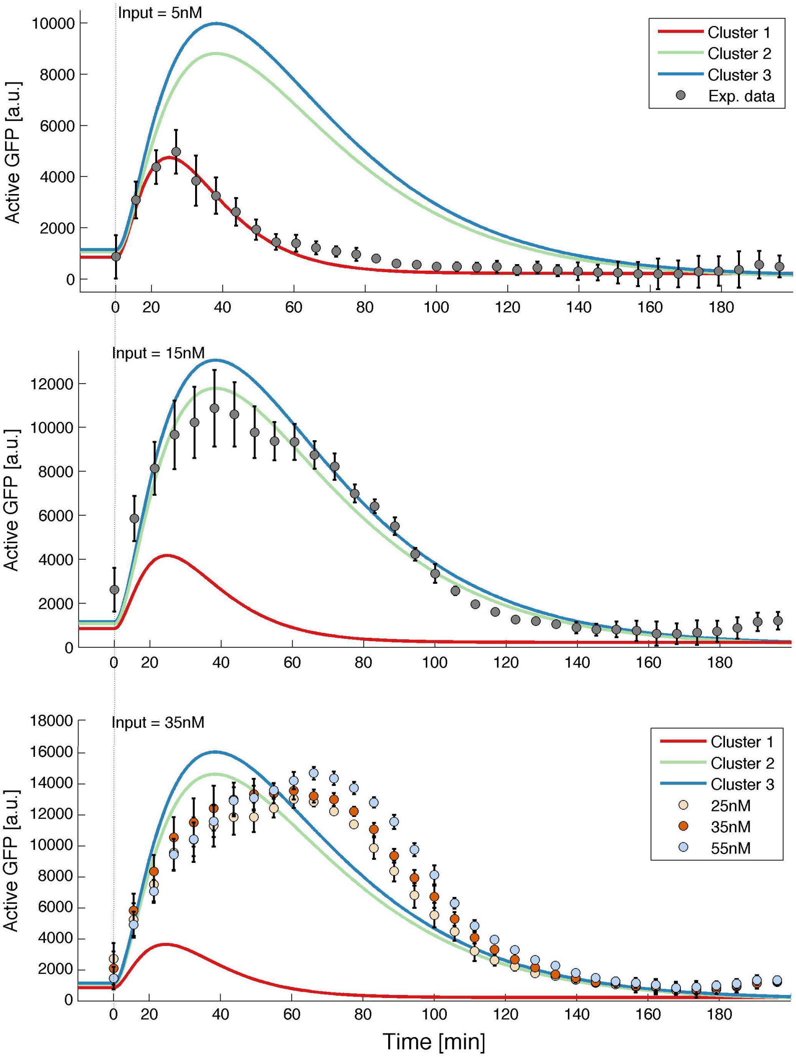


Figure 5.6. Comparison of the predicted and experimental data for different inductions. Dots correspond to mean values of experimental data (different data sets that the ones used for identification), with its variance as vertical bars. Predictions (continuous line) obtained for the three cluster representatives.

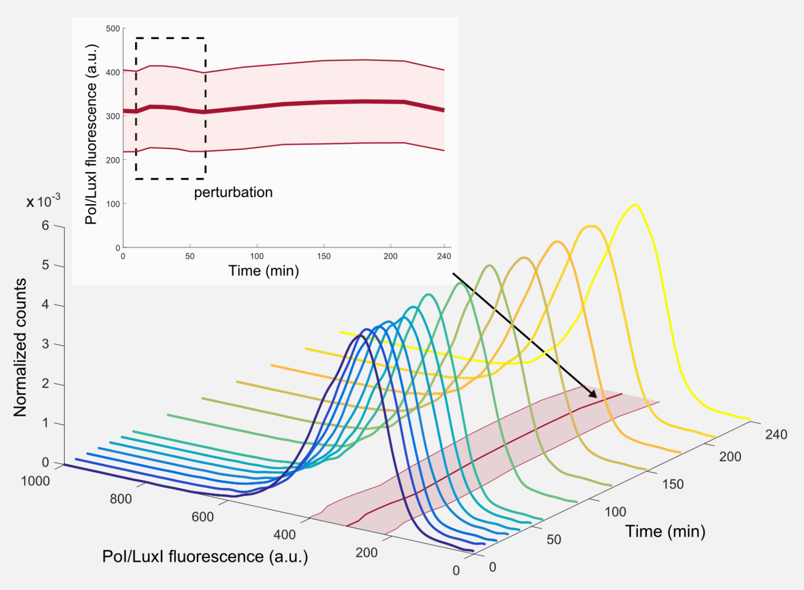


Figure 5.7. QS/Fb circuit rejects perturbations. Output Pol/LuxI fluorescence recovers to its previous level after adding a concentration of $\mathrm{AHL}_{\mathrm{ext}} = 10$nM as a perturbation signal. Perturbation is an impulse-like signal added at $t = 10$min, and its effect lasted around 1 hour.

in the previous section is extremely powerful, and its applicability was demonstrated with deterministic models and using bulk time course data at the population level. In this section, we address the problem of estimating the parameters of a stochastic model (4.42) corresponding to the QS/Fb circuit.

Recall the circuit is a feedback one devised to regulate the mean expression level of a protein of interest, while minimizing its variance. It is not only for an individual cells, but also across the population of cells. The strong feedback regulation embedded in the circuit hampers the possibility of obtaining enough experimental data for parameter estimation purposes. Indeed, perturbations can be induced in the circuit, i.e. by adding external $\mathrm{AHL}_{\mathrm{ext}}$. Yet, as seen in figure 5.7, the circuit reacts with very fast dynamics, so the data is not sufficiently exciting to be used for parameter estimation.

To address this problem, this Thesis propose an approach based on a two-stage estimation. Recall from section 3.3 the QS/Fb circuit (hereinafter referred to as the *closed-loop* circuit) has its counterpart NoQS/NoFb circuit (hereinafter the *open-loop* circuit). As depicted in figure 5.8, the QS/Fb circuit results from introducing extra feedback dynamics on the NoQS/NoFb one.

In the two-stage estimation described in more detail later, first the NoQS/NoFb model parameters are obtained using time-series data of population averaged values. note

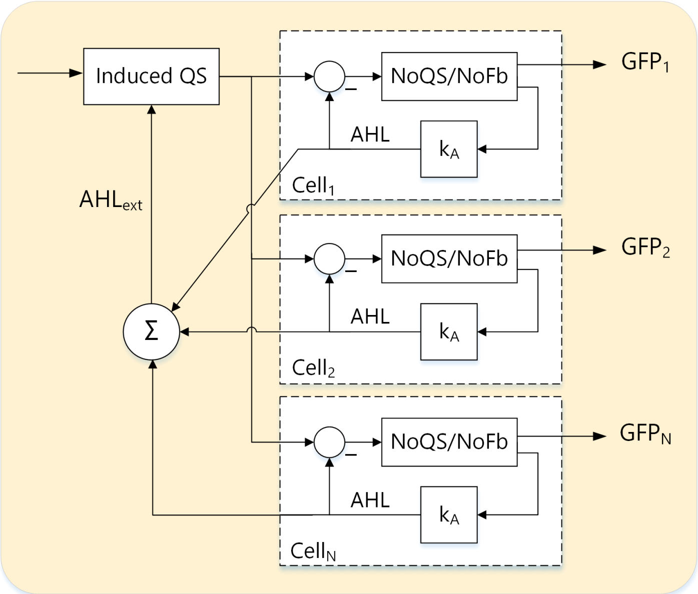


Figure 5.8. The NoQS/NoFb circuit with the external input $\mathrm{AHL}_{\mathrm{ext}}$, and using the feedback gain k_{A} becomes the QS/Fb circuit.

the open-loop circuit does not present problems when excited externally, so sufficiently rich experimental data can be gathered. In a second stage, the NoQS/NoFb stochastic model obtained from the results in the first stage, and steady-state flow cytometry data giving information at the individual cell level are used to estimate the parameters of the extra feedback dynamics.

5.3.1 Obtaining experimental data

Figure 5.9 explains how the two types of measurements were taken: *i)* bulk time-series data of the population from the open-loop NoQS/NoFb circuit, and *ii)* flow cytometry data of many individual cells across the population corresponding to the closed-loop QS/Fb circuit.

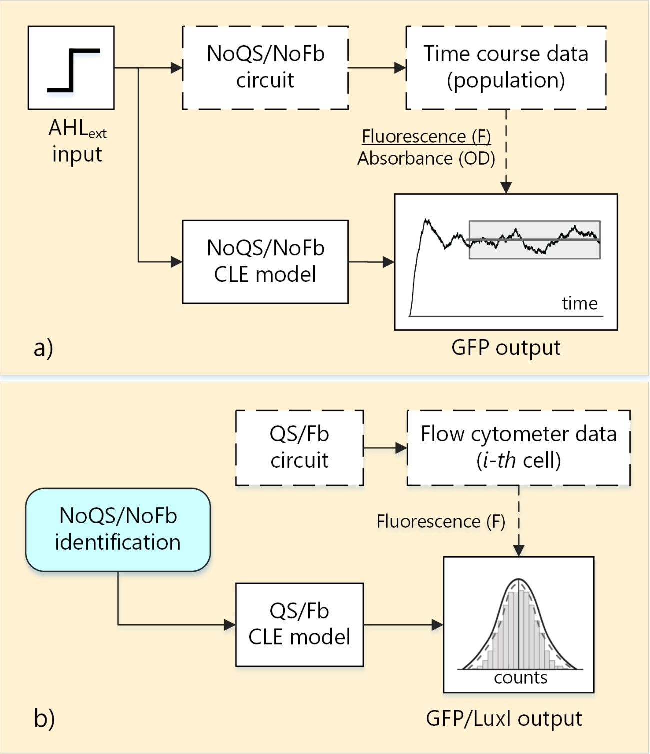


Figure 5.9. **a)** Procedure for the QS/Fb identification using bulk data (time-series) and model predictions. **b)** Parameters estimation using population snapshot (flow cytometry) and the stochastic simulation results.

Time-series data at the population level were measured in the culture samples that followed the same protocol for bacterial growth as in section 5.2.1. Remember that the input and output of the open-loop is the number of molecules of the inducer $\mathrm{AHL}_{\mathrm{ext}}$, and the fluorescence signal of GFP protein (see Fig.5.9a). *E.coli* cells (Top 10, NEB) carrying the pCB2tc together with the pAV02ta plasmids for the open-loop circuit, and the pCB2tc with the pYB06ta plasmids for the closed loop circuit were grown overnight in LB medium with the appropriate antibiotics. Then, 96 well-plates were inoculated at $\mathrm{OD}_{600} \approx 0.05$ and incubated to reach an optical density of $\mathrm{OD}_{600} \approx 0.1$. At this point, selected wells were induced with appropriate concentrations of AHL (N-3-Oxohexanoyl-L-homoserine lactone, Santa Cruz Biotecnology Catalog Number SC205396) and incubated for 400 minutes. Measurements were taken with a Biotek Cytation3 Imaging Plate Reader with the following protocol: 7 min of shaking, absorbance (OD, 600 nm) and fluorescence (F) measurements. Each condition of induction has 4 replicates, that is 4 time data sets for each one. Therefore, the input $\mathrm{AHL}_{\mathrm{ext}}$ ranges from 0 nM to 50 nM, generating time course data with OD and F measurements every 10 minutes during 400 minutes of incubation after induction (the full protocol is in annex C.1). Afterwards the data was analyzed using custom scripts implemented in Matlab as in ssection 5.2.

For the flow cytometry data, both QS/Fb and NoQS/NoFb circuits were measured. But since the QS/Fb circuit is an auto-regulated system, there is no input reference and the output represents the GFP/LuxI dynamics together (see Fig.5.9b). *E. coli* cells (cloning strain DH-5α) carrying the closed-loop and the open-loop (QS/Fb and NoQS/NoFb circuits, respectively) were inoculated from -80^{o}C stocks into 3 mL of LB with appropriate antibiotics, followed by an overnight incubation at 37 ^{o}C and 250 rpm in 14 mL culture tubes. When the culture samples reached an OD of 4 (600 nm, Eppendorf BioPhotometer D30), the overnight cultures were diluted 500-fold (OD_{600} of 0.02) into M9 medium broth[2] (M9CA amresco, Code J864-100G) with appropriate salts and antibiotics. These were used to inoculate new cultures, which latter were incubated for 7 hours (37^{o}C , 250 rpm,14 mL culture tubes) until they reached an OD_{600} between 0.2–0.3. At this point, cell growth and protein expression were interrupted by transferring the culture into an ice-water bath for 10 min. Next, 50 μL of each tube were transferred into 1 mL of phosphate-buffered saline with 500 μg/mL of the transcription inhibitor rifampicin (PBS + Rif) in one 5 mL cytometer tube, and incubated during 1 hour in a water bath at 37^{o}C, so that transcription kept blocked and GFP had time to mature and fold properly (see section 5.2.2). After that, culture samples were measured using the BD FACSCalibur flow cytometer (original default configuration parameters), and data were analyzed using custom scripts in Matlab (further details in annex C.2).

The time-series data represent the open-loop NoQS/NoFb circuit dynamics across the cell population for $\mathrm{AHL}_{\mathrm{ext}}$ inducer input values $[0,\ 1,\ 10,\ 50]$ nM. As in sections 2.1.3 and 5.2.1, in order to compare these measurements with the NoQS/NoFb model simulations, first the effect of *background* from the culture medium absorbance OD_{b},

[2] M9 medium broth is a highly-referenced microbial growth medium used for the cultivation of *E. coli*at slower growth rate than other ones corresponding to typical mediums like LB.

Table 5.3. Species of the QS/Fb and NoQS/NoFb circuits.

Variable	Species		Unit
	QS/Fb circuit	NoQS/NoFb circuit	
n_1	GFP/LuxI protein	GFP protein	molecules
n_2	LuxR protein		molecules
n_3	Dimer of (R.A)		molecules
n_4	AHL intracellular inducer		molecules
n_5	$\mathrm{AHL_{ext}}$ extracellular inducer	$\mathrm{AHL_{ext}}$ input	molecules
n_6	Monomer (R.A)		molecules

and the auto-fluorescence $\mathrm{F_b}$ of the cells was eliminated by subtracting these measurements from their corresponding temporal data. Therefore, $\mathrm{OD} = \mathrm{OD_{raw}} - \mathrm{OD_b}$, and $\mathrm{F} = \mathrm{F_{raw}} - \mathrm{F_b}$. Additionally, the Plate Reader gain was also removed. Then, the fluorescence per cell FOD = F/OD for each induction was obtained following the same procedure from 5.3.1. At $\mathrm{AHL_{ext}} = 0$ nM, the NoQS/NoFb circuit shows higher fluorescence (higher FOD) than the remaining data sets, as expected. Therefore, it was defined as the input reference $\mathrm{FOD_r}$. The corresponding experimental data sets for $\mathrm{AHL_{ext}}$ concentrations of 1, 10 and 50 nM were normalized with respect to the reference before they were used in the identification process. For example, the FOD at $\mathrm{AHL_{ext}} = 1$ nM was computed as $\mathrm{FOD_1} = \mathrm{F1_{raw}} - \mathrm{FOD_r}$.

5.3.2 QS/Fb and NoQS/NoFb stochastic models

Recall the QS/Fb CLE-based model (4.42) with the species listed in Table 4.6. Synthesis of intracellular AHL molecules by LuxI protein does not take place in the NoQS/NoFb circuit. Therefore, the model (4.42) also represents the NoQS/NoFb CLE-based model, if the synthesis rate is $\mathrm{k_A} = 0$ (refer section 4.4.1). Table 5.3 recalls the species for both circuits. Table 5.4 enumerates the model parameters for both the QS/Fb and the NoQS/NoFb circuits with 24 and 23 parameters, respectively. Out of them, 18 parameters are known from the literature, and they were kept fixed (top of Table 5.4). The parameters marked with an asterisk * in Table 5.4 refer to the QS/Fb system, where the gene *pol/luxI* is actually *gfp/luxI*. Yet, they also work for the NoQS/NoFb circuit with only the gene *gfp*.

The parameters at the bottom of Table 5.4 are the ones to be estimated. Recall for the open-loop NoQS/NoFb circuit $\mathrm{k_A}$ is set to zero.

Table 5.4. Parameters of the QS/Fb and NoQS/NoFb models.

Fixed Parameter	Description	Value
$\mathrm{C_I}^*$	Plasmid copy number times *luxI* transcription rate	17.5 molecules·min^{-1}
$\mathrm{C_R}$	Plasmid copy number times *luxR* transcription rate	7.9 molecules·min^{-1}
α	$\mathrm{P_{luxR}}$ promoter basal expression	0.01
$\mathrm{k_{-1}}$	Dissociation rate of (R.A)	10 min^{-1}
$\mathrm{k_{-2}}$	Dissociation rate of dimer (R.A)$_2$	1 min^{-1}
$\mathrm{k_{d1}}$	Dissociation constant of (R.A)	100 molecules
$\mathrm{k_{d2}}$	Dissociation constant of (R.A)$_2$	20 molecules
$\mathrm{d_R}$	R degradation rate	0.2 min^{-1}
$\mathrm{d_A}$	A degradation rate	0.057 min^{-1}
$\mathrm{d_{A_e}}$	A degradation rate in culture medium	0.04 min^{-1}
$\mathrm{d_{RA}}$	(R.A) degradation rate	0.156 min^{-1}
$\mathrm{d_{RA_2}}$	(R.A)$_2$ degradation rate	0.017 min^{-1}
$\mathrm{dm_I}^*$	mPI degradation rate	0.247 min^{-1}
$\mathrm{dm_R}$	mR degradation rate	0.247 min^{-1}
D	Diffusion rate of AHL through the cell membrane	2 min^{-1}
$\mathrm{V_{cell}}$	Typical volume of *E. coli.*	1.1×10^{-9} μL/cell
$\mathrm{V_{ext}}$	Typical volume of microfluidic device	1×10^{-3} mL

Unknown Parameter	Description	Range of values
$\mathrm{p_I}^*$	*gfp/luxI* messenger RNA translation rate	[0.1 10] min^{-1}
$\mathrm{d_I}^*$	PI degradation rate	[0.005 0.1] min^{-1}
$\mathrm{k_{dlux}}$	Dissociation constant of (R.A)$_2$ to the $\mathrm{P_{luxR}}$ promoter	[1 1000] molecules
$\mathrm{p_R}$	*luxR* messenger RNA translation rate	[0.1 10] min^{-1}
$\mathrm{K_{pr}}$	Fluorescence to number of molecules ratio (plate reader gain only)	[0.04 5]
$\mathrm{K_{fc}}$	Fluorescence to number of molecules ratio (flow cytometer gain only)	[0.04 5]
$\mathrm{k_A}$	Synthesis rate of AHL by LuxI (QS/Fb circuit only)	[0.005 0.1] min^{-1}

5.3.3 MO-based identification of the QS/Fb circuit

As in 5.2.3, the following subsections describe the multi-objective optimization design (MOOD) to perform the MO-based identification. First, open-loop characterization of the NoQS/NoFb circuit using time-course data was carried out. Next, the best resulting parameters were replaced in the QS/Fb CLE-based model, in order to perform closed-loop identification of the QS/Fb circuit using flow cytometry.

Open-loop identification using time-course data

For the open-loop characterization, the mean square error (MSE) between the temporal profile of fluorescence per cell (FOD) and the model predictions at different input values was minimized, as having **three different objectives**, one for each of the three input stimulus evaluated

$$J_{[a=1,\ldots,3]}(\theta) = \frac{1}{n}\sum_{q=1}^{n}\frac{1}{m}\sum_{k=1}^{m}\left(\hat{n}_{1_{aq}}(k) - n_{1_{aq}}(kT)\right)^2 \tag{5.4}$$

where $\hat{n}_1$ is the experimental observation of FOD at the instant k, a is the objective, $q = [1, \ldots, n]$ is the observation replicate at time k for each objective, m is the total number of temporal samples. The input stimulus is applied at $t_0 = 0$. The predicted observation n_1 is the result of the model simulation.

The first five parameters from the bottom of Table 5.4 are the ones to be identified for the NoQS/NoFb CLE-based model (4.42), that is, the decision variables θ for the MO-based identification. Their corresponding range is also shown in Table 5.4.

Finally, the MO-based identification looked for the values of θ that minimize all objectives $J(\theta)$. These three objectives are in conflict if one tries to identify a single acceptable set of parameters for all induction levels. So, a trade-off between them must be reached. This problem can be formulated as a multi-objective one

$$\begin{aligned} \min_{\theta \in \Re^5} J(\theta) = & \qquad [J_1(\theta), ..., J_3(\theta)] \in \Re^3 \\ \text{subject to:} & \qquad \text{NoQS/NoFb} \ \ \text{model (4.42)} \end{aligned} \tag{5.5}$$

As in section 5.2.3, the selection of the preferable solution according to designer's criteria takes place in an *a-posteriori* multi-criteria analysis of the Pareto front approximation. Again, the visualization tool Level Diagram (LD) already described in section 2.5.3 allows users to correlate design objectives with decision variables by providing two graphs. Remember that the first graph contains each objective, where its Y-axis is the p-norm $\|J(\theta)\|_p$ of the objectives vector, and the X-axis corresponds to each objective value $J_a(\theta)$ (see Figure 5.10A). The second graph provided by the LD-Tool shows $\|J(\theta)\|_p$ with respect to each decision variable (see Figure 5.10B). Thus, a given solution will have the same y-value in all graphs. In addition, the solutions were clustered using the *kmeans* algorithm and all the graphs were colored by the resulting clusters.

Closed-loop identification using flow cytometer data

After the selection of the preferable solution obtained from the open-loop identification process, the QS/Fb stochastic model was used with the flow cytometer data to perform another optimization in order to obtain values for the closed-loop parameters, i.e. feedback gain or the intracellular AHL synthesis rate $\mathrm{k_A}$ (Table 5.4).

In this context, flow cytometry data provide the distributions of each species coming from many individual cells. These distributions are steady-state measurements of the population fluorescence at a given time t. In the QS/Fb case, the output fluorescence measured is the GFP/LuxI content. The measurements were taken with the BD FACSCalibur flow cytometer with the protocol shown in annex C.2. Thus, the absolute value of the relative **errors for the mean** e_{μ}, and the **noise strength** e_{η^2} were optimized. The new objectives are derived by the indexes

$$
\begin{aligned}
J_4(\theta) = \mathrm{e}_{\mu(\mathrm{QS/Fb})} &= \left| \frac{\mu_{(QS/Fb)} - \mathrm{K_{fc}}\, \widehat{\mu}_{(QS/Fb)}}{\mu_{(QS/Fb)}} \right| \\
J_5(\theta) = \mathrm{e}_{\eta^2(\mathrm{QS/Fb})} &= \left| \frac{\eta^2_{(QS/Fb)} - \widehat{\eta}^2{}_{(QS/Fb)}}{\eta^2_{(QS/Fb)}} \right|
\end{aligned} \tag{5.6}
$$

where $\mu_{(QS/Fb)}$ is the mean of the experimental data obtained by flow cytometry for the protein GFP/LuxI, and $\eta^2_{(QS/Fb)}$ is its corresponding noise strength. The simulated mean and noise strength ($\widehat{\mu}$ and $\widehat{\eta^2}$, respectively) of GFP/LuxI (n_1 in the model (4.42)) were computed from the QS/Fb stochastic model simulation, using the previously selected preferable solution obtained in the open-loop identification stage, and the corresponding identified parameter values.

The GFP/LuxI mean $\widehat{\mu}$, and its total noise strength $\widehat{\eta^2}$, are obtained from the steady-state of the GFP/LuxI dynamics over the population of cells, where the law of total expectation, and the law of total variance (Basak and Chabakauri, 2010) are used. As we will see, this procedure is also used in Chapter 6. The result is the set of equations (5.7) where $n_1^i(kT)$ is the value of protein GFP/LuxI (in number of molecules) at time instant kT for the i-cell , $k \in \mathcal{N}$, k_0T is the time instant when the steady-state is reached, k_fT is the end of the simulation, and N is the total number of cells in the population. The mean of GFP/LuxI over the population at the time kT is $m(kT)$ ant is variance is $s^2(kT)$, the long-term mean of GFP/LuxI is $\widehat{\mu}$ and the variance $\widehat{\eta^2}$.

$$
\begin{aligned}
m(kT) =& \frac{1}{N} \sum_{i=1}^{N} n_1^i(kT) \\
s^2(kT) =& \frac{1}{N} \sum_{i=1}^{N} \left(n_1^i(kT) - m(kT) \right)^2 \\
\widehat{\mu} =& \frac{1}{(k_f - k_0)T} \sum_{k=k_0}^{k_f} m(kT) \\
\widehat{\sigma^2} =& \frac{1}{(k_f - k_0)T} \sum_{k=k_0}^{k_f} s^2(kT) + \frac{1}{(k_f - k_0)T} \sum_{k=k_0}^{k_f} \left(m(kT) - \widehat{\mu} \right)^2 \\
\widehat{\eta^2} =& \frac{\widehat{\sigma^2}}{\widehat{\mu^2}}
\end{aligned} \tag{5.7}
$$

The MO-based identification process looked for the values of the closed-loop decision variables $\theta_{\mathrm{cl}} = [\mathrm{k_A},\, \mathrm{K_{fl}}]$ that minimize all objectives $J_{4,5}(\theta_{\mathrm{cl}})$. These two objectives are in conflict, if one tries to identify a single value of parameters. Thereby, achieving a trade-off between them can be formulated as a multi-objective problem

$$\begin{aligned} &\min_{\theta_{cl} \in \Re^2} J(\theta_{cl}) = [J_4(\theta_{cl}), J_5(\theta_{cl})] \in \Re^2 \\ &\text{subject to: QS/Fb model(4.42)} \end{aligned} \tag{5.8}$$

Taking the above into consideration, identification of the open-loop parameters for the NoQS/NoFb model obtained before, together with the closed-loop parameters for the QS/Fb model is completed.

A first optimization to obtain an initial estimation of the Pareto front and the unknown parameters of the open-loop circuit was carried out using the equation (5.5). From this preliminary optimization, the appropriate minimum and maximum limits for each objective, the so-called *pertinency* of $J_{[i=1,\ldots,3]}(\theta)$ were found. They can be used to enhance the search of the Pareto front in a narrower region of the parameters space θ. In both cases, the optimization was done using spMODE starting with an initial population of candidate solutions chosen randomly from a uniform distribution in the parameters space.

In the next step, an approximation of the Pareto front with 6 solutions was obtained (Figure 5.10A), together with the Pareto set (Figure 5.10B) containing their corresponding parameters. These solutions were classified using the *kmeans* algorithm into three clusters showing a trade-off among the different objectives. This clustering can help to choose better parameters for different cases. The solutions are well distributed in the pertinency range. All of them were found with less than 8% noise error, and some of them (in particular the red cluster) even with less than 10% mean error leading to an overall good estimation.

The Pareto front (Figure 5.10A) analysis depicts the classical trade-offs. For instance, red points, which are best solutions for J_3 correspond to one of the worst solutions for J_1. The blue solutions, are the best solutions for J_1 but not so good for J_2 and J_3. However, in the Pareto set (Figure 5.10B) some trends can be seen. GFP/LuxI translation rate $\mathrm{p_I}$ has an opposite trend to the one of translation rate of LuxR p_R. The degradation of the measured protein $\mathrm{d_I}$ has consistent values for both clusters (for all the solutions). This value is in the limit of the initial interval for this parameter, and it is compatible with the slowest measured growth rate of the microorganisms 140 min (see section 2.1), and equivalent to a degradation rate of $\mathrm{d_I} = 0.005\,\mathrm{min}^{-1}$.

Figure 5.11 illustrates the resulting range of values for the 5 parameters with different values in each cluster. Out of the 5 estimated parameters, 2 parameters had approximately the same value in both clusters: $\mathrm{d_I} = 0.005$ min^{-1}, $\mathrm{K_{pr}} = 0.05$. Notice that even the parameters with different values in both clusters share a common order of magnitude: $\mathrm{p_I}$ is around 2 min^{-1}, and $\mathrm{p_R}$ is in the order of 8 min^{-1}. Also, $\mathrm{k_{dlux}}$ has a range in the hundreds of molecules (500 to 1000 molecules), which is a smaller interval than the initial considered optimization range (1 to 1000 molecules).

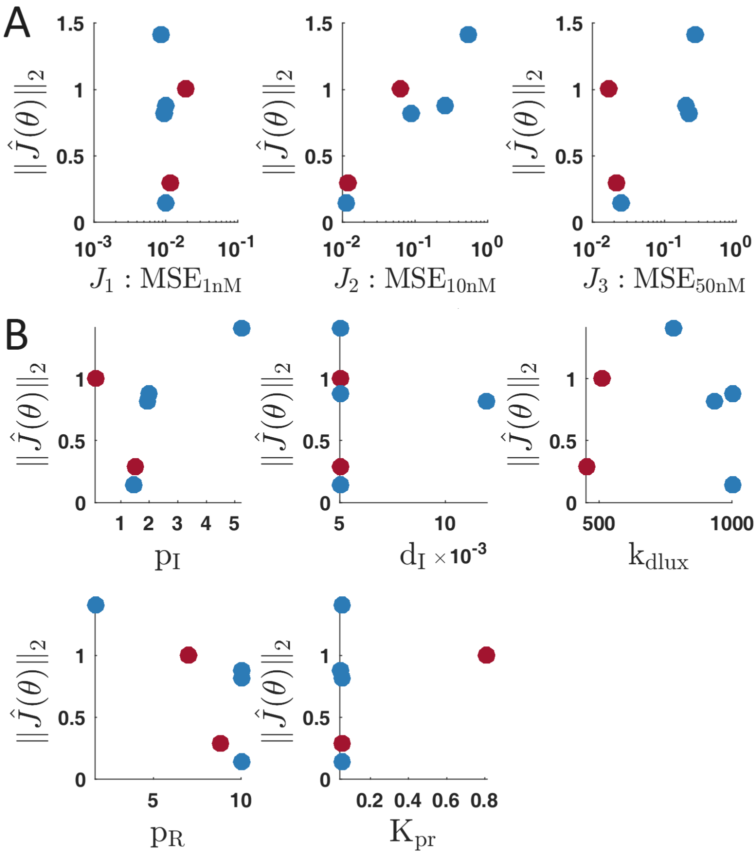


Figure 5.10. Identification results for the open-loop stage. (A) LD-modified representation of the Pareto Front for each objective colored with the two resulting clusters. LD-modified representation of the Pareto Set (B) for the 5 parameters colored with the two resulting clusters.

As a first validation of the open-loop circuit identification, one solution was chosen from the resulting clusters (red cluster) as a preferred solution. As we can see in figures 5.10 and 5.11, the point is suitable for minimizing the MSE in all objectives J_1, J_2, and J_3. The values corresponding to this solution are listed in Table 5.5. Figure 5.12 plots the comparison between predictions from the model and experimental temporal data for different $\mathrm{AHL_{ext}}$ induction levels (see section 5.3.1), showing good agreement. This validation was performed with time-course data not previously used for identification.

As was mentioned in section 5.3.3, the parameters $\theta_{\mathrm{cl}} = [\mathrm{k_A}, \mathrm{K_{fc}}]$ of the closed-loop circuit were estimated using equation (5.8), and the solution in Table . The GFP/LuxI protein estimations coming from long-term distributions generated by the QS/Fb sto-

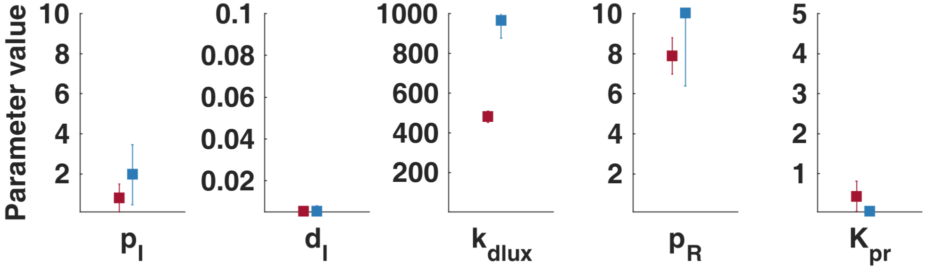


Figure 5.11. Clustering results for the open-loop stage. Values of the estimated parameters in the three different clusters.

Table 5.5. Selected solution for open-loop circuit.

Parameter	Description	Value
$\mathrm{p_R}$	*luxR* messenger RNA translation rate	10 min^{-1}
$\mathrm{p_I}$	*gfp/luxI* messenger RNA translation rate	1.98 min^{-1}
$\mathrm{k_{dlux}}$	Dissociation constant of (LuxR.AHL)$_2$ to the $\mathrm{P_{luxR}}$ promoter	1000 molecules
$\mathrm{d_I}$	GFP/LuxI degradation rate	0.005 min^{-1}
$\mathrm{K_{pr}}$	Fluorescence to number of molecules ratio (plate reader only)	0.04

chastic model with the previously estimated parameters were used to optimize the best values for θ_{cl}. Particularly, we used the estimation errors for both mean and noise strength formulated in equation (5.6). The Pareto front and set obtained are depicted in Fig.5.13. As it seen, one of these optimal solutions is the ideal one (lowest norm to the ideal point). Following a MCDM process, the values for both the feedback and the flow cytometer gain were optimized to

$$\begin{aligned} \mathrm{k_A} &= 0.048\ \mathrm{min}^{-1} \\ \mathrm{K_{fc}} &= 0.1 \end{aligned} \tag{5.9}$$

Finally, we used the values in equation (5.9) for validation. The resulting histograms are shown in Figure 5.14. Histograms corresponding to the simulations of the stochastic model are shown in solid colors. The purple histogram corresponds to the open loop circuit and the orange histogram to the closed loop circuit. Experimental data obtained via flow cytometry are plotted in black dashed lines and black lines, for the open and closed-loop circuits respectively. As it is evident, the superposition of the histograms are in agreement, validating both the open-loop identification process and most importantly the reduced model obtained for the QS/Fb synthetic gene circuit. The NoQS/NoFb stochastic model was simulated to validate the population distribution using the flow cytometer experimental data of steady-state GFP fluorescence content. In the open-loop circuit NoQS/NoFb , the GFP fluorescent output at dif-

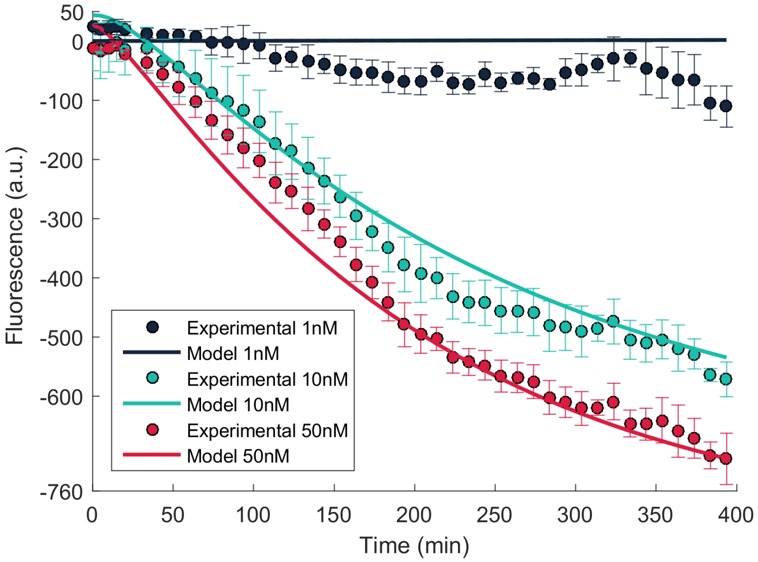


Figure 5.12. Comparison of predicted and experimental temporal data of the open-loop circuit (NoQS/NoFb) using a solution from the Pareto front.

ferent AHL induction levels were compared with their corresponding stochastic GFP model distributions.

5.4 Summary

We have proposed a methodology based on multi-objective optimization design (MOOD) to perform model parameter estimation of synthetic gene circuits. Minimizing the errors between experimental data and model predictions, the MOOD found sets of optimal values for the parameters in a Pareto front sense. Particularly, for the I1-FFL circuit the MOOD led to an important hint when one has to calibrate biological circuit models using experimental data: 'an ensemble of local models'. On the other hand, for the QS/Fb circuit one of strongest feature of the MOOD technique was dealing with large number of parameters to be identified, as well as simultaneously including measurements from different nature like time-course data from a cell population, or distributions data from a individual cells. The values estimated for the model parameters will be used in Chapter 6 to perform stochastic analysis of the feedback control included in the QS/Fb circuit.

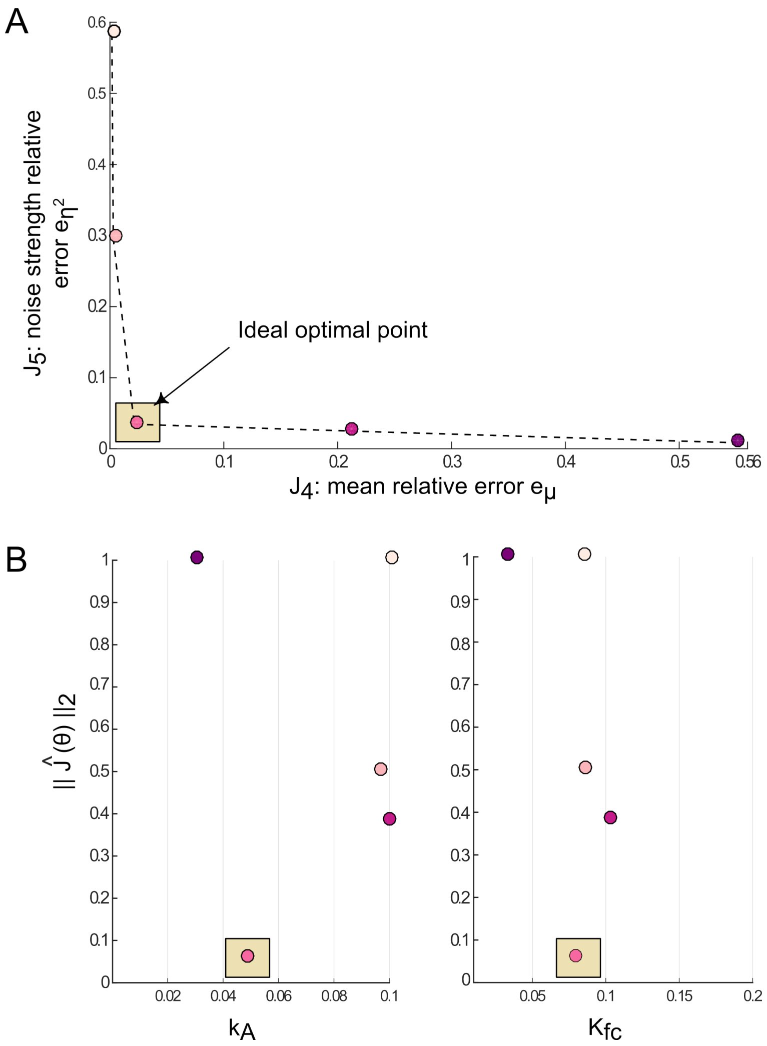


Figure 5.13. A) Pareto front showing the typical trade-off between both relative errors J_4 and J_5. B) In the Pareto set θ_{cl} the best optimal value for the feedback gain $\mathrm{k_A} = 0.048$ min^{-1} was chosen.

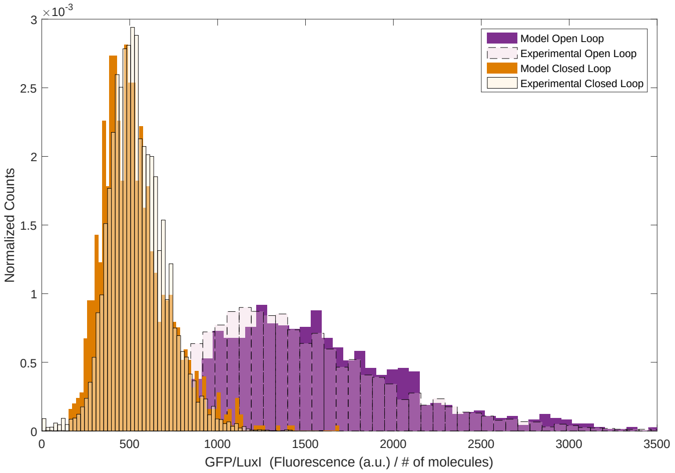


Figure 5.14. Validation of predicted and experimental population distribution data for the open loop circuit and the closed-loop circuit. Open-loop circuit, experimental histogram in black dashed lines, simulation in purple. Closed-loop circuit, experimental histogram in black lines, simulation in orange.

Chapter 6

Stochastic analysis of a feedback control synthetic gene circuit

6.1 Introduction

In Chapter 3, we discussed about the QS/Fb synthetic gene circuit, which was engineered to control noise in protein expression across a cell population, leading to robust protein production in industrial biotechnology. Chapters 4 and 5 proposed two methodologies to model and fully characterize the QS/Fb circuit via Chemical Langevin Equation for a whole population, and Multi-objective Optimization for system identification respectively.

Once the stochastic model has been validated, this Chapter focuses on the analysis of the QS/Fb gene circuit. Stochastic simulations of the QS/Fb CLE-based model for a cell population will be used to analyze the capability of the circuit to reduce noise in protein expression as a function of its model parameters. Eventually, a sensitivity analysis of these parameters can elucidate if some of them will require fine tuning to achieve better performance. This will be done in Chapter 7.

Many bioprocesses aim to develop efficient production systems for heterologous protein expression. As we have seen, heterologous protein production starts by introducing an exogenous protein-coding gene in the cell. Traditionally, for the design, optimization and control of bioprocesses, the population of microorganisms has been typically considered as an aggregate quantity, characterized by averaged properties (Carlquist et al., 2012). Yet, it is a fact that even isogenic (i.e. with the same genetic content) microbial populations have certain degree of **heterogeneity**. Indeed, individual microorganisms, even if part of a 'clonal' or isogenic population, may differ greatly in terms of genetic composition, physiology, biochemistry, or behavior (Elowitz et al.).

In particular, the phenomenon of phenotypic noise is described as variation within an isogenic population due to fluctuations in gene expression of single cells (Toni and Tidor, 2013). This heterogeneity at the population level has been shown to be one of the causes of productivity decrease in bioprocesses, when protein expression is scaling-up to industrial production, as mentioned in 3.3. Characterization and control of protein expression moments (mean and variance) across the cells population is, thus, a challenging topic (Sánchez and Kondev, 2008; Weber and Buceta, 2011; Vignoni et al., 2013b; Mélykúti et al., 2014; Oyarzún et al.) of relevance also for synthesis of commodities through synthetic pathways (Oyarzún, 2011).

Dealing with the problems above requires both appropriate dynamic predictive models, and designing dynamic controls of synthetic pathways and protein expression systems (Vignoni et al., 2013b; Menolascina et al., 2011; Singh; Holtz and Keasling). Therefore as said earlier, the QS/Fb circuit was designed to control the mean and variance of protein expression across a population of cells, using an interplay between an intracellular negative feedback loop, and an external loop based on quorum sensing (QS) for cell-to-cell communication (see Fig.6.1).

Part of the contents of this Chapter appeared in the following journal and congress articles

- Y. Boada, A. Vignoni, and J. Picó. Engineered control of genetic variability reveals interplay among quorum sensing, feedback regulation, and biochemical noise. *ACS Synthetic Biology*, 6(10):1903–1912, 2017a. doi: 10.1021/acssynbio.7b00087.
- E. Picó-Marco, Y. Boada, J. Picó, and A. Vignoni. Contractivity of a genetic circuit with internal feedback and cell-to-cell communication. *IFAC-PapersOnLine*, 49(26):213 – 218, 2016. ISSN 2405-8963. Foundations of Systems Biology in Engineering FOSBE.

This Chapter is organized as follows. Section 6.2 revolves around the methods used to perform *in silico* and *in vivo* experiments, respectively. Section 6.3 demonstrates how the proposed QS/Fb circuit attenuates gene expression noise of the protein of interest. In both sections 6.4 and 6.5, the benefits of having feedback and/or QS are analyzed with respect to both intrinsic and extrinsic noise. In turn, sections 6.6 and 6.7 describe how to improve noise reduction by tuning some circuit parameters. Finally, in section 6.8 some conclusions are exposed.

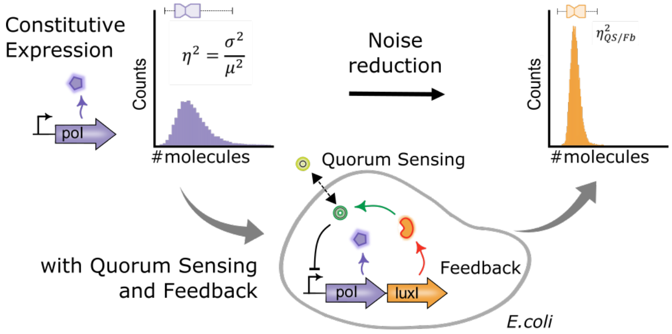


Figure 6.1. QS/Fb circuit combines an intracellular negative feedback loop and quorum sensing based cell-to-cell communication system to attenuate gene expression noise of the protein of interest.

6.2 QS/Fb gene circuit analysis

Once the QS/Fb stochastic model was characterized and validated, it is important to highlight some aspects. The circuit uses the repressible promoter $\mathrm{P_{luxR}}$ to implement a **negative feedback loop** over the gene that codes for the protein of interest (Pol), and adds a **QS mechanism** based on N-acyl-L-homoserine lactone (AHL) to induce population consensus, as was mentioned in 3.3. Using the CLE-based model for a cell population stochastic described in 4.4.1 and 5.3.2, the impact on noise strength of some key circuit parameters such as promoters or RBS was explored. Remember the promoter is a biopart that can restrict the transcription rate of a gene into its messenger RNA, and the RBS is another biopart which defines the translation rate of the messenger RNA into its corresponding protein (see Fig.6.2).

Recalling the five species for both gene circuits are properly distinguish in Table 6.2. According to figure 6.2a) to assess the role played by feedback and QS and, the proposed circuit was compared with:

1. NoQS/NoFb circuit. Recall this circuit has no quorum sensing to control expression of the protein of interest (see Fig.6.2b). Its model was characterized in section 5.3.2 with ($\mathrm{k_A} = 0$ min^{-1}),

2. NoQS/Fb circuit. It has a feedback loop but no quorum sensing, making the diffusion rate of AHL molecules null (D=0 min^{-1}).

The stochastic model (4.42) of the QS/Fb circuit (see section 4.4.1) was used to explore the impact of some key circuit parameters on noise. As control circuit to compare with, the second model of the NoQS/NoFb circuit, which removes both QS and the feedback loop was considered. For the computational analysis, this accounts to setting the synthesis of AHL to zero ($\mathrm{k_A} = 0\,\mathrm{min}^{-1}$) in the model (4.42). This condition is achieved in the lab experimental implementation by taking out the gene

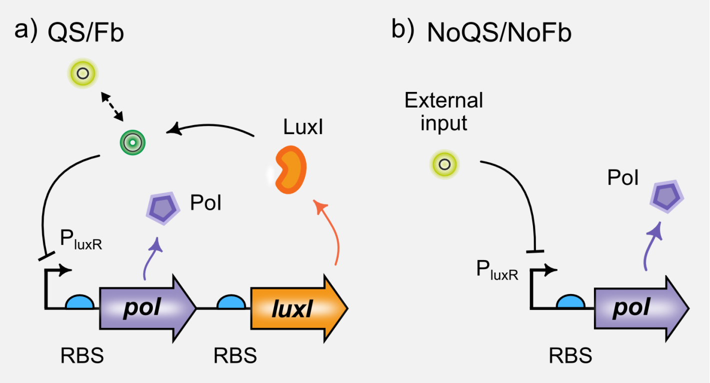


Figure 6.2. (A) QS/Fb scheme has each Pol and LuxI protein with its corresponding promotor and two different RBS downstream of LuxR protein (not represented). (B) NoQS/NoFb circuit has no LuxI protein, so one RBS of the protein of interest is the only difference with respect to the QS/Fb system.

Table 6.1. QS/Fb and NoQS/NoFb species.

Variable	Species		Unit
	QS/Fb circuit	**NoQS/NoFb circuit**	
n_1	GFP/LuxI protein	GFP protein	molecules
n_2	LuxR protein		molecules
n_3	Dimer of (R.A)		molecules
n_4	AHL intracellular inducer		molecules
n_5	$\mathrm{AHL_{ext}}$ extracellular inducer	$\mathrm{AHL_{ext}}$ input	molecules
n_6	Monomer (R.A)		molecules

coding for LuxI, as mentioned in subsection 3.3.2. To asses the effect of cell-to-cell communication, a hypothetical circuit with feedback but without quorum sensing (NoQS/Fb, $\mathrm{D} = 0\,\mathrm{min}^{-1}$) was also analyzed. Notice the circuit NoQS/Fb cannot actually be implemented for it assumes there is no diffusion of the autoinducer molecule across the cell membrane. Yet, it is useful as a computational thought experiment to account for the contribution of the cell-to-cell communication.

6.2.1 Getting statistical moments and minimising stochastic realizations

Gene expression noise was evaluated using the squared coefficient of variation, i.e. the noise strength measure that was described in section 2.4.1. This measure properly captures the contributions of both intrinsic and extrinsic noise, and allows comparisons for different expression rates.

The general procedure illustrated in Fig.6.3 was followed, using the guidelines defined in 4.4.2 for efficient simulations to obtain the noise strength. In this way, we performed for different combinations of model parameters temporal simulations of each species in the QS/Fb circuit (in number of molecules) inside every *i*-th cell in the population. Again, extrinsic noise was modeled by randomizing the values of the model parameters using a normal distribution with a variance of 15%. The models were implemented using OpenFPM (refer to 4.4.2). The corresponding code is available in annex D.2.

The procedure goes as follows:

1. First, a simulation was ran with a population of N = 240 cells in a culture volume of $10^{-3}\,\mu\mathrm{L}$, corresponding to an optical cell density OD_{600} = 0.3, for 400 minutes. Cell density variations did not appreciably change the results, confirming the results in (Tanouchi et al., 2008). The defined value of N provided a good representative of a cell population, as confirmed by comparing with cell populations up to N=12000 without significant variations in the population distributions obtained for each species. It was described in section 4.4.1. From this simulation, 240 time courses corresponding to the protein expression levels in time for each one of the 240 cells in the population were obtain. The first 134 minutes of simulation were discarded to ensure the system has reached steady-state, using the time samples corresponding to the last 266 minutes (around 100.000 time samples). This is depicted in Fig.6.3A.

2. With these 240 time courses, the first two statistical moments mean (μ) and the variance (σ^2) for each species in the cell across the population at every time instant t_k were calculated. From these moments, long-term distributions were computed to infer the noise strength of each species (see Fig.6.3B).

3. Then, using the time-mean across the population, the temporal mean was calculated. This gave us representative long-term means of the protein levels in the population (see Fig.6.3C). We used the law of total expectation (Basak and Chabakauri, 2010).

4. The long-term variance was also calculated by using the law of total variance, that is, the total variance is the sum of the mean of the variance plus the variance of the mean (Basak and Chabakauri, 2010). In particular, each stochastic model was checked if one realization of the population of N cells is enough to obtain unbiased values of the long-term moments of the population as in section 4.4.2. The results confirmed that 400 min is enough time to perform the corresponding average among time described in Fig.6.3C.

5. Finally, the noise strength $(\eta^2 = (\sigma/\mu)^2)$ is then calculated with the total mean and total variance of the system. In this way we incorporate and aggregate all the noise (intrinsic) coming from the different cells in the population (extrinsic).

The last two steps were performed to obtain the long-term statistics using only one realization of the simulation, so the computational burden was reduced. One can do this if the system is *ergodic*, that is, if enough time averaging along one realization is equivalent to getting statistics from many realizations at each time instant. Theoretically proving ergodicity is difficult for the proposed system, so ergodicity was computationally assessed. Consequently, both QS/Fb and NoQS/NoFb circuits are ergodic systems unless their stochastic simulations have not enough time.

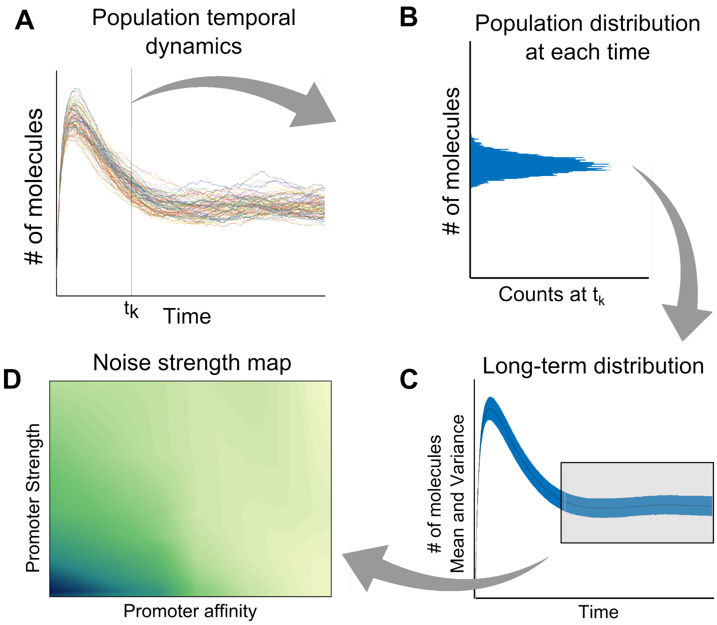


Figure 6.3. Methodological procedure to obtain the statistical moments from stochastic simulations of the circuit. (A) Temporal evolution of one species in the population of cells. (B) Distribution of the number of molecules across the population at each time instant. (C) Acquisition of the long-term distribution for each species. (D) Noise strength map for varying model parameters.

6.2.2 Computational analysis

Noise strength maps for different sets of varying model parameters were generated (see Fig.6.3D). Hence, the effect of variations in parameters associated to expression of LuxI and LuxR, as they are as key parameters in the QS/Fb circuit were explored. Table 6.2 lists all the parameters of the QS/Fb CLE-based model, including the six ones for LuxI and LuxR expression. In the case of LuxI production, there are three parameters whose values were changed: the dissociation constant $\mathrm{k_{dlux}}$ between the transcription factor (LuxR.AHL)$_2$ and the repressible $\mathrm{P_{luxR}}$ promoter, the translation rate $\mathrm{p_I}$, and the leakage α of the $\mathrm{P_{luxR}}$ promoter. They were sampled in the ranges

Table 6.2. Parameters of the QS/Fb and NoQS/NoFb models.

Fixed Parameter	**Description**	**Value**
$\mathrm{C_I}$	Plasmid copy number times *luxI* transcription rate	17.5 molecules·min^{-1}
$\mathrm{C_R}$	Plasmid copy number times *luxR* transcription rate	7.9 molecules·min^{-1}
$\mathrm{k_A}$	Synthesis rate of AHL by LuxI (QS/Fb circuit only)	0.04 min^{-1}
$\mathrm{k_{-1}}$	Dissociation rate of (R.A)	10 min^{-1}
$\mathrm{k_{-2}}$	Dissociation rate of dimer (R.A)$_2$	1 min^{-1}
$\mathrm{k_{d1}}$	Dissociation constant of (R.A)	100 molecules
$\mathrm{k_{d2}}$	Dissociation constant of (R.A)$_2$	20 molecules
$\mathrm{d_A}$	A degradation rate	0.057 min^{-1}
$\mathrm{d_{A_e}}$	A degradation rate in culture medium	0.04 min^{-1}
$\mathrm{d_{RA}}$	(R.A) degradation rate	0.156 min^{-1}
$\mathrm{d_{RA_2}}$	(R.A)$_2$ degradation rate	0.017 min^{-1}
$\mathrm{dm_I}$	mPI degradation rate	0.247 min^{-1}
$\mathrm{dm_R}$	mR degradation rate	0.247 min^{-1}
$\mathrm{d_I}$	PI degradation rate	0.027 min^{-1}
D	Diffusion rate of AHL through the cell membrane	2 min^{-1}
$\mathrm{V_{cell}}$	Typical volume of *E. coli.*	1.1×10^{-9} μL/cell
$\mathrm{V_{ext}}$	Typical volume of microfluidic device	1×10^{-3} mL
Changed parameter	**Description**	**Range of values**
α	$\mathrm{P_{luxR}}$ promoter basal expression	[0.01 0.1]
$\mathrm{p_I}^*$	*gfp/luxI* messenger RNA translation rate	[0.2 10] min^{-1}
$\mathrm{k_{dlux}}$	Dissociation constant of (R.A)$_2$ to the $\mathrm{P_{luxR}}$ promoter	[10 2000] molecules
$\mathrm{p_R}$	*luxR* messenger RNA translation rate	2 or 10 min^{-1}
$\mathrm{d_R}$	R degradation rate	[0.02 0.2] min^{-1}

$\mathrm{k_{dlux}} = [10 - 2000]$ molecules, $\alpha = [0.01 - 0.1]$, and $\mathrm{p_I} = [0.2 - 10]$ min^{-1} selected from the literature (Salis et al., 2009a; Egbert and Klavins, 2012; Schmidl et al., 2014), and experimentally achievable in the lab.

As for LuxR, two values were considered for the translation rate $\mathrm{p_R}$: a strong RBS ($\mathrm{p_R} = 10$ min^{-1}), and a medium-weak one ($\mathrm{p_R} = 2$ min^{-1}). In addition, the effect of different degradation rates $\mathrm{d_R}$ in the range $[0.02 - 0.2]$ min^{-1} was analyzed. For the case of the low mean scenario in Fig.6.7, simulations in the following range $\mathrm{p_I} = [0.004 - 0.02]$ min^{-1} were also included.

Notice from the ODE model (4.36) that although only variations in the translation rates $\mathrm{p_I}$ and $\mathrm{p_R}$ were considered, these are tantamount to considering variations in the lumped values $\frac{\mathrm{C_I p_I}}{\mathrm{dm_I}}, \frac{\mathrm{C_R p_R}}{\mathrm{dm_R}}$ corresponding to the products of protein burst size, transcription rate and gene copy number. Variations in translation rates were assumed, just because they are relatively simple to modify in a graded way by tuning the RBS (Salis et al., 2009a), though also transcription rates could be easily tuned (Brewster et al., 2012).

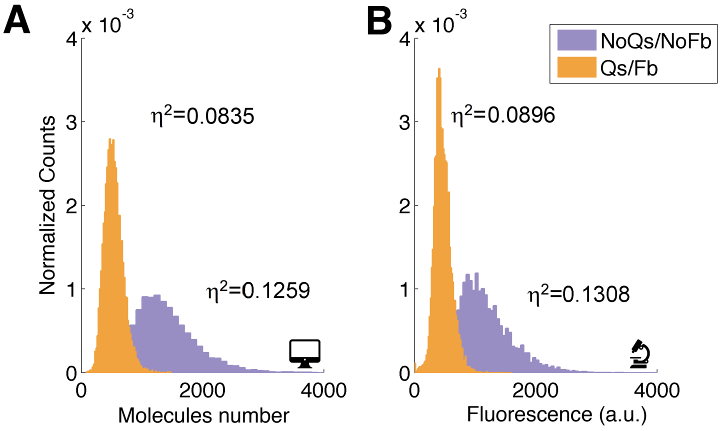


Figure 6.4. Representative computational (A) and experimental (B) population histograms of LuxI noise strength for QS/Fb (orange) showing a narrower gaussian-like distribution as compared to the Poisson-like one of NoQS/NoFb (purple).

6.2.3 Experimental analysis

To validate the *in silico* computational results, the QS/Fb and NoQS/NoFb circuits were implemented *in vivo* according the protocol Strains and plasmids already mentioned in section 3.3.2. There are two types of data that were collected on the QS/Fb and NoQS/NoFb circuits operating *in vivo*: *i)* time-series data for the whole population, and *ii)* flow cytometer distributions for every cell in the population (refer section 5.3.1).

6.3 Quorum sensing and negative feedback attenuate gene expression noise

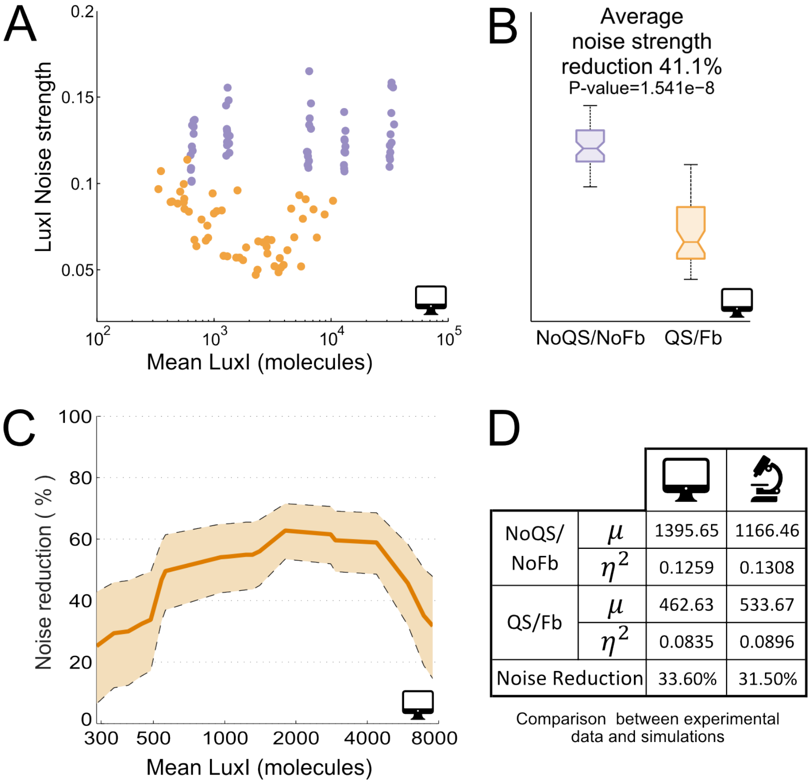
In the Thesis, the question whether the proposed QS/Fb circuit effectively reduces noise strength with respect to the circuit NoQS/NoFb (Fig.6.2B) was addressed. The last one consists of the LuxR expression on the one hand, and the protein of interest (PoI) downstream the P_{luxR} repressible promoter, without the *luxI* gene coding for LuxI protein, on the other. Since no autoinducer AHL is neither produced nor externally introduced, there is no repression, so the expression of PoI is essentially a constitutive one. This corresponds to the Poisson distribution observed in the purple population histogram in the left panel of Fig.6.5C. Conversely, the QS/Fb histogram departs from the Poisson distribution to become a narrow Gaussian-like one in the orange population histogram in the left panel of Fig.6.4. This fact, and the reduction in the mean expression value, indicate the strong presence of regulation. In both cases the nominal circuit parameters defined in Table 6.2 were used. Since the protein PoI is co-expressed together with LuxI in the QS/Fb system, this PoI/LuxI co-expression will be referred hereinafter LuxI.

Reduction in noise strength was not due to a particular choice of the circuit parameter values, but a property of the proposed topology. Figure 6.5A depicts LuxI noise strength *vs.* mean expression for 60 different combinations of the $\mathrm{P_{LuxR}}$ characteristics for both QS/Fb (orange points) and NoQS/NoFb (purple points). The points in the figure correspond to the mean values across the cells population for each combination of parameters using both QS/Fb and NoQS/NoFb CLE-based models from section 6.2.2. The magnitude of noise strength reduction was larger for medium values of mean protein expression. Noise strength levels were similar for all mean expression values in the case of the NoQS/NoFb circuit. Mean expression values in this case depend only on the translation rate $\mathrm{p_I}$ for which five discrete values were used, inducing the five mean values seen in the figure. In contrast, the QS/Fb circuit showed lower values of noise strength and more graded values of the mean expression level, as it depends on the combination of all three parameters varied.

More importantly, noise strength was consistently lower for the QS/Fb circuit. Taking together all the different combinations of promoter parameters for each circuit, and the average noise strength was significantly reduced by 41% in the presence of quorum sensing and negative feedback as shown in Fig.6.5B.

For the given fixed LuxR expression parameters, the noise strength reduction in LuxI showed a clear dependence on its mean expression level. In Figure 6.5C the minimum and maximum values of LuxI noise reduction are plotted as a function of its mean value. In the range between 600 and 6000 LuxI molecules it was possible to reduce the noise variance at least in 35% in the worst case scenario, with a maximum reduction of around 70% for means between 2000 and 3000 molecules.

In this context, the LuxI expression parameters in the ranges of $\mathrm{k_{dlux}} = [10-2000]$ nM, $\alpha = [0.01-0.1]$, and $\mathrm{p_I} = [0.2-10]$ min^{-1}, together with the LuxR parameters in the intervals of $\mathrm{d_R} = [0.02-0.2]\,\mathrm{min}^{-1}$, and $\mathrm{p_R} = [0.2-10]$ min^{-1} were also sampled. The simulation results are showed in Fig.6.6. In the top panel, $\mathrm{d_R} = 0.02\,\mathrm{min}^{-1}$ and the different colors code for several values of $\mathrm{p_R}$. The same, for the central panel with $\mathrm{d_R} = 0.07\,\mathrm{min}^{-1}$ and the bottom panel $\mathrm{d_R} = 0.2\,\mathrm{min}^{-1}$. Thereby, changing the parameters of LuxR protein expression showed a trend consistent with the findings in (Tanouchi et al., 2008), that is, the higher values of translation $\mathrm{p_R}$ and degradation $\mathrm{d_R}$ are, the larger the noise reduction (see Fig.6.6, and Fig.6.10).



		🖥	🔬
NoQS/ NoFb	μ	1395.65	1166.46
	η^2	0.1259	0.1308
QS/Fb	μ	462.63	533.67
	η^2	0.0835	0.0896
Noise Reduction		33.60%	31.50%

Comparison between experimental data and simulations

Figure 6.5. (A) Sampled combinations of LuxI expression characteristics for fixed LuxR ones show larger values of LuxI noise strength *vs.* mean for NoQS/NoFb (purple dots) than for QS/Fb (orange dots). (B) The QS/Fb circuit significantly reduces the average noise strength for the sampled parameters space by 41%, from $\langle\eta^2_{\mathrm{NoQS/NoFb}}\rangle = 0.1263$ down to $\langle\eta^2_{\mathrm{QS/Fb}}\rangle = 0.0744$. (C) For varying LuxI parameters the average reduction of noise strength in LuxI ranges from 30 % up to 60 % and shows dependence on the mean expression level. (D) Comparison of experimental data with computational results.

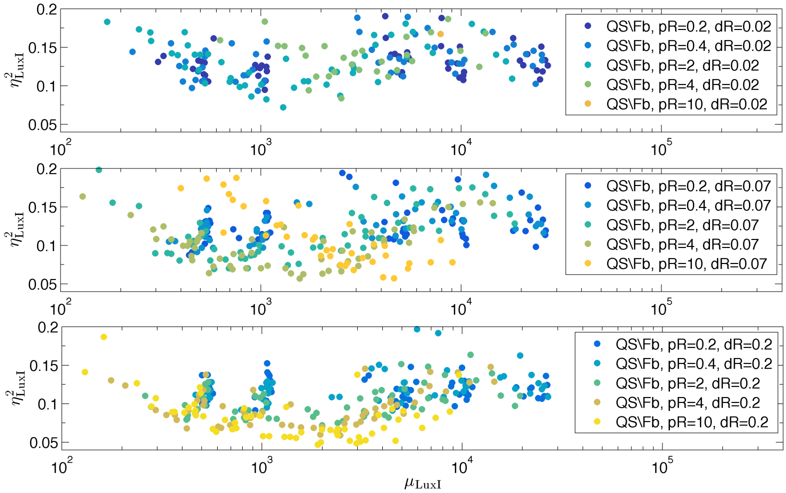


Figure 6.6. LuxI noise strength *vs.* LuxI mean. Increasing the LuxR turnover as a function of the degradation rate attenuates LuxI noise strength: $\mathrm{d_R} = 0.02\,\mathrm{min}^{-1}$ (top panel). $\mathrm{d_R} = 0.07\,\mathrm{min}^{-1}$ (central panel). $\mathrm{d_R} = 0.2\,\mathrm{min}^{-1}$ (bottom panel).

6.3.1 Experimental results confirm computational predictions

Experimental implementation of the proposed QS/Fb circuit would not only allow a preliminary experimental validation of its capability to reduce noise strength, but would also further validate the model parameters used throughout this study, as was already demonstrated in section 5.2.1 for system identification. The experimentally implemented NoQS/NoFb and QS/Fb circuits and their data were compared with the computational ones using the ad hoc Matlab scripts for gating the appropriate culture samples (refer annex D.1 for further details). As model parameters, $\mathrm{p_I} = 0.4\,\mathrm{min}^{-1}$, $\mathrm{k_{dlux}} = 200$ molecules, $\alpha_I = 0.01$, $\mathrm{d_R} = 0.07\,\mathrm{min}^{-1}$, $\mathrm{p_R} = 4\,\mathrm{min}^{-1}$ were used, and nominal values in Table 6.2 otherwise. Notice that from the closed-loop identification in section 5.3.3, we obtained the feedback gain (synthesis rate of AHL) as $\mathrm{k_A} = 0.048$ min^{-1} (see Fig.5.13). In practice, this value is the same to the one showed in literature $\mathrm{k_A} = 0.04$ min^{-1}. Thereby, this last one was used in all simulations.

The steady-state population histograms of LuxI for the circuits QS/Fb (orange) and NoQS/NoFb (purple) under the same experimental conditions are depicted in Fig.6.5C. The computational predictions are in the left panel, while the right panel shows flow cytometry experimental results. Both results were qualitatively comparable without any tuning, fitting or change in the model parameters. This only required a common scaling factor to convert from relative units of fluorescence to number of molecules (see annex D.1 for further details). The experimental results showed LuxI noise strength re-

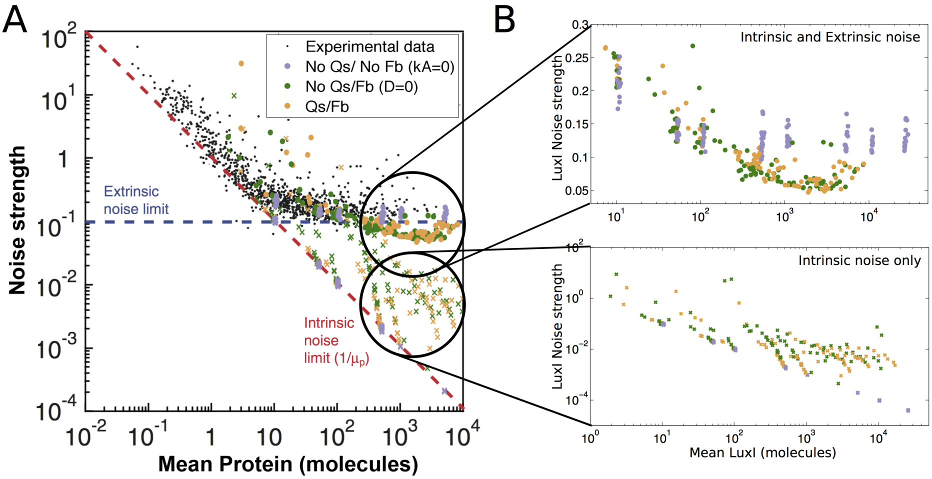


Figure 6.7. Comparison between experimental data and different scenarios evaluated computationally. (A) Experimental data of protein abundance and noise in *E. coli* taken from (Taniguchi et al., 2010) is plotted as black dots. The dashed red and blue lines are the intrinsic noise limit and the extrinsic noise limits respectively, taken from the same reference. Simulations of the gene circuits in our study, including both intrinsic and extrinsic noise, are plotted using purple dots (NoQS/NoFb), green (NoQS/Fb) and orange ones (QS/Fb). Simulations including only intrinsic noise are plotted as crosses: violet (NoQS/NoFb), green (NoQS/Fb) and orange (QS/Fb). (B) Zoom of the scenarios considering both intrinsic and extrinsic noise (top) and only intrinsic noise (bottom).

duced by 31.5% meanwhile the computational simulations predicted a 33.6% reduction (Fig.6.5D).

6.4 Feedback pays-off when extrinsic noise dominates

At this point the question arises as to what are the roles of quorum sensing and feedback in noise strength reduction, and what are their effects in view of both intrinsic and extrinsic noise.

To answer this question, first of all, the computational results using available experimental data of noise strength and protein abundance in *E. coli* were contextualized. Experimental data taken from (Taniguchi et al., 2010) were plotted against the computational results in three scenarios: *i)* base control circuit with no quorum sensing or feedback (NoQS/NoFb, $\mathrm{k_A} = 0$), *ii)* only the QS/Fb circuit , and *iii)* the hypothetical circuit with feedback but without quorum sensing (NoQS/Fb, D $= 0$). For each scenario, different combinations of parameters were considered. The values of the mean protein number are in the range $\left[1 - 10^5\right]$, as was enumerated earlier.

Figure 6.7A shows the experimental data plotted as black dots. The dashed red and blue lines are the intrinsic and extrinsic noise limits respectively, taken from the same reference above. Simulations including both intrinsic and extrinsic noise are plotted as purple dots (NoQS/NoFb), green (NoQS/Fb) and orange ones (QS/Fb) using the same data as in Fig. 6.5C. These computational results closely match with the experimental data and derived limits in (Taniguchi et al., 2010). The results corresponding to the base control circuit NoQS/NoFb clearly were over the noise limits.

Unexpectedly, noise strength of both circuits QS/Fb and NoQS/Fb showed very similar behavior. As shown in the upper panel of Fig. 6.7B, the QS/Fb and NoQS/Fb points lay in the same region. For medium and high mean protein expression values noise strength in QS/Fb and NoQS/Fb decreased just below the reported extrinsic noise limit, and well below the noise strength for the base NoQS/NoFb circuit. Though high protein expression are of main interest for the intended application of the QS/Fb circuit in an industrial biotechnological context of heterologous protein production, it is also interesting to analyze the performance of these circuits at low mean protein numbers. Essentially, the situation in this region was reversed. The open-loop circuit NoQS/NoFb showed consistent lower noise strength values than QS/Fb and NoQS/Fb. Therefore, feedback contributed to reducing noise strength for medium-high protein expression where extrinsic noise dominates.

6.5 Quorum sensing helps feedback to cope with intrinsic noise

The last result was inconclusive about the contribution of quorum sensing to reduce noise strength. To settle this issue, I concentrated the analysis in the medium-high protein expression region where feedback contributed to reduce noise strength and extrinsic noise dominates.

Here it is important to elucidate whether QS mainly contributed reducing the intrinsic component of noise. If this was the case, its effect could be masked by the dominant extrinsic noise. To that end, simulations for the same combinations of parameters as before, but suppressing extrinsic noise, and considering the three scenarios NoQS/NoFb, QS/Fb, and NoQS/Fb were carried out. The results are shown in Fig. 6.7A, plotted as violet (NoQS/NoFb, $\mathrm{k_A} = 0$), green (NoQS/Fb, D $= 0$) and orange crosses (QS/Fb). The bottom panel of Fig. 6.7B shows a zoom into the relevant region. Introducing either feedback alone or feedback plus quorum sensing increased noise strength values with respect to the minimal base control circuit representing plain constitutive protein expression. The results for this base NoQS/NoFb circuit were along the intrinsic noise limit (Taniguchi et al., 2010). These results were consistent with the findings at low mean protein values where intrinsic noise dominates. The circuit NoQS/Fb with feedback and no cell-to-cell communication showed higher

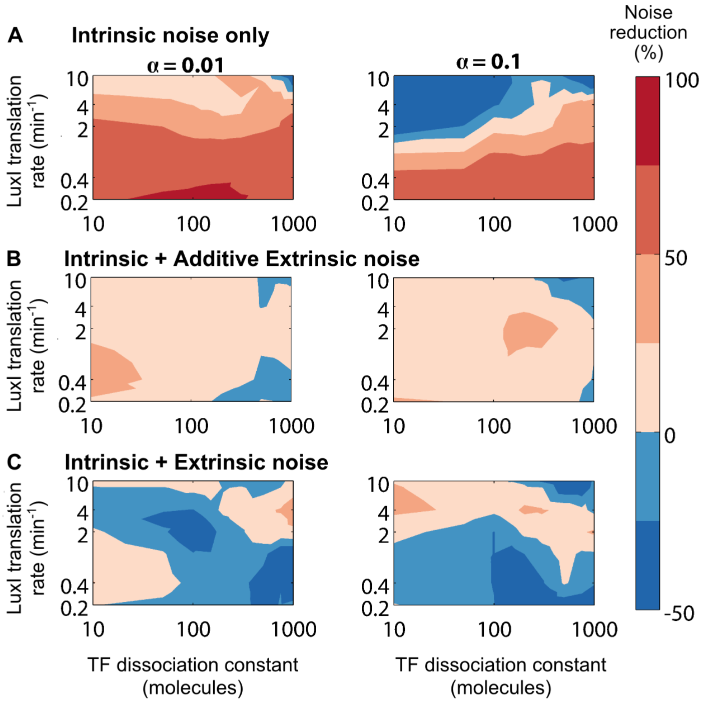


Figure 6.8. LuxI noise strength reduction as a function of circuit parameters. Color map of the reduction of LuxI noise strength when QS is added to Fb w.r.t. the dissociation constant k_{dlux} and the LuxI translation rate p_I. All other parameters were set to their values from Table 6.2. Left) Tight promoter $\alpha = 0.01$. Right) Leaky promoter $\alpha = 0.1$. (A) Only intrinsic noise is present. (B) Intrinsic noise and additive extrinsic noise. (C) Intrinsic and parametric extrinsic noise.

values of noise strength, specially for lower values of mean protein number. Finally, reintroducing quorum sensing (QS/Fb) was able to slightly improve noise strength.

To confirm this outcome, the difference between the noise strength in LuxI between the circuits QS/Fb and NoQS/Fb when only intrinsic noise is present, was evaluated as a function of circuit parameters associated to LuxI expression. As the LuxR parameters used before were close to be a best case scenario (see Fig.6.10), this time a smaller translation rate $p_R = 2$ min^{-1} was used, corresponding to an average scenario. Figure 6.8A depicts the noise strength map difference for different combinations of the dissociation constant k_{dlux} *vs.* the LuxI translation rate p_I for a tight promoter P_{luxR}, $\alpha = 0.01$ and a leaky one $\alpha = 0.1$ in both noise scenarios. The noise strength reduction when QS was added reached a 100% for low values of p_I. Increasing the dissociation constant improved noise reduction, specially for a leaky promoter.

This previous result suggested that the effects reported in the literature showing a reduction in noise strength as function of QS, are a consequence of modeling extrinsic noise as an additive signal. In the Thesis, this hypothesis was confirmed when besides intrinsic noise, an additive extrinsic noise was added to the QS/Fb system. Extrinsic noise has variance independent of the system states. Figure 6.8B illustrates that, in this case there was also a generalized noise strength reduction for most parameter combinations.

Finally when extrinsic noise as parametric variability was restored, the results showed that adding QS may increase or decrease noise strength (see Fig.6.8C) strongly depending on the values of the circuits parameters, and suggesting that getting benefit of QS for medium-large mean expression values requires fine-tuning and optimizing the circuit parameters. This will be further developed in Chapter 7.

6.6 Tuning LuxI expression allows minimising noise strength

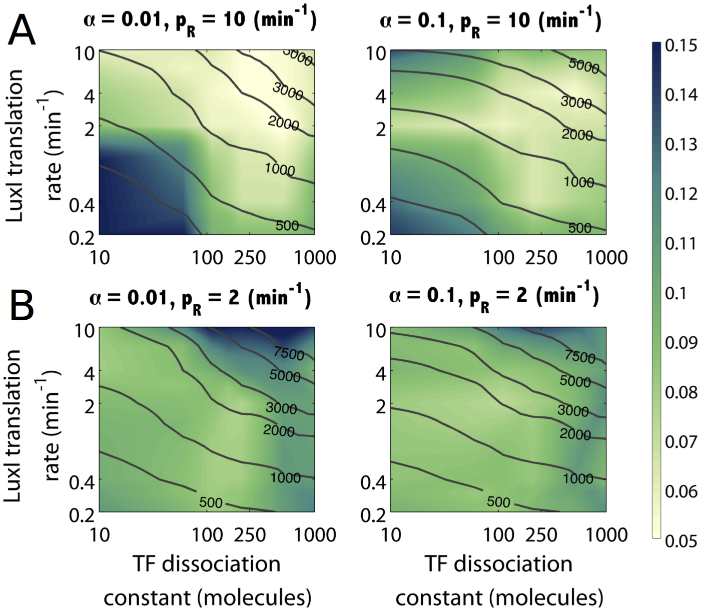


Figure 6.9. LuxI noise strength and mean as a function of circuit parameters. Color map of LuxI noise strength w.r.t. the dissociation constant k_{dlux} and the LuxI translation rate p_I. The level curves correspond to the mean number of LuxI molecules. (A) Strong LuxR RBS with $p_R = 10$ (1/min). (B) Medium-weak LuxR RBS with $p_R = 2$ (1/min).

Dependence of mean expression and noise strength on the Qs/Fb circuit parameters is a key factor to understand for the circuit to be of potential practical usage. In this perspective, *in silico* experiments were ran in order to estimate the noise strength and mean expression value of LuxI, as a proxy of the protein of interest (PoI), for different sets of the circuit parameters associated to LuxI expression, as we saw in section 6.2.2. Only two values for the basal expression, corresponding to a tight $\mathrm{P_{luxR}}$ promoter ($\alpha = 0.01$), and a leaky one ($\alpha = 0.1$) were evaluated. As for LuxR, I also considered two values corresponding to the two scenarios before: *i)* a strong RBS ($\mathrm{p_R} = 10$ min^{-1}) close to a best scenario for noise reduction, and *ii)* a medium-weak one ($\mathrm{p_R} = 2$ min^{-1}). All other parameters kept their nominal values from Table 6.2. Notice that although some variations in the translation rates $\mathrm{p_I}$ and $\mathrm{p_R}$ have been considered, in the QS/Fb model these are *tantamount*, making equivalent variations in the lumped values of the corresponding products of protein burst size, transcription rate and gene copy number (refer section 6.2.2).

Figure 6.9 shows the noise strength map for different combinations of the dissociation constant $\mathrm{k_{dlux}}$ *vs.* the LuxI translation rate $\mathrm{p_I}$ considering both a tight promoter $\mathrm{P_{luxR}}$ ($\alpha = 0.01$, Fig.6.9A), and a leaky one ($\alpha = 0.1$, Fig.6.9B). Different LuxI expression levels computed in protein number are shown as contour lines.

The mean expression levels of LuxI presented general monotonous trends in all cases. It increased for simultaneous rising of the dissociation constant and the LuxI translation rate. On the other hand, increasing leakiness of the LuxI promoter did tend to lower mean expression levels of LuxI for low values of the dissociation constant. Finally, using a weaker RBS controlling the translation of LuxR (see Fig. 6.9B) produced a steeper increasing of the mean expression level as the dissociation constant, and the LuxI translation rate increased.

Noise strength did not show simple patterns as a function of the circuit parameters. Larger variations between high and low noise strength values were observed for stronger LuxR RBS (Fig. 6.9A) independently of the leakiness from the $\mathrm{P_{luxR}}$ promoter. In this case, the lowest values of noise strength were achieved for values of the dissociation constant $\mathrm{k_{dlux}}$ in the range $[100 - 500]$ molecules, and values of LuxI translation rate $\mathrm{p_I}$ in the range $[2-10]$ min^{-1}. The mean expression levels in this region were between $2\cdot10^3$ and $4\cdot10^3$ proteins, in agreement with the results shown in Fig. 6.5. Decreasing the LuxR RBS strength kept the the values of minimal noise strength essentially in the same region, but with higher values (Fig. 6.9B). The same trend towards higher values of noise strength was observed when the tight promoter $\mathrm{P_{luxR}}$ was changed for a leaky one. This was more evident when a stronger LuxR RBS was used (Fig. 6.9A).

6.7 Fast LuxR turnover reduces LuxI noise strength

Finally, the effect of LuxR expression parameters over LuxI mean expression level and its noise strength we analyzed. In particular, we were interested in the effect of the LuxR translation rate $\mathrm{p_R}$, as the main tuning knob of LuxI mean expression level, and the one of the degradation rate $\mathrm{d_R}$.

On the one hand, LuxR synthesis rates proved to be a good sensitive tool to tune the desired LuxI mean expression level, with larger values of the last as the former decreased. Fig. 6.10 plots the LuxI noise strength maps and mean expression level curves as a function of the LuxR translation rates in the range 0.2 to 10 min^{-1}, and LuxR degradation rate in the range 0.02 to 0.2 min^{-1}. The LuxI translation rate was fixed to two values $\mathrm{p_I} = 2$ min^{-1} and $\mathrm{p_I} = 4$ min^{-1} around its nominal value , and considered both a tight $\mathrm{P_{luxR}}$ promoter ($\alpha = 0.01$) and a a leaky one ($\alpha = 0.1$). All other parameters were kept to their nominal values described in Table 6.2.

On the other hand, LuxI noise strength decreased with LuxR fast turnover was confirmed. Unlike suggested in (Tanouchi et al., 2008), the decrease is not uniform, having optimal values for $\mathrm{d_R}$ in the range 0.07 to 0.2 min^{-1} when LuxR translation rate $\mathrm{p_R}$ had medium to high values in the range 2 to 10 min^{-1}, as was described in section 6.2.2. The mean expression level was not very sensitive to the LuxR degradation rate, with a slight increase as the degradation rate increased.

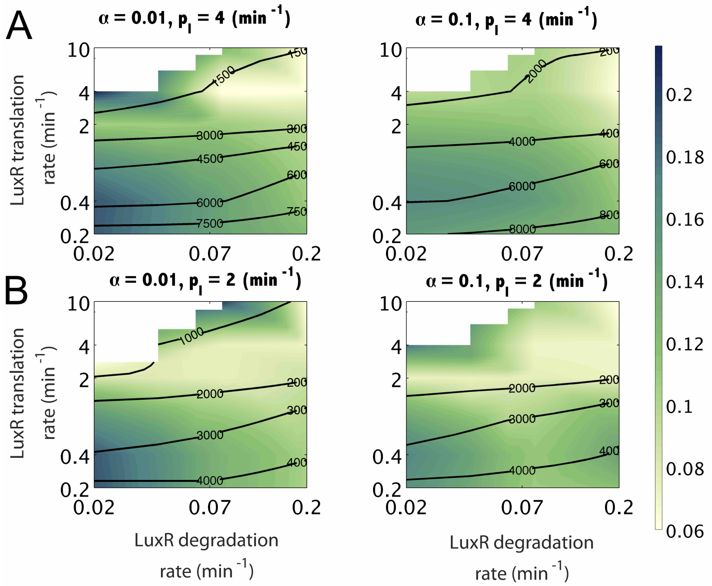


Figure 6.10. LuxI noise strength *vs.* LuxR parameters. LuxI noise strength maps and mean expression level curves for a tight $\mathrm{P_{luxR}}$ promoter ($\alpha = 0.01$, top) and a a leaky one ($\alpha = 0.1$, bottom) with LuxI translation rates $\mathrm{p_I} = 2\,\mathrm{min}^{-1}$ (left) and $\mathrm{p_I} = 4\,\mathrm{min}^{-1}$ (right) around its nominal value.

6.8 Discussion

The results show that the proposed QS/Fb synthetic gene circuit benefits from the interplay between feedback and cell-to-cell communication, allowing us to control of the mean expression level and noise strength of a protein of interest. A few circuit parameters easy to tune in the wet-lab can be used to achieve noise strength reductions up to a 60% with respect to constitutive expression of the protein of interest PoI.

Mean expression level and noise strength are not independent goals. At low mean values intrinsic noise dominates and sets the minimum noise strength attainable. At high mean values extrinsic noise dominates. Thus, there is a trade-off between expression level and noise strength, as revealed both by system-wide experimental data and theoretical analysis reported in the literature. Our computational results fitted well in this scenario, and suggest that tuning synthetic gene circuits to minimize noise while achieving a desired expression level will require a multi-objective optimization approach.

For high mean expression values, a clear benefit of having feedback as compared to constitutive expression was observed. Even if achieving best noise suppression requires an optimal feedback tuning, as already seen e.g. in (Toni and Tidor, 2013), noise reduction due to feedback was essentially structural, i.e. independent of its parameters, in this high mean expression region. Yet, adding quorum sensing on top of feedback did not decrease noise strength unless the circuit parameters are tuned. That is, the benefit from adding cell-to-cell communication is not structural, but depended on proper choice of the circuit parameters. This result is somewhat counter-intuitive and does not fully agree with previous works reporting a reduction of extrinsic noise in quorum sensing-based gene circuits, e.g. (Tanouchi et al., 2008), that reported a structural benefit. This may be explained by the different approaches to model extrinsic noise.

While in the Thesis, extrinsic noise has been modeled as parametric variability, most often in the literature extrinsic noise has been added as an additive stochastic signal. This is essentially analogous to the intrinsic noise term. Thus, if we considered a scenario with intrinsic noise and no extrinsic one while keeping medium-high expression means, our results also showed an important reduction of noise strength when quorum sensing was added to feedback. Although the amount of reduction depended on the circuit parameters, noise reduction was observed for almost any combination of them. Moreover, considering additive extrinsic noise, then qualitatively similar results to the ones when only intrinsic noise is present were obtained.

Two different gene circuits (QS/Fb and NoQS/NoFb) will result in different levels of noise because their physiological effects on the cell will be different. Yet, the chances that adding an extra random structure on top of a given circuit will result in a extremely low noise reduction. Thus, in the hypothetical scenario with no extrinsic noise, we also found that adding either feedback or feedback and quorum sensing increased the noise strength with respect to the open-loop (NoQS/NoFb) constitutive gene expression circuit. This result might be explained by the increased complexity introduced by these circuits (Potvin-Trottier et al., 2016). However, circuit complexity is not the only factor contributing. On the one hand, the circuit with quorum sensing and feedback (QS/Fb) achieved lower average noise strength values than the less complex only-feedback one in this scenario (NoQS/Fb). On the other hand, when extrinsic noise was present constitutive expression was clearly noisier than any of the more complex QS/Fb and NoQS/Fb circuits for high protein mean expression values, though not for low ones where intrinsic noise dominates. Thus, the circuit complexity contribution to noise depends not only on its size, but in the interplay between size and noise structure. Also, at the mean expression of the protein of interest for industrial biotechnology (medium-high range), tuning circuit parameters with both quorum sensing and feedback clearly allows both intrinsic and extrinsic noise to be coped with, independently of the circuit structure.

The experimental results, though preliminary, showed a high concordance with the computational ones and confirmed the capability of the proposed QS/Fb circuit to reduce noise strength.

Chapter 7

Performance tuning via multi-objective optimization

7.1 Introduction

Up to now, we have discussed about how to extract information from a gene circuit model to design and construct this circuit in the lab. Nevertheless, several problems arise when building up biological devices by combining parts. First, composing different biological parts and devices together can be difficult, even if assuming a synthetic circuit structure has been properly designed to have a pre-specified dynamic behavior. This is because the desired input and output levels of a module are often unknown, difficult to measure quantitatively, or difficult to compare. Next, the ratio part/device performance may be altered due to the interaction of loads in the combined system, the so-called retroactivity (Jayanthi et al., 2013). Along with this, there is an ever-growing appreciation for biological complexity, which requires new circuit modeling and design principles to overcome barriers such as metabolic load, cross-talk, resource sharing, and gene expression noise (Church et al., 2014; Mélykúti et al., 2014; Oyarzún et al.; Picó et al., 2015). Finally, one must never forget the gap between computational or ***dry-lab* design, and *wet-lab* implementation**. In practice, biological parts are subject to uncertainty. Circuit structure design and parameters tuning methods must cope with this uncertainty in the biological parts and context to narrow the gap.

To this end, the modular and systematic design of gene circuits saw in Chapter 6, i.e. the systematic way of finding combinations of components from a library of standard parts allowing to optimally perform a pre-defined function, can be formulated using an optimization framework (Feng et al., 2004; Dasika and Maranas, 2008; Rodrigo et al., 2007). Advanced optimization-based methods, capable of handling high levels of complexity and multiple design criteria have been proposed for the modular and systematic structural design of genetic circuits (Otero-Muras and Banga, 2014). These new approaches combine the efficiency of global Mixed Integer Nonlinear Programming solvers with multi-objective optimization techniques (Banga, 2008; Sendin et al., 2010).

A natural approach to model-based tuning of synthetic circuits consists of the analysis of the effect of key parameters that can be used as tuning knobs in the experimental implementation. In this approach, selection of biological parts is understood as choice of the values of key parameters that yield the device's desired dynamical behavior. A current challenge is to devise methods to provide the set of circuit parameters that satisfies a specified circuit behavior in a way that can be readily used for their wet-lab implementation (Miller et al., 2012). Thus, for instance, in (Ellis et al., 2009), the authors synthesize regulatory promoter libraries, characterize key parameters, and use them to guideline the construction of synthetic networks with different predicted input-output characteristics. Global sensitivity analysis is used in (Feng et al., 2004). The sensitivity information is used to guide the selection of circuit components and thereby reduce the wet-lab implementation effort. In (Koeppl et al., 2013) the authors express the desired behavior as a functional cost index of the desired circuit trajectories. Then, the inverse sensitivity of the mapping between parameters and cost index is obtained after linearising the functional cost index around an initial value of the model parameters. This local inverse mapping is used to map a region of specifications into a one of parameters.

Although the specification of the desired circuit's dynamic is most often naturally expressed as a multi-objective global optimization problem, this approach has not been used so far. Instead, current approaches define independent thresholds set *a priori* for each of the functional goals characterizing the desired behavior of the circuit. Then, global Monte Carlo-like approaches are used, sampling the parameters space and simulating the circuit time response. The result of these simulations is used to assess the circuit behavior, so as to profile the subset of the parameters space that result in circuit behavior fulfilling all thresholds. After this, some statistical post-treatment of the results is used, like clustering or correlation analysis or global sensitivity analysis to draw conclusions between the distribution of the parameters, and the circuit behavior (Chiang and Hwang, 2013). This Monte Carlo based framework has a huge computational cost. Given a defined search space in the parameters space, the Monte Carlo sampling does not ensure that a solution will be found, thus requiring a large number of samples to find solutions. This problem increases as the thresholds defining the acceptable circuit behavior are more stringent. Also, the solution space obtained weighs, either equally or *ad hoc*, all the functional goals of the circuit. Thus, besides missing many possible optimal solutions, there may be little variability among the different solutions in the parameters space, making the statistical post-treatment less sensitive.

Regarding the multi-objective optimization design (MOOD) framework described in section 2.5, this Thesis proposes this approach to build a given functional device with desired dynamic behavior, and obtain a model-based set of guidelines for the selection of its biological parts. Thus, the optimal design of both I1-FFL and QS/Fb gene circuits was developed by using the MOOD methodology. Some of the contents of this Chapter have appeared in the following journal and congress publications

- Y. Boada, G. Reynoso-Meza, J. Picó, and A. Vignoni. Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices: an adaptive network case. *BMC Syst Biol*, 10(1):27, 2016b.

- Y. Boada, A. Vignoni, and J. Picó. Multi-objective optimization for gene expression noise reduction in a synthetic gene circuit. *IFAC-PapersOnLine*, 50(1): 4472 – 4477, 2017b. ISSN 2405-8963. 20th IFAC World Congress.

- Y. Boada, J. Pitarch, A. Vignoni, G. Reynoso-Meza, and J. Picó. Optimization alternatives for robust model-based design of synthetic biological circuits. *IFAC-PapersOnLine*, 49(7):821 – 826, 2016a. ISSN 2405-8963. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB.

The outline of this Chapter goes as follows: The first section 7.2 is dedicated to the MOOD framework and its three steps (1) Defining the circuit behavioral specifications, (2) optimization of the cost function, and (3) deducing guidelines for the wet-lab implementation (see Fig.7.1). In section 7.3, the MOOD is applied to the I1-FFL gene circuit trying to achieve the important biological function so-called *adaptation*. The advantages of the proposed framework are analyzed by comparing two different optimization algorithms with the Monte Carlo approach. Then, section 7.4 focuses on how to minimize the protein noise but maintaining a desired mean level in the QS/Fb circuit by using MOOD. In the final section 7.5, the main conclusions of the Chapter are drawn.

7.2 Multi-objective optimization design framework (MOOD)

Achieving a synthetic biological circuit to fulfill some behavioral specifications requires in practice an iterative process through three main steps: *i)* choosing a gene circuit structure capable to perform the desired behavior after proper tuning of its parameters, *ii)* tuning the circuit parameters, and *iii)* validating the circuit with the selected tuned components. The use of models to solve the first two subproblems *in silico*, before attempting the wet-lab implementation to validate the circuit, reduces the wet-lab effort and speeds-up the design process. This part of the Thesis focuses on the second subproblem: *in silico* tuning of the circuit model parameters, so as to achieve the desired behavioral specifications.

First, a topology for the functional module or gene circuit is needed, capable to accomplish the desired behavior after the suitable tuning of its parameters. This will provide the circuit model structure. Although currently there are no *catalogues* as such functional modules, there is a vast literature in the systems biology area on network motifs producing a variety of dynamic behaviors, as was described in section 3.2. Alternatively, one may find the potential circuit structure casting the problem as an optimization one, starting from coarse-grained models of the potential circuit structural components, and looking for the optimal circuit topology (Otero-Muras and Banga, 2014).

Models may have different degrees of detail. The key goal of this Chapter is to tune the model parameters using a degree of detail in the model amenable to serve as basis to provide guidelines for the experimental implementation of the circuit. That is, the parameters to be tuned should correspond to biological tuning knobs that can be modified experimentally (Arpino et al., 2013). In certain cases such as the reduced-order models for both I1-FFL and QS/Fb gene circuits, even if aggregated, the parameters clearly match with experimental biological tuning knobs (Hancock et al., 2015; Anderson et al., 2011; Prescott and Papachristodoulou, 2014).

From this starting point, one can proceed to tune the model parameters so that eventually the circuit fulfills the behavioral specifications. In this Thesis, the general case when a set of specifications is desired will be considered, and thus leading to a *multi-objective problem*. A usual approach to face a multi-objective problem consists of building an aggregate function in order to assemble the design objectives in a unique index, normally by means of a weighting vector. This approach is followed for example in (Chiang and Hwang, 2013). However, the solution obtained depends too much on the correct selection of the weighting factors, and it might not possibly reflect with enough clarity the designer's preferences in relation with the desired balance of requirements.

An alternative option is to use multi-objective optimization (section 2.5). This is a natural choice to face this kind of problems. Chapter 5 showed that in multi-objective optimization all design objectives are important to the designer, so all of them are optimized simultaneously. Thus, the solution rarely is unique, but a set of solutions is obtained called the *Pareto Front* (already described in section 2.5). In this sense all solutions are Pareto-optimal and differ from each other in the trade-off of objectives each one represents. Again from Chapter 5, this overall **multi-objective optimization design (MOOD)** procedure enables to analyze design objectives trade-offs to implement a preferable solution (see Fig.7.1). Furthermore, it may provide a better understanding of the problem at hand by the so-called process of *innovization* through optimization as stated by (Deb et al., 2014). Next, the MOOD steps involved in tuning model parameters are described.

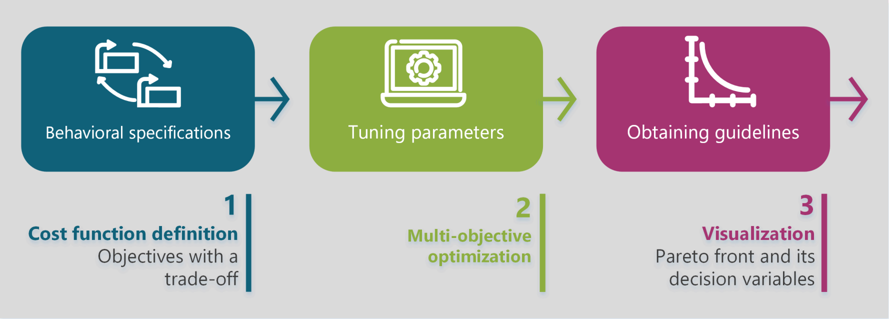

Behavioral specifications
Tuning parameters
Obtaining guidelines
1
Cost function definition
Objectives with a trade-off
2
Multi-objective optimization
3
Visualization
Pareto front and its decision variables

Figure 7.1. MOOD framework for optimal design and wet-lab implementation of a synthetic gene circuit.

7.2.1 Defining the circuit behavioral specifications

The first step of the proposed methodology is the **multi-objective problem definition**, that is, the specification of the desired dynamical behavior for the circuit to be designed (see Fig.7.1). This can be done in several ways. From the designer's point of view, specifying the circuit behavior in terms of the desired output signal profile for a given input signal profile is a natural approach (Ang et al., 2010b). The input signal is chosen as the one that is going to be used in working conditions, or as simple standard probing input-signals (e.g. step-like, sinusoidal, or pulse ones). Once the desired input-output relationship is defined, the set of circuit parameters achieving it can be obtained by optimization-based system identification (Banga, 2008).

This approach is useful for linear dynamical systems, as their time-response to these probing signals fully characterizes the circuit dynamical behavior. This is not the case for nonlinear circuits as the ones typically encountered in synthetic biology. Thus, the particular signal to be used in working conditions should be chosen. Yet, this may be very restrictive. Indeed, usually the input signal to a circuit will have varying characteristics. In the best case, it will belong to a given class (e.g. step-like signal with varying amplitude). Therefore, the dynamical behavior, i.e. the desired circuit time-response to a given input signal, is better given as a set of input-output performance indexes to be optimized.

Specifying the desired circuit behavior in terms of performance indexes to be optimized has many advantages. In the general case, the indexes will take the form of functionals mapping the circuit trajectories to the reals. Thus, for a circuit with dynamics given by the model

$$\begin{aligned} \dot{x} &= f(x,\theta) \\ 0 &= g(x,\theta) \end{aligned} \tag{7.1}$$

where $x \in \mathrm{I\!R}^n$ is the state, $\theta \in \mathrm{I\!R}^p$ the parameters, and function $g(.)$ represents algebraic constraints in the system. The indexes can be expressed as

$$J_i(\theta) = \int_{t_0}^{t_f} h(x(\tau,\theta),\tau)d\tau \tag{7.2}$$

for some function possibly time-dependent function $h(.)$ of the system trajectories during a time interval of interest $[t_0, t_f]$, being $i = 1 \ldots n_i$ is the number of indexes. These can be made valid for a whole class of input signals. The indexes may consider other signals in the circuit besides the input and output ones, robustness with respect to uncertainty in the circuit parameters can be included, etc. They will typically consider the desired performance at steady-state (*precision*), and some measure of the quality of the transient. Proper definition of the optimization indexes representing the desired behavior is a key point. An incorrectly specified objective, which does not properly represent the actual desired behavior will lead the optimization in a wrong direction. It will return a parameter set that will give misleading design guidelines. Moreover, for the proper interpretation of results by the designer, one must pose meaningful design objectives.

7.2.2 Multi-objective parameters tuning

As mentioned above, representing the desired behavior will eventually lead to several objectives to be optimized . That is, the optimization problem will be a multi-objective one in the general case. Typically, some of the objectives will be in conflict, so a trade-off among solutions is required. In the Thesis, the problem is addressed as a truly MOOD problem, following the same principles used in Chapter 5.

Section 2.5 exposed that the multi-objective optimization process seeks to approximate the best parameters θ_P^* that give the best Pareto front approximation J_P^*. Such search could be done through a random Monte Carlo sampling in the decision variables space θ –the set of parameters determining our biological model–, followed by filtering of the solutions in order to obtain the θ_P^* that defines the Pareto front approximation J_P^*. This could be a good option for problems with few decision variables. But, for problems with a large number of decision variables, as the I1-FFL or the QS/Fb case, it is more efficient to use an appropriate multi-objective optimization algorithm to approximate this solution.

7.2.3 Obtaining tuning guidelines for implementation

Figure 7.1 shows the three steps of the MOOD. in the second one, a set of solutions is obtained: values for the kinetic parameters that represent a trade-off between the objectives. Then, the third and final step consists of deducing tuning guidelines and select the values of the kinetic parameters of the model, which also enable the gene circuit construction in the lab.

In this Thesis, two interconnected tools are used: *i)* an optimized clustering of the solutions, and *ii)* a visualization of the Pareto front and set using suitable tools. In both cases the goal is to provide guidelines on the tuning of those parameters that have been identified as proper **tuning knobs** for experimental lab implementation. In other words, the kind of information extracted represents qualitative levels for the kinetic parameters that can be commonly modified in the wet-lab as was explained in section 2.1.2. For instance:

- **Plasmid copy number** . It can be tuned by selecting the appropriate replication origin of the plasmid.
- **Promoter strength**. It can be modified by selecting the appropriate promoter with predicted strength; for example from the Anderson Promoter library (Anderson, 2006) available at the iGEM Parts Registry.
- **Ribosome Binding Site strength**, and is one of the easiest parameters to tune in the wet-lab using RBS libraries, the RBS Calculator from Sallis Lab (Salis et al., 2009b), or nucleotides repetition (Egbert and Klavins, 2012).
- **Protein degradation rate**. It also can be tuned globally by changing the growth rate of the microorganism, or by adding a protein degradation tag to include the protein in an active degradation pathway.

In order to facilitate obtaining the guidelines, a *hierarchical clustering* is performed with the solutions (see Matlab script in annex E.1), including the values of the objectives and also the kinetic parameters of each solution. This process is achieved by using a *cluster tree based on the Euclidean distance* among the vectors containing the attained values of the objectives for all points along the Pareto front. The distance among clusters is obtained by means of the weighted center of mass distance. Then, the number of clusters is set in an iterative manner from 10 to 2. After that, a *Kruskal-Wallis* (Kruskal and Wallis, 1952) test is performed for each iteration in order to analyze the correlation between the kinetic parameters and the clusters. With this process the optimal number of clusters is selected by choosing the one that maximizes the number of significantly correlated parameters with the clusters.

All the resulting correlated parameters has different value ranges in each cluster, which in turn represents a guideline for this parameter. For example, one parameter can range around low values (w.r.t. the initial interval for that parameter) for some clusters, and high values for other clusters. These parameters with strong correlation between their values and the cluster they belong to are used as guideline ones.

The parameters that do not exhibit a significative correlation between their values and the clusters are parameters whose range of values will not affect to which cluster the solution belongs. This range may be ideally a narrow one, but could be wide in some cases. The width of the solutions range will be an indicator of the sensitivity of the optimal solutions with respect to that parameter.

It is accepted that visualization techniques are valuable in order to analyze the trade-off among competing objectives. As in Chapter 5, Level Diagrams (LD) was the visualization tool chosen in this Thesis (further details in section 2.5.3) to analyze the different Pareto front and Pareto set approximations. Remember that LD allows to correlate design objectives with decision variables.

A graph for each objective is displayed (see section 2.5.3), where the Y-axis is the p-norm $\|\hat{J}(\theta)\|_p$ of the objectives vector, and the X-axis corresponds to the objective value or decision variable depending on the case. A second graph displays $\|\hat{J}(\theta)\|_p$ with respect to each decision variable. These characteristics make it helpful in order to share the clustering information from thedesign objectives space and the decision variables space.

In order to incorporate the information obtained from the clustering, the Y-axis of the LD plot was modified to show the membership of a solution to a cluster, therefore improving completeness for this problem. The solutions were also color-coded in all graphs, improving persistence and simplicity. This correspondence of colors helps to evaluate general tendencies along the Pareto front and compare solutions according to the clusters they belong to. Additionally, a complete analysis was achieved by plotting the dynamic response of each species in the system using the same color code.

All these steps were performed using Matlab scripts that are described in annex E.1[1]. Finally, it is interesting to note that the selection of the preferable solution according to designer's criteria, or equivalently the extraction of qualitative levels for the parameters, takes place in an *a-posteriori* multi-criteria analysis of the Pareto Front approximation.

7.3 I1-FFL tuning using MOOD

Using the framework above, the kinetic model parameters of the I1-FFL gene circuit were tuned to achieve adaptation behaviour. The idea is to apply the three steps of the MOOD considering the I1-FFL model (4.27) presented in section 4.2, with the same species listed in Table 7.1. The desired behavior of the output protein x_8 depends on the input level x_9, which are highlighted in Table 7.1. Later I will show two scenarios related with the lab-implementation and usability of the obtained guidelines.

[1] Matlab scripts are also publicly available at http://sb2cl.ai2.upv.es/content/software

Table 7.1. Variables of the I1-FFL model.

Variable	Species	Description	Units
x_1	mR	*luxR* messenger RNA	nM
x_2	R	LuxR protein	nM
x_3	A	AHL intracellular inducer	nM
M	(R.A)	LuxR and AHL monomer	nM
x_4	$(R.A)_2$	dimer of (R.A)	nM
x_5	mI	*cI* messenger RNA	nM
x_6	I	cI protein	nM
x_7	mG	*gfp* messenger RNA	nM
x_8	G	GFP protein (output)	nM
x_9	$\mathrm{AHL_{ext}}$	AHL extracellular inducer (input)	nM

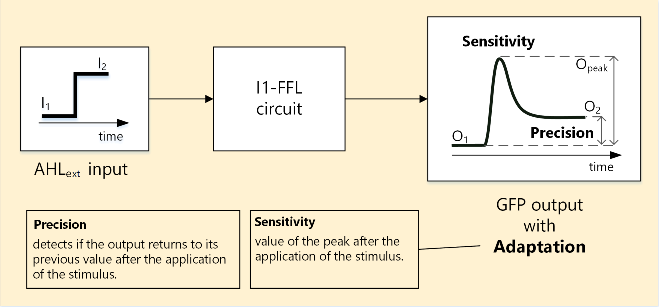

I₂
I₁
time
AHLext input
I1-FFL circuit
Sensitivity
Opeak
O₂
O₁
Precision
time
GFP output with
Adaptation
Precision
detects if the output returns to its previous value after the application of the stimulus.
Sensitivity
value of the peak after the application of the stimulus.

Figure 7.2. Input-output adaptive behavior. Adaptation is an important property of biological systems, related to homeostasis. After an input stimulus the output signal responds by first quickly reaching a peak value, after which it returns to its previous value even if the stimulus persists.

7.3.1 Multi-objective problem definition

The first step of the MOOD framework is to formulate the circuit specifications as design objectives to be optimized. Recall the desired input-output behavior for the I1-FFL circuit, depicted in Fig.7.2.

Two basic objectives can be considered for this circuit (Ma et al., 2009; Ang et al., 2010a; Chiang and Hwang, 2013; Rodrigo and Elena, 2011):

- **Sensitivity:** after input stimulation, a clear transient peak value is desired for the output. Sensitivity can be defined in relative terms as the relationship between the input and output variation during the transient. In this case, sensitivity was defined as the ratio between the absolute total variation of the output signal –the GFP protein concentration x_8–, and the variation of the input signal –the external AHL inducer x_9– (both highlighted in Table 7.1).
- **Precision:** after the peak transient, the output must go back to its value previous to circuit stimulation. Thus, precision can be defined as the inverse of

Table 7.2. I1-FFL model parameters selected for optimization.

Parameter	Wet-lab implication	Range of values
$\mathrm{k_{mI}C_I}$, $\mathrm{k_{mG}C_G}$	Promoter strength and Plasmid origin of replication	[1 200] min^{-1}
$\mathrm{d_I}$, $\mathrm{d_G}$	Degradation tag sequence	[0.01 0.3] min^{-1}
$\mathrm{k_{pI}}$, $\mathrm{k_{pG}}$	RBS Strength	[1 100] min^{-1}
γ_1	$\mathrm{P_{lux}}$ promoter Hill constant	[50 200]
γ_3	$\mathrm{P_{lux/cI}}$ promoter coefficient	[1×10^{-4} 0.5] nM
γ_4	$\mathrm{P_{lux/cI}}$ promoter coefficient	[5×10^{-4} 5]
γ_5	$\mathrm{P_{lux/cI}}$ promoter coefficient	[1 100] nM^{-1}

the normalized output error. The lower the steady-state error, the higher the precision.

Let θ denote the parameters selected for optimization from the I1-FFL model (4.27) in Chapter 4. $\theta := [k_{mB}C_{gB}, k_{mC}C_{gC}, k_{pB}, k_{pC}, d_B, d_C, \gamma_1, \gamma_3, \gamma_4, \gamma_5]$ is the set of decision variables. All these parameters are suitable of wet-lab modification (further information in Table 7.2).

The two design objectives for the I1-FFL circuit can be mathematically expressed by means of the indexes

$$\begin{aligned} J_1(\theta) = & \quad \frac{2\,(x_9(t_f) - x_9(t_0))}{\int_{t_0}^{t_f} |\frac{dx_8}{dt}| dt} \\ J_2(\theta) = & \quad \frac{x_8(t_f) - x_8(t_0)}{x_9(t_f) - x_9(t_0)} \end{aligned} \tag{7.3}$$

where t_f is the time length of the experiment, and the input stimulus is applied at t_0.

Sensitivity is the inverse of $J_1(\theta)$. Notice the total absolute variation of the GFP protein concentration is obtained as half the accumulated absolute value of the time derivative of the GFP concentration (x_8). The lower $J_1(\theta)$ (larger output peak w.r.t. input variation), the higher the sensitivity.

Precision is the inverse of $J_2(\theta)$, i.e. the inverse of the ratio between the variation of the GFP protein concentration between t_0 and t_f, and the variation of the external inducer concentration between t_0 and t_f. If the GFP protein concentration x_8 at time t_f is the same as the initial one at time t_0, precision is infinite.

Note that both objectives are defined as the inverses of Sensitivity and Precision in order to use them in the *minimization problem*, as it is the standard for optimization problems (Miettinen et al., 2008).

Additionally, other objectives could be considered. For instance, fulfillment of constraints on the species. In the I1-FFL case, in order to obtain realistic solutions regarding the concentration values of the cI protein (x_6), its absolute total variation was taken into account as a constraint. This can be expressed as

$$P(\theta) \quad = \quad \int_{t_0}^{t_f} |\frac{dx_6}{dt}|dt,$$

Hence, the considered constraint is

$$1 < P(\theta) < 10000 \tag{7.4}$$

To make precision higher (that is, low output error) the easiest option is to have very high concentration values of the protein cI, which acts as repressor of the protein GFP. To avoid this unrealistic solution, it is possible to force the concentration of the protein cI to have an upper bound. In case of not having this restriction, the solutions may have higher precision at the cost of unrealistically high values of cI concentration. The restriction penalizes this fact and drives the search to a different region of the parameter space (going away from the undesired region, corresponding to high values of protein cI).

Another relevant issue is the definition of the limits for $J_1(\theta)$ and $J_2(\theta)$ beyond which precision and sensitivity degrade and adaptive behavior is not achieved anymore (Ma et al., 2009). This is the so-called *pertinency* range of the objectives. The limits established in this Thesis are: $J_1(\theta) \in [1 \times 10^{-3}\ ,\ 200]$, and $J_2(\theta) \in [1 \times 10^{-4}\ ,\ 20]$.

Finally, optimization looks for a set of values for the 10 decision variables θ that minimize both objectives. Yet, precision and sensitivity are conflicting objectives. So a trade-off must be reached. Therefore, this problem can be formulated as a multi-objective problem (MOP)

$$\begin{aligned} \min_{\theta \in \Re^{10}} J(\theta) = \qquad & [J_1(\theta),\ J_2(\theta)] \in \Re^2 \\ \text{subject to:} \qquad & \text{I1} - \text{FFL dynamics (4.27)} \\ & 1 \times 10^{-3} < J_1(\theta) < 200 \\ & 1 \times 10^{-4} < J_2(\theta) < 20 \\ & 1 < P(\theta) < 1 \times 10^{4} \end{aligned} \tag{7.5}$$

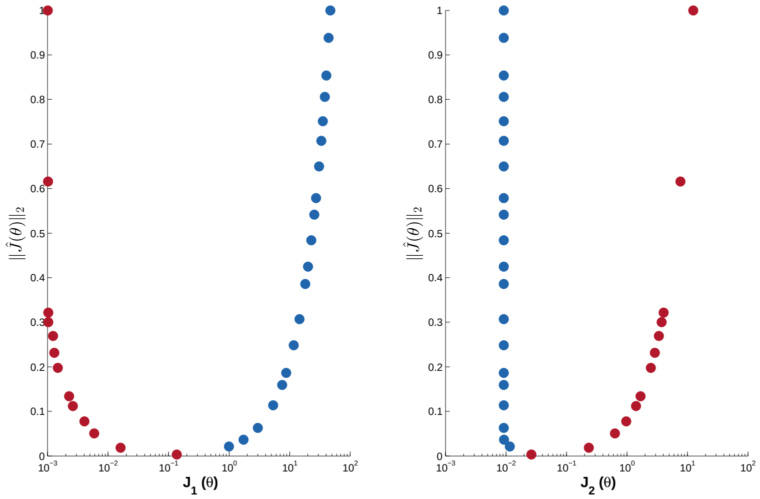


Figure 7.3. Pareto Front as the distance to ideal point. $J_1(\theta)$ is the sensitivity and $J_2(\theta)$ is the precision objectives. Both objectives show a typical trade-off. Cluster 1 is plotted in red circles and cluster 2 is plotted in blue circles.

7.3.2 Optimization

As a second step, dynamic optimization of (7.5) using the multi-objective differential evolutionary algorithm spMODE (see section 2.5) was carried out. Starting from an initial random population of candidate solutions, 15.000 iterations as the maximum number of evaluations of the objective functions were set. A Pareto front containing 33 solutions that achieve adaptation, together with the Pareto set θ containing the model parameters corresponding to the Pareto front solutions resulted from the optimization. The original Level Diagrams of the Pareto front and set are illustrated in Fig.7.3. these diagrams are useful in case the designer needs to get more insight for the guidance of its multi-criteria decision-making.

As expected, the solutions show as expected, a trade-off. Solutions range from high sensitivity (low values of J_1) and low precision (high values of J_2) ones to low sensitivity (high values of J_1) and high precision (low values of J_2) ones. Note in all cases these solutions are the optimal ones, in the sense of Pareto.

Additionally, a Monte Carlo sampling (MCS) and a Latin Hypercube sampling (data not shown) with the same computational cost were performed for the sake of comparison. In both cases, the solutions must be selected with a dominance filter so as to detect the ones actually fulfilling the constraints and yielding adaptive dynamics (Chiang and Hwang, 2013). Note this functional association step is not required in our approach, as the optimal sets of parameters obtained already correspond to functi-

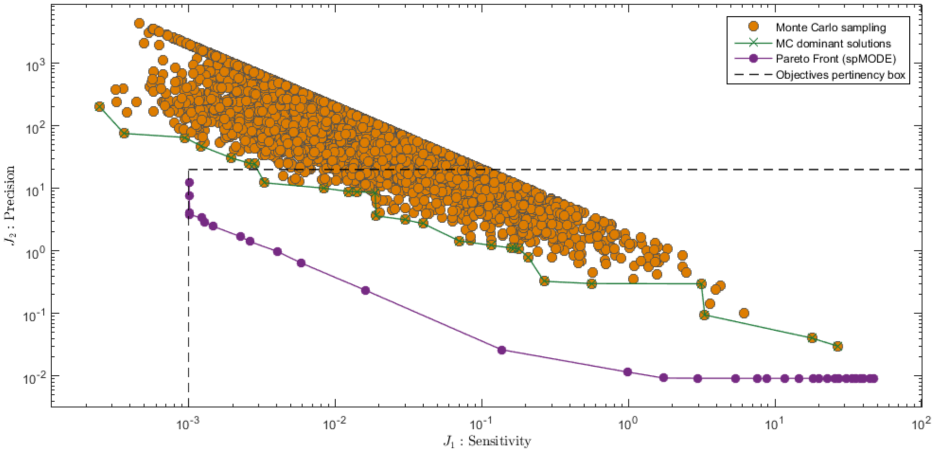


Figure 7.4. Pareto Front comparison. Pareto Front representation for J_1 and J_2 obtained with the spMODE algorithm for the MOO (blue line). Monte Carlo random sampling results are colored in red and the dominant solutions are in green. The time response of the C protein concentration for three representative points are shown.

onal ones. From the functional solutions obtained with these sampling techniques, the corresponding Pareto front was approximated. Figure 7.4 shows the results achieved. The Pareto front from the MCS (dominant solutions in green) covers a larger region of the objectives space, but outside of our region of interest (pertinency box), and it is far away behind the optimal one obtained with spMODE.

7.3.3 Guidelines for implementation

The third step consists on obtaining guidelines and guidance for the implementation of the circuit. To do this, the solutions gathered from the optimization were clustered hierarchically in an agglomerative tree (see annex E.1). The optimal number of clusters were obtained with the procedure explained in section 7.2.3. The parameter intervals corresponding to each cluster are enumerated in Table 7.3.

The intervals for the I1-FFL model parameters from Table 7.3 can be expressed as **general guidelines** that are necessary for achieving *adaptation*:

- $\mathrm{d_I}$: the degradation rate of the protein cI has to be the lowest possible in all cases.
- $\mathrm{k_{pI}}$: the RBS strength of gene *cI* has to be the lowest possible in all cases.
- γ_1: the promoter strength (activation strength), has to be the high in general, but it does not has an apparent effect.

Table 7.3. Optimal parameter intervals providing design guidelines. Each one of the optimized parameters either has a common range for all clusters, or is a trade-off tuning knob determining specific clusters.

Parameter	**Initial parameter range**	**Design Guideline**		
		Common range	Cluster 1	Cluster 2
$\mathrm{km_R C_R}^*$	[1 200]	-	[1 171.91]	1 *
$\mathrm{km_I C_I}$	[1 200]	-	1	[1 200]
$\mathrm{km_G C_G}$	[1 200]	-	[1 171.91]	1
$\mathrm{k_{pI}}$	[1 100]	1	-	-
$\mathrm{k_{pG}}$	[1 100]	-	[1 15.68]	1
$\mathrm{d_I}$	[0.01 0.3]	[0.01 0.0792]	-	-
$\mathrm{d_G}$	[0.01 0.3]	-	[0.2784 0.3]	0.3
γ_1	[50 200]	[78.93 200]	-	-
γ_3	[1×10^{-4} 0.5]	-	[1×10^{-4} 0.013]	[1×10^{-4} 0.0141]
γ_4	[5×10^{-4} 5]	-	[5×10^{-4} 1.4424]	[0.0697 5]
γ_5	[1 100]	-	[1 9.2546]	[12.125 100]

* $\mathrm{km_R C_R}$ is the same as $\mathrm{km_G C_G}$ as they physically in the same plasmid.

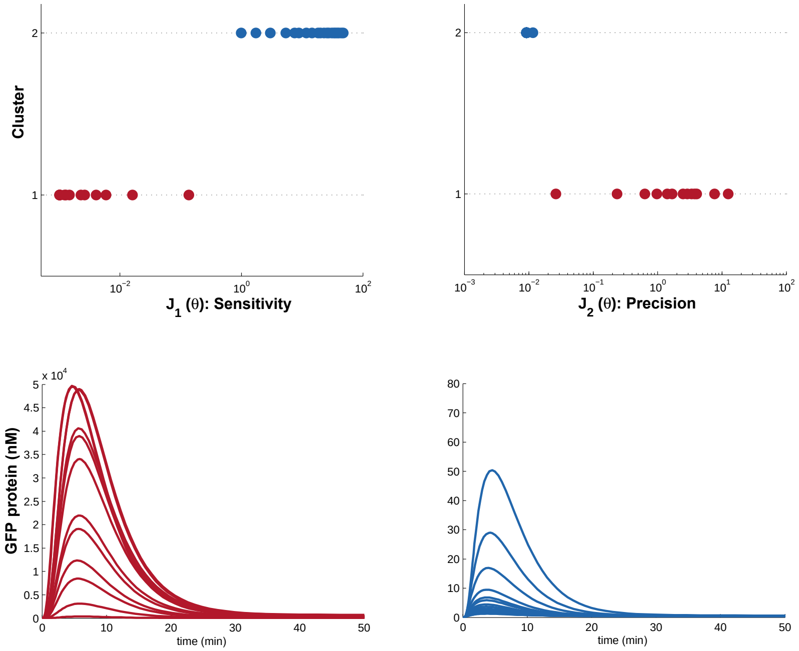


Figure 7.5. Pareto front representation in the cluster-modified LD tool. A. Value of the objectives J_1 and J_2 for each solution where each cluster is identified by a different color. Clusters range from high sensitivity-low precision (red) to low sensitivity-high precision ones (blue). **B.** Time courses of protein C concentration for the different solution in the clusters.

- γ_3: the hybrid promoter strength (activation strength) has to be the lowest possible in all cases.

Depending on whether high sensitivity or high precision are chosen, specific guidelines (see Table 7.3) can be given for the tuning knobs to be modified in the wet-lab. In turn, these guidelines could tune the behavior of the circuit:

High Sensitivity Strategy

- $\mathrm{k_{mG}C_G}$ and $\mathrm{k_{pG}}$: increasing values of the promoter strength and the plasmid copy number of the gene *gfp*, and the RBS strength of the same gene lead to increasing values of sensitivity (higher peak values). These are tuning knobs for sensitivity.
- $\mathrm{d_G}$: the degradation of the GFP protein has to be slightly lower for high sensitivity.
- γ_4 and γ_5: the hybrid promoter strengths (repression, and activation - repression cross combined strength) must be kept low.
- $\mathrm{k_{mI}C_I}$: the promoter strength and the plasmid copy number of the gene *cI* have to be kept the lowest possible.

High Precision Strategy

- $\mathrm{k_{mI}C_I}$: the promoter strength and the plasmid copy number of the gene *cI* is a tuning knob for Cluster 2, increasing precision proportionally to its value.
- γ_4 and γ_5: increasing values of the hybrid promoter strengths lead to increasing values of precision (lower error).
- $\mathrm{k_{mG}C_G}$ and $\mathrm{k_{pG}}$: the promoter strength and the plasmid copy number of gene *gfp*, and the RBS strength of the same gene must be keep low.
- $\mathrm{d_G}$: the degradation of the protein GFP has to be the highest.

The results show that the degradation rate $\mathrm{d_G}$ of the GFP protein is a key parameter to correctly achieve adaptation. With high values of this parameter, the concentration of the GFP protein will to return faster to its original level. Some parameters such as γ in the hybrid promoter of GFP are also required to take certain values for the system to attain the adaptive behavior. In particular, it is interesting to note that the repression strength γ_4 plays an important role, which is in agreement with the analysis in (Basu et al., 2004), where a mutation was performed on the hybrid promoter so as to affect the same parameter. In the case the designed needs more insight, the

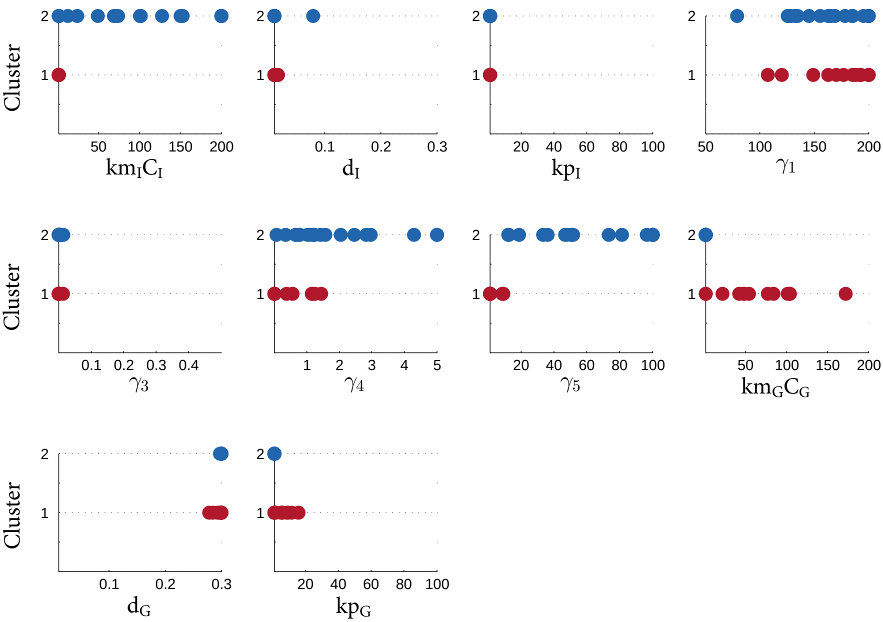


Figure 7.6. Representation of the Pareto set. Cluster-modified LD representation for decision variables (kinetic parameters) in the High Sensitivity Strategy (cluster 1, red dots) and in the High Precision Strategy (cluster 2, blue dots).

tools for visualization to allow a proper decision making procedure and selection of the appropriate parameters for the design were provided.

A clusterized representation of the Pareto front together with the time response of GFP protein concentration for each point are shown in Fig.7.5. Clusters range from high sensitivity and low precision (cluster 1) to low sensitivity-high precision ones (cluster 2). In Figure 7.6 the Pareto set is plotted, including the value of each parameter and its membership to the corresponding cluster. This way is easy to directly find the implication of each parameter in the design. Finally, analyzing the Pareto set plot, it is possible to find that parameters $\mathrm{d_I}$, $\mathrm{k_{pI}}$ and γ_3 have uniform (and tight) values for both clusters, and γ_1 has a uniform and wide range of values also for both clusters.

In case the designer needs further information and insight for guidance of its multi-criteria decision making, the figures in annex E.2 show the original LD of the Pareto front and set.

7.3.4 Application scenario I: Selecting parameters for an implementation

As a proof of concept, and also to validate the guidelines obtained for the I1-FFL system, we will proceed as we would do in the lab. Let us suppose we have built two implementations according to the guidelines proposed earlier: (1) one designed with the High Sensitivity Strategy (Case A), and (2) one with a High Precision Strategy (Case B).

High Sensitivity Strategy

This case is a solution with low precision, but high sensitivity as it belongs to cluster number 1. It is located in the low extreme of J_1, and in the high end of J_2 in Figure 7.5. For this design will use the High Sensitivity Strategy and we will choose, for example, $\mathrm{k_{pG}}$ as a tuning-knob. Changing the value of this parameter will affect the position of the solution in the Pareto front. Although moving exactly along the Pareto front requires modifying more parameters as shown in the guidelines before, we can see (by looking at the reddish dots in Figure 7.7) how the initial solution chosen moves almost on top of the Pareto front. This shows that the obtained guidelines are robust so that we can use the selected parameter as a tuning knob in the wet-lab implementation.

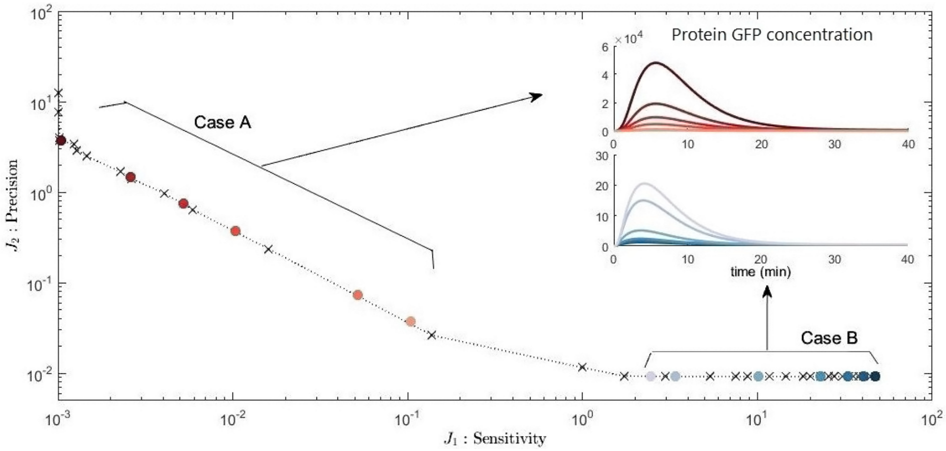


Figure 7.7. Application scenario I. Pareto Front in blue line connected dots. **A.** Dots with reddish color are obtained when using the RBS strength of gene *gfp* as a trade-off tuning knob and represented by modifying $\mathrm{k_{pG}} \in [5 \;\; 0.05]$ starting at the extreme solution. Notice, that decreasing only $\mathrm{k_{pG}}$ it is possible to increase the sensitivity, almost without losing optimality (without getting away from the Pareto front). Inset shows the time course of GFP protein. As expected, sensitivity of the solution is increased, i.e. the peak of protein concentration after stimulus is higher. **B.** Dots blueish color are obtained when using the promoter strength and plasmid copy number gene B by modifying $\mathrm{k_{mI}C_I} \in [200 \;\; 1]$.

High Precision Strategy

For the high precision implementation, it is shown how changing one of the tuning-knobs from our High Precision Strategy (for example $\mathrm{k_{mI}C_I}$) one can almost move along the Pareto front and obtain higher sensitivity solutions without losing precision, as shown by the blueish dots. In the insets of Figure 7.7 is possible to see the temporal behavior of the obtained solutions. Conversely to this, changing values of key parameters like $\mathrm{d_G}$ completely destroy the adaptation behavior independently of the selected solution (see Fig.E.1 in annex E.2).

7.3.5 Application scenario II: Output robustness analysis

This framework is also useful to analyze the output performance of the designed functional device when connecting it to other devices.

Here, a simple binding reaction is used as a load to demonstrate the procedure (see Fig.7.8) is used. This is one of the most common types of load. For example, GFP protein (or a generic protein C) could be a transcription factor and bind to a promoter region in the DNA. The next equations model this load binding reaction as

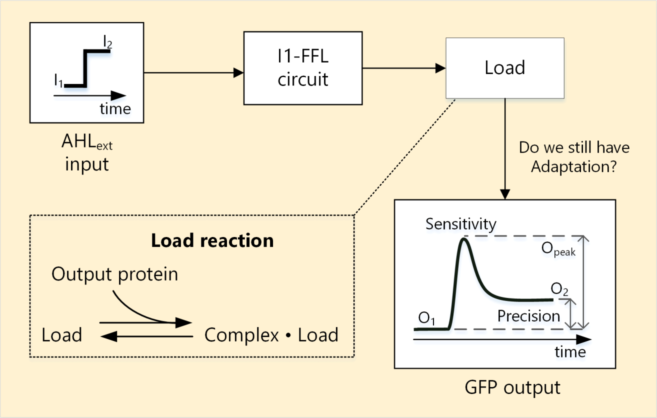


Figure 7.8. Application scenario II. Depiction of the incorporation of information on the context. Connecting our module to a load.

$$
\begin{aligned}
\dot{x}_8 &= \mathrm{k_{pG}} x_7 - \mathrm{d_G} x_8 - \mathrm{K}_1 x_8 x_{10} + \mathrm{K}_2 x_{11} \\
\dot{x}_{10} &= -\mathrm{K}_1 x_8 x_{10} + \mathrm{K}_2 x_{11} \\
\dot{x}_{11} &= \mathrm{K}_1 x_8 x_{10} - \mathrm{K}_2 x_{11}
\end{aligned}
\tag{7.6}
$$

where x_{10} represents the empty load species (e.g. an unbound promoter or protein), x_{11} represents the complex GFP bound to the load species, and K_1 and K_2 are the binding constants. For this case, $\mathrm{K}_1 = 40\,\mathrm{nM}^{-1}\mathrm{min}^{-1}$ and $\mathrm{K}_2 = 20\,\mathrm{min}^{-1}$, which correspond to a mildly fast binding. The initial condition is $x_{10}(t_0) + x_{11}(t_0) = 800$ nM. Since the model (7.6) did not consider degradation terms, the initial condition represents the total amount of available loading species.

In Figure 7.9, the Pareto front of the loaded device is shown in pink colored diamonds, and the original Pareto front in green circles. Notice that the analysis needs to be performed only along the Pareto front solutions. Thus, it is computationally very efficient. As it is shown for the I1-FFL circuit, solutions with low sensitivity are more affected by the load effect at high values of J_1, i.e. lower peak values of GFP protein. This happens when the concentration of GFP protein is in the order of 800 nM, which is the total amount of loading species concentration in this example.

Finally in the inset of Fig.7.9, the loaded time courses of the protein GFP concentration after stimulus (pink line) are shown and compared with the original ones (green line) for values of the parameters corresponding to solutions 1, 2 and 3. As we see, solution 1 is practically not influenced; but solution 2 is considerably affected. Besides, solution 3 is way out from its location and actually looses adaptation behavior. Consequently, it is possible to use this framework to evaluate the output performance of our designed circuit.

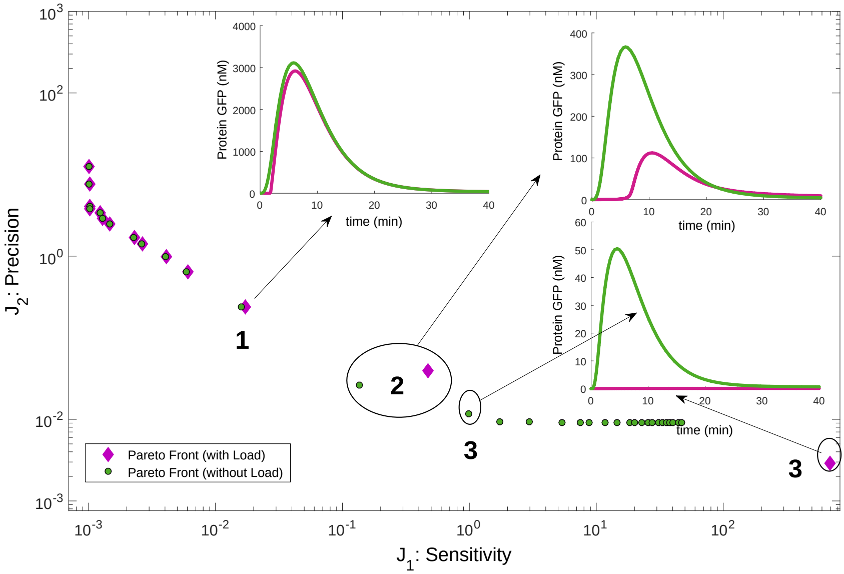


Figure 7.9. Application scenario II. Pareto front of the functional module without load (green circles) and with load (pink diamonds). Inset: temporal responses of the solutions 1, 2 and 3 with (pink line) and without load (green line).

7.3.6 Optimization alternative algorithms

In this section, different optimization algorithms are compared. In particular the spMODE algorithm based on differential evolution, and the large-scale NLP solver (Wächter and Biegler, 2006) embedded in simulation and automatic-differentiation frameworks (Andersson, Joel and Åkesson, Johan and Diehl, Moritz, 2012). Optimizers based on nonlinear programming like IPOPT, or genetic algorithms like spMODE have been used in the past to solve problems with sizes including 15 objectives, and hundreds of decision variables with reasonable computational cost. Their main features are:

- spMODE is a MOEA based on the differential evolution algorithm, which uses a spherical pruning to approximate the Pareto front. Given the stochastic nature of multi-objective differential evolution algorithms like the spMODE, the search for all the possible solutions in the parameters space along the Pareto front is feasible. However, convergence cannot be guaranteed and the tuning the algorithm parameters setting to obtain good performances may be a non-trivial task too.

- IPOPT is an interior-point NLP solver with an automatic-differentiation framework (CasADi[2]) for numerical optimal control which, following the direction of the cost-function gradient, finds an efficient and suitable path from the initial guess to the (possible local) optimum. This has been successfully used in dynamic programming with sequential and simultaneous approaches (Andersson, Joel and Åkesson, Johan and Diehl, Moritz, 2012; Martí et al., 2014).

In contrast to differential evolutionary algorithms, deterministic algorithms are very robust and can guarantee local convergence. Unfortunately, they are very sensitive to the initial guess required for the optimization process, so they may be stuck in a local optimum. Recent developments in *gradient-based nonlinear programming*, which implement automatic differentiation algorithms have provided a good alternative to compute an approximation of the Pareto front by means of these NLP solvers. The main advantage of such tools is feeding the solver (SQP-type (Gill et al., 2005), or the interior-point ones (Wächter and Biegler, 2006)) with the exact Jacobians and Hessians of the objective function and the constraints. This provides a fast and accurate convergence, contrarily to what happens for instance with finite-differences approximations of these derivatives. Also memory handling in the mentioned algorithms is very efficient, as the only information that propagates from generation to generation is the population.

Consequently, here the spMODE and IPOPT optimizers will be tested using the I1-FFL gene circuit to achieve adaptation (the desired behaviour), but comparing performance, and pointing out particular advantages and drawbacks of both alternatives.

The dynamic optimization of equation (7.5) using the spMODE and IPOPT tools was carried out. As in section 7.3.2, in order to evaluate the I1-FFL dynamic behavior in (4.27), the system initial conditions correspond to the equilibrium, that is

$$x(0) = [\frac{k_{mB}C_{gB}}{d_{m_A}}, \frac{k_{mB}C_{gB} \cdot k_{pA}}{d_{m_A} \cdot d_A}, 0, 0, 0, 0, 0, 0, 0]$$

Recall the external inducer concentration $\mathrm{AHL}_{\mathrm{ext}}$ is added as a pulse input to the culture in the lab (see section 3.2.1). However, the I1-FFL circuit inside the cell senses the input as a step-like function due to the passive diffusion process of the $\mathrm{AHL}_{\mathrm{ext}}$ molecules across the cellular membrane. Therefore, $x_9(0) = 50$ nM acts as a pulse input until the I1-FFL gene circuit is relaxed, and reaches again the steady-state.

The spMODE optimization started with an initial population of candidate solutions, chosen randomly within a normal distribution in the parameters search space (provided in Table 7.4). An approximation of the Pareto front with 46 solutions of the multi-objective problem (7.5) was obtained (green curve in Fig.7.10), together with the Pareto set containing their corresponding kinetic model parameters θ. These solutions

[2] Tool available in `https://github.com/casadi/casadi/wiki`

show as in section 7.3.3 a trade-off between good sensitivity (low values of $J_1(\theta)$) and good precision (low values of $J_2(\theta)$).

In order to allow a parallel implementation of the IPOPT algorithm, equation (7.5) is slightly modified. An additional constraint with an user-defined upper bound in $J_1(\theta)$ denoted by $\bar{J}_1$ was set, whereas only $\mathbf{J_2}(\theta)$ is in the objective function. In this way the original multi-objective problem is cast as a set of single-objective optimization problems given by

$$\begin{aligned} \min_{\theta \in \mathbb{R}^{10}} \; & J_2(\theta) \in \mathbb{R} \\ \text{subject to:} \quad & \mathrm{I1-FFL} \;\; \text{dynamics (4.27)} \\ & 1 < P(\theta) < 1 \times 10^4 \\ & J_1(\theta) \leq \bar{J}_1 \end{aligned} \tag{7.7}$$

First, a common initial guess for all the independent optimizations of (7.7) were randomly chosen as $\theta_0 = [5, 20, 0.02, 0.02, 150, 0.005, 1, 10, 10, 10]$. Besides, a well distributed grid of 46 points within the pertinency range for $J_1(\theta)$ was defined, so an approximation of the Pareto front was computed (squared curve in Fig.7.10b). As it can be seen, both solutions found by IPOPT and spMODE are practically the same. Note that all these dominant solutions are the “optimal ones” in the Pareto sense.

The output protein evolution (GFP concentration in this case) for some characteristic optimal points is also depicted in Fig.7.10a. The different responses go from high sensitivity-low precision (H peak in red) to low sensitivity-high precision ones (L peak in blue). Table 7.4 shows the parameters obtained by the two algorithms for two different types of solutions within the Pareto front: one with high sensitivity, and other with high precision.

The spMODE computed its Pareto front approximation in 42.7 minutes when it was running in parallel in an Intel Core i7-4510U machine. In contrast, the IPOPT solver took 2.69 minutes to obtain its Pareto front, also running in parallel in the same machine (5.84 minutes in a single core). The number of objective function evaluations performed by the spMODE was 3100. The IPOPT algorithm ran 3624 (in total for all optimizations) plus 2231 evaluations of the objective function gradient, 2277 constraints Jacobian evaluations, and 2185 Lagrangian Hessian evaluations.

To complete the spMODE and IPOPT comparison, a random Monte Carlo (MC) sampling was performed (orange points in Fig.7.10b). In this case, a dominance filter is required in order to select the best solutions that will allow to the Pareto front approximation (green dotted curve in Fig.7.10b). Results demonstrate the MC sampling covers a large region in the objectives space, but sometimes outside of the pertinency box, as expected, because there is no simple way to focus a random search in it. In addition, the Pareto front approximation obtained from the MC sampling is

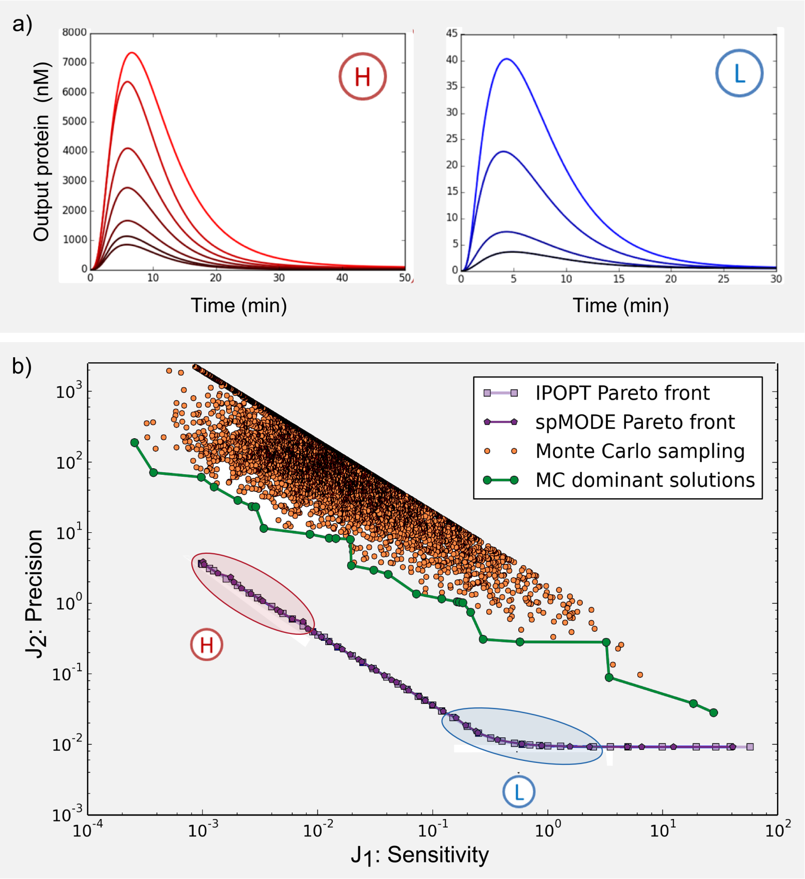


Figure 7.10. **a)** Time evolution of the protein GFP concentration for two sets of Pareto optimal solutions (H ≡ high sensitivity, L ≡ high precision). **b)** Pareto front estimation obtained with the MC sampling (orange dots), spMODE (purple dotted curve), and with the NLP solver IPOPT (purple squared curve).

Table 7.4. Pareto set optimal results.

Parameter	Initial	Sensitivity		Precision	
	Range	IPOPT	spMODE	IPOPT	spMODE
$\mathrm{k_{mR}C_R}^*$	[1 200]	13.041	19.123	1.0012	1
$\mathrm{k_{mI}C_I}$	[1 200]	1	1	14.913	1
$\mathrm{k_{mG}C_G}$	[1 200]	13.041	19.123	1.0012	1
$\mathrm{k_{pI}}$	[1 100]	1	1	10.565	19.742
$\mathrm{k_{pG}}$	[1 100]	9.5174	5.8916	1.0012	1
$\mathrm{d_I}$	[0.01 0.3]	0.01	0.01	0.0221	0.01
$\mathrm{d_G}$	[0.01 0.3]	0.2611	0.3	0.2996	0.3
γ_1	[50 200]	161.6	200	146.92	107.16
γ_3	[1×10^{-4} 0.5]	0.0001	0.0001	0.0205	0.0001
γ_4	[5×10^{-4} 5]	0.8886	0.0005	0.5703	0.0005
γ_5	[1 100]	1	1	4.162	12.121

$^*\mathrm{k_{mR}C_R}$ takes the same value as $\mathrm{k_{mG}C_G}$ because gene *luxR* and gene *gfp* are physically in the same plasmid.

clearly worse than the ones obtained using optimization algorithms, both in accuracy (in the considered pertinency region), and in computational effort (the MC sampling time computing was 2 hours and 10 min).

In summary, all these results show that using a NLP solver (IPOPT) with automatic differentiation to estimate the actual Pareto front is more efficient than a MOEA algorithm (spMODE). Nevertheless, an evolutionary algorithm is a *global* optimizer, which means that it may obtain better approximations of the Pareto fronts in other cases (or allowing a higher number of objective function evaluations). In addition, the performance of a *gradient-based optimizer* highly depends on the provided initial guess, and depends on the particular system "smoothness", so it may be stuck in a local optimum. If this is the case, **a combined evolutionary gradient-based approach** would be a good option, where the fast NLP solver computes a preliminary set of suboptimal solutions to be used later as the initial population for the MOEA. The approach described can be extended to the analysis of interconnection of several devices. However, this will be led in further work, as evident difficulties arise when dealing with larger networks.

7.3.7 Discussion

Computer-aided model-based methods and tools are being increasingly used in synthetic biology to guide the design of synthetic biochemical pathways so as to achieve user-defined functions and behaviors Marchisio and Stelling (2011); Rodrigo et al. (2012); Crook and Alper (2013).

In this work, in order to obtain a set of guidelines to aid the design of synthetic genetic networks with a predefined functionality (functional modules), we developed a framework using a multi-objective optimization design (MOOD) procedure. Compared to previous studies (Chiang and Hwang, 2013), a novel feature of our framework is that the result of the optimization is already a set of parameters that optimally achieve the desired function and dynamics, as encoded in the objective indexes. Specifying the desired circuit behavior in terms of performance indexes to be optimized has many advantages. The indexes or objectives can be made valid for a whole class of input signals, they may consider other signals in the circuit apart from input and output, the robustness with respect to uncertainty in the circuit parameters can be included, etc. The proper definition of the optimization indexes representing the desired behavior is a key point. An incorrectly specified objective, not properly representing the actual desired behavior, will lead the optimization in a wrong direction, thus returning a parameters set that will give misleading design guidelines. This is a drawback, but easier to handle than setting the thresholds defining the acceptable circuit behavior after a Monte Carlo sampling, for these do not ensure that a solution will be found (Chiang et al., 2014; Chiang and Hwang, 2013).

The solutions obtained, i.e. the design objectives together with the respective parameter sets, may be clustered hierarchically, or post-processed with any multivariate statistical analysis tool in order to get further insight into the role of the different parameters. The importance of this, is that the spMODE and LD-tools already order the Pareto front solutions with respect to the objective functions. The LD-tool, as a matter of fact, already provides insight into the role of the different solutions. Further statistical processing is very efficient, as only a small set of data has to be processed (the solutions at the Pareto front), and this set is already ordered. This allows us to reveal and understand associations of parameters and functionality. For example, cluster 1 (red) in the Results Section has the highest sensitivity together with the lowest precision. To implement in the wet-lab a system with this functionality, the RBS in gene *cI* has to be weak, and it should be cloned in a low copy plasmid, as reflected by the guidelines obtained for parameters kp_I and Km_ICgI, respectively. On the contrary, to implement a cluster 2 (blue) system, the guidelines obtained for the same parameters tell us to put gene *cI* with also in a weak RBS and but in a high copy plasmid (Figure 7.6).

For a given circuit design with a desired functionality, the guidelines for the kinetic parameters (Figure 7.6, Table 7.3) are very useful to decide which biological components to use out of the ones available from a library of biological parts, such as the MIT Registry of Standard Biological Parts (Biobrick Foundation, 2006) by iGEM Foundation, the BIOSS Toolbox (BIOSS, 2006), or BioFab (BioFab, 2006). In particular, for the I1-FFL, we showed that important tuning knobs are:

- Km_XCgX is the lumped value of plasmid copy number and promoter strength. It can be tuned by selecting the appropriate replication origin of the plasmid and the promoter; for example from the Anderson Promoter library (Anderson, 2006) available at the iGEM Parts Registry.

- kp_X represents the Ribosome Binding Site (RBS) strength. It is one of the easiest parameters to tune in the wet-lab using, for instance, RBS libraries, the RBS Calculator from Sallis Lab (Salis et al., 2009b), or nucleotides repetition (Egbert and Klavins, 2012).
- d_X is the protein degradation rate. It can be tuned globally by changing the growth rate of the microorganism. It also can be tuned by adding a protein degradation tag to include the protein in an active degradation pathway.

As more and more parts are deposited and characterized in these libraries, frameworks providing guidelines for the design and wet-lab implementation, like the ones presented here, will gain more applicability and the design of synthetic genetic circuits will become more rationale-based than intuition-based.

The analysis performed in the Application Scenario I, shows that it is possible to use only one parameter to move from the Pareto front to a sub-optimal solution. For example, starting from a solution with high precision and low sensitivity, one can move to a solution with higher sensitivity and lower precision; with the almost no losing optimality. This is very useful in the wet-lab, because it means that once you have the system implemented in the wet-lab, it is possible to change the output of your system in a controlled way by performing the minimum amount of changes to it. The methodology easily allows to check how the initial solution will deteriorate by changing the value of only one parameter (see Figure 7.7). Of course, moving along the Pareto front solutions requires modifying more parameters, i.e. changing the values of the parameters from a cluster to another one; however we showed that the obtained guidelines are really robust and that we can use a particular parameter as a tuning knob in the wet-lab implementation.

In the Application Scenario II, we saw that it is straightforward to have an idea of how much the functionality of the system can be compromised by loading it, i.e. by connecting it to another module. The proposed methodology allows to design the system taking this into account. The analysis is computationally efficient, as it has to be performed only for the Pareto front solutions, and not for the whole objective space. Thus, we foresee that extending the approach to the analysis of interconnecting several devices will not be difficult. In a way, as advocated in (Church et al., 2014), the approach is less like highly modular electrical engineering, and more like civil and mechanical engineering in its use of optimization of modeling of whole system-level taking into account loads and flows.

Notice that the analysis needs to be performed only along the Pareto front solutions. In this case, we are performing a robustness analysis *a posteriori* with the Pareto optimal solutions approximated. That is, the decision making process is carried out at the end of the MOOD process using additional information, in order to select a *robust* configuration. This is congruent with similar analysis of uncertainties and decision making (Vallerio et al., 2015).

In case the decision maker wants to seek actively for a robust set of solutions, a different approach will be required. In this case, in order to get such solutions the robustness measure analysis should be included *a priori* within the optimization process. This leads to different optimization approaches known as robust design optimization (RDO) and reliability based design optimization (RBDO) (Frangopol and Maute, 2003). The former seeks to minimize the sensitivity of a solution; the latter to provide a measure of risk failure. In any case, such optimization approaches are out of the scope of this work and are proposed as future work.

The general applicability of the framework allows us to use it with different functional modules and topologies, as soon as the ODE model can be obtained from reactions, although evidently difficulties will arise when dealing with larger networks. In that sense it is interesting to note the difference between the problem of expensive computation and the one of large-scale optimization. Expensive computation arises when the complexity of the system makes the evaluation of the objective function an expensive task. On the contrary, large-scale is related with the amount of decision variables and the size of the objective space. In the cases we are dealing with, these two problems are coupled. For a larger network, there will be more kinetic parameters (decision variables) and more expensive computation of the dynamics of the system to evaluate the objectives. Nevertheless, one of the key issues will be to obtain a reasonable reduced model to be used by the optimization algorithms. As for these task, genetic algorithms like spMODE have been used in the past with problems with sizes including 15 objectives and hundreds of decision variables with reasonable computational cost, and related research is a hot topic (Lozano et al., 2011; Santana-Quintero et al., 2010). Also memory handling in the mentioned algorithms is very efficient, as the only information that propagates from generation to generation is the population.

In the Thesis we have compared the spMODE (differential evolution) and the IPOPT (gradient-based) optimization algorithms. The performance of the last one highly depends on the provided initial guess, and is related with the particular system "smoothness", so it may be stuck in a local optimum. To avoid this, a combined evolutionary gradient-based approach would be a good option, where the fast NLP solver computes a preliminary set of suboptimal solutions to be used later as the initial population for the MOEA.

7.4 QS/Fb tuning using MOOD

The multi-objective optimization design (MOOD) methodology (see section 7.2) will now be used for tuning the kinetic model parameters of the QS/Fb gene circuit, which was designed in Chapter 6. Recall that the QS/Fb circuit aims to reduce protein expression variability by using a cell-to-cell communication via quorum sensing and a negative feedback loop (see 3.3).

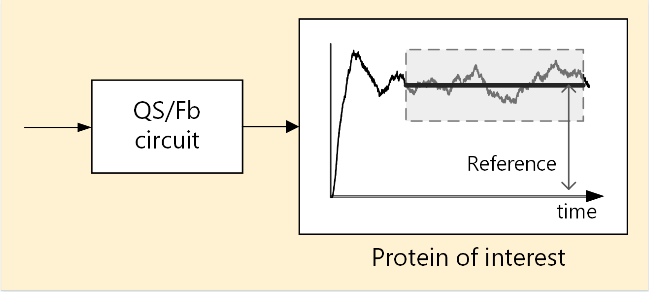


Figure 7.11. Output desired behavior. Tuning the QS/Fb parameters, the desired mean level of the protein of interest can be achieve with the minimum noise strength.

Table 7.5. Variables of the QS/Fb CLE-based model.

Variable	Species	Unit
n_1	PoI/LuxI protein	molecules
n_2	LuxR protein	molecules
n_3	Dimer of (R.A)	molecules
n_4	AHL intracellular inducer	molecules
n_5	$\mathrm{AHL}_{\mathrm{ext}}$ extracellular inducer	molecules
n_6	Monomer (R.A)	molecules

The main species of this synthetic gene circuit are enumerated in Table 7.5, where n_1 is the protein of interest PoI co-expressed with the protein LuxI in the *i*-th cell of the population (see section 3.3). Chapter 6 described how different sets of the QS/Fb model parameters can modify gene expression noise of the protein of interest, by affecting its mean (μ), variance (σ^2) and noise strength ($\eta^2 = \sigma^2/\mu^2$).

Now, we will look for the best kinetic model parameter combinations that **minimize the noise strength for a desired mean** (see Fig.7.11). As in section 7.3, this problem can be formulated using an optimization framework. Again, the MOOD framework is applied to address the Pareto optimal outline of the QS/Fb system, regulating gene expression noise due to both extrinsic and intrinsic fluctuations in a cell population.

The problem of finding the model parameters to minimize gene expression noise will be cast as a multi-objective problem. A *global* multi-objective evolutionary algorithm (spMODE) and a multi-criteria decision making (MCDM) strategy will be used to select the most suitable solutions that minimize the noise strength for a given mean value of the protein level. The spMODE will be applied to find the best approximation to the Pareto front of model parameters corresponding to two scenarios: one with only intrinsic noise, and a second one with both extrinsic and intrinsic noise. Then, MCDM will be performed to analyze how model parameters affect the noise level in each scenario. The Pareto sets obtained in the parameters space from the two scenarios are clustered together, allowing not only to capture parameter groups corresponding to both, but also to find differentiated groups within the scenarios.

7.4.1 Multi-objective problem definition

As a first step, intrinsic noise (I) and the intrinsic plus extrinsic noise (E/I) are defined as two different scenarios, where the mean and noise strength of protein n_1 : Pol/LuxI (μ and η^2, respectively) are formulated as objectives to be optimized as follows

$$J_1(\theta) = \mu_{n_1}, \ \ J_2(\theta) = \eta^2_{n_1} \tag{7.8}$$

The cost function (7.8) computes the mean and noise strength of protein Pol/LuxI for the population of N cells following the expressions

$$\begin{aligned}
m_{n_1}(kT) =& \frac{1}{N}\sum_{i=1}^{N} n_1^i(kT) \\
s^2_{n_1}(kT) =& \frac{1}{N}\sum_{i=1}^{N} \left(n_1^i(kT) - \mu_{n_1}(kT)\right)^2 \\
\mu_{n_1} =& \frac{1}{(k_f-k_0)T}\sum_{k=k_0}^{k_f} m_{n_1}(kT) \\
\sigma^2_{n_1} =& \frac{1}{(k_f-k_0)T}\sum_{k=k_0}^{k_f} s^2_{n_1}(kT) \\
&+ \frac{1}{(k_f-k_0)T}\sum_{k=k_0}^{k_f} \left(m_{n_1}(kT) - \mu_{n_1}\right)^2 \\
\eta^2_{n_1} =& \frac{\sigma^2_{n_1}}{\mu^2_{n_1}}
\end{aligned} \tag{7.9}$$

For both scenarios I and E/I, the mean of protein Pol/LuxI μ_{n_1}, and its total noise strength $\eta^2_{n_1}$ are obtained from the steady-state of the LuxI dynamics over the population of cells. The laws of the total expectation and the total variance (Basak and Chabakauri, 2010) were used. The resulting set of equations in (7.9) implies that $n_1^i(kT)$ is the value of protein Pol/LuxI (in number of molecules) at time instant kT for the *i*-th cell, $k \in \mathcal{N}$, k_0T is the time instant when the steady-state is reached and k_fT is the end of the simulation, and N is the total number of cells in the population.

The goal is to obtain the QS/Fb model parameters yielding a given mean with minimum noise strength. Notice that μ_{n_1} and $\eta^2_{n_1}$ are interrelated magnitudes that are in

Table 7.6. QS/Fb parameters to be optimized.

Parameter	Wet-lab implication	Range of values
α	$\mathrm{P_{luxR}}$ promoter leakage	[0.01 0.2]
$\mathrm{C_R}$	Plasmid copy number times *luxR* transcription rate	[4 60] molecules · min^{-1}
$\mathrm{p_I}$	*pol/luxI* messenger RNA translation rate	[0.2 10] min^{-1}
$\mathrm{d_R}$	LuxR degradation rate	[0.02 0.2] min^{-1}
$\mathrm{k_{dlux}}$	Dissociation constant of $(\mathrm{R.A})_2$ to the $\mathrm{P_{luxR}}$	[10 2000] molecules

conflict. In other words, when one tries to minimize them by finding a single ensemble of parameters, μ_{n_1} improves but $\eta^2_{n_1}$ worsens. That means a trade-off must be reached. Thus, this problem can be solved by minimizing both $\boldsymbol{\mu_{n_1}}$ and $\boldsymbol{\eta^2_{n_1}}$ as two competing objectives. The Pareto front will provide for a given mean, the parameters achieving minimum noise strength. Thus, the optimization problem is defined as

$$\begin{array}{ll} \min\limits_{\theta\in\Re^5} J(\theta) = & [J_1(\theta), J_2(\theta)] \in \Re^2 \\ \text{subject to:} & \text{QS/Fb dynamics (4.42)} \end{array} \tag{7.10}$$

Table 7.6 enumerates the five parameters of the QS/Fb CLE-based model (4.42), which were used as decision variables for optimization, and collectively denoted as θ. Also, Table 7.6 describes the wet-lab implication of these parameters and the ranges wherein they can be modified.

7.4.2 Optimization

As in the I1-FFL case, for this second step two separate optimizations of (7.10) to estimate the Pareto fronts for both scenarios were carried out. Then, the solutions were analyzed to deduce how changes in the decision variables modify the mean and the noise strength of protein Pol/LuxI (μ and η^2, respectively). In both cases, optimization was done using spMODE, starting with an initial population of candidate solutions chosen randomly from a uniform distribution in the parameters space, and setting 15.000 iterations as the maximum number of evaluations of the objective functions were set.

Two approximations of Pareto fronts together with the Pareto sets containing the corresponding parameters for the scenarios I and E/I were obtained. As in section 7.3.2, the solutions in figures 7.12 and Fig.7.13 were plotted using the LD-tool. Since the interest for the QS/Fb gene circuit is in decreasing the variability of Pol/LuxI production, the solutions are illustrated depending on the noise strength results for both scenarios. Each solution from the objectives space or the decision variables space (X-axis) has the same noise strength (Y-axis) in all graphs. The Pareto front analysis in section 7.4.2 shows the classical trade-off. Figure 7.12 depicts the Pol/LuxI noise strength η^2 as a function of its mean μ. The solutions of the Pareto front for intrinsic

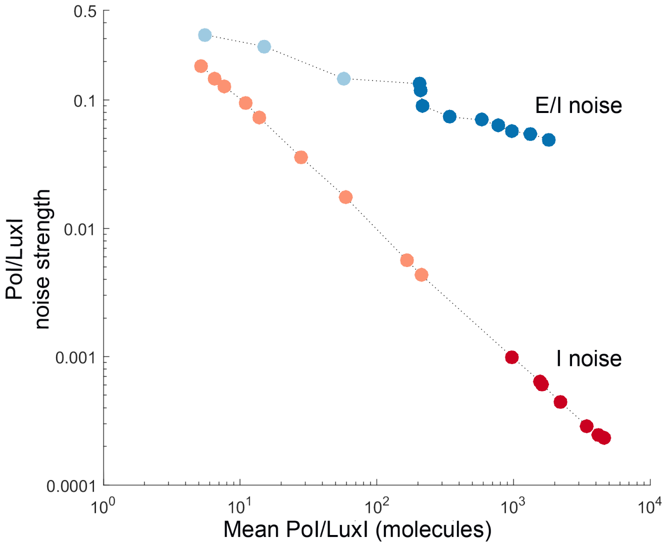


Figure 7.12. Pareto front of both scenarios: considering only intrinsic noise (I) and considering both intrinsic and extrinsic noise (E/I). Each Pareto front is split into two groups: high noise (H) and low noise (L). The Y-axis showing the noise strength is kept in the Pareto set representations following the Level Diagram philosophy, and it is useful to correlate one solution point in both Pareto front and Pareto set spaces.

noise (I) are in red/orange dots, and the ones for the extrinsic and intrinsic (E/I) Pareto front are in blue/light blue. The Pareto front for the I scenario contains 16 solutions that minimize noise with a defined mean, while 12 solutions were obtained for the Pareto front of the E/I noise case. As expected, the E/I front further away from the ideal point (origin X-Y axis of Fig.7.12) than the I one.

7.4.3 Guidelines for implementation

Next, the Pareto sets from both scenarios were gathered and clustered using the *kmeans* algorithm. Interestingly, the clustering is able to capture both scenarios and also finds two subgroups in each scenario: **a high noise, and a low noise** group. Note that the noise strength of each solution was not given to the clustering algorithm. This highlights a correlation between the noise level and the value of the decision variables (i.e. parameters values). Analyzing the solutions obtained, it is seen that the noise strength for E/I case is at least $\sim$ 2-fold larger than the one in the case I. It was expected since the E/I scenario exhibits additional fluctuations from both extrinsic and intrinsic sources. These findings are in total agreement with the experimental evidence (see Fig.7.14) of genome-wide proteomics and transcriptomics in *E. coli* (Taniguchi et al., 2010).

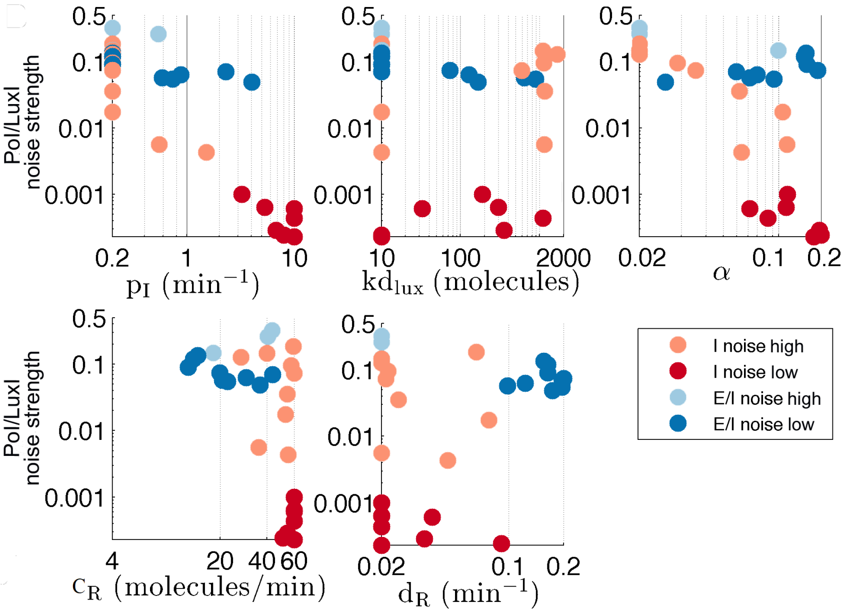


Figure 7.13. Pareto set representation with the value of the model parameters according each scenario I in red/orange and E/I in blue/light blue.

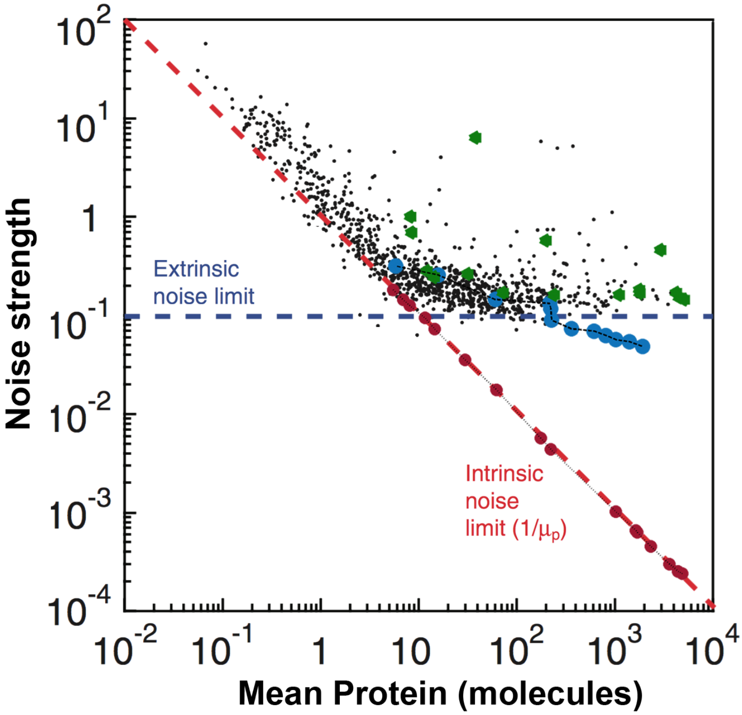


Figure 7.14. Comparison with experimental data and test of noise reduction design strategies. Experimental data of protein abundance and noise in *E. coli* taken from (Taniguchi et al., 2010) is plotted in black dots. Dashed red and blue lines are the intrinsic noise limit and the extrinsic noise limits respectively. The Pareto front obtained in the I scenario is plotted in red dots, and the one in the E/I scenario in blue dots. The green dots are the solutions from the I scenario evaluated in the presence of both extrinsic and intrinsic noise.

Note that the results obtained when only intrinsic noise is taken into account coincide with the experimental limit (red dots lay over red dashed line). Conversely, for the extrinsic noise E/I some parts of the Pareto front (in blue) lay below the experimental limit obtained in (Taniguchi et al., 2010). This is related with the fact that the MOOD is only optimizing 5 parameters (see Table 7.6), and also that the extrinsic noise was fixed to 15 % while in reality it might be larger than that.

Figure 7.13 illustrates the range of values obtained for 5 parameters with different noise strength values in each scenario. These parameters represent the dynamics of the two principal species in our system: Pol/LuxI (top of Fig.7.13), and LuxR (bottom of Fig.7.13). There are several interesting observations deduced from the optimization results and clustering. On the one hand, within the I scenario the kinetic model parameters p_I and α show a clear segregation in the high and low noise groups. They also present a tendency within these groups: **the large value for the parameter**

the lower noise. Recall $\mathrm{p_I}$ is the *pol/luxI* messenger RNA translation rate and α is the $\mathrm{P_{luxR}}$ promoter leakage. On the contrary, the plasmid copy number times *luxR* transcription rate $\mathrm{C_R}$ and the LuxR protein degradation rate $\mathrm{d_R}$ have not a explicit impact on the reduction of the fluctuations in the protein Pol/LuxI. But they do have a preferred ranges of values: $\mathrm{C_R} \in [30\ 60]$ molecules·min^{-1}, and $\mathrm{d_R} \in [0.02\ 0.09]$ min^{-1}.

On the other hand, within the E/I scenario the parameters $\mathrm{p_I}$, α and $\mathrm{d_R}$ present a separation in the high and low noise groups (see Fig.7.13). As before, $\mathrm{C_R}$ has not a clear impact on the reduction of the Pol/LuxI fluctuations, although it does have a preferred range of values: $\mathrm{C_R} \in [12\ 43]$ molecules·min^{-1}. In both scenarios, the dissociation constant of the transcription factor $\mathrm{kd_{lux}}$ can take different values ($\mathrm{kd_{lux}} = 10$ or $\mathrm{k_{dlux}} = 880$ molecules), and still the noise strength η^2_{LuxI} decreases.

Focusing in the low noise groups (I low and E/I low), two different and opposite strategies arise. For a low intrinsic noise (I low, red) a high level of transcription for Pol/LuxI and LuxR are required, together with a slow degradation rate of LuxR protein. On the contrary, for a low extrinsic/intrinsic noise (E/I low, blue) the opposite strategy is needed. E/I low demands small levels of transcription for both proteins Pol/LuxI and LuxR, together with a fast degradation rate of LuxR.

Finally, in order to test these two strategies obtained, a computational experiment was performed. The solutions obtained from the I low strategy were evaluated in the E/I scenario to see how they perform. Figure 7.14 shows the results of this experiment. The original Pareto fronts are in dotted (I) and dashed (E/I) lines. The I solutions evaluated in the E/I scenario (green crosses) are sub-optimal (i.e. they lay above the dotted line), and moreover its performance deteriorates considerably. The noise strength increases in approximately 3-fold for solutions in the low noise group.

To sum up, the MOOD approach finds the best set of model parameters that minimizes the noise strength for each given mean of the target protein expression. This was performed in two different scenarios corresponding to the design taking into account only intrinsic noise, and accounting for both extrinsic and intrinsic noises. The approach allowed us to identify the most relevant parameters that affect both scenarios. The parameters obtained are indeed the ones achieving the minimum possible noise in accordance with genome-wide experimental data available from the literature. The design strategies obtained are not transferable, that is, the strategy obtained considering only intrinsic noise becomes sub-optimal, and its performance decreases when it is evaluated in an extrinsic plus intrinsic scenario. We emphasize the fact that the Pareto front obtained computationally for our circuit fully matches the available genome-wide experimental values of noise strength and mean protein expression for *E. coli*. This strongly supports the hypothesis that the *E. coli* gene network has evolved to minimize noise strength for each protein expression level, reaching a Pareto-optimal solution.

7.5 Summary

The proposed multi-objective optimization design framework (MOOD) was able to provide effective guidelines to tune biological parameters so as to achieve a desired circuit behavior. The MOOD was applied to gene circuits of different nature and with different kinds of models such as the I1-FFL or the QS/Fb circuit. In both circuits, we obtained guidelines that successfully allow us to implement the circuits in the wet-lab with confidence. Moreover, it is easy to analyze the impact of the context on the synthetic device to be designed. Therefore, we analyzed how the presence of a downstream load influences the performance of the designed circuit, and assessed . The results of the multi-objective optimization approach suggest that –although system dynamics actually put constraints on the possible values of the kinetic parameters– robust design guidelines can be obtained to build a biological systems with a desired functionality.

Chapter 8

Conclusions

> - *Good night, and good luck.*
>
> Edward R. Murrow

The contributions of this work were listed in Chapter 1, and particular conclusions can be found at the end of each main chapter. Here some general conclusions are drawn and discussed together. Besides, some proposed lines for future work are discussed. The main contributions of the thesis were contained in Chapters 4 to 7. Thus:

- **Chapter 4** addressed the modeling of synthetic gene circuits using first-principles to obtain an appropriate model and a computationally efficient simulation. We have seen that for both deterministic and stochastic models both aspects are intertwined. The presented model reduction approach relieved the fact that gene circuits models often contain many parameters and species to describe the system dynamics. The resulting reduced-order nonlinear models have incomplete parameter identifiability. Nevertheless, they helped the system identification process to be more tractable and obtain reliable values for the parameters.

- In **Chapter 5** a methodology based on multi-objective optimization design for model parameter estimation of synthetic gene circuits has been proposed. This methodology ensures all values obtained for the model parameters are the optimal ones that fit the system model with the experimental data collected. If it is necessary, the solutions obtained can be post-processed with a multivariate analysis statistical tool to get more details about the role played by the different parameters. The multi-objective optimization methodology successfully handled the experimental data integration coming from the most common type of measurements for gene circuits, that is, cells population bulk data and single-cell snapshot data. The thesis adapted a practical approach. Model parameters were estimated using the ensemble modeling framework implicit in the multi-objective formulation. The parameter intervals instead of crisp values were used in most cases as proper representation of the estimated values.

- In **Chapter 6** a stochastic feedback control synthetic gene circuit for noise regulation of protein of interest under extrinsic and intrinsic fluctuations was designed. The stochastic differential model was formulated using the Chemical Langevin Equation, generating more accurately random paths at a reasonable computational cost. The controller within each cell benefits from the interplay between a feedback loop and cell-to-cell communication, having direct impact at the cells population level. *In silico* results showed that few easy to tune in the laboratory gene circuit parameters can be used to achieve noise strength reductions up to a 60% with respect to unregulated expression of the protein of interest. Extrinsic noise disturbing the controller was modeled as parametric variability. This differs from the approach found in the literature, i.e. an additive stochastic signal, analogous to the intrinsic noise term. This consideration allowed us to foresee an important reduction of noise strength when quorum sensing was added to feedback. Even knowing that the amount of reduction depended on the circuit parameters, noise reduction was observed for almost any combination of them. Finally, we concluded there is a trade-off between protein expression and its noise level as revealed both the system-wide experimental data and the theoretical analysis of the stochastic feedback control synthetic gene circuit.

- Finally in **Chapter 7** a multi-objective optimization framework to provide effective guidelines for tuning the gene circuits parameters when their desired function has trade-off has been proposed. The guidelines drawn from the obtained Pareto front obtained are all optimal ones, and the final solution is chosen depending on the designer's criteria. It was seen that parameters con be split in two broad groups. Those that must attain a common range of values for all solutions at the Pareto front, and those parameters that must take different values, thus acting as practical tuning knobs. The relevance of the thesis approach lies in the fact that both groups are not set *a-priori* but found by the optimization process. Although synthetic gene circuits dynamics actually put constraints on the possible values of their kinetic parameters, the multi-objective optimization framework provides design guidelines that lead to the confidently construction of those circuits in the laboratory.

Future work

The study carried out in this Thesis opens new research avenues. Additional work would be required to extend dynamic models of synthetic circuits with consideration of metabolism, use of cellular resources and environmental conditions. Detailed first-principles dynamic models of the synthetic circuits can be combined with the constraint-based metabolic models of the organisms used as hosts, and models of use of shared resources. In the latter case, mechanistic approaches as in (Weiße et al., 2015) take into account cell growth and substrates uptake. They could be used along with approaches like (Qian et al., 2017) that capture the dynamic loading effects on the circuit. This will improve the characterization process of a circuit-host interaction, elucidating how its insertion affects the growth of the host, and what are the capacities of the synthetic circuit in the metabolic context of the receiving microorganism.

The stochastic analysis of the feedback control synthetic genetic circuits may require a supplementary stability analysis. Stability analysis of complex genetic circuits via Lyapunov methods is in general not feasible. An alternative is contractivity analysis, that can in addition be applied to analyze cell population consensus. One can take advantage of the quasi-polynomial, degree two, structure of the models and their positivity. In addition, an interpretation in terms of graphs allows in a manageable way to establish a connection with the qualitative vision of the system. The contractive analysis can be applied reducing the problem to determining the consistency of a system of inequalities. To do this, the Fourier-Motzkin method can be applied, focusing on finding sufficiency conditions that are easy to apply but not at the price of being excessively conservative, drastically reducing the set of parameters that ensure compliance with certain specifications. The set of conditions obtained depend on the values of the system parameters. Thus, one could outline a method for stability analysis of gene circuits providing a region of their parameters where sufficient conditions for contractivity are fulfilled. Interestingly, the stability analysis via contractivity could be cast as additional inequality constraints within the thesis framework of multi-objective optimization for determining parameters achieving a desired performance.

The groundwork established in Chapters 5 and 7 agrees to go beyond, combining the results there, and developing a multi-objective optimization-based closed-loop tuning of synthetic gene circuits. This standard model-based circuit tuning as done in this thesis assumes that first an *in silico* parameter tuning is carried out. Then, in a second stage, the resulting gene circuit is constructed *in vivo.* if the model is good enough, only minor additional fine tuning may be required. This, in many practical cases is too optimistic. Closed-loop tuning iterates between the *in silico* optimization and the *in vivo* construction. So, that at each iteration the *in vivo* optimal Pareto front becomes closer to the *in silico* computational one. In essence the idea is to perform a multi-objective standard optimization with evaluation on the model at each iteration. From the theoretical Pareto front obtained a subpopulation can be selected that can be evaluated experimentally taking into account the material and temporal restrictions involved. The result of the evaluation will be used, on the one hand for the improvement of the model (subset of nominal parameters not used in the optimization) and, on the other, as initial seed population for the next iteration. The objective is to obtain effective closed-loop parametric adjustment algorithms of synthetic biological circuits in order to achieve desired output specifications. These algorithms should provide parametric tuning guidelines in the sense of qualitative rules about parameter values that can be effectively modified in the laboratory; continuing, completing and extending the work performed in this Thesis. To this end, recalling that solutions along the Pareto front define circuits with different behaviors, the appropriate way to translate this information on practical tuning guidelines could be improved by using not only clustering techniques as done in the thesis, but also local sensitivity analysis along the points of the Pareto front. This is an inverse problem, in which you have to move from the value space of the cost functions to the parameter space. The inverse problem can be solved using several tools, like SIVIA interval inversion algorithms. Finally, the procedure will be integrated into the tuning methodology through mixed closed-loop optimization.

Appendices

Appendix A

Foundations

A.1 BioBrick assembly

The most used library of biological parts in synthetic biology is the Registry of Standard Biological Parts (Biobrick Foundation, 2006). It has physical samples of specified parts in its database. This parts are available for request or they can been sent through a DNA kit. All parts in here are compatible with the BioBrick RFC[10] assembly method (Knight, 2007). An assembly method defines how part samples will be assembled together by the engineer. An assembly standard ensures compatibility between parts, allowing part samples to be assembled together creating new longer and more complex parts (e.g. a basic gene circuit), while still maintaining the structural elements of the assembly standard.

Figure A.1 illustrates the traditional assembly done through cutting and ligating (use of restriction sites) biological parts. The following description uses BioBrick assembly method to put 2 parts of the genetic circuit together:

1. Parts and backbone in circular plasmids must be compatible with the BioBrick method (see Fig.A.1a)
2. Restriction digests (cutting). The upstream part sample is cut out with EcoRI and SpeI enzymes. The downstream part is cut out with XbaI and PstI enzymes. The plasmid backbone is a linear piece of DNA previously cut with EcoRI and PstI enzymes. Both parts and the plasmid backbone have 3 different antibiotics (A, B and C respectively) to eliminate unwanted background cells (see Fig.A.1b).
3. Ligation. It is a reaction with an equimolar quantity of all 3 restriction digest products that produces a new composite part in a new plasmid (see Fig.A.1c).

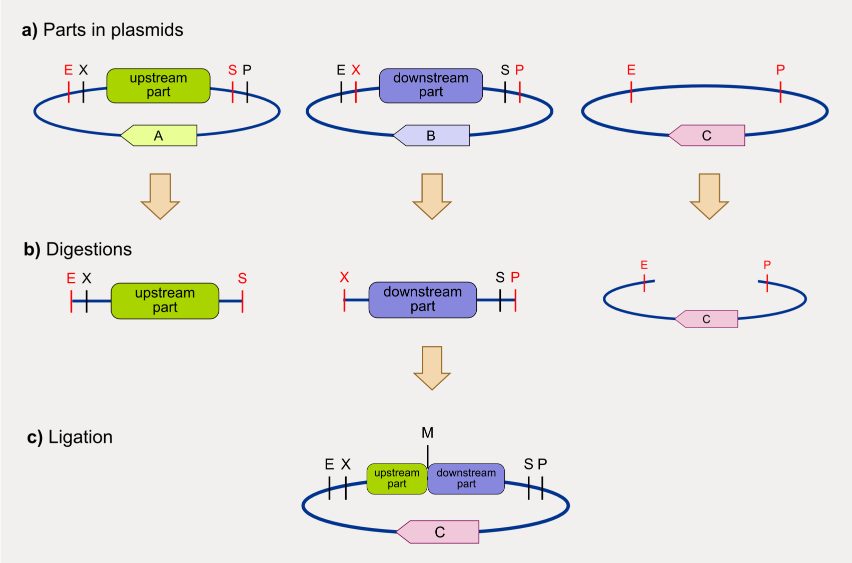


Figure A.1. 3A assembly and BioBrick assembly methods. a) Two parts contained in the plasmids and the linearized plasmid backbone with their 3 different antibiotics. **b)** Three digestions made with four restriction enzymes as the BioBrick assembly method. Upstream part cut out with EcoRI/SpeI enzymes. Downstream part cut with XbaI/PstI enzymes. Backbone digestion is made with EcoRI/PstI enzymes. **c)** A new composite part is the product of the mix and ligation of the three digestions.

4. The desired result is the upstream part sample's SpeI overhang ligated with the downstream part sample's XbaI overhang resulting in a scar that cannot be cut with any of our enzymes.

5. When the ligation is inserted into cells and grown with antibiotic C, only cells with the correct construction survive. This are represented as circular plasmids like Fig.2.4a.

In this way, any newly composed part will adhere to its assembly standard without need for manipulation, and can be used in future assemblies without problems to obtain the functional designed gene circuit. But synthetic gene circuits can be designed by simple or more elaborate control theoretic principles making their behaviour within a single cell or across a cell population more reliable, predictable and robust to perturbations.

A.2 Gibson assembly

Few years back, Daniel G. Gibson, of the J. Craig Venter Institute, described a robust exonuclease-based method to assemble DNA seamlessly and in the correct order, eponymously known as Gibson Assembly. The reaction is carried out under isothermal conditions using three enzymatic activities: a 5′ exonuclease generates long overhangs, a polymerase fills in the gaps of the annealed single strand regions, and a DNA ligase seals the nicks of the annealed and filled-in gaps. This method has been widely adopted and is a major workhorse of synthetic biology projects worldwide. Applying this methodology, the 16.3 kb mouse mitochondrial genome was assembled from 600 overlapping 60-mers (Gibson et al., 2009).

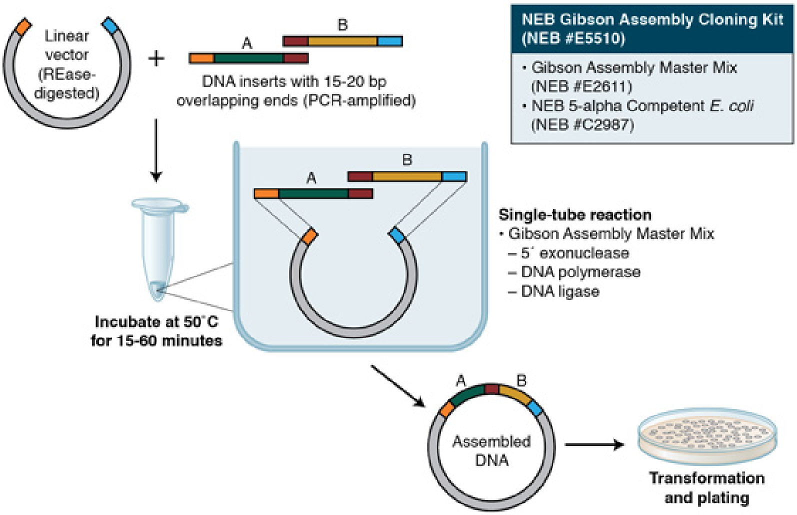

Linear vector (REase-digested)
+
A
B
DNA inserts with 15-20 bp overlapping ends (PCR-amplified)
NEB Gibson Assembly Cloning Kit (NEB #E5510)
• Gibson Assembly Master Mix (NEB #E2611)
• NEB 5-alpha Competent *E. coli* (NEB #C2987)
A
B
Single-tube reaction
• Gibson Assembly Master Mix
– 5´ exonuclease
– DNA polymerase
– DNA ligase
Incubate at 50˚C for 15-60 minutes
A
B
Assembled DNA
Transformation and plating

Figure A.2. Gibson Assembly method employs three enzymatic activities in a single-tube reaction: 5′ exonuclease, the 3′ extension activity of a DNA polymerase and DNA ligase activity. The 5′ exonuclease activity chews back the 5′ end sequences and exposes the complementary sequence for annealing. The polymerase activity then fills in the gaps on the annealed regions. A DNA ligase then seals the nick and covalently links the DNA fragments together. The overlapping sequence of adjoining fragments is much longer than those used in Golden Gate Assembly, and therefore results in a higher percentage of correct assemblies. Adapted from New England Biolabs®

A.3 Transformation into cells

Transformation is the process by which foreign DNA is introduced into a cell. Transformation of bacteria with plasmids is important not only for studies in bacteria but also because bacteria are used as the means for both storing and replicating plasmids. Because of this, nearly all plasmids (even those designed for mammalian cell expression) carry both a bacterial origin of replication and an antibiotic resistance gene for use as a selectable marker in bacteria.

Scientists have made many genetic modifications to create bacterial strains that can be more easily transformed and that will help to maintain the plasmid without rearrangement of the plasmid DNA. Additionally, specific treatments have been discovered that increase the transformation efficiency and make bacteria more susceptible to either chemical or electrical based transformation, generating what are commonly referred to as 'competent cells'.

Many companies sell competent cells, which come frozen and are prepared for optimal transformation efficiencies upon thawing. However, there are several protocols to make different kinds of competent cells in the lab, without the need for buying them. The different kinds of competence are electrocompetent, chemical competent, mix and go (Zymo Research) and RT electrocompetent. They differ in the efficiency of the transformation and the elements required. The majority of the transformations needed for this Thesis were performed with lab-made electrocompetent cells using the following adapted and improved protocol from the Tabor Lab, the Georgiev Lab, and the Knust Lab.

A.3.1 Electrocompetent E. coli preparation protocol

1. Streak an LB agar plate from a 80°C stock, grow overnight at 37°C.

2. Pick a well isolated colony with a sterile toothpick or pipet tip and inoculate 3mL LB + appropriate antibiotics in a 14mL culture tube.

3. Shake at 37°C, 250rpm until culture reaches $\mathrm{OD}_{600} \approx 1$ (6 to 10 hours, depending on the strain).

4. Dilute the culture 1:1000 or 1:2000 in 100mL to 1L (depending the desired number of aliquots) of fresh LB + antibiotics.

5. Shake at 37°C, 250-300rpm until OD_{600} = 0.4-0.8 (≈12 hours).
   (Optional) Shake at 19°C, 250-300rpm until OD_{600} = 0.6-0.8
   Note: If the culture overgrows, backdilute 1:10 and allow to reach the appropriate OD_{600}.

6. When the culture is at $\mathrm{OD}_{600} = $ 0.6-0.8 chill the cells in an ice/water slurry for 15 minutes.
   It is critical that the cells stay on ice or at 4 ° C for the entire remainder of the procedure.
7. Pellet the cells by spinning them at 3250g, 4°C for 8 minutes.
8. Discard the supernatant and resuspend pellet in 50 ml ice cold sterile MilliQ water. Do not vortex (it can damage the cells).
9. Pellet the cells by spinning them at 3250g, 4°C for 8 minutes.
10. Discard the supernatant and wash the cells in 10% glycerol for 1 hour.
11. Pellet the cells by spinning them at 3250g, 4°C for 8 minutes.
12. Carefully remove supernatant, and resuspend pellet in 1/100 volumes 10% glycerol.
13. Cells can now be aliquoted into 50 μL aliquots in 1.7mL microcentrifuge tubes, flash frozen in liquid nitrogen and stored at -80°C for ≈1 year.
14. Alternatively cells can be electroporated immediately. This results in ≈10x higher transformation efficiency.

A.3.2 Protocol for electroporation of *E. coli*

1. Place the electrocompetent aliquot, the electroporation cuvette and the DNA on ice.
2. Preheat at 37°C the necessary LB agar plates carrying appropriate antibiotics (2 plates per transformation).
3. Add 10 - 50ng of plasmid DNA (1 - 5 μL) to the 50 μL of electrocompetent *E.coli* (in its own tube) with and mix by flicking with your finger.
4. Transfer the mix into a cold electroporation cuvette.
5. Sit on ice for 5 minutes.
6. Clean the cuvette to remove moist and place cuvette in electroporator chamber, then pulse at 2.5kV.
7. Immediately add 750 μL ice cold SOB to cuvette, pipette up and down gently to mix and transfer the entire 800 μL to a labeled 14mL culture tube.
8. Recover by shaking at 37°C, 250rpm for 1 hour to permit expression of antibiotic resistance gene.

9. Plate on LB agar + appropriate antibiotics. If transforming circular plasmid, plate $10\,\mu\mathrm{L}$. and $100\,\mu\mathrm{L}$ of culture (or even less). If transforming a DNA assembly reaction plate $50\,\mu\mathrm{L}$ on one plate and the remaining $750\,\mu\mathrm{L}$ on a second plate.

A.4 Linear Noise Approximation (LNA)

The LNA tries to deal with noise in a deterministic setting, where analytical solutions are locally valid close to macroscopic trajectories (deterministic reaction rates) plus an additive noise called *fluctuation* term. This section follows the arguments from (Ullah and Wolkenhauer, 2011) to derive the LNA from the Chemical Master Equation (CME). An alternative notation more suited for Taylor expansion using the step operator E_j for the *j*-th in the CME.

$$\begin{aligned} \mathrm{E}_j f(n) &= f(n + \mathbf{S}_j) \\ \frac{\partial}{\partial_t} P(\mathbf{n}, t) &= \sum_{j=1}^{J} \left(\mathrm{E}_j^{-1} - 1\right) a_j(\mathbf{n})\, P(\mathbf{n}, t) \end{aligned} \tag{A.1}$$

The LNA can anticipate the way in which the solution of the CME $P(\mathbf{n}, t)$ will depend on the system size Ω. Assuming that the continuous approximation $\mathbf{n}(t)$ of the system fluctuates around a macroscopic trajectory (deterministic reaction rate) of order Ω with a fluctuation of order $\sqrt{\Omega}$

$$\begin{aligned} \mathbf{n}(t) &= \Omega\phi(t) + \sqrt{\Omega}\boldsymbol{\Xi}(t) \\ n_i &= \Omega\phi_i + \sqrt{\Omega}\xi_i \end{aligned} \tag{A.2}$$

where n_i is the molecules number of species i, ϕ is the macroscopic concentration defined in section **??** $(n_i(t) = \Omega x_i(t))$, and ξ is a random variable from the random matrix $\Xi(t)$, which models the fluctuations around $\phi(t)$. The probability distribution $P(\mathbf{n}, t)$ transforms into the probability distribution $\Pi(\xi, t)$ of $\Xi(t)$

$$P(\mathbf{n}, t) = P\left(\Omega\phi_i + \sqrt{\Omega}\xi_i\right) = \Pi(\xi, t) \tag{A.3}$$

The time derivative of (A.2) at constant number of molecules n implies that $\frac{d\xi}{dt} = -\Omega^{1/2}\dot{\phi}$. It can be used in the time derivative of (A.3)

$$\frac{\partial P(\mathbf{n}, t)}{\partial t} = \frac{\partial \Pi}{\partial t} - \Omega^{1/2} \sum_{i=1}^{I} \frac{d\phi_i}{dt} \frac{\partial \Pi}{\partial \xi_i} \tag{A.4}$$

Before comparing equation (A.4) with the CME (A.1), it is necessary to define the propensity function $a_j(\mathbf{n})$ in terms of the fluctuation ξ, the deterministic rate $\nu_j(x)$, and the operator E_j for each *j*-th reaction through

$$a_j(\mathbf{n}) = \Omega\left[\nu_j\left(\phi + \Omega^{-1/2}\xi\right) + O\left(\Omega^{-1}\right)\right]$$

.

Replacing $a_j(\mathbf{n})$ in the CME (2.31)

$$\frac{\partial}{\partial t}P(\mathbf{n},t) = \Omega\sum_{j=1}^{J}\left(\mathrm{E}_j^{-\Omega^{-1/2}} - 1\right)\left[\nu_j\left(\phi + \Omega^{-1/2}\xi\right) + O\left(\Omega^{-1}\right)\right]\Pi(\xi,t) \tag{A.5}$$

where $O(x)$ is the first neglected order with respect to x in an expansion. Taylor expansion of the reaction rates ν_j and the operator $\mathrm{E}_j^{-\Omega^{-1/2}}$ around ϕ in several dimensions

$$\begin{aligned}\nu_j\left(\phi + \Omega^{-1/2}\xi\right) =&\nu_j(\phi) + \Omega^{-1/2}\sum_i \frac{\partial \nu_j}{\partial \phi_i}\xi_i + O\left(\Omega^{-1}\right)\ ,\\ \mathrm{E}_j^{-\Omega^{-1/2}} =&1 - \Omega^{-1/2}\sum_i \mathbf{S}_{ij}\frac{\partial}{\partial \xi_i} + \frac{1}{2}\Omega^{-1}\sum_{i,k}\mathbf{S}_{ij}\mathbf{S}_{kj}\frac{\partial^2}{\partial \xi_i \partial \xi_k} + O\left(\Omega^{-3/2}\right)\end{aligned} \tag{A.6}$$

Inserting the above two expressions into (A.5), and comparing the results with (A.4) leads to the linear *Fokker-Planck equation* (Ullah and Wolkenhauer, 2011)

$$\begin{aligned}\frac{\partial \Pi}{\partial t} &= \sum_{j=1}^{J}\left(-\sum_{i,k}\mathbf{S}_{ij}\frac{\partial \nu_j}{\partial \phi_k}\frac{\partial(\xi_k \Pi)}{\partial \xi_i} + \frac{1}{2}\nu_j(\phi)\sum_{i,k}\mathbf{S}_{ij}\mathbf{S}_{kj}\frac{\partial^2 \Pi}{\partial \xi_i \partial \xi_k}\right)\\ &= -\sum_{i,k}\mathbf{A}_{ik}\frac{\partial(\xi_k \Pi)}{\partial \xi_i} + \frac{1}{2}\sum_{i,k}\mathbf{B}\mathbf{B}^T_{ik}\frac{\partial^2 \Pi}{\partial \xi_i \partial \xi_k}\end{aligned} \tag{A.7}$$

where $\mathbf{A} = \sum_{j=1}^{J}\mathbf{S}_{ij}\frac{\partial \nu_j}{\partial \phi_k}$ is the Jacobian matrix, and $\mathbf{B} = \sum_{j=1}^{J}\nu_j\mathbf{S}_{ij}\mathbf{S}_{kj}$ is the diffusion matrix. Both of these matrices depend on time thorough the deterministic rate concentration $\phi(t)$. The terms of order $\Omega^{-1/2}$ are proportional to $\frac{\partial \Pi}{\partial \xi_i}$, and ϕ was choosing as $\frac{d\phi_i}{dt} = \sum_{j=1}^{J}\mathbf{S}_{ij}\nu_j(\phi)$.

The stationary solution of (A.7) is a multidimensional Normal distribution $P(\xi) = \left((2\pi)^{I/2}\sqrt{\det \boldsymbol{\Xi}}\right)^{-1}\exp\left(-\xi^T\boldsymbol{\Xi}\xi/2\right)$, which has a covariance matrix $\boldsymbol{\Xi} = \langle\xi\xi^T\rangle$, and follows a Lyapunov equation $\mathbf{A}\mathbf{\Xi} + \mathbf{\Xi}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = 0$. Recall I is the total number of species in the system. The correlation matrix of the stationary process is $\langle \xi(t)\xi^T(s) \rangle = \mathbf{\Xi} \exp(\mathbf{A} \mid t - s \mid)$. Thereby, we can determine the symmetric **covariance matrix** as $\mathbf{C} = \Omega \mathbf{\Xi}$. The LNA solutions together with the matrix $\mathbf{C}$ often give very accurate descriptions of the size of molecule number fluctuations and how they are correlated.

Appendix B

Two case studies: incoherent feedforward and quorum sensing/feedback gene circuits.

B.1 I1-FFL Strains and plasmids

Figures B.1 and B.2 illustrate the two low copy plasmids that were built and co-transformed into the same cell to implement the I1-FFL gene circuit.

B.2 QS/Fb Strains and plasmids

Figures B.3 and B.4 depict the two low copy plasmids that were built and co-transformed into the same cell in order to implement the QS/Fb gene circuit.

B.3 NoQS/NoFb Strains and plasmids

The plasmid in figure B.5 together with the previous one pCB2tc (Fig.B.3) were also co-transformed into the same cell to obtain the NoQS/NoFb circuit. This synthetic gene circuit requires the inducer $\mathrm{AHL}_{\mathrm{ext}}$ as external input.

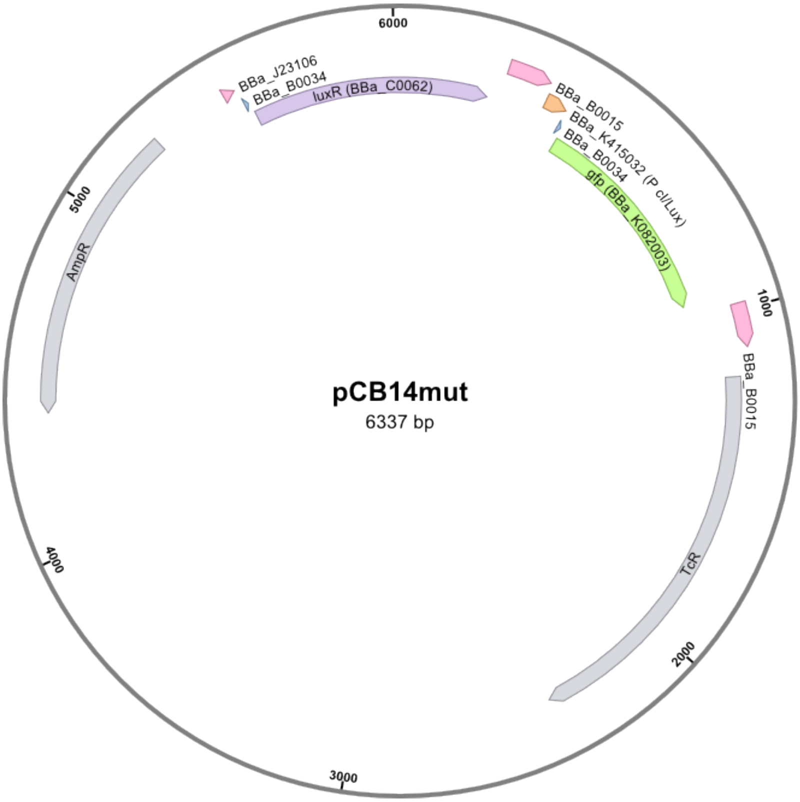


Figure B.1. Plasmid pCB14mut carrying both genes the *luxR* and the *gfp+LVA*. LVA tag sequence speeds up the GFP degradation. Both genes have their respective transcriptional units assembled in the same direction.

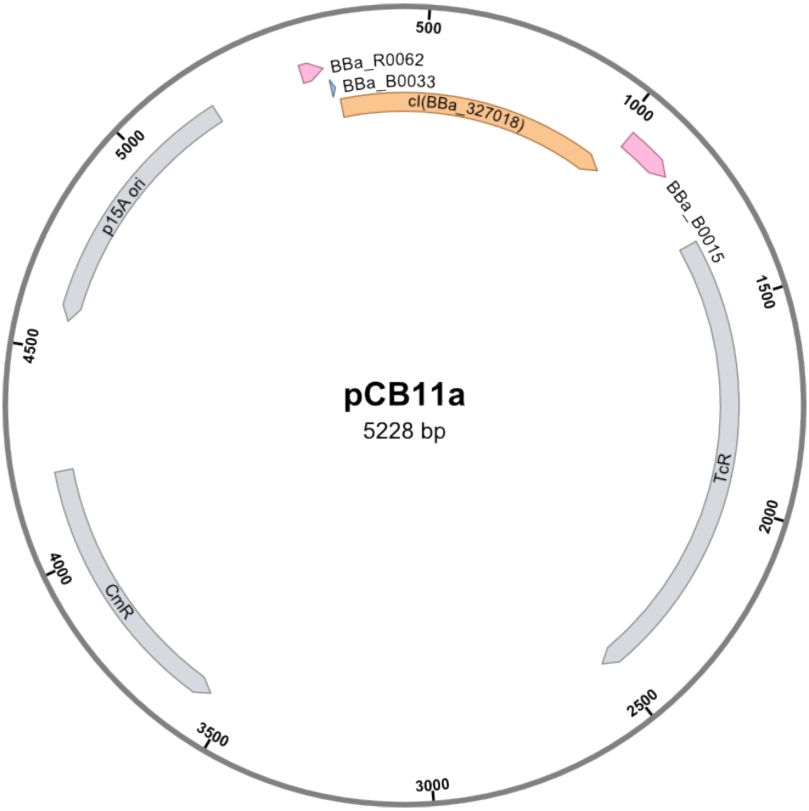


Figure B.2. Plasmid pCB11a carrying the transcriptional unit to express protein cI+LVA tag.

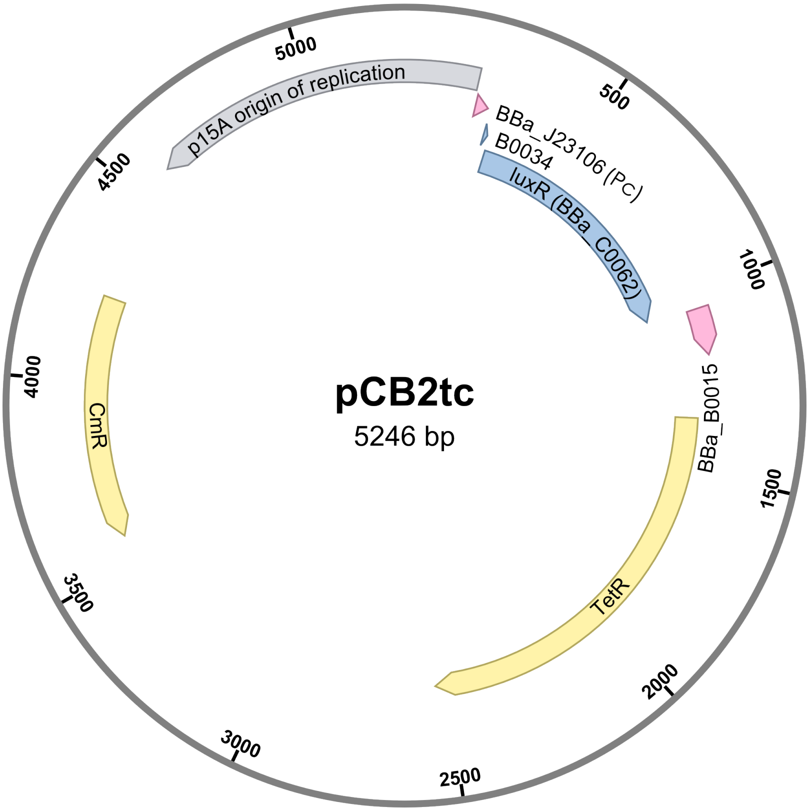


Figure B.3. Plasmid pCB2tc carrying the transcriptional unit to express protein LuxR.

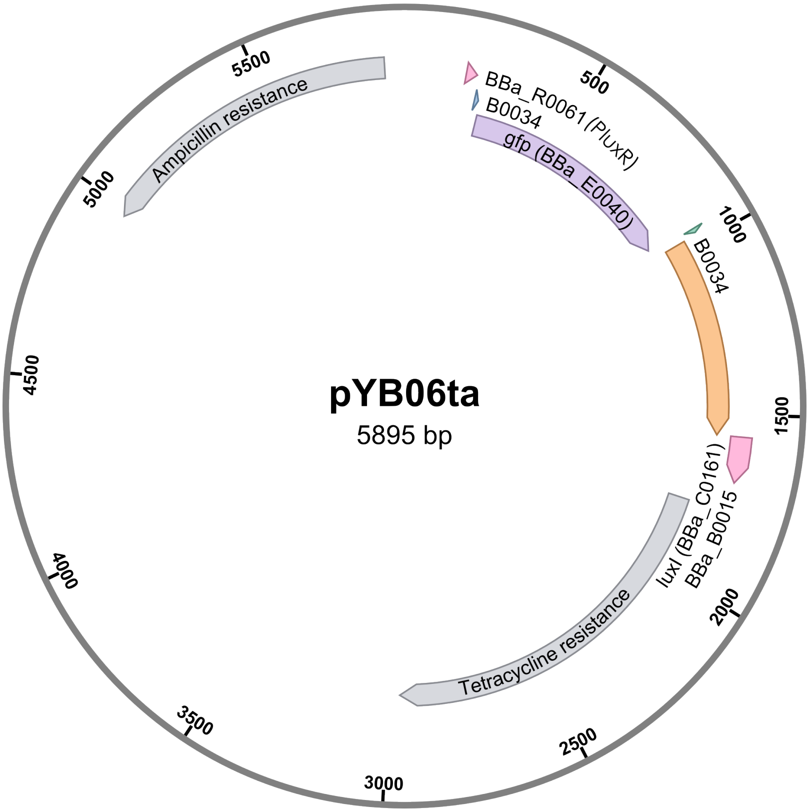


Figure B.4. Plasmid pYB06ta has two coding sequences to co-expresses the protein of interest (PoI) and the LuxI protein.

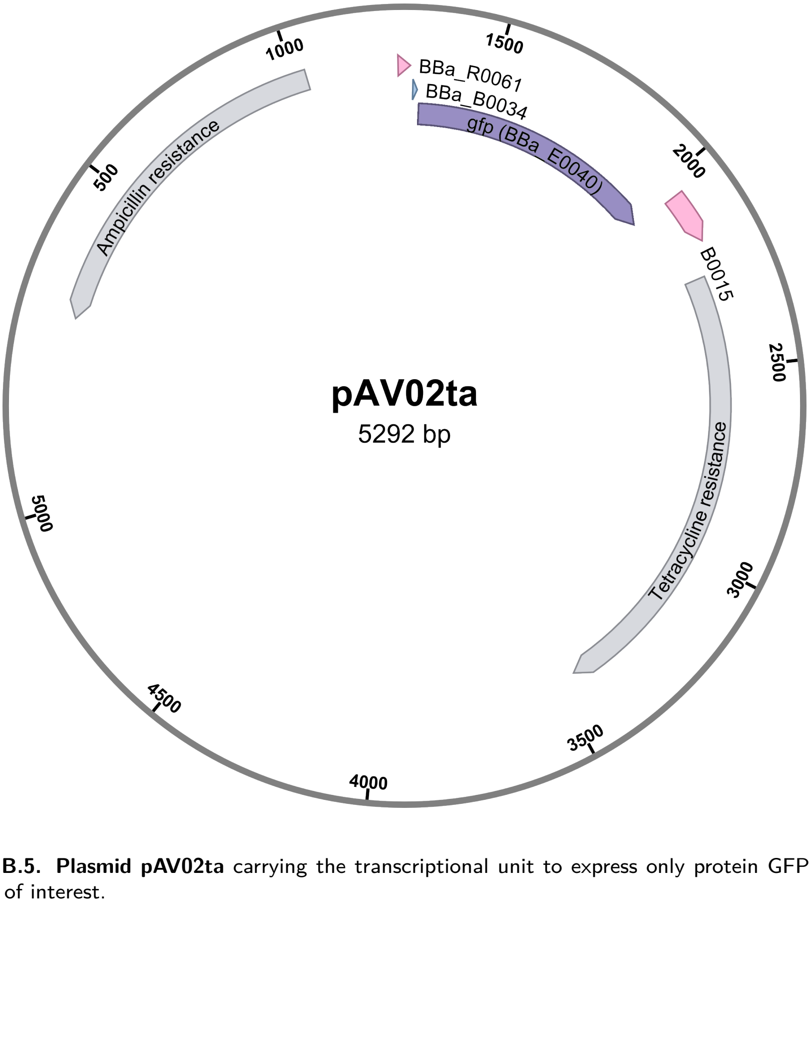


Figure B.5. Plasmid pAV02ta carrying the transcriptional unit to express only protein GFP as the protein of interest.

Appendix C

Model parameter estimation

The protocols presented in this section have been adapted from (Olson et al., 2014).

C.1 Time-course experimental conditions

This protocol was used to collect time-course data for the I1-FFL , QS/Fb and NoQS/NoFb synthetic gene circuits.

1. Start a 37 °C, shaking overnight culture from a -80 °C stock in a tube containing 3 mL LB medium and the appropriate antibiotics (100 μg/mL ampicillin, 12.5 μg/mL tetracycline and 34 μg/mL chloramphenicol for either I1-FFL , QS/Fb or NoQS/NoFb circuit in 14 mL culture tubes).

2. After the overnight culture has grown for 12-16 h, prepare M9 medium (200 mL is made with: 151.58 mL autoclaved, distilled H2O, 40 mL 5x M9 salts, 4 mL 10 % amino acids, 4 mL 20 % glucose, 400 μL 1 M MgSO4, 20 μL CaCl2). Add appropriate antibiotics to medium and stir the container to ensure the antibiotics are mixed well in the medium.

3. Measure the OD$_{600}$ of the overnight culture.

4. Dilute the overnight culture into the M9 + antibiotics, bringing the OD$_{600}$ to 0.004. Shake the container to ensure the cells are mixed well in the medium.

5. Distribute 3 mL of inoculated medium into each 8 BD Falcon round-bottom 14 mL polypropylene test tubes (BD Biosciences catalog ♯352006).

6. Incubate tubes at 37 °C with shaking at 250 rpm for 3 h.

7. Dilute 5 mg of AHL (N-3-Oxohexanoyl-L-homoserine lactone, Santa Cruz Biotecnology Catalog Number SC205396) into 468.98 μL of DMSO to reach a solution 50 mM. This stock was stored at -20 °C until use.

8. Successively dilute AHL 50 mM into M9 reaching different AHL concentrations to induce the culture samples. Take into account the final desired AHL concentration and the total test volume ($\approx 200\,ul$ each well).

9. After 2 h of growth, quickly induce the test tubes at the defined AHL concentrations.

10. Incubate tubes at 37 °C with shaking at 250 rpm for 4 h.

11. After 4 h of induction and growth, harvest all test tubes by immediately transferring them into an ice-water bath. Wait 10 min for the cultures to equilibrate to the cold temperature and for gene expression to stop.

12. Transfer the samples in a 96-well plate ($\approx 200\,ul$ each well), and measure absorbance and fluorescence (Biotek Cytation 3 plate reader) for 1 h before induction.

13. Induced as quickly aas possible each sample at the desired AHL concentration from step 8.

14. Continue measuring absorbance and fluorescence for around ≈ 6 h.

C.2 QS/Fb and NoQS/NoFb flow cytometer conditions

This protocol was used to measure snapshot data from individual cells via flow cytometry in both the QS/Fb and the NoQS/NoFb gene circuits.

1. Start a 37 °C, shaking overnight culture from a -80 °C stock in a tube containing 3 mL LB medium and the appropriate antibiotics (100 μg/mL ampicillin, 12.5 μg/mL tetracycline and 34 μg/mL chloramphenicol for both Qs/Fb and NQs/NFb systems in 14 mL culture tubes).

2. After the overnight culture has grown for 12-16 h, prepare M9 medium (200 mL is made with: 151.58 mL autoclaved, distilled H2O, 40 mL 5x M9 salts, 4 mL 10 % amino acids, 4 mL 20 % glucose, 400 μL 1 M MgSO4, 20 μL CaCl2). Add appropriate antibiotics to medium and stir the container to ensure the antibiotics are mixed well in the medium.

3. Measure the OD$_{600}$ of the overnight culture.

4. Dilute the overnight culture into the M9 + antibiotics, bringing the OD$_{600}$ to 0.004. Shake the container to ensure the cells are mixed well in the medium.

5. Distribute 3 mL of inoculated medium into each 8 BD Falcon round-bottom 14 mL polypropylene test tubes (BD Biosciences catalog ♯352006).

6. Incubate tubes at 37 °C with shaking at 250 rpm for 3 h.

7. Dilute 5 mg of AHL (N-3-Oxohexanoyl-L-homoserine lactone, Santa Cruz Biotecnology Catalog Number SC205396) into 468.98 μL of DMSO to reach a solution 50 mM. This stock was stored at -20 °C until use.

8. Successively dilute AHL 50 mM into M9 reaching different AHL concentrations to induce the culture tubes. Take into account the final desired AHL concentration and the total test tube volume. We induce AHL 10 and 50 nM to measure final repression levels Egland and Greenberg (2000).

9. After 2 h of growth, quickly induce the test tubes at AHL 0, 10 and 50 nM.

10. Incubate tubes at 37 °C with shaking at 250 rpm for 4 h.

11. After 4 h of induction and growth, harvest all test tubes by immediately transferring them into an ice-water bath. Wait 10 min for the cultures to equilibrate to the cold temperature and for gene expression to stop.

12. Prepare a solution of phosphate-buffered saline (PBS: 137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 2 mM KH2PO4, pH to 7.4) + 500 μg/mL of the transcription inhibitor rifampicin (Rif, Tokyo Chemical Industry, cat. ♯R0079). Prepare at least 1 mL for each culture tube to be measured via flow cytometry. Rif dissolves slowly, so allow 45 - 60 min of stirring. Also at this time, begin preparing a 37 °C water bath.

13. Filter the dissolved solution of PBS + Rif through a 0.22-μm 20-mL syringe filter.

14. Transfer 1 mL of the filtered PBS + Rif into one 5 mL cytometer tube per culture sample, and chill tubes in an ice-water bath.

15. Transfer 50 μL of each chilled culture from step 7 into the chilled PBS + Rif solution.

16. Incubate the PBS + Rif + culture tubes in a 37 °C water bath for 1 h.

17. Transfer the tubes back into ice-water bath.

18. Wait 15 min, and then begin measuring each tube on a flow cytometer.

C.2.1 Measuring different inductions at defined time instants via flow cytometry

1. Repeat steps 1 to 8.
2. Induce AHL 10nM (4 μL of AHL 7.5 μM in 3mL of the sample culture) every 30 min during 3 h, and then every 10 min during the last 1 h (See Note 1 at the end of this protocol). After each induction, incubate all tubes at 37 °C with shaking at 250 rpm.
3. Repeat steps 11 to 18.

Flow cytometry data acquisition. Cytometry acquisition and analysis was performed using a BD FACSCalibur (Serie Nr. E14600085) flow cytometer with the laser system blue (488 nm) and red (635 nm). The FL1 (GFPmut3b) acquisition channel emission filter has a 510/21-nm filter. Acquisition is performed with typical count rates of 1,000-2,000 events/sec. Approximately 40,000 events are stored for each sample.

Appendix D

Stochastic analysis of a feedback control synthetic gene circuit

D.1 Experimental data post-processing

Data from flow cytometry were processed with our scripts (see SI Section D.2). First, cytometry data were read using the ***fca_readfcs*** Matlab function. Then, the first 250 and last 100 events were removed from the data set to avoid transient errors introduced owing to uneven pressurization of the sample tube. After this, the highest and lowest measured histogram channel for each of the measured values (FSC, SSC, and FL1) were removed, as the events in these channels have an undetermined fluorescence value. All this was done using the ***trim*** Matlab function. Next, the 2D binning of FSC and SSC was performed together with a smooth representation of the 2D histogram using the function ***smoothing***, shown in Fig. D.1A. The fluorescence histogram from FL1 raw data, corresponding to the all this events is plotted in Fig. D.1B. From this, the normalized and smoothed representation of the histogram was used to obtain contour level curves. Then, it is possible to use them as gate to select the events that are enclosed by the desired contour level using the function ***contour_gating***. This contour level curve was used to isolate a uniformly sized population of cells, and it is naturally aligned with the observed cell population. The gating procedure leaves $N = 15000 - 20000$ events shin in Fig. D.1C. This events were then scaled back to linear (detectors were set to log scale) using the parameters from the header of the FCS file. Next, a trim was performed on FL1 to remove a small number of apparent noncellular events with very low and very high fluorescence, the fluorescence corresponding to the gated events is shown in Fig. D.1D.

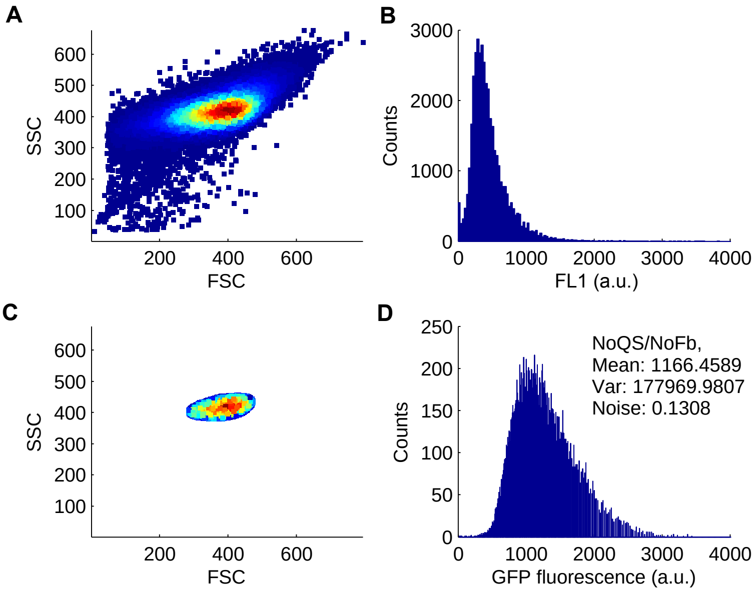


Figure D.1. Flow cytometry experimental data postprocessing. (A) Forward scatter vs. side scatter plot of the raw data. Colormap show numbers of events (from low in blue to high in dark red). (B) Fluorescence histogram from FL1 raw data, corresponding to the events plotted in panel A. (C) Forward scatter vs. side scatter plot showing the contour gated events. (D) Fluorescence histogram corresponding to the gated events plotted in panel C.

Finally experimental data were multiplied by a scale factor of 2.72 to obtain the histograms shown in section 6.3.1 of the main text of this Thesis.

D.2 Matlab and OpenFPM CODE

A short description of the main functions integrating the code used to simulate the model and process experimental data is given below. It has been divided in three groups: files related to parameters setting (Matlab), files to simulate the model (OpenFPM client in C++), and files used to process experimental data from flow cytometry (Matlab). All them can be downloaded from `http://sb2cl.ai2.upv.es/content/software`.

The stochastic simulation of our synthetic circuit is implemented using *langevin*, an OpenFPM client in C++. Information about OpenFPM installation can be found in its webpage `http://openfpm.mpi-cbg.de/`. The best option for a system that is natively supported (i.e. Linux based systems, Mac, etc.) is to run the code: *clone* `https:`

`//github.com/incardon/openfpm_pdata.git&&\cdopenfpm_pdata&&./install`, and follow the installation instructions therein.

C++ code - OpenFPM client

- ***main.cpp*** is the OpenFPM client *langevin* code. It implements the main body and two auxiliary functions. The first function opens the file param.dat created by the Matlab script and sets the parameters values for each cell. The second function is called at each simulation time step to update the system states (number of molecules of species) using the Euler-Maruyama algorithm.
- ***Makefile*** has the information for *make* to compile the C++ source code.
- ***langevin.mk*** has the information for *Makefile* to obtain all the paths and libraries. Should be replaced by the *example.mk* file generated by the OpenFPM instalation.

Computational cost Execution of 120 parameter sets takes around 20 minutes when performed in a Intel XEON Server with 8 cores and 32 Gb of RAM Memory.

Model code - Parameter setting

- ***Evaluate_CLE_Extrinsic.m*** is a script to set the parameters for the model and run the *langevin* OpenFPM client. It generates a matrix with all required parameters,runs *langevin* and saves the results obtained both as a variable in the Matlab workspace and as a Matlab .mat file.
- ***struct2csv_append.m*** is a function to convert a Matlab structure into a csv file that can be open with the *langevin* OpenFPM client.

Model code - Flow cytometry data postprocessing

- ***plot_tubes.m*** is the main script used to read, trim, smooth and gate the data. It plots the FSC vs SSC scatter and the FL1 histogram before and after the gating procedure. Then it calculates the mean and noise strength of the gated data.
- ***fca_readfcs.m*** is a function obtained from Matlab Central by Laszlo Balkay[1]. The function reads the raw data and returns the header of the file with information about the acquisition and the raw data (FSC, SSC, FL1).

[1] Download available at `www.mathworks.com/matlabcentral/fileexchange/9608-fcs-data-reader`

- ***trim.m*** is a function that trims the raw data. First, the first 250 and last 100 events are removed from the data set to avoid transient errors introduced owing to uneven pressurization of the sample tube. Then each channel is trimmed to the user defined limits. In general this limits are the highest and lowest measured histogram channel for each of the measured values (FSC, SSC, and FL1), as these events have an undetermined fluorescence value.

- ***smoothing.m*** is a function that binds the 2D (FSC,SSC) raw data and returns a smoothed version of the 2D histogram.

- ***contour_gating.m*** is a function that gates FSC and SSC data based on the contour obtained from *smoothing.m*. The user can select the contour level. Then all the events inside the contour are gated in.

Appendix E

Performance tuning via multi-objective optimization

E.1 Matlab CODE for optimization of the I1-FFL parameters

A short description of the main functions integrating this code, a description of the value sets, and supplementary results are given below. It has been divided in two groups: files related to the model computational characterization, and files used by the optimizer, which link to the first set.

Model code

- ***model_3genes.m*** is a function for the ODEs of the reduced model. Receives the value of the state vector x at time t, the parameters, initial conditions and time point; and returns a vector with the derivatives defined in it. When used with the command `ode23s` in the function *objective_func.m* one obtains the solution of the ODEs system for the given parameters.

- ***objective_func.m*** is the objective function. It receives the parameters and returns the **objectives values vector**, after calculating J_1 and J_2 for the corresponding dynamic response obtained with the given parameters.

  The 14 variables are initialized, and the 10 parameters are not because the optimizer will work with a given range in its code.

The *ode23s* algorithm gives the variables values Y for each t, using *model_3genes.m*. This *ode* algorithm was selected because our system model is what it is known as *stiff*, in terms of the numerical solution of ordinary differential equations, i.e. it has both slow and fast dynamics. An ordinary differential equation problem is stiff if the solution being sought is varying slowly, but there are nearby solutions that vary rapidly, so the numerical method must take small steps to obtain satisfactory results. For our integration problem we use an absolute tolerance of 1×10^{-8}, and a relative tolerance of 1×10^{-6}.

For computational simulation, we start from the equilibrium initial conditions (pre-computed) and give a jump of 50nM to the concentration of x_9 to simulate the induction. With respect to the simulation parameters, the simulation sampling time (δt) was fixed to 1×10^{-3} minutes, and a total simulation time $\mathrm{T_{sim}} = 300$ minutes was used.

- ***eval_obj_fun.m*** is the function that receive a population of parameters, evaluates the objective functions in this population, and accumulates the results in a matrix to return it. It is executed at each iteration of the spMODE algorithm.

MOO code First, highlight that we use the script ***Tutorial.m*** to run all the functions used to obtain the results shown in the main paper.

The first step is to run the *spMODEparam* file to build the variable 'spMODEDat' with the variables required for the optimization. Here the number of objectives are defined, also the number of decision variables and the *'Cost Function'*, which brings the objectives matrix after previous *ode* simulations (by means of interlinked functions mentioned above, constituting in essence the problem 'nucleus' or characterization). The field of search, and bounds to improve pertinency of solutions in the objective space so as to cut solutions with no interest to the DM, are defined here too. Also other aspects, such as maximum Pareto optimal solutions required and a bound on the number of function evaluations.

Once the Pareto set and the Pareto front are found by the optimizer, results can be plot with optional features through the *Leveltool*. This tool provides the LD visualization for the MCDM.

- ***spMODEparam.m*** generates the required parameters to run the spMODE optimization algorithm. In this file the variables regarding the multi-objective problem are defined. The values of interest for our problem are:

  1. Number of objectives.
     spMODEDat.NOBJ = 2

  2. Number of decision variables.
     spMODEDat.NVAR = 10

3. Cost Function.
   spMODEDat.mop = str2func('CostFunction')

4. Problem Instance.
   spMODEDat.CostProblem = 'modelo3genes'

5. Maximum and minimum values for the parameters or decision variables are fixed in order to give a range to the optimizer to search the optimal solutions, (spMODEDat.FieldD). $\mathrm{k_d}$ and $\mathrm{d_{Ie}}$ were fixed to avoid the optimizer to modify the model input I_e, as I want an step input determined by $K_e(t)$.

6. Bounds on objectives.
   spMODEDat.Pertinency=$[1 \times 10^{-3}\,200; 1 \times 10^{-4}\,20]$; A row for each objective, with the minimum and maximum values desired.

- ***CostFunction.m*** calls the cost function of your own multi-objective problem. In this case *eval_obj_fun.m*. It also includes a default mechanism to improve pertinency (Objective space bounded).

Clustering and Visualization

- ***clustering.m*** is a script that performs the hierarchical clustering with the solutions obtained from the spMODE optimization algorithm, and uses the modified LD-tool to plot the LD plots with the cluster number as Y-axis.

Computational cost Execution of our MOO using the spMODE algorithm (15.000 evaluations of the objective function) took around 10 hours and 25 minutes and was performed in a Intel XEON Server with 12 cores and 32 Gb of RAM Memory.

E.2 Supplementary results of the I1-FFL optimization

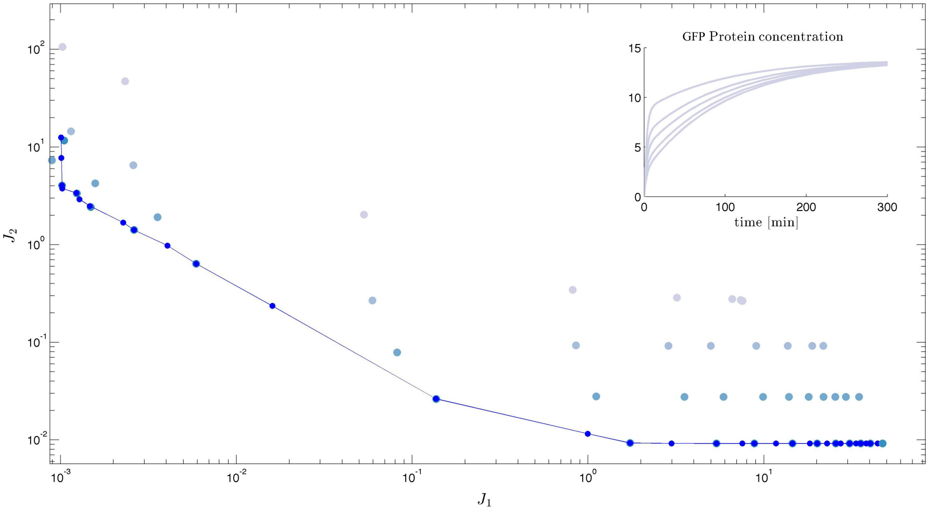


Figure E.1. Pareto Front of J_1 : Sensitivity vs. J_2 : Precision in blue line connected dots. Dots changing from blue to light blue are obtained by changing the degradation rate of protein GFP represented by $\mathrm{d_G} \in [0.3\,0.1\,0.03\,0.01]$ starting at the extreme solution. Notice, that decreasing $\mathrm{d_G}$ leads to a complete lose of optimality and moreover of the adaptation behavior, as we can see in the temporal profile of GFP protein concentration in the inset.

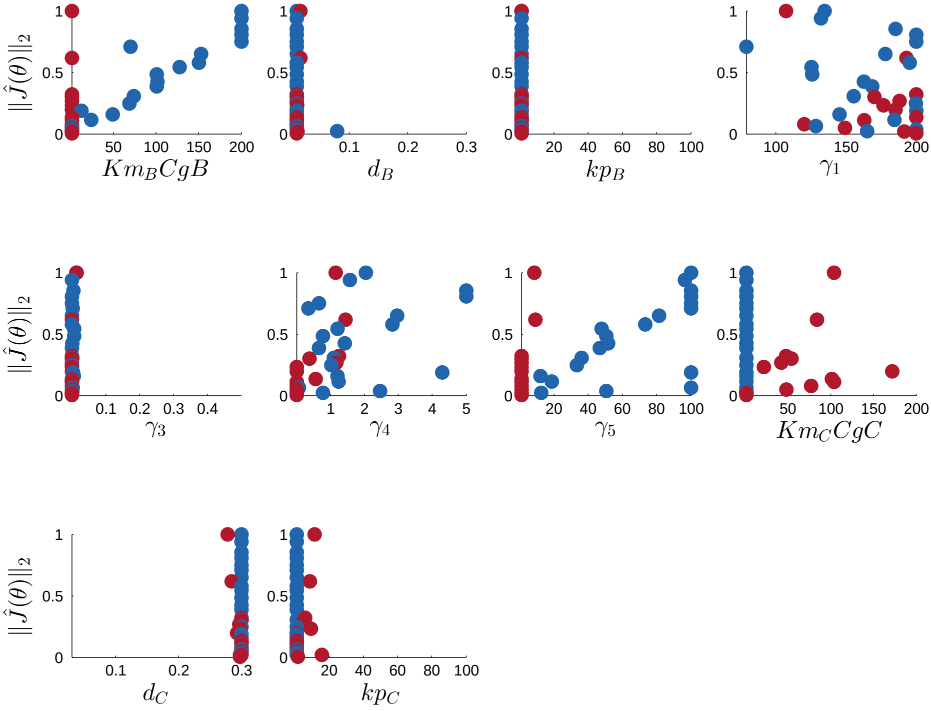


Figure E.2. Pareto set of decision variables. Each parameter of the model was plotted according its distance to ideal point. Red circles represent the cluster 1 and blue circles allow to the cluster 2.

Bibliography

M. Acar, J. T. Mettetal, and A. van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature genetics*, 40(4):471–475, 2008.

M. Adler and U. Alon. Fold-change detection in biological systems. *Current Opinion in Systems Biology*, 2017.

B. Alberts, D. Bray, K. Hopkin, A. D. Johnson, A. Johnson, K. Roberts, J. Lewis, M. Raff, and P. Walter. *Essential Cell Biology*. Garland Science, 3rd edition, 2009.

U. Alon. *An Introduction to Systems Biology. Desing Principles of Biological Circuits*. Champan and Hall/CRC, Edition, 2007.

C. Anderson. Anderson promoter collection [online]. `http://parts.igem.org/Promoters/Catalog/Anderson`, 2006. Accesed: 20/02/2015.

J. Anderson, Y.-C. . C. Chang, and A. Papachristodoulou. Model decomposition and reduction tools for large-scale networks in systems biology. *Automatica*, 47(6): 1165–1174, 6 2011.

Andersson, Joel and Åkesson, Johan and Diehl, Moritz. Casadi: A symbolic package for automatic differentiation and optimal control. In *Recent Advances in Algorithmic Differentiation*, volume 87. 2012. ISBN 978-3-642-30022-6.

J. Ang, S. Bagh, B. P. Ingalls, and D. R. McMillen. Considerations for using integral feedback control to construct a perfectly adapting synthetic gene network. *J Theor Biol*, 266(4):723–38, 2010a.

J. Ang, B. Ingalls, and D. McMillen. Probing the input-output behavior of biochemical and genetic systems: System identification methods from control theory. *Computer Methods*, (Part C):279–317, 2010b.

J. Ang, E. Harris, B. J. Hussey, R. Kil, and D. R. McMillen. Tuning response curves for synthetic biology. *ACS synthetic biology*, 2(10):547–567, 2013.

D. Angeli, P. De Leenheer, and E. D. Sontag. A petri net approach to the study of persistence in chemical reaction networks. *Mathematical biosciences*, 210(2): 598–618, 2007.

J. A. J. Arpino, E. J. Hancock, J. Anderson, M. Barahona, G.-B. V. B. Stan, A. Papachristodoulou, and K. Polizzi. Tuning the dials of synthetic biology. *Microbiology*, 159(Pt 7):1236–53, 7 2013.

M. Ashyraliyev, J. Jaeger, and J. G. Blom. Parameter estimation and determinability analysis applied to drosophila gap gene circuits. *BMC Systems Biology*, 2(1):83, 2008.

E. Balsa-Canto, A. A. Alonso, and J. R. Banga. An iterative identification procedure for dynamic modeling of biochemical networks. *BMC systems biology*, 4(1):11, 2010.

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

S. Basak and G. Chabakauri. Dynamic mean-variance asset allocation. *The Review of Financial Studies*, 23(8):2970–3016, 2010.

S. Basu, R. Mehreja, S. Thiberge, M.-T. Chen, and R. Weiss. Spatiotemporal control of gene expression with pulse-generating networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6355–6360, 2004.

S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss. A synthetic multicellular system for programmed pattern formation. *Nature*, 434(7037):1130–4, 4 2005.

M. Beguerisse-Díaz, G. Bosque, D. Oyarzún, J. Picó, and M. Barahona. Flux-dependent graphs for metabolic networks. *arXiv preprint arXiv:1605.01639*, 2016.

M. Behar, N. Hao, H. G. Dohlman, and T. C. Elston. Mathematical and computational analysis of adaptation via feedback inhibition in signal transduction pathways. *Biophysical journal*, 93(3):806–821, 2007.

J. M. Berg, J. L. Tymoczko, L. Stryer, and N. D. Clarke. Biochemistry. 2002. *New York, New York*, 10010, 2002.

Biobrick Foundation. Part registry [online]. `http://partsregistry.org/`, 2006. Accesed: 20/02/2015.

BioFab. International open facility advancing biotechnology [online]. `http://www.biofab.org/`, 2006. Accesed: 20/02/2015.

BIOSS. Bioss toolbox [online]. `http://www.bioss.uni-freiburg.de/cms/toolbox-home.html`, 2006. Accesed: 20/02/2015.

X. Blasco, J. M. Herrero, J. Sanchis, and M. Martínez. A new graphical visualization of n-dimensional pareto front for decision-making in multiobjective optimization. *Information Sciences*, 178(20):3908 – 3924, 2008.

M. L. Blinov, J. R. Faeder, B. Goldstein, and W. S. Hlavacek. Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.

Y. Boada, A. Vignoni, and J. Picó. Model reduction and multi-objective identification of a feedback synthetic gene circuit. *IEEE Transactions on Control Systems Technology*.

Y. Boada, A. Vignoni, J. L. Navarro, and J. Picó. Improvement of a cle stochastic simulation of gene synthetic network with quorum sensing and feedback in a cell population. In *2015 European Control Conference (ECC)*, pages 2274–2279, 2015.

Y. Boada, J. Pitarch, A. Vignoni, G. Reynoso-Meza, and J. Picó. Optimization alternatives for robust model-based design of synthetic biological circuits. *IFAC-PapersOnLine*, 49(7):821 – 826, 2016a. ISSN 2405-8963. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB.

Y. Boada, G. Reynoso-Meza, J. Picó, and A. Vignoni. Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices: an adaptive network case. *BMC Syst Biol*, 10(1):27, 2016b.

Y. Boada, A. Vignoni, G. Reynoso-Meza, and J. Picó. Parameter identification in synthetic biological circuits using multi-objective optimization. volume 49, pages 77 – 82, 2016c. Foundations of Systems Biology in Engineering FOSBE.

Y. Boada, A. Vignoni, and J. Picó. Engineered control of genetic variability reveals interplay among quorum sensing, feedback regulation, and biochemical noise. *ACS Synthetic Biology*, 6(10):1903–1912, 2017a. doi: 10.1021/acssynbio.7b00087.

Y. Boada, A. Vignoni, and J. Picó. Multi-objective optimization for gene expression noise reduction in a synthetic gene circuit. *IFAC-PapersOnLine*, 50(1):4472 – 4477, 2017b. ISSN 2405-8963. 20th IFAC World Congress.

Y. Boada, A. Vignoni, and J. Picó. Multi-objective identification of synthetic circuits stochastic models using flow cytometry data. *Proceedings 25th Mediterranean Conference on Control and Automation MED*, pages 1077–1082, 2017c.

Y. Boada, A. Vignoni, D. Oyarzún, and J. Picó. Host-circuit interactions explain unexpected behaviours of a feedforward gene circuit. 2018. Foundations of Systems Biology in Engineering FOSBE.

H. Bremer, P. Dennis, and M. Ehrenberg. Free rna polymerase and modeling global transcription in escherichia coli. *Biochimie*, 85(6):597–609, 2003.

R. C. Brewster, D. L. Jones, and R. Phillips. Tuning promoter strength through rna polymerase binding site design in escherichia coli. *PLoS Comput Biol*, 8(12): e1002811, 2012.

N. E. Buchler, U. Gerland, and T. Hwa. Nonlinear protein degradation and the function of genetic circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9559–9564, 2005.

L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 3 2006.

Y. Cao, D. T. Gillespie, and L. R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of chemical physics*, 122(1):014116, 2005.

M. Carlquist, R. L. Fernandes, S. Helmark, A.-L. L. Heins, L. Lundin, S. J. Sorensen, K. Gernaey, and A. E. Lantz. Physiological heterogeneities in microbial populations and implications for physical stress tolerance. *Microb Cell Fact*, 11:94, 2012.

G. Chalancon, C. N. Ravarani, S. Balaji, A. Martinez-Arias, L. Aravind, R. Jothi, and M. Madan Babu. Interplay between gene expression noise and regulatory network architecture. *Trends in Genetics*, 28(5):221–232, 2012. ISSN 0168-9525.

V. Chellaboina, S. Bhat, M. Haddad, and D. Bernstein. Modeling and analysis of mass-action kinetics. *Control Systems, IEEE*, 29(4):60–78, Aug 2009. ISSN 1066-033X. doi: 10.1109/MCS.2009.932926.

W. W. Chen, M. Niepel, and P. K. Sorger. Classic and contemporary approaches to modeling biochemical reactions. *Genes & development*, 24(17):1861–1875, 2010a.

X. Chen, E. Pham, and K. Truong. Tev protease-facilitated stoichiometric delivery of multiple genes using a single expression vector. *Protein Sci*, 19(12):2379–88, 12 2010b.

A. W. T. Chiang and M.-J. J. Hwang. A computational pipeline for identifying kinetic motifs to aid in the design and improvement of synthetic gene circuits. *BMC Bioinformatics*, 14 Suppl 16:S5, 2013.

A. W. T. Chiang, W.-C. C. Liu, P. Charusanti, and M.-J. J. Hwang. Understanding system dynamics of an adaptive enzyme network from globally profiled kinetic parameters. *BMC Syst Biol*, 8:4, 2014.

O.-T. Chis, J. R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: a critical comparison of methods. *PloS one*, 6(11):e27755, 2011.

G. M. Church, M. B. Elowitz, C. D. Smolke, C. A. Voigt, and R. Weiss. Realizing the potential of synthetic biology. *Nature Reviews. Molecular Cell Biology*, 15(4): 289–294, 2014.

E. Cinquemani. Structural identification of biochemical reaction networks from population snapshot data. *IFAC-PapersOnLine*, 50(1):12629–12634, 2017.

R. S. Cox, C. Madsen, J. McLaughlin, T. Nguyen, N. Roehner, B. Bartley, S. Bhatia, M. Bissell, K. Clancy, T. Gorochowski, et al. Synthetic biology open language visual (sbol visual) version 2.0. *Journal of integrative bioinformatics*, 15(1), 2018.

G. Craciun and M. Feinberg. Multiple equilibria in complex chemical reaction networks: I. the injectivity property. *SIAM Journal on Applied Mathematics*, 65(5):1526–1546, 2005.

F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

N. Crook and H. S. Alper. Model-based design of synthetic, biological systems. *Chemical Engineering Science*, 103:2–11, 2013.

R. H. Dahl, F. Zhang, J. Alonso-Gutierrez, E. Baidoo, T. S. Batth, A. M. Redding-Johanson, C. J. Petzold, A. Mukhopadhyay, T. S. Lee, P. D. Adams, et al. Engineering dynamic pathway regulation using stress-response promoters. *Nature biotechnology*, 31(11):1039, 2013.

M. S. Dasika and C. D. Maranas. Optcircuit: An optimization based method for computational design of genetic circuits. *BMC Systems Biology*, 2:24, 2008.

H. De Jong, S. Casagranda, N. Giordano, E. Cinquemani, D. Ropers, J. Geiselmann, and J.-L. Gouzé. Mathematical modelling of microbes: metabolism, gene expression and growth. *Journal of the Royal Society Interface*, 14(136):20170502, 2017.

K. Deb, S. Bandaru, D. Greiner, A. Gaspar-Cunha, and C. C. Tutum. An integrated approach to automated innovization for discovering useful design principles: Case studies from engineering. *Applied Soft Computing*, 15(0):42 – 56, 2014.

D. Del Vecchio, A. J. Dy, and Y. Qian. Control theory meets synthetic biology. *Journal of The Royal Society Interface*, 13(120):20160380, 2016.

I. B. Dodd, A. J. Perkins, D. Tsemitsidis, and J. B. Egan. Octamerization of λ ci repressor is needed for effective repression of p rm and efficient switching from lysogeny. *Genes & development*, 15(22):3013–3022, 2001.

Y. Dublanche, K. Michalodimitrakis, N. Kümmerer, M. Foglierini, and L. Serrano. Noise in transcription negative feedback loops: simulation and experimental analysis. *Mol Syst Biol*, 2:41, 2006.

R. G. Egbert and E. Klavins. Fine-tuning gene networks using simple sequence repeats. *Proceedings of the National Academy of Sciences*, 109(42):16817–16822, 2012.

K. A. Egland and E. P. Greenberg. Conversion of the vibrio fischeri transcriptional activator, luxr, to a repressor. *Journal of Bacteriology*, 182(3):805–811, 2000.

A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467 (7312):167–173, 2010a.

A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467 (7312):167–173, 2010b.

J. Elf and M. Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, 13(11):2475–2484, 2003.

T. Ellis, X. Wang, and J. J. Collins. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat Biotechnol*, 27(5):465–71, 2009.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. *Science*.

M. Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors-i. the deficiency zero and deficiency one theorems. *Chemical Engineering Science*, 42(10):2229–2268, 1987.

X.-j. . J. Feng, S. Hooshangi, D. Chen, G. Li, R. Weiss, and H. Rabitz. Optimizing genetic circuits by global sensitivity analysis. *Biophysical journal*, 87(4):2195–2202, 2004.

R. L. Fernandes, M. Nierychlo, L. Lundin, A. E. Pedersen, P. P. Tellez, A. Dutta, M. Carlquist, A. Bolic, D. Schäpper, A. C. Brunetti, et al. Experimental methods and modeling techniques for description of cell population heterogeneity. *Biotechnology advances*, 29(6):575–599, 2011.

G. Fiore, F. Menolascina, M. Di Bernardo, and D. Di Bernardo. An experimental approach to identify dynamical models of transcriptional regulation in living cells. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(2):025106, 2013.

T. Folliard, H. Steel, T. P. Prescott, G. Wadhams, L. J. Rothschild, and A. Papachristodoulou. A synthetic recombinase-based feedback loop results in robust expression. *ACS synthetic biology*, 6(9):1663–1671, 2017.

D. M. Frangopol and K. Maute. Life-cycle reliability-based optimization of civil and aerospace structures. *Computers & structures*, 81(7):397–410, 2003.

C. Fuqua, M. Parsek, and E. Greenberg. Regulation of gene expression by cell-to-cell communication: acyl-homoserine lactone quorum sensing. *Annual review of genetics*, 35(1):439–468, 2001.

A. Gábor and J. R. Banga. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC systems biology*, 9(1):74, 2015.

K. A. Geiler-Samerotte, C. R. Bauer, S. Li, N. Ziv, D. Gresham, and M. L. Siegal. The details in the distributions: why and how to study phenotypic variability. *Curr Opin Biotechnol*, 24(4):752–9, 8 2013.

D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison III, and H. O. Smith. Enzymatic assembly of dna molecules up to several hundred kilobases. *Nature methods*, 6(5):343, 2009.

P. E. Gill, W. Murray, and M. A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Review*, 47(1):99–131, 2005.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1):404–425, 1992.

D. T. Gillespie. The chemical langevin equation. *The journal of Chemical Physics*, 113:297–306, 2000.

D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*, 58: 35–55, 2007.

N. Giordano, F. Mairet, J.-L. Gouzé, J. Geiselmann, and H. De Jong. Dynamical allocation of cellular resources as an optimal control problem: novel insights into microbial growth strategies. *PLoS computational biology*, 12(3):e1004802, 2016.

L. Goentoro, O. Shoval, M. W. Kirschner, and U. Alon. The incoherent feedforward loop can provide fold-change detection in gene regulation. *Molecular cell*, 36(5): 894–899, 2009.

A. B. Goryachev, D. J. Toh, and T. Lee. Systems analysis of a quorum sensing network: design constraints imposed by the functional requirements, network topology and kinetic constants. *Biosystems*, 83(2-3):178–87, 2006.

R. Grima, D. Schmidt, and T. Newman. Steady-state fluctuations of a genetic feedback loop: An exact solution. *The Journal of chemical physics*, 137(3):035104, 2012.

A. Gupta, C. Briat, and M. Khammash. A scalable computational framework for establishing long-term behavior of stochastic reaction networks. *PLoS computational biology*, 10(6):e1003669, 2014.

A. Gyorgy and D. Del Vecchio. Modular composition of gene transcription networks. *PLoS computational biology*, 10(3):e1003486, 2014.

J. F. Hair and M. G. Suárez. *Análisis multivariante*, volume 491. Prentice Hall Madrid, 1999.

E. J. Hancock, G.-B. B. Stan, J. A. J. Arpino, and A. Papachristodoulou. Simplified mechanistic models of gene regulation for analysis and design. *J R Soc Interface*, 12(108):20150312, 7 2015.

J. G. Harman. Allosteric regulation of the camp receptor protein. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1547(1):1–17, 2001.

R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Chapman and Hall, New York, 1996.

J. M. Herrero, X. Blasco, M. Martínez, and C. Ramos. Nonlinear robust identification using multiobjective evolutionary algorithms. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 231–241. Springer, 2005.

D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

D. J. Higham. Modeling and simulating chemical reactions. *SIAM Review*, 50(2): 347–368, 2008.

A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29): 12167–12172, 2011.

A. Hill. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J Physiol (Lond)*, 40:4–7, 1910. URL https://ci.nii.ac.jp/naid/10008305829/en/.

W. J. Holtz and J. D. Keasling. Engineering static and dynamic control of synthetic pathways. *Cell*.

F. Horn and R. Jackson. General mass action kinetics. *Archive for rational mechanics and analysis*, 47(2):81–116, 1972.

V. Hsiao, A. Swaminathan, and R. M. Murray. Control theory for synthetic biology: Recent advances in system characterization, control design, and controller implementation for synthetic biology. *IEEE Control Systems Magazine*, 38(3):32–62, 2018. doi: 10.1109/MCS.2018.2810459.

P. Incardona, A. Leo, Y. Zaluzhnyi, R. Ramaswamy, and I. F. Sbalzarini. Openfpm: A scalable open framework for particle and particle-mesh codes on parallel computers. *arXiv preprint arXiv:1804.07598*, 2018.

T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of mathematical biology*, 54(1):1–26, 2007.

A. S. Jarrah, R. Laubenbacher, B. Stigler, and M. Stillman. Reverse-engineering of polynomial dynamical systems. *Advances in Applied Mathematics*, 39(4):477–489, 2007.

S. Jayanthi, K. Nilgiriwala, and D. Del Vecchio. Retroactivity controls the temporal dynamics of gene transcription. *ACS synthetic biology*, 2(8):431–441, 2013.

D. L. Jones, R. C. Brewster, and R. Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–6, 12 2014.

J. Joo, S. J. Plimpton, and J.-L. L. Faulon. Statistical ensemble analysis for simulating extrinsic noise-driven response in nf-υb signaling networks. *BMC Syst Biol*, 7:45, 2013.

M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6(6):451–64, 2005.

H.-M. Kaltenbach, S. Dimopoulos, and J. Stelling. Systems analysis of cellular networks under uncertainty. *FEBS letters*, 583(24):3923–3930, 2009.

H. B. Kaplan and E. P. Greenberg. Diffusion of autoinducer is involved in regulation of the vibrio fischeri luminescence system. *Journal of bacteriology*, 163(3):1210–1214, 1985.

G. F. Kaufmann, R. Sartorio, S.-H. Lee, C. J. Rogers, M. M. Meijler, J. A. Moss, B. Clapham, A. P. Brogan, T. J. Dickerson, and K. D. Janda. Revisiting quorum sensing: discovery of additional chemical and biological functions for 3-oxo-n-acylhomoserine lactones. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2):309–314, 2005.

V. Kazeev, M. Khammash, M. Nip, and C. Schwab. Direct solution of the chemical master equation using quantized tensor trains. *PLoS computational biology*, 10(3): e1003359, 2014.

H. K. Khalil. *Nonlinear systems*, volume 2. 1996.

J. Kim, I. Khetarpal, S. Sen, and R. M. . Synthetic circuit for exact adaptation and fold-change detection. *Nucleic acids research*, page gku233, 2014.

K. H. Kim and H. M. Sauro. Adjusting phenotypes by noise control. *PLoS Comput Biol*, 8(1):e1002344, 2012.

D. J. Kiviet, P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans. Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522): 376, 2014.

S. Klumpp and T. Hwa. Growth-rate-dependent partitioning of rna polymerases in bacteria. *Proceedings of the National Academy of Sciences*, 105(51):20245–20250, 2008.

T. Knight. Draft standard for biobrick biological parts. 2007.

H. Koeppl, M. Hafner, and J. Lu. Mapping behavioral specifications to model parameters in synthetic biology. *BMC bioinformatics*, 14(Suppl 10):S9, 2013.

P. Kokotovic, H. Khalil, and J. O'Reilly. *Singular perturbation methods in control: analysis and design*. Academic Press, 1986.

A. Koseska, A. Zaikin, J. Kurths, and J. García-Ojalvo. Timing cellular decision making under noise via cell-to-cell communication. *PloS one*, 4(3):4872, 2009.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

P. Labhsetwar, J. A. Cole, E. Roberts, N. D. Price, and Z. A. Luthey-Schulten. Heterogeneity in protein expression induces metabolic variability in a modeled escherichia coli population. *Proceedings of the National Academy of Sciences*, 110(34):14006–14011, 2013.

J. H. Leveau and S. E. Lindow. Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *Journal of bacteriology*, 183(23): 6752–6762, 2001.

D. Lewis, P. Le, C. Zurla, L. Finzi, and S. Adhya. Multilevel autoregulation of λ repressor protein ci by dna looping in vitro. *Proceedings of the National Academy of Sciences*, 108(36):14807–14812, 2011.

G. Lillacci and M. Khammash. Parameter estimation and model selection in computational biology. *PLoS Comput Biol*, 6(3):e1000696, 2010.

J. Lippincott-Schwartz, E. Snapp, and A. Kenworthy. Studying protein dynamics in living cells. *Nature reviews Molecular cell biology*, 2(6):444, 2001.

M. Lozano, D. Molina, and F. Herrera. Editorial scalability of evolutionary algorithms and other metaheuristics for large-scale continuous optimization problems. *Soft Computing*, 15(11):2085–2087, 2011.

W. Ma, A. Trusina, H. El-Samad, W. A. Lim, and C. Tang. Defining network topologies that can achieve biochemical adaptation. *Cell*, 138(4):760–773, 2009.

M. A. Marchisio and J. Stelling. Automatic design of digital synthetic gene circuits. *PLoS Computational Biology*, 7(2), 2011.

R. Martí, T. Rodriguez, J. L. Pitarch, D. Sarabia, and C. De Prada. Dynamic optimization by automatic differentiation using EcosimPro CasADi. In *XXXV Jornadas de Automática*, pages 354–361. CEA-IFAC, 2014.

C. A. Mattson and A. Messac. Pareto frontier based concept selection under uncertainty, with visualization. *Optimization and Engineering*, 6(1):85–115, 2005.

B. Mélykúti, J. a. P. Hespanha, and M. Khammash. Equilibrium distributions of simple biochemical reaction systems for time-scale separation in stochastic reaction networks. *J R Soc Interface*, 11(97):20140054, 8 2014.

P. Mendes and D. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

P. Mendes, S. Hoops, S. Sahle, R. Gauges, J. Dada, and U. Kummer. Computational modeling of biochemical networks using copasi. In *Systems Biology*, pages 17–59. Springer, 2009.

F. Menolascina, M. Di Bernardo, and D. Di Bernardo. Analysis, design and implementation of a novel scheme for in-vivo control of synthetic gene regulatory networks. *Automatica*, 47(6):1265–1270, 2011.

F. Menolascina, G. Fiore, E. Orabona, L. De Stefano, M. Ferry, J. Hasty, M. di Bernardo, and D. di Bernardo. In-vivo real-time control of protein expression from endogenous and synthetic gene networks. *PLoS computational biology*, 10(5):e1003625, 2014.

G. R. Meza. *Controller Tuning by Means of Evolutionary Multiobjective Optimization: A holistic multiobjective optimization design procedure*. PhD thesis, Editorial Universitat Politècnica de València, 2014.

K. Miettinen. Nonlinear multiobjective optimization, volume 12 of international series in operations research and management science, 1999.

K. Miettinen, F. Ruiz, and A. P. Wierzbicki. Introduction to multiobjective optimization: interactive approaches. In *Multiobjective Optimization*, pages 27–57. Springer, 2008.

A. Milias-Argeitis, S. Summers, J. Stewart-Ornstein, I. Zuleta, D. Pincus, H. El-Samad, M. Khammash, and J. Lygeros. In silico feedback for in vivo regulation of a gene expression circuit. *Nature biotechnology*, 29(12):1114, 2011.

M. Miller, M. Hafner, E. Sontag, N. Davidsohn, S. Subramanian, P. E. Purnick, D. Lauffenburger, and R. Weiss. Modular design of artificial tissue homeostasis: robust control through synthetic cellular heterogeneity. *PLoS Comput Biol*, 8(7): e1002579, 2012.

R. Milo and R. Phillips. *Cell Biology by the Numbers*. first edition, 2015. ISBN 9780815345374.

R. Milo, R. Phillips, and N. Orme. *Cell Biology by the Numbers*. Garland Science, 2016.

A. Miyawaki, A. Sawano, T. Kogure, et al. Lighting up cells: labelling proteins with fluorophores. 2003.

C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research*, 13(11): 2467–2474, 2003.

S. Müller, H. Harms, and T. Bley. Origin and analysis of microbial population heterogeneity in bioprocesses. *Curr Opin Biotechnol*, 21(1):100–13, 2 2010.

B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.

E. M. Nelson, V. Kurz, N. Perry, D. Kyrouac, and G. Timp. Biological noise abatement: Coordinating the responses of autonomous bacteria in a synthetic biofilm to a fluctuating environment using a stochastic bistable switch. *ACS synthetic biology*, 3(5):286–297, 2013.

P. Nilsson, A. Olofsson, M. Fagerlind, T. Fagerström, S. Rice, S. Kjelleberg, and P. Steinberg. Kinetics of the ahl regulatory system in a model biofilm system: how many bacteria constitute a quorum? *Journal of molecular biology*, 309(3):631–640, 2001.

A. Novick and M. Weiner. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America*, 43(7): 553, 1957.

M. A. Nowak. *Evolutionary dynamics*. Harvard University Press, 2006.

E. J. Olson and J. J. Tabor. Optogenetic characterization methods overcome key challenges in synthetic and systems biology. *Nature chemical biology*, 10(7):502, 2014.

E. J. Olson, L. A. Hartsough, B. P. Landry, R. Shroff, and J. J. Tabor. Characterizing bacterial gene circuit dynamics with optically programmed gene expression signals. *Nature methods*, 11(4):449–455, 2014.

E. J. Olson, C. N. Tzouanas, and J. J. Tabor. A photoconversion model for full spectral programming and multiplexing of optogenetic systems. *Molecular systems biology*, 13(4):926, 2017.

I. Otero-Muras and J. R. Banga. Multicriteria global optimization for biocircuit design. *arXiv preprint arXiv:1402.7323*, 2014.

I. Otero-Muras and J. R. Banga. Design principles of biological oscillators through optimization: forward and reverse analysis. *PloS one*, 11(12):e0166867, 2016.

I. Otero-Muras and J. R. Banga. Automated design framework for synthetic biology exploiting pareto optimality. *ACS synthetic biology*, 6(7):1180–1193, 2017.

D. Oyarzún. Optimal control of metabolic networks with saturable enzyme kinetics. *IET systems biology*, 5(2):110–119, 2011.

D. A. Oyarzún, J.-B. Lugagne, and G.-B. Stan. Noise propagation in synthetic gene circuits for metabolic control. *ACS synthetic biology*.

M. S. Paula, P. J. F. de Córdoba Castellá, A. M. Aquino, and G. R. Meza. Modelling and multi-objective optimisation for simulation of cyanobacterial metabolism. 2017.

J. Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–8, 1 2004.

J. Picó, F. Garelli, H. De Battista, and R. Mantz. Geometric invariance and reference conditioning ideas for control of overflow metabolism. *Journal of Process Control*, 19(10):1617–1626, 2009.

J. Picó, A. Vignoni, E. Picó-Marco, and Y. Boada. Modelling biochemical systems: from mass action kinetics to linear noise approximation. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, 12(3):241–252, 7 2015.

E. Picó-Marco, Y. Boada, J. Picó, and A. Vignoni. Contractivity of a genetic circuit with internal feedback and cell-to-cell communication. *IFAC-PapersOnLine*, 49(26): 213 – 218, 2016. ISSN 2405-8963. Foundations of Systems Biology in Engineering FOSBE.

R. Porreca, S. Drulhe, H. d. Jong, and G. Ferrari-Trecate. Structural identification of piecewise-linear models of genetic regulatory networks. *Journal of Computational Biology*, 15(10):1365–1380, 2008.

L. Potvin-Trottier, N. D. Lord, G. Vinnicombe, and J. Paulsson. Synchronous long-term oscillations in a synthetic gene circuit. *Nature*, 538(7626):514–517, 2016.

T. P. Prescott and A. Papachristodoulou. Layered decomposition for the model order reduction of timescale separated biochemical reaction networks. *Journal of theoretical biology*, 356:113–122, 2014.

L. Pronzato and A. Pázman. Design of experiments in nonlinear models: Asymptotic normality, optimality criteria and small-sample properties. *Lecture Notes in Statistics*, 2013.

Y. Qian, H.-H. Huang, J. I. Jiménez, and D. Del Vecchio. Resource competition shapes the response of genetic circuits. *ACS synthetic biology*, 6(7):1263–1272, 2017.

E. R Dougherty and M. L Bittner. Causality, randomness, intelligibility, and the epistemology of the cell. *Current genomics*, 11(4):221–237, 2010.

S. J. Rahi, J. Larsch, K. Pecani, A. Y. Katsov, N. Mansouri, K. Tsaneva-Atanasova, E. D. Sontag, and F. R. Cross. Oscillatory stimuli differentiate adapting circuit topologies. In *Nature Methods*, 2017.

A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

C. V. Rao and A. P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the gillespie algorithm. *The Journal of chemical physics*, 118(11):4999–5010, 2003.

J. M. Raser and E. K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

G. Reynoso-Meza, J. Sanchis, X. Blasco, and M. Martínez. Design of continuous controllers using a multiobjective differential evolution algorithm with spherical pruning. *Applications of Evolutionary Computation*, pages 532–541, 2010.

G. Reynoso-Meza, J. Sanchis, X. Blasco, and J. M. Herrero. Multiobjective evolutionary algortihms for multivariable PI controller tuning. *Expert Systems with Applications*, 39:7895 – 7907, 2012.

G. Reynoso-Meza, X. Blasco, J. Sanchis, and J. M. Herrero. Comparison of design concepts in multi-criteria decision-making using level diagrams. *Information Sciences*, 221:124–141, 2013a.

G. Reynoso-Meza, S. García-Nieto, J. Sanchis, and X. Blasco. Controller tuning using multiobjective optimization algorithms: a global tuning framework. *IEEE Transactions on Control Systems Technology*, 21(2):445–458, 2013b.

G. Reynoso-Meza, J. Carrillo-Ahumada, Y. Boada, and J. Picó. Pid controller tuning for unstable processes using a multi-objective optimisation design procedure. *IFAC-PapersOnLine*, 49(7):284 – 289, 2016. ISSN 2405-8963. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB 2016.

C. Roberts, K. L. Anderson, E. Murphy, S. J. Projan, W. Mounts, B. Hurlburt, M. Smeltzer, R. Overbeek, T. Disz, and P. M. Dunman. Characterizing the effect of the staphylococcus aureus virulence factor regulator, sara, on log-phase mrna half-lives. *Journal of bacteriology*, 188(7):2593–2603, 2006.

G. Rodrigo and S. F. Elena. Structural discrimination of robustness in transcriptional feedforward loops for pattern formation. *PLoS One*, 6(2):16904, 2011.

G. Rodrigo, J. Carrera, and A. Jaramillo. Genetdes. *Bioinformatics*, 23(14):1857–1858, 2007.

G. Rodrigo, J. Carrera, T. E. Landrain, and A. Jaramillo. Perspectives on the automatic design of regulatory systems for synthetic biology. *FEBS letters*, 586(15):2037–2042, 2012.

M. Rodriguez-Fernandez, J. A. Egea, and J. R. Banga. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC bioinformatics*, 7 (1):483, 2006.

G. Russo and J. J. E. Slotine. Global convergence of quorum-sensing networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 82(4 Pt 1):041919, 10 2010.

H. M. Salis, E. A. Mirsky, and C. A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27(10):946–950, 2009a.

H. M. Salis, E. A. Mirsky, and C. A. Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27(10):946–950, 2009b.

M. A. H. Samee, B. Lim, N. Samper, H. Lu, C. Rushlow, G. Jiménez, S. Shvartsman, and S. Sinha. A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. *Cell Systems*, 1(6):396–407, 12 2015. ISSN 2405-4712.

M. Samoilov, S. Plyasunov, and A. P. Arkin. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (7):2310–2315, 2005.

A. Sánchez and J. Kondev. Transcriptional control of noise in gene expression. *Proceedings of the National Academy of Sciences*, 105(13):5081–5086, 2008.

L. V. Santana-Quintero, A. A. Montano, and C. A. C. Coello. A review of techniques for handling expensive functions in evolutionary multi-objective optimization. In *Computational intelligence in expensive optimization problems*, pages 29–59. Springer, Berlin, 2010.

M. Santillán and M. C. Mackey. Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data. *Proceedings of the National Academy of Sciences*, 98(4):1364–1369, 2001.

I. F. Sbalzarini, J. H. Walther, M. Bergdorf, S. E. Hieber, E. M. Kotsalis, and P. Koumoutsakos. Ppm- a highly efficient parallel particle-mesh library for the simulation of continuum systems. *Journal of Computational Physics*, 215(2):566–588, 2006.

A. L. Schaefer, D. L. Val, B. L. Hanzelka, J. E. Cronan, and E. P. Greenberg. Generation of cell-to-cell signals in quorum sensing: acyl homoserine lactone synthase activity of a purified vibrio fischeri luxi protein. *Proceedings of the National Academy of Sciences*, 93(18):9505–9509, 1996.

S. R. Schmidl, R. U. Sheth, A. Wu, and J. J. Tabor. Refactoring and optimization of light-switchable escherichia coli two-component systems. *ACS synthetic biology*, 3 (11):820–831, 2014.

D. Schnoerr, G. Sanguinetti, and R. Grima. Approximation and inference methods for stochastic biochemical kinetics-a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9):093001, 2017.

M. Schwarz-Schilling, L. Aufinger, A. Mückl, and F. Simmel. Chemical communication between bacteria and cell-free gene expression systems within linear chains of emulsion droplets. *Integrative Biology*, 8(4):564–570, 2016.

L. A. Segel and M. Slemrod. The quasi-steady-state assumption: a case study in perturbation. *SIAM review*, 31(3):446–477, 1989.

J. Sendin, O. Exler, and J. R. Banga. Multi-objective mixed integer strategy for the optimisation of biological networks. *IET systems biology*, 4(3):236–248, 2010.

T. Shopera, L. He, T. Oyetunde, Y. J. Tang, and T. S. Moon. Decoupling resource-coupled gene expression in living cells. *ACS synthetic biology*, 6(8):1596–1604, 2017.

A. Singh. Negative feedback through mrna provides the best control of gene-expression noise. *NanoBioscience, IEEE Transactions on*.

F. Siso-Nadal, J. F. Ollivier, and P. S. Swain. Facile: a command-line network compiler for systems biology. *BMC systems biology*, 1(1):36, 2007.

G. M. Skinner, C. G. Baumann, D. M. Quinn, J. E. Molloy, and J. G. Hoggett. Promoter binding, initiation, and elongation by bacteriophage t7 rna polymerase a single-molecule view of the transcription cycle. *Journal of Biological Chemistry*, 279 (5):3239–3244, 2004.

S. Srinath and R. Gunawan. Parameter identifiability of power-law biochemical system models. *Journal of biotechnology*, 149(3):132–140, 2010.

M. Srinivas and L. M. Patnaik. Genetic algorithms: A survey. *Computer*, 27(6):17–26, 1994.

J. I. Steinfeld, J. S. Francisco, and W. L. Hase. *Chemical kinetics and dynamics*, volume 3. Prentice Hall Englewood Cliffs (New Jersey), 1989.

R. Steuer, T. Gross, J. Selbig, and B. Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences*, 103(32):11868–11873, 2006.

S. Sutton. Measurement of cell concentration in suspension by optical density. In *Pharmaceutical Microbiology Forum Newsletter*, volume 12, pages 3–13, 2006.

P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

N. Tabareau, J.-J. J. Slotine, and Q.-C. C. Pham. How synchronization protects from noise. *PLoS Comput Biol*, 6(1):e1000637, 2010.

O. P. Tabbaa, G. Nudelman, S. C. Sealfon, F. Hayot, and C. Jayaprakash. Noise propagation through extracellular signaling leads to fluctuations in gene expression. *BMC Syst Biol*, 7:94, 2013.

Y. Taniguchi, P. J. Choi, G.-W. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8, 2010.

Y. Tanouchi, D. Tu, J. Kim, and L. You. Noise reduction by diffusional dissipation in a minimal quorum sensing motif. *PLoS Comput Biol*, 4(8):e1000167, 2008.

The Royal Academy of Engineering. Synthetic biology:scope, applications and implications. Technical report, The Royal Academy of Engineering, 2009.

T. Toni and B. Tidor. Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS Comput Biol*, 9(3):e1002960, 2013.

M. Ullah and O. Wolkenhauer. *Stochastic approaches for systems biology*. Springer Science & Business Media, 2011.

M. L. Urbanowski, C. P. Lostroh, and E. P. Greenberg. Reversible acyl-homoserine lactone binding to purified vibrio fischeri luxr protein. *Journal of Bacteriology*, 186 (3):631–637, 1 2004.

M. Vallerio, J. Hufkens, J. Van Impe, and F. Logist. An interactive decision-support system for multi-objective optimization of nonlinear dynamic processes with uncertainty. *Expert Systems with Applications*, 42(21):7710–7731, 2015.

N. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library. Elsevier Science, 2011. ISBN 9780080475363. URL `https://books.google.es/books?id=N6II-6HlPxEC`.

A. Vignoni, F. Garelli, and J. Picó. Sliding mode reference coordination of constrained feedback systems. *Mathematical Problems in Engineering*, 2013, 2013a.

A. Vignoni, D. A. Oyarzún, J. Picó, and G. B. Stan. Control of protein concentrations in heterogeneous cell populations. In *2013 European Control Conference (ECC)*, 2013b.

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

A. F. Villaverde, S. Bongard, K. Mauch, D. Müller, E. Balsa-Canto, J. Schmid, and J. R. Banga. A consensus approach for estimating the predictive accuracy of dynamic models in biology. *Comput Methods Programs Biomed*, 119(1):17–28, 4 2015. ISSN 1872-7565.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1), 2006. ISSN 0025-5610.

J. D. Watson, F. H. Crick, et al. A structure for deoxyribose nucleic acid. 1953.

J. C. Way, J. J. Collins, J. D. Keasling, and P. A. Silver. Integrating biological redesign: where synthetic biology came from and where it needs to go. *Cell*, 157(1):151–61, 2014a.

J. C. Way, J. J. Collins, J. D. Keasling, and P. A. Silver. Integrating biological redesign: where synthetic biology came from and where it needs to go. *Cell*, 157(1):151–161, 2014b.

M. Weber and J. Buceta. Noise regulation by quorum sensing in low mrna copy number systems. *BMC Syst Biol*, 5:11, 2011.

M. Weber and J. Buceta. Dynamics of the quorum sensing switch: stochastic and non-stationary effects. *BMC Syst Biol*, 7:6, 2013. ISSN 1752-0509. doi: 10.1186/1752-0509-7-6.

J. N. Weiss. The hill equation revisited: uses and misuses. *The FASEB Journal*, 11 (11):835–841, 1997.

T. F. Weiss. *Cellular biophysics*, volume 1. MIT press Cambridge, Mass:, 1996.

A. Y. Weiße, D. A. Oyarzún, V. Danos, and P. S. Swain. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proceedings of the National Academy of Sciences*, page 201416533, 2015.

S. A. Westermayer, G. Fritz, J. Gutiérrez, J. A. Megerle, M. P. Weißl, K. Schnetz, U. Gerland, and J. O. Rädler. Single-cell characterization of metabolic switching in the sugar phosphotransferase system of escherichia coli. *Molecular microbiology*, 100(3):472–485, 2016.

D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. Champan and Hall/CRC. Mathematical and computational Biology Series, 2006.

D. J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet*, 10(2):122–33, 2 2009.

A. Wittmann and B. Suess. Engineered riboswitches: Expanding researchers' toolbox with synthetic rna regulators. *FEBS letters*, 586(15):2076–2083, 2012.

M. L. Woods, M. Leon, R. Perez-Carrasco, and C. P. Barnes. A statistical approach reveals designs for the most robust stochastic gene oscillators. *ACS Synth Biol*, 2 2016.

A. Zagaris, H. G. Kaper, and T. J. Kaper. Analysis of the computational singular perturbation reduction method for chemical kinetics. *Journal of Nonlinear Science*, 14(1):59–91, 1 2004. ISSN 0938-8974. doi: 10.1007/s00332-003-0582-9.

A. Zargar, D. N. Quan, and W. E. Bentley. Enhancing intercellular coordination: Rewiring quorum sensing networks for increased protein expression through autonomous induction. *ACS Synth Biol*, 5(9):923–8, 2016.

C. Zechner, G. Seelig, M. Rullan, and M. Khammash. Molecular circuits for dynamic noise filtering. *Proceedings of the National Academy of Sciences*, page 201517109, 4 2016.

C. Zhang, R. Tsoi, and L. You. Addressing biological uncertainties in engineering gene circuits. *Integrative Biology*, 8(4):456–464, 2016.

S. Zucca, L. Pasotti, N. Politi, M. G. Cusella, and P. Magni. A standard vector for the chromosomal integration and characterization of biobrick$^{\mathrm{TM}}$parts in *Escherichia coli*. *Journal of Biological Engineering*, 2013.