Computational design and designability of gene regulatory networks

Guillermo RODRIGO TÁRREGA

Thesis Advisors:
Prof. Santiago F. ELENA FITO
Dr. Alfonso JARAMILLO ROSALES

MINISTERIO
DE CIENCIA
E INNOVACIÓN

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

UNIVERSIDAD
POLITECNICA
DE VALENCIA

INSTITUTO DE BIOLOGIA MOLECULAR Y CELULAR DE PLANTAS

D. Santiago F. Elena Fito, Doctor en Ciencias Biológicas y Profesor de Investigación del Consejo Superior de Investigaciones Científicas (CSIC) en el Instituto de Biología Molecular y Celular de Plantas (IBMCP), centro mixto del CSIC y la Universidad Politécnica de Valencia,

CERTIFICA

Que D. Guillermo Rodrigo Tárrega, Ingeniero Industrial, ha realizado bajo mi supervisión la tesis doctoral titulada "*Computational design and designability of gene regulatory networks*".

Y para que así conste, firmo la presente en Valencia, a 25 de Noviembre de 2011.

Santiago Elena Fito

Dr. Alfonso JARAMILLO
Alfonso.Jaramillo@issb.genopole.fr

***i*SSB** (EA 4527)
University of Évry-Val-d'Essonne
Genopole®

D. Alfonso Jaramillo Rosales, Doctor en Ciencias Físicas y Chargé de Recherche (CNRS) en el Institute of Systems and Synthetic Biology (iSSB), Université d'Évry-Val-d'Essonne – CNRS UPS3201 – Genopole®,

CERTIFICA

Que D. Guillermo Rodrigo Tárrega, Ingeniero Industrial, ha realizado bajo mi co-supervisión la tesis doctoral titulada "*Computational design and designability of gene regulatory networks*".

Y para que así conste, firmo la presente en Évry, a 25 de Noviembre de 2011,

To my parents
and Maribel

Computational design and designability of gene regulatory networks

Guillermo Rodrigo Tárrega

Abstract

Our understanding of molecular interactions has now conducted to an engineering perspective, where designs and implementations of artificial regulatory systems are attempted to provide instrumental insights for cell reprogramming. We here addressed the design of regulatory networks as a way to further understand the natural regulations. We also tackled the designability problem provided a library of interoperable elements. For that, we applied heuristic optimization methods that implement routines to solve inverse problems and mathematical analysis tools to quantitatively study the dynamics of gene expression. Because the engineering of transcription networks has mostly relied on the assembly of few characterized regulatory elements using rational design principles, we developed a computational framework to exploit such a design approach. Libraries of models of regulatory elements were examined to screen the genotypic space associated to a given phenotypic behavior. Additionally, we developed a fully automated procedure to design small non-coding RNAs with regulatory ability, based on a physicochemical model and exploiting allosteric regulation. The resulting RNA devices implemented a mechanism of post-transcriptional control of protein expression that could be combined with transcription regulation. We also applied heuristic techniques to study the designability of metabolic pathways. Certainly, computational design methods can also learn from natural mechanisms to exploit their underlying principles. In that way, such studies would allow us to go deeper in our ability of engineering artificial systems. Of relevance, integral control and incoherent regulations are ubiquitous strategies that organisms employ and we here analyzed. Moreover, genomic techniques allow us to study the multiple and complex interactions of the global network of the cell. In particular, we studied the transcription reprogramming upon viral infection. In sum, our results demonstrate that computational methods can be applied to *de novo* design genetic networks and characterize the designability of desired functions,

noting that a quantitative optimization algorithm has resulted useful in diverse regulatory frameworks. The mathematical study of natural systems has also served to reveal instrumental mechanisms to manage biological functions. All together, this thesis provides further quantitative insights about the natural control mechanisms implemented by gene regulatory networks.

Diseño computacional y diseñabilidad de redes de regulación genética

Guillermo Rodrigo Tárrega

Resumen

Nuestro conocimiento de las interacciones moleculares nos ha conducido hoy hacia una perspectiva ingenieril, donde diseños e implementaciones de sistemas artificiales de regulación intentan proporcionar instrucciones fundamentales para la reprogramación celular. Nosotros aquí abordamos el diseño de redes de genes como una forma de profundizar en la comprensión de las regulaciones naturales. También abordamos el problema de la diseñabilidad dada una genoteca de elementos compatibles. Con este fin, aplicamos métodos heurísticos de optimización que implementan rutinas para resolver problemas inversos, así como herramientas de análisis matemático para estudiar la dinámica de la expresión genética. Debido a que la ingeniería de redes de transcripción se ha basado principalmente en el ensamblaje de unos pocos elementos regulatorios usando principios de diseño racional, desarrollamos un marco de diseño computacional para explotar este enfoque. Modelos asociados a genotecas fueron examinados para descubrir el espacio genotípico asociado a un cierto fenotipo. Además, desarrollamos un procedimiento completamente automatizado para diseñar moleculas de ARN no codificante con capacidad regulatoria, basándonos en un modelo fisicoquímico y aprovechando la regulación alostérica. Los circuitos de ARN resultantes implementaban un mecanismo de control post-transcripcional para la expresión de proteínas que podía ser combinado con elementos transcripcionales. También aplicamos los métodos heurísticos para analizar la diseñabilidad de rutas metabólicas. Ciertamente, los métodos de diseño computacional pueden al mismo tiempo aprender de los mecanismos naturales con el fin de explotar sus principios fundamentales. Así, los estudios de estos sistemas nos permiten profundizar en la ingeniería genética. De relevancia, el control integral y las regulaciones incoherentes son estrategias generales que los organismos emplean y que aquí analizamos. Además, las técnicas genómicas nos permiten el estudio de las múltiples y complejas interacciones de la red global de la célula. En particular, estudiamos la reprogramación trancripcional bajo una infección viral. En suma, nuestros resultados demuestran que los métodos computacionales pueden ser aplicados para el diseño y la caracterización de redes de genes, resaltando que los algoritmos de optimización han resultado muy útiles en diferentes contextos de regulación. El estudio matemático de los sistemas naturales también ha servido para revelar mecanismos instrumentales de gestión de funciones biológicas. Con todo, esta tesis proporciona conclusiones cuantitativas sobre los mecanismos de control naturales implementados por redes de regulación genética.

Diseny computacional i disenyabilitat de xarxes de regulació gènica

Guillermo Rodrigo Tárrega

Resum

La nostra comprensió de les interaccions moleculars ens ha conduït hui a un punt de vista d'enginyeria, on dissenys i implementacions de sistemes artificials de regulació tracten de proporcionar instruccions fonamentals per a la reprogramació cel·lular. Nosaltres ací abordem el disseny de xarxes de gens com una forma d'entendre millor les regulacions naturals. També abordem el problema de la disenyabilitat donada una genoteca d'elements compatibles. Amb aquest objectiu, apliquem mètodes d'optimització heurística que implementen rutines per resoldre problemes inversos i eines d'anàlisi matemàtic per estudiar la dinàmica de l'expressió gènica. Com que l'enginyeria de xarxes de transcripció s'ha basat principalment en l'acoblament de pocs elements de regulació utilitzant els principis de disseny racional, vam desenvolupar un marc computacional per explotar aquest enfocament. Models associats a genoteques van ser examinats per detectar l'espai genotípic associat a un cert fenotip. A més, vam desenvolupar un procediment totalment automatitzat per dissenyar molecules d'ARN no codificants amb capacitat de regulació, basat en un model fisicoquímic i aprofitant la regulació al·lostèrica. Els circuits d'ARN que van resultar implementaven un mecanisme de control post-transcripcional de l'expressió de proteïnes que podia combinar-se amb elements de transcripció. També vam aplicar les tècniques heurístiques per a l'estudi de la disenyabilitat de vies metabòliques. Certament, els mètodes computacionals de disseny també poden al mateix temps aprendre dels mecanismes naturals per explotar els seus principis fonamentals. D'aquesta manera, els estudis de sistemes naturals ens permeten aprofondir en l'enginyeria genètica. De rellevància, el control integral i les regulacions incoherents són estratègies generals que els organismes utilitzen i que hem analitzat. D'altra banda, les tècniques genòmiques ens permeten estudiar les interaccions múltiples i complexes de la xarxa global de la cèl·lula. En particular, vam estudiar la reprogramació transcripcional deguda a una infecció viral. En resum, els nostres resultats demostren que els

mètodes computacionals es poden aplicar al disseny i caracterització de xarxes de gens, tenint en compte que els algorismes d'optimització han sigut útils en diversos marcs de regulació. L'estudi matemàtic dels sistemes naturals també ha servit per revelar els mecanismes instrumentals de gestió de funcions biològiques. En conjunt, aquesta tesi proporciona una visió més quantitativa sobre els mecanismes de control naturals implementats per xarxes de regulació gènica.

Contents

Objectives

The objectives of this thesis are

- Understanding and characterization of the designability of regulatory networks from a library of models of genetic elements. Application of optimization methods to explore the combinatorial space.

- Understanding ubiquitous mechanisms found in natural gene regulatory networks. Application of optimization methods to unravel design principles. Associative analysis of network architecture, function and robustness. Application of mathematical analysis techniques to disentangle the role of a given regulation. Comparative analysis of different regulatory modes. Application of control theory perspectives.

- Understanding by a design approach the mechanisms of riboregulation. Application of optimization methods to evolve sequences of nucleic acids. Development of a statistical mechanics model.

- Understanding and characterization of the designability of metabolic pathways. Application of heuristic methods.

- Understanding the cellular organization of gene regulations and how this global network serves to respond to external perturbations. Application to the case of plant viruses.

Chapter 1

Introduction

...they succeed in adapting themselves
best to their environment.
– Charles Darwin

About one century and a half ago, Darwin stated that the ability of organisms to adapt ultimately results in their ability to survive to environmental changes [1]. In a fluctuating environment, it would not be the strongest that would survive but the most adaptable to change. These fluctuations would affect the available resources that organisms would use for growth, development, survival and reproduction; in addition to impose new conditions for what they would not be used to. Currently we know, with the advances in Molecular Biology, that gene regulations are instrumental to process external signals and trigger the expression of several genes accordingly [2]. This complex sensor-actuator machinery, even in the simplest organisms, is the result of millions of years of evolution subjected to environmental fluctuations. Not surprisingly therefore, organisms present certain degree of tolerance to perturbations in their natural niche. The study of the gene regulatory networks of the cell is hence promising to provide fundamental insights about the life-driving principles.

As our understanding of molecular interactions increases [3], we start to develop an engineering perspective, where designs of artificial regulatory systems are aimed. At the same time, this perspective, from the observation of the engineered systems, does not only assess the accumulated knowledge but provides further comprehension, and eventually new issues, of the molecular interactions. Whether we can construct mathematical models of molecular interactions able to predict the expected behavior of a genetic circuit, the design of these circuits will require borrowing appropriate methodologies from hard-core engineering [4]. In this thesis, we adopt a computational design perspective to further understand and even reveal the underlying mechanisms of gene regulatory networks. To this end, we develop computational methods to support the *de novo* design and implementation of biological regulatory systems with a desired behavior, the goal of Synthetic Biology [5]. We also apply analysis techniques to conceptualize different natural regulatory mechanisms, which can be then exploited for the rational design of artificial systems [6].

For the computational design of regulatory networks, we adopt a strategy consisting in the design by optimization. Over the last decades, the algorithms based on probabilistic schemes that mimic natural evolution have been used to address optimization problems. Depending on the implementation of the iterative process of mutation-then-selection, we can describe a wide range of evolutionary algorithms [7], a popular family of optimization methods. These are inspired in the biological principles that govern the evolution of a finite population through certain selective pressure. A system evolves from a defined starting state by successive steps of variation (random or directed) and selection. The variation steps correspond to small modifications to the model of the system by using unary operators (the modification only depends on the former state), binary (the modification depends on two former states, usually called parents in analogy with sexual reproduction) or higher order operators. The selection is performed by means of an objective function (or fitness) that measures, for instance, how close the dynamics is to a target function. Of note, the difference between different algorithms is more a question of implementation. The reproduction operators and the selection procedure can be parameterized obtaining a wide range of algorithms depending on the values of a set of control parameters. The selection method constitutes the major element in the algorithm, as the designed system will be the result for what we have selected. Here, to approach the computational design of gene regulation circuits, we use algorithms based on Monte Carlo Simulated Annealing [8]. With the Boltzmann criterion, the initial evolutionary dynamics closely resembles a random walk (and hence a more efficient exploration of the fitness landscape), whereas as time goes on the dynamics resembles an adaptive walk. However, could automated design methods reach the required degree of reliability for their application in biology and thus provide useful insights over the problem of designing regulatory systems? If so, what are the application domains of such a design strategy?

We apply computational methods to design transcription networks, small regulatory RNAs, and metabolic pathways. Although we abstract each problem, it is expected the use of computational techniques to design more sophisticated networks involving several regulatory mechanisms, as certainly it occurs in natural systems. We approach the design of networks as the inverse problem of finding the right sequence of nucleic acids given a desired functionality. The computational design framework allows us, in addition, to tackle the problem of designability. This measures the ability we have to design functional networks provided a library of composable elements. Once a design is obtained, it is also possible to compare it with an eventual natural analog and infer design principles. Do artificial systems implement the control mechanisms found in natural ones? Moreover, computational design can also learn as rational design does from natural examples to create new ones. This is sometimes beneficial to accelerate the design process at the cost of a biased approach. Noting also that the designed networks will be then integrated into a cellular background, the incorporation into the design process of cellular factors would enhance the reliability of the networks. Certainly, the network components (nucleic acids and encoded proteins) can establish multiple and complex interactions not only among themselves but also with certain components of the host cell. As a result in the not so long term, the novel functional networks prospect to serve for reprogramming the future synthetic cells [9].

Bibliography

[1] Darwin C (1859) On the origin of species. John Murray, London.

[2] Ptashne M, Gann A (2002) Genes and signals. Cold Spring Harbor Laboratory Press, New York.

[3] Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3: 318-356.

[4] Lu TK, Khalil AS, Collins JJ (2009) Next-generation synthetic gene networks. *Nat Biotechnol*, 27: 1139-1150.

[5] Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet*, 6: 533-543.

[6] Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genet*, 5: 34-42.

[7] Holland JH (1992) Adaptation in natural and artificial systems. MIT Press, Cambridge MA.

[8] Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science*, 220: 671-680.

[9] Gibson DG, Glass JI, *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329: 52-56.

Chapter 2

Designability of transcription regulatory networks

Limited in his nature,
infinite in his desires...
– Alphonse de Lamartine

One approach to unserstand the operational mechanisms imposed by gene regulations consists in designing functional networks. Whether we can design such networks, we gain quantitative insights about the underlying principles that govern the biological behaviors. For the implementation of the networks, different genetic elements are combined. However, there is a limitation in the number of interoperable and well-characterized regulatory elements we can use to implement the designed networks. Hence, in this chapter, we study how a set of regulatory elements can be assembled to implement functional networks. We address intriguing questions that arise from this design approach, such as the number of functional circuits we can engineer with a given library of elements and the diversity of possible behaviors.

2.1 Assembly of networks

Over the last decade, we have witnessed the expansion of Synthetic Biology [1], where the attempts for cell reprogramming to perform new tasks have fructified in the engineering of several synthetic gene regulatory networks [2–20]. Usually, the design of synthetic networks has been inspired on the use of mathematical models [21, 22] and empirical engineering rules inferred from natural examples [23–24], although requiring in many cases a genetic fine-tuning to achieve the desired behavior [25]. It is expected that the widespread use of libraries of previously well-characterized genetic regulatory elements [26–29], together with the ability of engineering combinatorially those elements [30], will allow avoiding trial-and-error procedures, which are not efficient for optimizing and implementing complex systems. Those designed circuits may be later fine-tuned with directed evolution techniques, although there is no a general methodology for the *de novo* network engineering. In fact, this bottom-up approach is commonly used in other areas of engineering where a set of off-the-shelf parts with precise specifications of their operating points can be used to engineer sophisticated systems, and has been already successful to engineer novel biological circuits [12, 19].

Large efforts in generating genetic diversity, especially libraries of promoters [19, 31–36] but also post-transcriptional regulatory elements [6, 14, 37–39] and synthetic transcription factors [40, 41], encourage to use a combinatorial approach to design artificial circuits. In addition, the quantitative characterization of these regulatory elements allows inferring simple phenomenological mathematical models, which could be used to construct the model of a system that assembles different elements. In that way, several Synthetic Biology-oriented design tools have been developed to make available a library of mathematical models created from that genetic diversity, together with an interface to create gene networks by wiring elements [42–47]. Notably, such a genetic diversity is translated into a functional diversity when assembling networks, and these networks could be readily compiled into nucleic acid sequences. However, the design is reduced to examine one-by-one all possible combinations (*e.g.*, simulating the dynamical behavior), resulting in a tedious design process. Thereby, the evolutionary algorithms and optimization techniques [48–52] allow us to automate this process to find the desired circuits and finally depict the functional diversity of a library of regulatory elements. Our

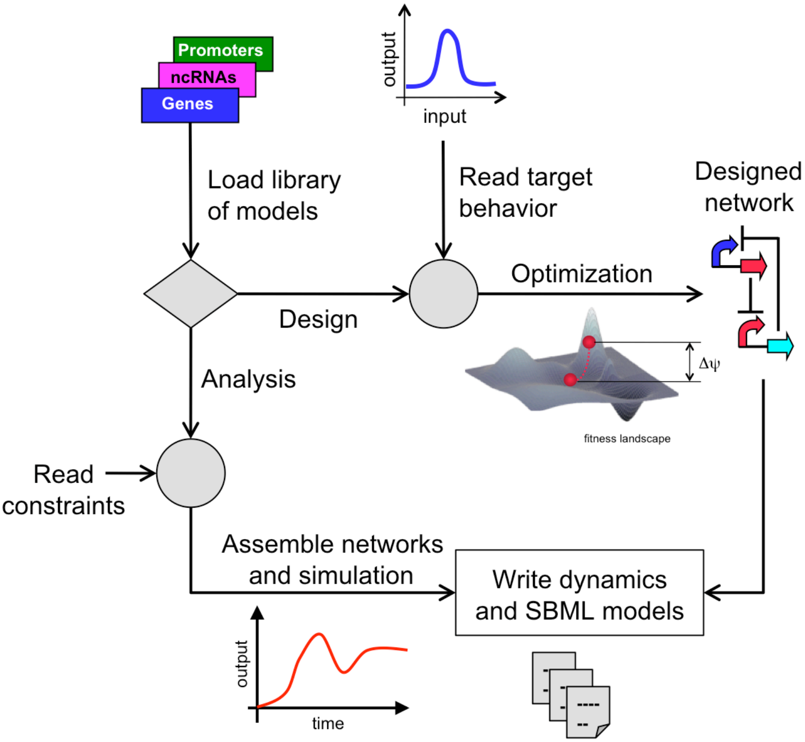


Figure 2.1: Scheme of the design platform adopted by harnessing a library of models of composable regulatory elements. We explore the functional networks that can be engineered either by exhaustive combinatorial assembly or by heuristic optimization.

novel approach allows assembling models of regulatory elements from a library and couples this with an automated design strategy.

In this chapter, we tackle fundamental questions that naturally emerge from that approach. What functional circuits can we engineer with a given library of regulatory elements? What is the diversity of possible behaviors and what is the designability (defined as the fraction of assembled circuits that follow a given behavior) of each one? Is one behavior easier to design than others? Certainly, these features depend on the employed library. We also wonder what is the sensitivity of the results to the regulatory elements; in other words, how many functional circuits involve a given regulatory element? In addition, we look at the robustness of a circuit by locally perturbing its parameters and evaluating the resulting fitness. At fixed network topology, we further analyze the whole parameter space that provides the targeted functionality, which accounts for the robustness of all operative points and asymptotically tends to a value that we call asymptotic robustness. Indeed, this property accounts for the ability to design such a circuit given the limitation of the number of genetic elements, and it could be important to analyze the natural occurrence of certain genetic architectures. All in all, to solve these questions, we developed a computational framework to assemble, simulate and design circuits, and that allowed us to explore the functional diversity that came from the assembled circuits with certain behavior (see Fig. 2.1). The design of circuits was accomplished by a selection step according to a dynamical behavior-based fitness function that can also account for robustness. Because the composability of genetic elements is simpler, we focused on bacterial systems. Initially, we applied the methodology to design several functional circuits with unlimited genetic diversity (given by the parameter space) and study their asymptotic robustness. Then, we designed complex circuits by plugging functional modules. Subsequently, we dissected the whole dynamical spectrum of a limited library of regulatory elements and analyzed the properties of the resulting circuits. We also analyzed the dependence of these results on the constituent library and how they could change when the stochasticity of the cell is taken into account. Finally, we discussed the reliability and implementability of the designed circuits.

2.2 Mathematical modeling and optimization method

For modeling genetic networks, we used a coupled system of differential equations. We considered three different types of species: mRNA (it can also be non-coding), proteins (mainly transcription factors) and small molecules that interact with proteins to activate or inhibit their regulatory ability (*e.g.*, isopropyl β-D-1-thiogalactopyranoside –IPTG– inhibits the activity of LacI). Likewise, the production of the i^{th} mRNA (x_i) from a regulated promoter follows

$$\frac{dx_i}{dt} = Cf(y_j, u_j) - (\delta_i + \mu)x_i, \tag{2.1}$$

where the term $f(y_j, u_j)$ is the transcription rate (y_j and u_j represent the concentrations of the j^{th} protein and its regulating chemical, respectively), C the gene copy number, δ_i the mRNA degradation coefficient, and μ the growth rate of the cell (dilution term). $C = 1$ is assumed to be constant in this work. For the computational design of a circuit, we did not impose variability on μ but we assumed a constant value (*e.g.*, $\mu = 0.02\ \mathrm{min}^{-1}$). For simplicity, we assumed that all genes in an operon (*i.e.*, controlled by the same promoter) have the same mRNA expression. The term $f(y_j, u_j)$ accounts for protein-DNA and protein-molecule interactions [22], and for constitutive promoters it is constant. Importantly, our approach is independent of the choice of this function, thus giving a big degree of freedom to the kinetic characterization from experimental data. Afterwards, the production of i^{th} protein (y_i) is given by

$$\frac{dy_i}{dt} = g(x_i, x_j) - (\beta_i + \mu)y_i, \tag{2.2}$$

where the term $g(x_i, x_j)$ is the translation rate and β_i the protein degradation coefficient. The term $g(x_i, x_j)$ accounts for post-transcriptional regulatory mechanisms, such as riboswitches, allowing a further genetic element, such as a *trans*-RNA, to control translation [6]. In case of no post-transcriptional elements, the translation rate is proportional to the mRNA concentration. In addition, in this work we only considered first-order degradation kinetics. For the stochastic simulation, we adopted a Langevin model [72] accounting for intrinsic and extrinsic noise, resulting in

$$\begin{aligned} \frac{dx_i}{dt} &= Cf(y_j, u_j) - (\delta_i + \mu)x_i + \sqrt{Cf(y_j, u_j) + (\delta_i + \mu)x_i}\ \xi_{x_i}(t) + q_g\xi_g(t), \\ \frac{dy_i}{dt} &= g(x_i, x_j) - (\beta_i + \mu)y_i + \sqrt{g(x_i, x_j) + (\beta_i + \mu)y_i}\ \xi_{y_i}(t) + q_g\xi_g(t), \end{aligned} \tag{2.3}$$

where ξ_i are Wiener processes with statistics $\langle \xi_i(t) \rangle = 0$ and $\langle \xi_i(0)\xi_i(t) \rangle = \delta(t)$ (Dirac delta) for intrinsic noise, and $\langle \xi_g(t) \rangle = 0$ and $\langle \xi_g(0)\xi_g(t) \rangle = \frac{\mu}{2} exp(-\mu|t|)$ for extrinsic noise. Also q_g gives the amplitude of the extrinsic noise.

To construct a library, each genetic regulatory element was modeled by transfer functions that related the output to the input values. These functions can be fitted from experimental data. As DNA fragments, mathematical models can be assembled in a standard way to simulate the behavior of circuits. Here, we only allowed joining promoter and genes, or genes and genes (*i.e.*, two consecutive promoters was not allowed; such a construction should be specified as a whole part). One useful format to store a mathematical model (molecular species, kinetic parameters and DNA sequence) is SBML [53]. Hence, we had a single SBML file for the model of each biological part; similarly as crystallographic data is stored in PDB format. The models for promoter parts only account for the transcription rate, whereas for gene parts the model accounts for the translation rate and the degradation and dilution rates of mRNA and proteins. We selected Hill function models because their overwhelming use in current characterization of transcription regulation works. In the future, when more advanced models may be used to fit characterization data, they could be readily used with our computational design procedure. A range of variation can be specified for some kinetic parameters; likewise the corresponding value will be susceptible to be changed during the design process.

Multiple circuits can be constructed by harnessing the available regulatory elements. To computationally explore the functional diversity that offers such a library and the designability of certain behaviors, two different strategies can be adopted, and our approach provides an automated implementation of them. On the one hand, following the exhaustive design strategy, all possible circuits, up to a maximal number of elements, are constructed and simulated. Having the large collection of dynamics, a post-processing step is applied to find those circuits that behave according to the design specifications. This approach allows obtaining the whole functional diversity and designability. On the other hand, a heuristic design strategy provides a probabilistic sampling frame of the functional diversity. It allows iteratively assembling models of existing elements and evaluating the performance of the resulting circuit according to a dynamical behavior-based fitness function [54]. For that, we used Monte Carlo Simulated Annealing (MCSA) as optimization scheme [55]. A movement in the fitness landscape consists in a replacement, addition or deletion of a given regulatory element. To evaluate the fitness function, we firstly calculated the average distance (metric function) for all genes i considered as outputs, for a given target behavior k, between the current circuit dynamics (y_{ik}) and the target one (z_{ik}) according to

$$\phi_k = \sum_i \frac{\displaystyle\int_0^T |log(y_{ik}(t)) - log(z_{ik}(t))|\chi_{ik}(t)dt}{\displaystyle\int_0^T \chi_{ik}(t)dt}, \qquad (2.4)$$

where T is the final time (*e.g.*, the time to reach the steady state). The metric is in logarithmic scale to properly balance the species concentrations, since they can vary in several orders of magnitude in biological systems. The function $\chi(t)$ is a weighting factor to only evaluate the circuit dynamics in a specified temporal domain ($\chi(t) : [0, T] \to [0, 1]$). Subsequently, the fitness function that aggregates all targets we used reads

$$\psi = \prod_k \left(1 - \frac{\phi_k}{\phi_0}\right)^{\gamma_k}, \tag{2.5}$$

where ϕ_0 is a normalization constant to adjust the fitness value to the metric function (*e.g.*, $\phi_0 = 3$), and γ_k gives the scalar weight in logarithmic scale for optimizing target k (*e.g.*, $\gamma_k = 10\gamma_l$ indicates that target k has 10 times more priority than l). If $\phi_k > \phi_0$ for one k then we assumed $\psi = 0$. Importantly, this fitness function (ψ belongs to the interval [0,1]) penalizes those circuits that do not satisfy simultaneously all targets. Being $\Delta\psi$ the fitness update after a movement, this is accepted with probability $\max(1, \exp(\Delta\psi/T_{MCSA}))$, where T_{MCSA} is the MCSA temperature. T_{MCSA} is continuously adjusted during the optimization process following an exponential cooling scheme.

2.3 Network design and modularity

Initially, we constructed a library of artificial regulatory elements, including all types of logic combinatorial promoters of two entries. Additionally, the kinetic parameters characterizing those elements were specified as a range of variation. This feature allows that the genetic sequence of many biological parts could be easily modified to create a new part with diminished binding affinity or stability by a single mutation. Otherwise, it is much more difficult to find a suitable mutation that would increase the binding or stability. Therefore, by allowing this range in the parameter space, we would enlarge the search space while still maintaining the linking with the genotype, because the parts from an optimal solution could be readily engineered to follow a model agreeing with the designed parameters. The nominal values were taken from several experimental studies [11, 12, 27, 36]. Thus, the genetic diversity was almost unlimited, being the design space defined by topological and parameter modifications of the circuit. To explore this space we applied the heuristic design strategy to find

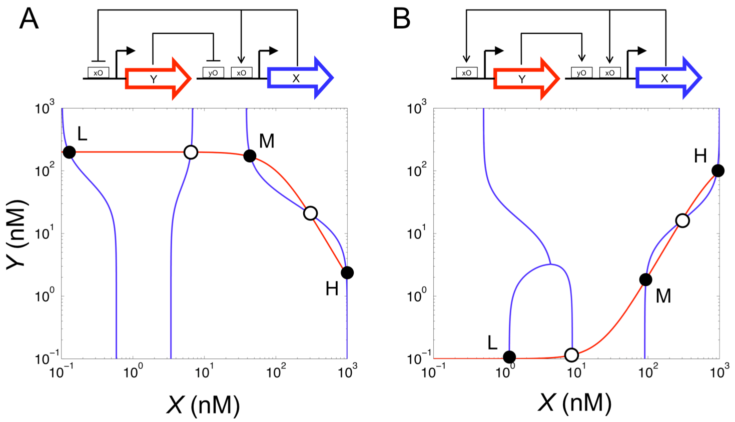


Figure 2.2: Schemes of two two-gene circuits designed to reach tristability, showing the corresponding phase diagrams. Filled and open circles represent stable and unstable steady states, respectively.

the optimal assemblies of elements and parameterizations that gave functional circuits. Firstly, we repeatedly applied the optimization method to design all possible circuits relying on a feed-forward loop (FFL) structure for one-stripe pattern formation [56–59]. We found six different architectures for working as an amplitude filter (see also next chapter), where five of them corresponded to incoherent FFLs (I1-FFL, I2-FFL$_{NOR}$, I2-FFL$_{XNOR}$, I3-FFL, I4-FFL$_{AND}$) and one to C1-FFL$_{XOR}$. In architectures I1-FFL and I3-FFL repression dominated over activation. Certainly, the two regulatory branches with opposite sign are responsible for such a behavior, and the combinatorial promoter of the downstream promoter is central to get a variety of functionally analogous circuits. Interestingly, some of those architectures have been found involved in developmental processes [60, 61]. Notably, we did not exhaustively construct all possible FFL circuits from the library for their scoring. Instead, we probabilistically sampled the fitness landscape and we always found a solution corresponding to one of the six FFL structures presented. Moreover, this approach can be applied to design functional circuits without accounting for the designability of the desired behavior, and then study the intrinsic properties of the circuit irrespective to the library, such as its asymptotic robustness.

We then investigated the asymptotic robustness of those FFL circuits, which functioned as amplitude filters with a fold-change

(F) of at least one order of magnitude at the detection point (F). By constraining the sign of the regulations (fixed topology), we obtained a parameter space of $\sim 2 \cdot 10^5$ different combinations for each topology. Accordingly, the highest asymptotic robustness was reached by the architecture I3-FFL with the 21.29%, followed by the architectures C1-FFL$_{XOR}$ with the 17.17% and I4-FFL$_{AND}$ with the 17.66%, indicating that those circuits have a one-stripe pattern-prone structure. Interestingly, this could be because the input gene has a non-monochromatic regulatory mode (*i.e.*, both activator and repressor) in these topologies. On the contrary, the architecture I1-FFL$_{NOR}$ was highly sensitive to parameter variations with asymptotic robustness of only the 2.54%. The architectures I1-FFL and I2-FFL$_{XNOR}$ with the 8.60% and 7.69%, respectively, were in between. However, despite of its low asymptotic robustness, the structural core I2-FFL$_{NOR}$ is broadly found in many natural systems. For instance, in the *Drosophila* patterning circuitry, gene *hb* represses both genes *kni* and *Kr* and *kni* also represses *Kr* [60]. In addition, the core I1-FFL is the motif most abundant within the regulatory map of bacteria and yeast [62]. That the genes involved in the structures I1-FFL and I2-FFL have a monochromatic regulatory mode could explain the increasing presence of these circuits. Moreover, from a synthetic perspective, promoters type NOR and IMPLIES could be engineered by placing contiguously the corresponding operators in the promoter region [7, 12, 36].

Subsequently, we used the optimization method to design a circuit able to count. This is an interesting example that could already unveil many of the issues we meet in more complex networks. Cells may take advantage of this sort of circuits to regulate fundamental processes, such us telomere length control [63], where a machinery to count molecules or events is required. Counters have different stable states and rely on memory-like architectures that allow retaining the initial state, unless a perturbation switches the system [3, 14, 18, 64]. Generally, the underlying mechanism of biological counters consists in overcome certain threshold after a specific number of consecutive pulse-like events. Herein, we attempted the design of a two-pulse counter, where we imposed that the system had to reach three different states. We applied the optimization method to design all possible two-gene circuits. We found that, within a delimited time domain, all possible circuits were functional and reached three states. However, those circuits based on an activator-repressor core had a meta-stable state, which falls into the basin of attraction of one of the two stable states after certain time. In Fig. 2.2, we show the phase diagrams for the circuits based on a monochromatic regulatory core and one self-activation. The addition of another self-activation on the buffer gene allows having a symmetric multistable device [64]. We then computed the asymptotic robustness of these two circuits, by exploring exhaustively all possible parameterizations (for that we discretized the parameter space into $\sim 2 \cdot 10^5$ different combinations). The double repression core allowed tristability in the 0.2422% of the cases, whereas the double activation core in the 0.1787% (relatively low in both cases).

Next, we attempted the design of a two-pulse counter relying on just two states. As design specifications, we imposed pulses of 10 minutes within an interval of 50 minutes with amplitude of 100-fold. The designed circuit presented a topology like in Fig. 2.2A. However, the functioning of this system (*i.e.*, number of pulses it is able to count) depended on the pulse length and interval. In addition, we attempted the automated design of a tunable genetic timer. These devices consist of memories that change the state of operation according to an external signal and the time to accomplish this transition (time to reach the steady state) can be modulated by another signal [19]. The designed circuit consisted in a coherent FFL coupled to a memory-like mechanism based on a self-activation, and it existed a threshold in the input concentration from which the circuit responded to different levels of it.

Once a functional genetic device is obtained, either from computational or rational design methods, and experimentally validated, it can be integrated in the library as a new element to be used in the construction of more complex systems. As a first approach, we included in our library of regulatory elements a circuit previously optimized to operate as a tristable. Remarkably, the incorporation into the design procedure of black-box modules enhances the optimization of the impedance matching, where the output of a device serves directly as the input of a downstream one, and could considerably enlarge the functional diversity of the library. Hence, following such a design approach, we were able to obtain complex functions with modular systems. In our particular case, we designed a system coupling a tristable, an amplitude filter, and a frequency-tunable oscillator (Fig. 2.3). Initially, this tristable gave a low concentration. After a pulse of 20 minutes with amplitude of 1000-fold in the inducer, the device switched its state to reach the intermediate concentration

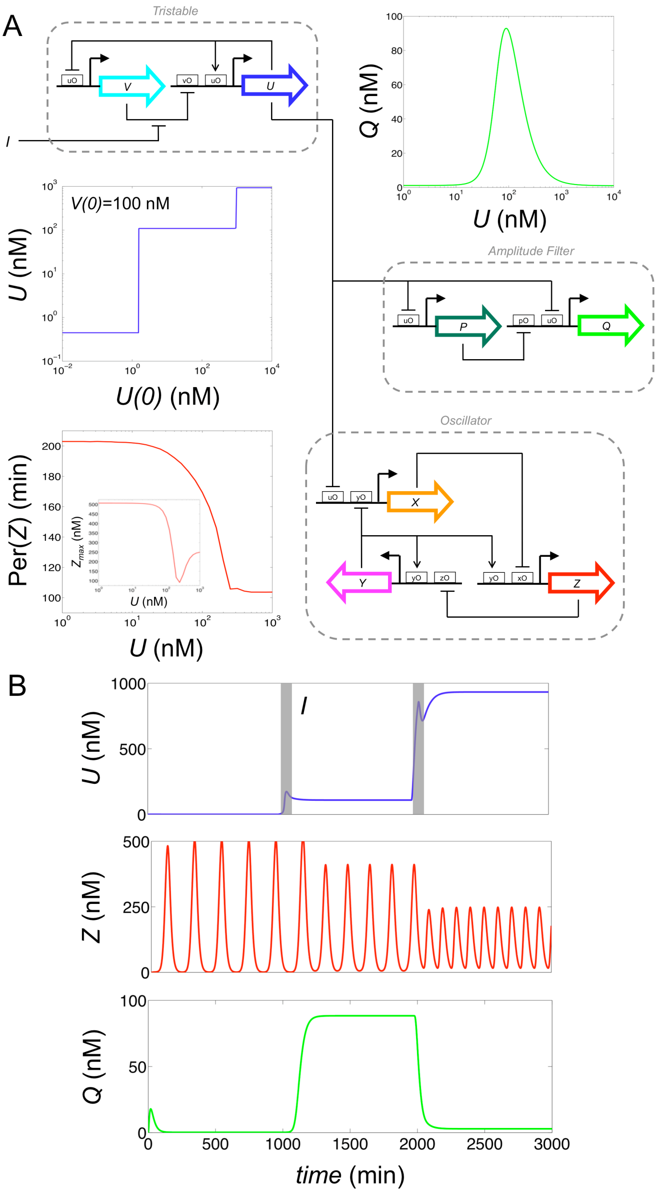


Figure 2.3: (A) Scheme of a complex regulatory system comprising a frequency-tunable oscillator and a state detector, designed by using the tristable device as an element of the library. Moreover, we show the transfer functions of the different devices that form the system. (B) Dynamics of the output genes of the complex system. Pulses in the input (I) of 20 min and 1000-fold of amplitude were applied at $t = 1000$ and $t = 2000$ min.

level, and subsequently the oscillator changed its frequency and the amplitude filter, which operates as a detector of the intermediate state, reached its ON state. After a second pulse, the device switched to its high concentration point, inducing a new change in the frequency of the oscillator and giving the detector back to its OFF state. Interestingly, the frequency-tunable oscillator evolved to couple two different regulatory mechanisms, and the external signal switched from one mechanism to another, then changing the frequency of the oscillations. In addition, with the consideration of delayed reactions (*e.g.*, due to translation and multimerization) we could obtain complex oscillations, which can drive to a route towards chaos [65]. In fact, this mechanism has been previously applied to design genetic oscillators with a minimal number of elements [16].

However, one important issue in such an approach is the possibility of the loss of function of a device when plugging it to a downstream module. This effect, usually called retroactivity [66], emerges when a transcription factor plays two different roles in both modules, and is indeed a consequence of the limited protein amount. This result may have significant consequences on the dynamics of the system, even when the stochasticity of the cell is taken into account [67]. Here, our modeling neglects this effect by assuming that the concentration of free protein is always much higher than the protein bound to DNA [22]; also as an imposition to ensure modularity in the design and to be able to combine different elements from the library. Although for many systems this approach is valid [9], it could be found some examples where such a model is not too accurate. To solve this problem in practice, one strategy would be to impose as a design constraint that the output gene had no regulatory effects on the circuit. Likewise, this output could be used as the input in further downstream modules with increased guarantees of a proper functioning. Thereby, in the system shown in Fig. 2.3, gene U could be split into two genes, one for working within the tristable device and another for setting the amplitude filter and the oscillator, although still it would exist a coupling between these two devices due to a common input.

2.4 Functional diversity and designability

We further studied the designability of a given dynamical behavior. For that, we constructed a library of SBML models of well-characterized regulatory elements previously implemented *in*

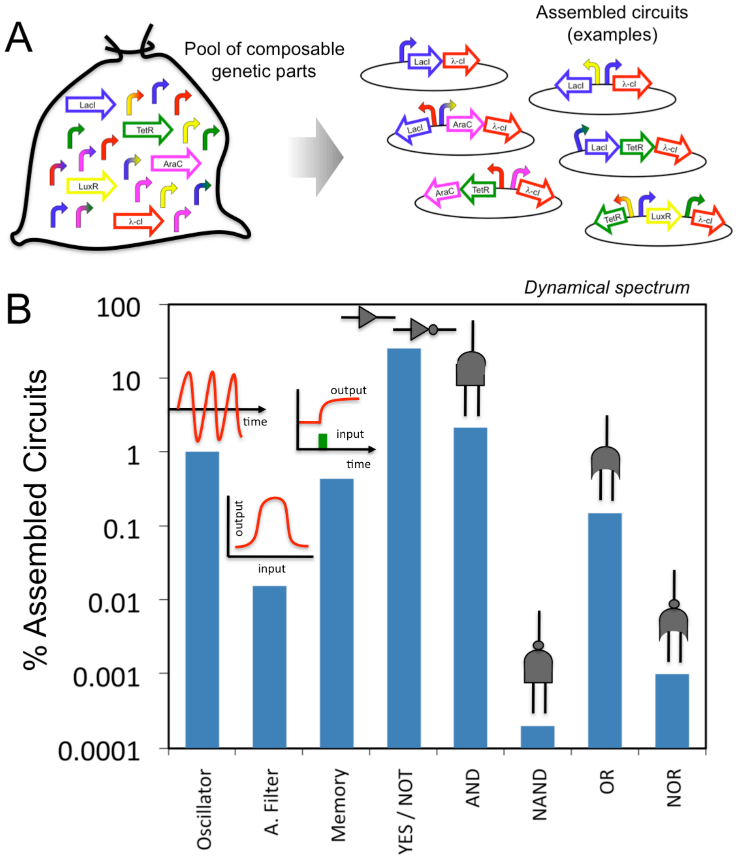


Figure 2.4: (A) Graphical representation of the exhaustive design strategy. Starting from a library of composable genetic regulatory elements (mathematical models in SBML format), we constructed all possible circuits up to three genes for simulation. (B) Dynamical spectrum of the library by exhaustive exploration (functional diversity). We represent the percentage of circuits that behave as oscillators, amplitude filters, memories, and logic gates (designability). To differentiate between two states of a circuit, we imposed at least one order of magnitude in concentration.

vivo. Likewise, the corresponding kinetic parameters were fitted from experimental data and kept fixed. Using this library, we constructed by *in silico* assembly all possible architectures up to three genes, giving 501,952 different circuits (see the different configurations in Fig. 2.4A). We systematically imposed λ-cI as output gene in all circuits. Thereby, we computed the dynamics of all circuits to perform an analysis of the behaviors that could be obtained with such a library. For this work, we considered a library of 36 elements, involving 5 genes and 31 synthetic promoters. As genes, we contemplated the classical repressors LacI, TetR and λ-cI, and the activators AraC and LuxR. Moreover, we built a library with 3 constitutive promoters with different transcription rates, 16 single promoters involving 4 lacO, 4 tetO, 2 araO, 2 luxO [36], 2 λ_RO [11], and 2 λ_{RM}O [5], and 12 combinatorial promoters involving 4 lacO-tetO, 3 araO-lacO, 2 araO-tetO [36], 1 λ_{RM}O-lacO [12], 1 luxO-λ_RO [7], 1 luxO-lacO [68]. The models also accounted for the external molecules (IPTG, anhydrotetracycline –aTc–, L(+)-arabinose, and acyl homoserine lactone –AHL–) that modify the regulatory ability of the transcription factors and represent the inputs of the circuits. Then, for each external inducer we considered three different states (low, intermediate, and high), giving 81 environmental conditions for all combinations, and four more conditions in which the inducers had a pulse-like dynamics.

By compiling all numerical results, we were able to dissect the dynamical spectrum of the library (*i.e.*, its functional diversity), which included circuits operating as oscillators, amplitude filters, memories, and different logic gates (Fig. 2.4B). As expected, the majority of the circuits functioned as logic gates, and because the external signals (IPTG, aTc, arabinose and AHL) always activated transcription, the set of NAND and NOR gates was highly reduced. In addition, approximately the 1% of the assembled circuits was able to exhibit oscillations. Furthermore, we found amplitude filters in the 0.016% of the cases, and memories in the 0.436%. Certainly, this spectrum depends on the value of F specified to differentiate between two concentration levels (F gives their ratio). Herein, we imposed $F > 10$, although we also performed a screening to see the effect of different values of F. Not surprisingly, as higher is F, the number of functional circuits decreases. In addition, we studied the effect of the initial condition on the output gene finding that the results were almost independent of this. Certainly, the initial condition only affects in memory-like circuits, but this effect was captured by imposing pulse-like dynamics on the input. Interestingly, the repertoire of designed circuits was essentially based on minimal cores that provided the required functional mechanism (Fig. 2.6). These cores illustrate the design principles in which the dynamical spectrum is based on. However, the use of a limited library and a partial set of input conditions, while allowing an exhaustive exploration,

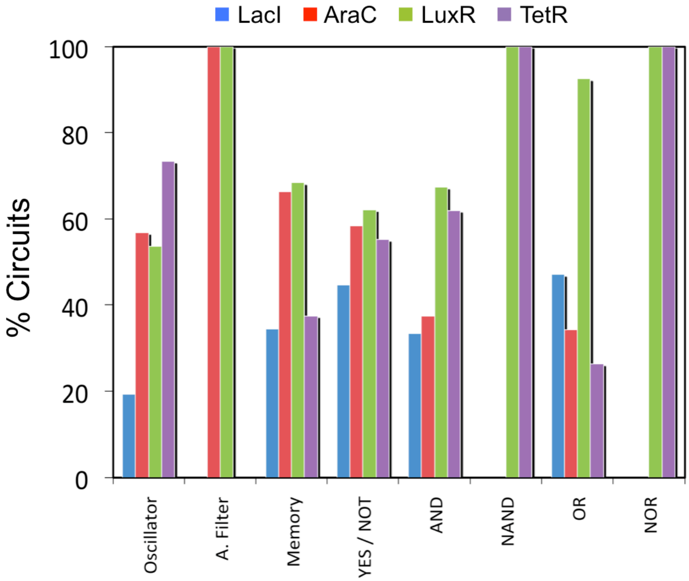


Figure 2.5: Sensitivity analysis of the dynamical spectrum. We release one regulatory element of the library (in particular, one gene) to analyze its contribution to the dynamical spectrum (we represent the remaining number of functional circuits relative to the total).

prevent obtaining a comprehensive analysis of the design principles. For instance, as we have shown above, a double activation core gives a memory-like mechanism but it was not found in the repertoire of circuits. In addition, all amplitude filters were based on the I2-FFL$_{NOR}$ architecture, although further circuits, not necessarily FFLs, can be employed to read morphogen gradients [59]. We did not obtain further topologies because the monochromatic regulatory mode and the lack of cooperation between transcription factors.

Furthermore, we investigated the dependence of the designability of a function on the existing elements of the library by calculating the degree of sensitivity of each regulator over the resulting dynamical spectrum (Fig. 2.5). Accordingly, LacI appeared to be the most important regulator, indeed for this particular case of study, since it participated in the majority of the functional circuits. In the specific case of the amplitude filters, since their mechanism relied on two different repressions (Fig. 2.6), LacI and TetR participated in all circuits. Certainly, the addition of more regulatory elements in the library would enlarge the designability of the different behaviors, and the identification of the regulatory cores in Fig. 2.6 would lead to rationally decide on the elements of more interest. In addition, we studied whether the designability could be estimated by sampling a small subset of assembled circuits instead of an exhaustive exploration. This would provide further support to the heuristic exploration by means of optimization methods. Interestingly, we found similar results for the dynamical spectrum of the library when analyzing the dynamics of about 1000 circuits (corresponding to the 0.2% of the total circuits). This suggests that even a small fraction of assembled circuits is representative of the whole population of circuits. By exploiting this fact, we could analyze the functional diversity and designability of several libraries of models at a minimal computational cost or we could study how to enrich the library with new regulatory elements.

We further studied the designability of the different behaviors when considering the stochasticity inherent to the cellular processes (we focused on intrinsic noise) [69]. Since the stochastic simulation entails a higher computational cost, we considered a subset of circuits as described above to perform this study, and because it is expected this will not strongly affect the results. For each condition of inputs, we considered the average value and standard deviation of the output (computed using the time dynamics after a transient period). In general, we found similar results as in the deterministic regime (Fig. 2.7). We could explain this by the fact that in most cases gene expression is high enough, which minimizes the effect of intrinsic noise, although in some cases a particular circuit topology could also help in such noise reduction [59]. However, we found an increase of almost a doubling in the number of oscillators. By examining the circuits, we realized that circuits based on an activation-repression mechanism with fast damped oscillations in the deterministic regime and that were identified as stable circuits were then selected as noise-induced oscillators. For the other behaviors, the designability results in the stochastic regime were slightly lower. The maximal reduction in designability was of about the 20% in the case of YES/NOT gates. In the circuits that were selected according to the deterministic solution but not to the stochastic one, there is an increase of noise in protein expression that prevents identifying different states of operation.

Afterwards, we wondered whether a unique circuit could exhibit different behaviors. Interestingly, we observed special circuits that

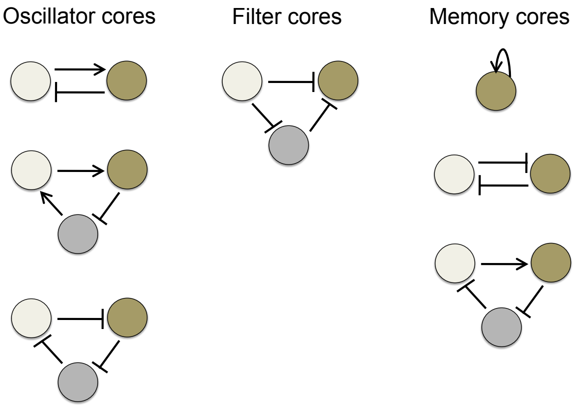


Figure 2.6: Genetic cores that define the design space of functional circuits provided the library of composable parts (promoters and coding regions) shown in Fig. 2.4.

displayed multifunctionality according to different input conditions (*e.g.*, oscillators working as amplitude filters, memories or logic gates). For instance, the 0.3% of the total set of circuits functioning as oscillators and memories held the two functions by properly setting the environmental factors (statistical significance assessed by bootstrapping). In addition, we calculated the number of circuits with multifunctionality showing a tendency log-normal in the distribution. This sort of circuits is appealing for cellular regulation and organization, because the rewiring required to change the function of the circuit is accomplished by means of external signals without genetic modification, likely as an on-the-fly reprogramming sentence [70]. As well as a single gene can attain several functions (*e.g.*, a protein with different enzymatic properties [71]), a multifunctional circuit can be exploited by the cell to exert a conditional control of different responses.

2.5 Discussion

In this chapter, we have tackled the problem of the designability of a given gene dynamics provided a library of composable regulatory elements, considering that the functional circuits come from combining different elements of the library. This measure of designability quantifies the entropy of a given dynamical behavior (number of possible states in Boltzmann usage). For that, we have developed a computational methodology that allows exploring the diversity of behaviors that can be obtained by assembling circuits by means of two different design strategies: one based on heuristic optimization and other based on exhaustive simulation of circuits. We have taken advantage of current characterizations of regulatory elements into libraries of mathematical models [26–29], allowing to rapidly select the regulatory element of interest for our circuit. Although the emergence of unexpected behaviors is always an issue in Synthetic Biology, it is anticipated that the use of standardized parts allows reducing the endless tweaking process when engineering a synthetic gene circuit [12, 19]. Using a proper mathematical formulation, we were able to generate a large collection of genetic circuits by assembling those regulatory elements, and identify the functional subset according to certain specifications. Initially, we constructed an artificial library of models to design circuits by optimization towards a configuration satisfying the specifications. We designed filters and counters of gene expression, which allowed us to find new regulatory mechanisms able to provide such behaviors. Sometimes the behavior requires a very precise genotype, making unlikely to get many cells with such behavior in a heterogeneous population. To investigate this, we have defined the concept of asymptotic robustness, which provides a measure of the maximum genotypic heterogeneity for a given of phenotypic behavior. In the long term, it is expected that Synthetic Biology projects will provide many examples of standardized circuits with a given dynamics, which could be incorporated into the available libraries. Then, one could extend our analysis to such cases. One issue here would involve the interfacing of such modules. To analyze this, we exploited one of our designed circuits as a single element of the library to obtain a complex system involving such a functional unit, illustrating a hierarchical design approach and allowing the design of plug-and-play devices with optimal impedance matching.

Given a library of regulatory elements, it is possible to construct many circuits with various dynamical behaviors. But some behaviors occur more often than others. To quantitatively analyze this, we computed the designability of a set of useful behaviors. There, we constructed a more reduced library of regulatory elements to assemble all possible circuits up to three genes and process their dynamics. Remarkably, the library involved promoters that had been previously characterized and even used for engineering various

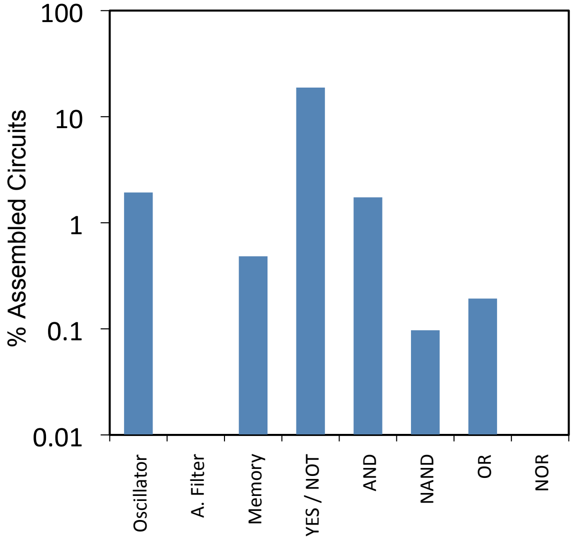


Figure 2.7: Dynamical spectrum of the library by exhaustive exploration of one sampling of assembled circuits (about the 0.2% of the circuits from Fig. 2.4) using stochastic simulation. We represent the percentage of circuits that behave as oscillators, amplitude filters, memories, and logic gates (designability). To differentiate between two states of a circuit, we imposed at least one order of magnitude in concentration and avoidance of overlapping in concentration due to the intrinsic noise.

synthetic circuits in the bacterium *E. coli.* Interestingly, we found that a limited library could encode a large number of behaviors. Certainly, our computational method allowed constructing and simulating the dynamics of this large set of circuits and assisted to dissect the spectrum of dynamical behaviors and study their designability. We found that a same genotype could have several functions depending on the external signals. Nevertheless, as the size of the circuits and the number of elements of the library increases, the exhaustive design strategy becomes unpractical, thus requiring heuristic methods. Since noise is an important factor that affects the dynamics of a circuit, we also included it in our analysis. The consideration of intrinsic noise slightly reduces the designability of digital circuits, but it increases the designability of oscillators. This is understandable from the fact that digital devices are steady-state based circuits, where noise could only spoil the behavior. On the other hand, oscillatory circuits are dynamical systems, where the noise could contribute to enhance the behavior. In addition, we expect that our results would be maintained when the library is enhanced by incorporating more accurate experimental measurements of the transcription regulation elements. That the new models could be more elaborated and accurate, they would not much change the fitness landscape and thus the designability of behaviors.

Interestingly, one possible extension to our work would be the development of a more complex, hierarchically distributed design platform [26]. Herein, more diverse, characterized regulatory elements would be considered, involving transcriptional, riboregulatory, metabolic and signaling elements. These different regulatory elements would be combined to yield complex functional genetic circuits, involving different regulatory mechanisms. In addition to new elements, inherent effects such as the variation of cell growth rate due to different culture media, the delay in the biochemical reactions and the parameter uncertainty of the models are important questions that would be explored. Furthermore, the design of circuits could be combined with tools for the design of synthetic DNA sequences. This would exploit the interactions between nucleic acids and the reengineering of natural proteins. Promoters with targeted transcription rates or multiple operators [35, 36], small RNAs with targeted secondary structures [6], or chimeric proteins acting as new transcription factors [73] are examples of what we could design computationally. All these elements would be modeled by transfer functions and these would be stored in a library. Importantly, it could be also specified a degree of evolvability, by which the value of the kinetic parameters characterizing that element would be susceptible to be changed after specific mutations in its sequence. Finally, the cellular chassis in which the circuit is going to be deployed could be also introduced as a generalized element by modeling the host elements that require the circuit for its expression [74]. This would allow to provide a prediction of the response of the engineered cell under the conditions for which the circuit was designed, and consequently improve the design process.

The following publication holds the contents presented in this chapter

- Rodrigo G, Carrera J, Jaramillo A (2011) Computational design of synthetic regulatory networks from a genetic library to characterize the designability of dynamical behaviors. *Nucl Acids Res*, 39: e138.

Further reading

- Rodrigo G, Carrera J, Jaramillo A (2007) Genetdes: automatic design of transcriptional networks. *Bioinformatics*, 23: 1857-1858.
- Rodrigo G, Carrera J, Elena SF (2010) Network design meets in silico evolutionary biology. *Biochimie*, 92: 746-752.

Bibliography

[1] Andrianantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol*, 2: 2006.0028.

[2] Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, 403: 335-338.

[3] Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403: 339-342.

[4] Atkinson MR, Savageau MA, Myers JT, Ninfa AJ (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell*, 113: 597-607.

[5] Isaacs FJ, Hasty J, Cantor CR, Collins JJ (2003) Prediction and measurement of an autoregulatory genetic module. *Proc Natl Acad Sci USA*, 100: 7714-7719.

[6] Isaacs FJ, Dwyer DJ, Ding C, Pervouchine DD, Cantor CR, Collins JJ (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nat Biotechnol*, 22: 841-847.

[7] Basu S, Mehreja R, Thiberge S, Chen M, Weiss R (2004) Spatiotemporal control of gene expression with pulse-generating networks. *Proc Natl Acad Sci USA*, 101: 6355-6360.

[8] Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R (2005) A synthetic multicellular system for programmed pattern formation. *Nature*, 434: 1130-1134.

[9] Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, Cantor CR, Collins JJ (2004) Programmable cells: interfacing natural and engineered gene networks. *Proc Natl Acad Sci USA*, 101: 8414-8419.

[10] Levskaya A, Chevalier AA, Tabor JJ, Simpson ZB, Lavery LA, Levy M, Davidson EA, Scouras A, Ellington AD, Marcotte EM, Voigt CA (2005) Synthetic biology: engineering *Escherichia coli*

to see light. *Nature*, 438: 441-442.

[11] Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. *Science*, 307: 1962-1965.

[12] Guido NJ, Wang X, Adalsteinsson D, McMillen D, Hasty J, Cantor CR, Elston TC, Collins JJ (2006) A bottom-up approach to gene regulation. *Nature*, 439: 856-860.

[13] Anderson J, Voigt C, Arkin A. (2007) Environmental signal integration by a modular AND gate. *Mol Syst Biol*, 3: 133.

[14] Deans TL, Cantor CR, Collins JJ (2007) A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells. *Cell*, 130: 363-372.

[15] Balagaddé FK, Song H, Ozaki J, Collins CH, Barnet M, Arnold FH, Quake SR, You L (2008) A synthetic *Escherichia coli* predator-prey ecosystem. *Mol Syst Biol*, 4: 187.

[16] Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J (2008) A fast, robust and tunable synthetic gene oscillator. *Nature*, 456: 516-519.

[17] Tigges M, Marquez-Lago TT, Stelling J, Fussenegger M (2009) A tunable synthetic mammalian oscillator. *Nature*, 457: 309-312.

[18] Friedland AE, Lu TK, Wang X, Shi D, Church G, Collins JJ (2009) Synthetic gene networks that count. *Science*, 324: 1199-1202.

[19] Ellis T, Wang X, Collins JJ (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat Biotechnol*, 27: 465-471.

[20] Cagatay T, Turcotte M, Elowitz MB, Garcia-Ojalvo J, Suel GM (2009) Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139: 512-522.

[21] deJong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9: 67-103.

[22] Bintu L, Buchler NE, Garcia H, Gerland U, Hwa T, Kondev J, Philips R (2005) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15: 116-124.

[23] Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genet*, 5: 34-42.

[24] Hasty J, McMillen D, Collins JJ (2002) Engineered gene circuits. *Nature*, 420: 224-230.

[25] Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. *Proc Natl Acad Sci USA*, 99: 16587-16591.

[26] Endy D (2005) Foundations for engineering biology. *Nature*, 438:

449-453.

[27] Canton B, Labno A, Endy D (2008) Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol*, 26: 787-793.

[28] Voigt CA (2006) Genetic parts to program bacteria. *Curr Opin Biotechnol*, 17: 548-557.

[29] Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, de Mora K, Glieberman AL, Monie DD, Endy D (2009) Measuring the activity of BioBrick promoters using an in vivo reference standard. *J Biol Eng*, 3: 4.

[30] Guet CC, Elowitz MB, Hsing W, Leibler S (2002) Combinatorial synthesis of genetic networks. *Science*, 296: 1466-1470.

[31] Dubendorff JW, Studier FW (1991) Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor. *J Mol Biol*, 219: 45-59.

[32] Edelman GM, Meech R, Owens GC, Jones FS (2000) Synthetic promoter elements obtained by nucleotide sequence variation and selection for activity. *Proc Natl Acad Sci USA*, 97: 3038-3043.

[33] Imburgio D, Rong M, Ma K, McAllister WT (2000) Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. *Biochemistry*, 39: 10419-10430.

[34] Mey M, Maertens J, Lequeux GJ, Soetaert WK, Vandamme E (2007) Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering. *BMC Biotechnol*, 7: 34.

[35] Murphy KF, Balazsi G, Collins JJ (2007) Combinatorial promoter design for engineering noisy gene expression. *Proc Natl Acad Sci USA*, 104: 12726-12731.

[36] Cox RS III, Surette MG, Elowitz MB (2007) Programming gene expression with combinatorial promoters. *Mol Syst Biol*, 3: 145.

[37] Beisel CL, Bayer TS, Hoff KG, Smolke CD (2008) Model-guided design of ligand-regulated RNAi for programmable control of gene expression. *Mol Syst Biol*, 4: 224.

[38] Che AJ, Knight TF Jr (2010) Engineering a family of synthetic splicing ribozymes. *Nucl Acids Res*, 38: 2748-2755.

[39] Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*, 27: 946-950.

[40] Isalan M, Klug A, Choo Y (2001) A rapid, generally applicable

method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nat Biotechnol*, 19: 656-660.

[41] Krueger M, Scholz O, Wisshak S, Hillen W (2007) Engineered Tet repressors with recognition specificity for the tetO-4C5G operator variant. *Gene*, 404: 93-100.

[42] Rodrigo G, Carrera J, Jaramillo A (2007) Asmparts: assembly of biological model parts. *Syst Synth Biol*, 1: 167-170.

[43] Marchisio MA, Stelling J (2008) Computational design of synthetic gene circuits with composable parts. *Bioinformatics*, 24: 1903-1910.

[44] Cai Y, Hartnett B, Gustafsson C, Peccoud J (2007) A syntactic model to design and verify synthetic genetic constructs derived from standard biological parts. *Bioinformatics*, 23: 2760-2767.

[45] Chandran D, Bergmann FT, Sauro HM (2009) TinkerCell: modular CAD tool for synthetic biology. *J Biol Eng*, 3: 19.

[46] Densmore D, Hsiau TH, Kittleson JT, DeLoache W, Batten C, Anderson JC (2010) Algorithms for automated DNA assembly. *Nucl Acids Res*, 38: 2607-2616.

[47] Cooling MT, Rouilly V, Misirli G, Lawson J, Yu T, Hallinan J, Wipat A (2010) Standard virtual biological parts: a repository of modular modeling components for synthetic biology. *Bioinformatics*, 26: 925-931.

[48] Francois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci USA*, 101: 580-585.

[49] Paladugu SR, Chickarmane V, Deckard A, Frumkin JP, McCormack M, Sauro HM (2006) In silico evolution of functional modules in biochemical networks. *IEE Proc Syst Biol*, 153: 223-235.

[50] Rodrigo G, Carrera J, Jaramillo A (2007) Genetdes: automatic design of transcriptional networks. *Bioinformatics*, 23: 1857-1858.

[51] Tagkopoulos I, Liu Y, Tavazoie S (2008) Predictive behavior within microbial genetic networks. *Science*, 320: 1313-1317.

[52] Dasika MS, Maranas CD (2008) OptCircuit: An optimization based method for computational design of genetic circuits. *BMC Syst Biol*, 2: 24.

[53] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.

Bioinformatics, 19: 524-531.

[54] Rodrigo G, Carrera J, Elena SF (2010) Network design meets in silico evolutionary biology. *Biochimie*, 92: 746-752.

[55] Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science*, 220: 671-680.

[56] Entus R, Aufderheide B, Herbert M, Sauro MH (2007) Design and implementation of three incoherent feed-forward motif based biological concentration sensors. *Syst Synth Biol*, 1: 119-128.

[57] Kaplan S, Bren A, Dekel E, Alon U (2008) The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol Syst Biol*, 4: 203.

[58] Kim D, Kwon YK, Cho KH (2008) The biphasic behavior of incoherent feed-forward loops in biomolecular regulatory networks. *Bioessays*, 30: 1204-1211.

[59] Cotterell J, Sharpe J (2010) An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol Syst Biol*, 6: 425.

[60] Ashe HL, Briscoe J (2006) The interpretation of morphogen gradients. *Development*, 133: 385-394.

[61] Reeves GT, Muratov CB, Schupbach T, Shvartsman SY (2006) Quantitative models of developmental pattern formation. *Dev Cell*, 11: 289-300.

[62] Mangan S, Alon U (2003) Structure and function of the feedforward loop network motif. *Proc Natl Acad Sci USA*, 100: 11980-11985.

[63] Marcand S, Gilson E, Shore D (1999) A protein-counting mechanism for telomere length regulation in yeast. *Science*, 275: 986-990.

[64] Guantes R, Poyatos JF (2008) Multistable decision switches for flexible control of epigenetic differentiation. *PLoS Comput Biol*, 4: e1000235.

[65] Mackey MC, Glass L (1977) Oscillation and chaos in physiological control systems. *Science*, 197: 287-289.

[66] Del Vecchio D, Ninfa AJ, Sontag ED (2008) Modular cell biology: retroactivity and insulation. *Mol Syst Biol*, 4: 161.

[67] Kim KH, Sauro HM (2011) Measuring retroactivity from noise in gene regulatory networks. *Biophys J*, 100: 1167-1177.

[68] Sayut DJ, Niu Y, Sun L (2009) Construction and enhancement of a minimal genetic and logic gate. *Appl Environ Microbiol*, 75: 637-642.

[69] Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6: 451-464.

[70] Segal ME, Frieder O (1993) On-the-fly program modification: systems for dynamic updating. *IEEE Software*, 10: 53-65.

[71] Stark GR (1977) Multifunctional proteins: one gene - more than one enzyme. *Trends Biochem Sci*, 2: 64-66.

[72] Wilkinson DJ (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet*, 10: 122-133.

[73] Hollis M, Valenzuela D, Pioli D, Wharton R, Ptashne M (1988) A repressor heterodimer binds to a chimeric operator. *Proc Natl Acad Sci USA*, 85: 5834-5838.

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

Chapter 3

Network design to identify robustness principles

Nature may reach the same result
in many ways.
– Nikola Tesla

A same biological function could be reached, in principle, using diverse designs. With a same regulatory mechanism, the difference between designs strives in the network architecture. Therefore, we should determine the functional consequences of alternative designs to predict which kind of network would be selected in a given context. In this chapter, we focus on a family of minimal gene regulatory structures able to provide spatial organization. Following a design approach, we obtained different networks satisfying the desired specifications, and we analyzed them to uncover a design principle to reach robustness.

3.1 Gradient-driven pattern formation

Complex organisms have evolved precise spatiotemporal control programs, by transducing the presence of signaling molecules to transcription factors, which lead to development and differentiation [1–3]. Within this framework, it is important to address the mechanisms by which cells are able to read a gradient of diffusing molecules (morphogens) to trigger the expression of genes that orchestrate spatial organization. The dissection of the minimal genetic architectures that control cell fate [4] will help to understand how a graded signal is transformed into a discrete sequence of states and how fluctuations are counteracted for a robust and precise development. In that way, the natural occurrence in *Drosophila melanogaster* embryos of different networks based on the FFL motif for reading morphogen gradients [2], together with the engineering in *Escherichia coli* of synthetic FFL circuits responding in a non-monotonic manner to a graded signal [5–7], suggests that this architecture is particularly suitable for pattern formation.

The FFL motif consists in a three-node network where the input regulates the output and a third element, which also regulates the output. FFLs are broadly found both in prokaryotes and eukaryotes and can be classified into eight different architectures depending on the sign of its regulations [8]. Notably, this particular structure has certain functionalities *per se*. Theoretical and experimental work on the incoherent FFL (I-FFL), mostly based on transcriptional regulations but also enzymatic reactions, has revealed its ability to work as an amplitude (concentration) filter [5–7, 9–11], to accelerate the output response [8, 12], for signal amplification and fold-change detection [13, 14], and to generate temporal pulses in response to a constant stimulus [8–10, 15, 16]. Interestingly, this last attribute can be interpreted in terms of adaptiveness, where after a transient behavior the system returns to the previous state, being the output steady state level independent of the input level [17–22].

In this chapter, we investigate, by dissecting the design space that contains all possible topological configurations (wiring) and kinetic parameter values, whether a single FFL circuit (a topology with certain parameterization) can accommodate both (*i*) the ability to read a gradient by means of an amplitude detection mechanism and (*ii*) the ability to achieve optimal adaptive response at high output levels. Certainly, the capacity for adaptive responses of living organisms (partial or absolute) is an intriguing question in biology, and previous work, mostly based on metabolic systems (bacterial chemotaxis), has pointed out that optimal adaptiveness is more a consequence of circuit topology than of the fine tuning of kinetic parameters [19–22]. Thus, although the different I-FFL configurations can yield *a priori* a palette of functionally analogous devices, they may display different robustness profiles against external perturbations (*i.e.*, structural discrimination of robustness).

3.2 Mathematical modeling

The FFL motif consists in three genes (x, y and z) and it can indeed appear as eight different architectures, four coherent and four incoherent, depending on the nature of the regulations [8]. In addition, we consider an external molecule (u) that modulates the active form of x (x^*) by post-translational inhibition. Our model parameterizes all these architectures following a Hill-like function formalism [23] and reads

$$\begin{aligned}
\frac{dx}{dt} &= \alpha - x, \; x^* = \frac{x}{1+u}, \\
\tau\frac{dy}{dt} &= \frac{\beta_0 + \beta_1 (x^*/\theta_0)^n}{1+(x^*/\theta_0)^n} - y, \\
\frac{dz}{dt} &= \frac{\gamma_0 + \gamma_1 (x^*/\theta_1)^n + \gamma_2 (y/\theta_2)^m + \gamma_3 \omega (x^*/\theta_1)^n (y/\theta_2)^m}{1 + (x^*/\theta_1)^n + (y/\theta_2)^m + \omega (x^*/\theta_1)^n (y/\theta_2)^m} - z,
\end{aligned} \tag{3.1}$$

where α is the synthesis rate of x (here $\alpha = 10^4$), β_0 and β_1 the synthesis rates of y from the unregulated and x-regulated promoter respectively, and γ_0, γ_1, γ_2, and γ_3 the synthesis rates of z from the unregulated, x-regulated, y-regulated and x, y-regulated promoter respectively. The regulatory coefficients (bindings protein-DNA) are θ_0, θ_1, and θ_2, and n, m are the Hill coefficients. Typically, the active form of a transcription factor to activate/repress the promoter consists of a dimmer, thus for simplicity we fix $n = m = 2$ otherwise specified, although it could be straightforward the exploration of higher order aggregations. The parameter ω accounts for the potential interaction in the promoter region of x and y, from competitive ($\omega \ll 1$) to cooperative binding ($\omega \gg 1$). For independent binding, $\omega = 1$. In addition, τ is a dimensionless parameter that accounts for the relative stability of the intermediate protein y (here $\tau = 10$), related to the transient behavior but not affecting the stationary value. In case of adaptation, this parameter, which can be viewed as a delay over the expression of y, controls the amplitude and duration of the transient response after which the system returns to the original state [8, 13]. This model could be enlarged to account for mRNA dynamics, although for FFL circuits this would not affect the steady state of the system, or slightly modified to account for post-transcriptional regulations, as miRNA-mediated FFLs are recurrently found in mammals [24]. For notation purposes, in steady state we have $y = g(u)$ and $z = f(u, y)$, being $x = \alpha$.

To quantitatively study the robustness of a circuit, we introduced the concept of susceptibility, that is, a measure that relates the change in the output (z) from a perturbation in the system (*i.e.*, a change in one variable of the model). Here, we considered two measures: the input susceptibility (H_u), which relates the output level to changes in the input, and the intrinsic susceptibility (H_k), which relates the output level to changes in the kinetic parameters of the model. We also introduced the geometric average output fold-change, $F_z = (\prod_i z_i/z_0)^{1/N}$ for N perturbations. Then, we indentified the input susceptibility according to $F_z = exp(H_u(F_u - 1))$, where the variable F_u denotes a change in the input of $u = F_u u_0$ or $u = u_0/F_u$. The fit was done by considering $F_u \in [1, 2]$. This definition of susceptibility turns out into $H_u = \frac{F_z - 1}{F_u - 1} = \frac{\partial ln(z)}{\partial ln(u)}$ (the logarithmic gain of the system) for small input perturbations. For the intrinsic susceptibility, we assumed that each parameter (k) was a Gaussian distributed random variable with mean its nominal value ($\langle k \rangle = k_0$) and standard deviation a percentage of it ($\Delta k = h_k k_0$). Then, we fit the intrinsic susceptibility to $F_z = exp(H_k h_k)$, with a range of variation of $h_k \in [0, 1]$.

The stochastic modeling was performed via Langevin formulation [25–28]. We assumed that noise in x is negligible due to its high synthesis rate. Therefore, noise in x^* comes from noise in the input (u), whose statistics are $\langle u(t) \rangle = u_0$ and $\langle u(0)u(t) \rangle = \nu u_0 exp(-|t|)$, where ν is the Fano factor. We assumed that the diffusion time is of the order of the half-life of protein x, which is assumed to be short-lived. For instance, the Bicoid protein diffuses about 0.3 mm^2/s in *D. melanogaster* embryos of about 100 mm^2 giving a diffusion time of about 5-6 min [29]. The stochastic model reads

$$\begin{aligned} \tau \frac{dy}{dt} &= g(u) - y + \xi_y(\sqrt{\tau} t)\sqrt{g(u) + y}, \\ \frac{dz}{dt} &= f(u, y) - z + \xi_z(t)\sqrt{f(u, y) + y}, \end{aligned} \tag{3.2}$$

where ξ_y and ξ_z are Wiener processes with statistics $\langle \xi_y(t) \rangle = \langle \xi_z(t) \rangle = 0$ and $\langle \xi_y(0)\xi_y(t) \rangle = \langle \xi_z(0)\xi_z(t) \rangle = \delta(t)$ (Dirac delta). Using perturbation theory (the mean field is deterministic and the perturbation amplitude only depends on the mean field) and Fourier analysis [25–28], it is straightforward to show that noise in the output reads $\eta_z^2 = \frac{1}{z_0} + c_1|\partial_u f(u_0, y_0)|^2 \frac{\nu u_0}{z_0^2} + c_2|\partial_y f(u_0, y_0)|^2 \frac{y_0}{z_0}$, where y_0 and z_0 are the stationary solutions at the state ON, being c_1 and c_2 two constants. By using the concept of susceptibility, with $H_u = \frac{\partial ln(z)}{\partial ln(u)} = \frac{u_0}{z_0}\partial_u f(u_0, y_0)$ and $H_y = \frac{y_0}{z_0}\partial_y f(u_0, y_0)$, we can write

$$\eta_z^2 = \frac{1}{z_0} + c_1 H_u^2 \frac{\nu}{u_0} + c_2 H_y^2 \frac{1}{y_0}. \tag{3.3}$$

3.3 Optimal FFL circuits for pattern formation

We aimed at designing FFL circuits able of generating one-stripe patterns. For that, we computationally explored the whole designing space (FFL architectures and kinetic parameters). Our mathematical model, simultaneously accounting for transcription and translation processes, contains ten parameters (β_0, β_1, γ_0, γ_1, γ_2, γ_3, θ_0, θ_1, θ_2, and ω) that define the design space. For an efficient exploration, and given that design space is vast for an exhaustive computation, we adopted a heuristic optimization-based approach [30, 31]. We simplified the spatial diffusion and focused our study on amplitude filtering systems where the output reaches a maximum at intermediate input levels. Analogous results could be obtained for inverse amplitude filters (existence of a minimum). The transfer function is in brief characterized by the input detection amplitude (or bandwidth) and by the output amplitude (ratio between the maximal and basal output concentrations). The shape of this function serves to classify the amplitude filters into those exhibiting precision, *i.e.*, the detection is accomplished at a very accurate position, and those being adaptive, *i.e.*, a wide detection range exists so the stationary output level is insensitive to variations in the input. Certainly, a reliable pattern requires perceptible output amplitude, at least one order of magnitude, to differentiate the two cell fates (ON/OFF). Here, we imposed the condition that the output amplitude must be 100-fold. Nevertheless, there is a clear tradeoff between the bandwidth and the output amplitude, in the sense that a given output amplitude constrains both the maximal and minimal bandwidths that the system can attain. Herein, we considered that the morphogen (the input) interacts at the genetic level by inhibiting post-translationally the regulatory ability of a sensory transcription factor [2]. Similar results can be obtained if the morphogen induces the degradation of that regulator (*e.g.*, proteasome-mediation) or activates it (*e.g.*, phosphorylation). In fact, such a regulatory mode is not very relevant because of the symmetry of the transfer function.

First, we sought for patterns with maximal precision (Fig. 3.1a). This entails a transfer function with a narrow bandwidth. In Fig. 3.1b, we show the histograms for the kinetic parameters that characterize all optimal solutions. Remarkably, these histograms are not dense, indicating that there are few optimal points. In fact, these histograms correspond to four solution modes, which are the four I-FFL architectures with a specific parameterization (Fig. 3.1c). We denote I1-FFL-P, I2-FFL-P, I3-FFL-P, and I4-FFL-P these four circuits (the P stands for optimized for precision). In Fig. 3.1d, we plot the transfer, $z(u)$, and sensitivity, $F_z(h_k)$, functions that characterize the behavior of each circuit. These circuits show no qualitative differences in the two functions, suggesting that the four architectures are equally good at precision. Indeed, these circuits rely on a mechanism based on a tradeoff between the two regulatory branches, which have opposite sign. At high input levels, both activation and repression branches are inactive (state OFF), and at low ones both branches are active, accomplishing the state OFF because repression is dominant. While, at intermediate input levels, the activation branch is active and the repression inactive (state ON). For circuits I1-FFL-P and I3-FFL-P, rAND means that the output gene is expressed in presence of the activator and absence of the repressor (also called IMPLIES). Interestingly, we found that the C1-FFL architecture with a combinatorial logic type XOR (*i.e.*, the activators inhibit each other) and a weak activation from the intermediary gene to the output is also a solution. With the exclusive logic, this is in fact an I1-FFL variant. In addition, circuit I2-FFL-P emerged with either a combinatorial logic type NOR (*i.e.*, the repressors act independently each other) or XNOR (*i.e.*, the repressors inhibit each other) and competitive binding. Moreover, circuit I4-FFL-P emerged with a combinatorial logic type AND (*i.e.*, both activators act synergistically) and independent binding. We did not obtained from the landscape exploration further combinatorial logics for circuits I2-FFL-P and I4-FFL-P, which suggests that such configurations would not be plausible because they would not introduce the required tradeoff between the opposite regulatory branches (this can be shown

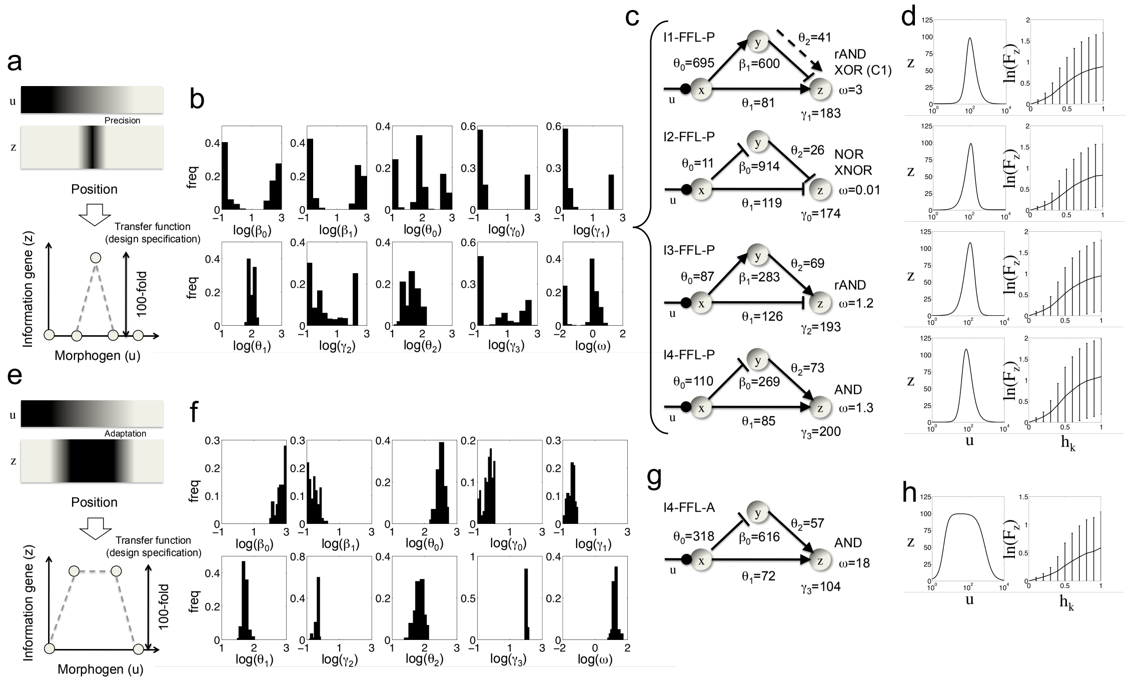


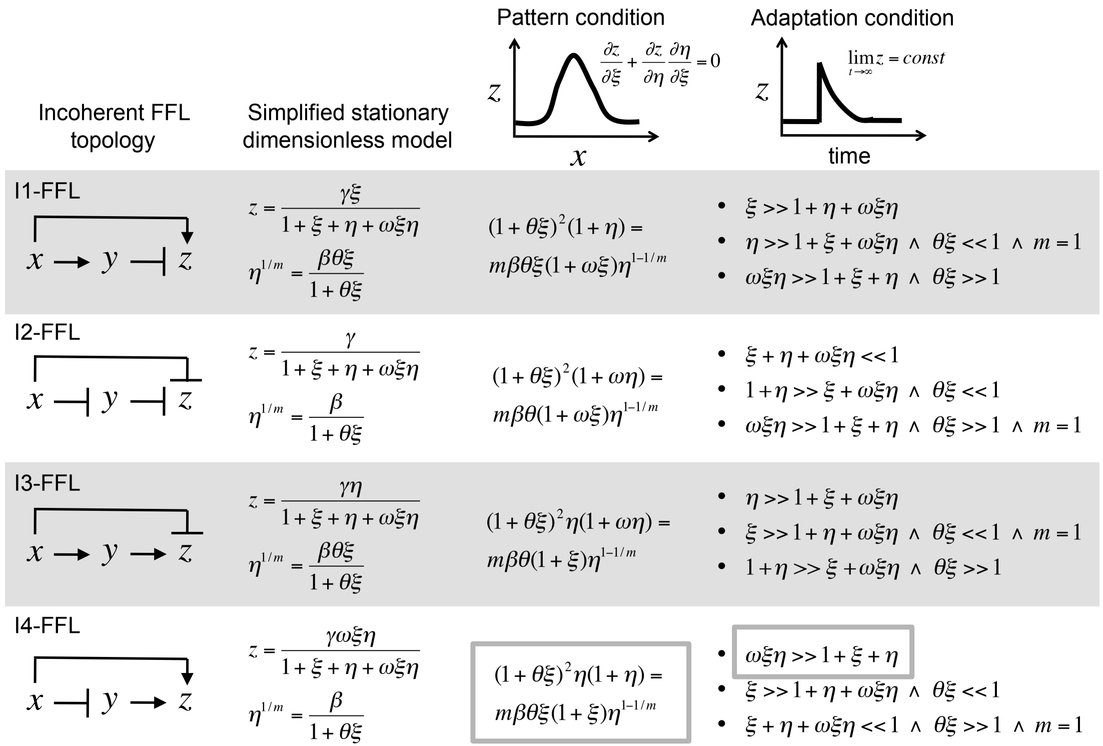
Figure 3.1: Landscape of FFLs for pattern formation. (a, e) A spatial gradient of an external molecule (morphogen) induces a particular cell fate depending on the position. The simplest pattern consists in a one-stripe composition with two cell fates, triggered by the expression level of one gene. The spatial information can be reduced to construct the transfer function of the system (relation output/input in steady state). (b, f) Histograms for the kinetic parameter values of the model resulting from multiple optimization runs with different initial guesses that explored the design space. The ordinates represents the value in logarithmic scale, while the abscises the frequency of each one produced by the heuristic procedure. (c, g) Incoherent FFL architectures together with the average value of the relevant kinetic parameters emerged from the landscape exploration. (d, h) Transfer (left) and sensitivity (right) functions characterizing each circuit. The sensitivity function is calculated at maximal output level.

mathematically).

Second, we sought for patterns with optimal adaptive response in the state ON (Fig. 3.1e). This entails a transfer function with a plateau, which gives definitively a wide bandwidth. In Fig. 3.1f, we show the histograms for the kinetic parameters that characterize all optimal solutions. Surprisingly, all kinetic parameters are highly constrained by the design specifications, which corresponds to just one solution mode, the I4-FFL architecture with a specific parameterization (Fig. 3.1g). We denote I4-FFL-A this circuit (here A stands for optimized for adaptation). In Fig. 3.1h, we plot its transfer and sensitivity functions. As it can be observed, this circuit has a wider bandwidth and presents a lower sensitivity to perturbations in the kinetic parameters at the state ON. The circuit emerged with a combinatorial logic type AND and cooperative binding, whose working principle also relies on the tradeoff between the two regulatory branches. On the light of these numerical results, it could be concluded that the optimal adaptive response (existence of a plateau) was structurally encoded by the I4-FFL topology and, in contrast to circuit I4-FFL-P, modulated by a strong binding cooperation (ω) between the two activators.

Motivated by the numerical results from the heuristic landscape exploration, we performed a theoretical analysis to elucidate the attribute that discriminates the I4-FFL as the central topology with adaptive performance in the state ON. On the one hand, mathematically, the one-stripe pattern condition implies that the output concentration reaches an optimum, which gives the equation $\partial_u f(u_0, y_0) + \partial_y f(u_0, y_0) \partial_u g(u_0) = 0$ where $f(u_0, y_0)$ is the production term of z and $g(u_0)$ of y in the steady state. Certainly, this can be satisfied in case of I-FFL circuits, where the sign of the direct regulatory branch (x to z) is opposite to that of the indirect branch (x to y to z). This condition only guarantees the presence of an optimum and not a reliable amplitude level. Together, the specification of a desired amplitude level (*e.g.*, 100-fold with respect to the basal state) entails a precise parameterization.

On the other hand, to achieve an absolute adaptive response the output concentration in steady state has to be input-independent regardless the values taken by the kinetic parameters. Only the transient behavior will be affected by such numerical values. For each topology three possibilities exist although for illustrative purposes we will focus on the I1-FFL topology. First, the system will show


z
x
$\frac{\partial z}{\partial \xi} + \frac{\partial z}{\partial \eta}\frac{\partial \eta}{\partial \xi} = 0$
z
time
$\lim_{t\to\infty} z = const$

Incoherent FFL topology	Simplified stationary dimensionless model	Pattern condition	Adaptation condition
I1-FFL $x \rightarrow y \dashv z$, $x \rightarrow z$	$z = \frac{\gamma\xi}{1+\xi+\eta+\omega\xi\eta}$ $\eta^{1/m} = \frac{\beta\theta\xi}{1+\theta\xi}$	$(1+\theta\xi)^2(1+\eta) = m\beta\theta\xi(1+\omega\xi)\eta^{1-1/m}$	• $\xi >> 1+\eta+\omega\xi\eta$ • $\eta >> 1+\xi+\omega\xi\eta \wedge \theta\xi << 1 \wedge m = 1$ • $\omega\xi\eta >> 1+\xi+\eta \wedge \theta\xi >> 1$
I2-FFL $x \dashv y \dashv z$, $x \dashv z$	$z = \frac{\gamma}{1+\xi+\eta+\omega\xi\eta}$ $\eta^{1/m} = \frac{\beta}{1+\theta\xi}$	$(1+\theta\xi)^2(1+\omega\eta) = m\beta\theta(1+\omega\xi)\eta^{1-1/m}$	• $\xi+\eta+\omega\xi\eta << 1$ • $1+\eta >> \xi+\omega\xi\eta \wedge \theta\xi << 1$ • $\omega\xi\eta >> 1+\xi+\eta \wedge \theta\xi >> 1 \wedge m = 1$
I3-FFL $x \rightarrow y \rightarrow z$, $x \dashv z$	$z = \frac{\gamma\eta}{1+\xi+\eta+\omega\xi\eta}$ $\eta^{1/m} = \frac{\beta\theta\xi}{1+\theta\xi}$	$(1+\theta\xi)^2\eta(1+\omega\eta) = m\beta\theta(1+\xi)\eta^{1-1/m}$	• $\eta >> 1+\xi+\omega\xi\eta$ • $\xi >> 1+\eta+\omega\xi\eta \wedge \theta\xi << 1 \wedge m = 1$ • $1+\eta >> \xi+\omega\xi\eta \wedge \theta\xi >> 1$
I4-FFL $x \dashv y \rightarrow z$, $x \rightarrow z$	$z = \frac{\gamma\omega\xi\eta}{1+\xi+\eta+\omega\xi\eta}$ $\eta^{1/m} = \frac{\beta}{1+\theta\xi}$	$\boxed{(1+\theta\xi)^2\eta(1+\eta) = m\beta\theta\xi(1+\xi)\eta^{1-1/m}}$	• $\boxed{\omega\xi\eta >> 1+\xi+\eta}$ • $\xi >> 1+\eta+\omega\xi\eta \wedge \theta\xi << 1$ • $\xi+\eta+\omega\xi\eta << 1 \wedge \theta\xi >> 1 \wedge m = 1$

Figure 3.2: Theoretical analysis of the four I-FFL topologies. The I2-FFL is assumed to have a combinatorial logic type NOR, and the I4-FFL a type AND. We considered a dimensionless model in steady state, where $\xi = (x^*/\theta_1)^n$ and $\eta = (y/\theta_2)^m$, and simplified it to only account for the higher synthesis rate. Moreover, θ, β and γ are dimensionless parameters. For each FFL topology, we mathematically derived the condition to achieve pattern formation (i.e., z must reach a maximum at intermediate levels of x^*) and adaptiveness (*i.e.*, z in steady state must be independent of x^*). For optimal adaptive response, there are three possible strategies that can be implemented with particular choices of kinetic parameters.

adaptiveness when $(x^*/\theta_1)^n$ is the dominant term in the denominator of $f(u, y)$, being θ_i a parameter for binding affinity. In this case, there is a strong activation of x^* that saturates the production of z, whereas the repression by y becomes negligible. Second, when the production of y is linear with x^* (*i.e.*, $x^* \ll \theta_0$ and $m = 1$) and $(y/\theta_2)^m$ is the dominant term in the denominator of $f(u, y)$. Now, since y is proportional to x^*, the activation of x^* on z is counteracted by the strong repression of y. Third, when the production of y saturates (*i.e.*, $x^* \gg \theta_0$) and the cooperative term $\omega(x^*/\theta_1)^n(y/\theta_2)^m$ dominates the denominator of $f(u, y)$. Analogous derivations can be done for the other I-FFL architectures.

Fig. 3.2 summarizes all pattern and adaptation conditions for the four I-FFL topologies. The optimality condition, together with a specific amplitude level, imposes a strict relation between some kinetic parameters of the model (mostly those binding-related) and the concentration values of the species. Nevertheless, only for the I4-FFL with a combinatorial logic type AND, that condition is independent of ω, which is free to adopt a given value. This fact is a direct consequence of the circuit topology and is instrumental to achieve adaptation at high output levels. By setting a high value of ω we can ensure the first adaptive condition for the I4-FFL circuit. In this case, the cooperative term $\omega(x^*/\theta_1)^n(y/\theta_2)^m$ dominates the denominator of $f(u, y)$, yielding a constant function, and hence the pattern condition is satisfied because $\partial_u f(u_0, y_0) = \partial_y f(u_0, y_0)$.

3.4 Robustness of FFLs: adaptiveness, parameter sensitivity and noise tolerance

We next explored the consequences of adaptiveness in the sense of congruent evolution to genetic robustness [32, 33]. For that, we calculated the susceptibility of the circuit under perturbations in the input level (H_u) and in the kinetic parameters of the model (H_k). We focused our study on circuits operating at the state ON. Here, to calculate the intrinsic susceptibility we just considered variations in the most important parameters, those related to the binding affinities between transcription factors and DNA [2]. Indeed, the amplitude detection mechanism exploits the differences in those binding affinities, and computational studies on the dorso-ventral gradient in *D. melanogaster* embryos have confirmed that these parameters mediate the major control on the expression of target genes [34]. Fig. 3.3 represents the four circuits optimized for precision (I1-FFL-P, I2-FFL-P, I3-FFL-P, and I4-FFL-P), the one optimized for adaptive response (I4-FFL-A), and four more suboptimal circuits (I1-FFL-S, C1-FFL-S, I2-FFL-S, and I3-FFL-S). Whereas I4-FFL-A achieves optimal adaptive response, it could be argued that the suboptimal circuits exhibit partial adaptation. In logarithmic scale, we show a strong correlation between the input and intrinsic susceptibilities. This fact suggests that the acquired ability of certain biological systems to be robust against mutations that change their kinetic properties is a direct consequence of their ability to respond to environmental perturbations (*i.e.*, environmental robustness). The I-FFL circuit by means of a tuned balance between the two regulatory branches allows counteracting by anticipation any perturbation in the input or in any element upstream the output. However, such a circuit cannot neutralize perturbations in the synthesis rate of the output gene. To do so, the circuit would need to introduce a negative feedback loop (N-FBL) implementing an integral control [18]. In fact, N-FBLs have been shown to provide robustness in transcription [35] and metabolic [22, 36] networks, and its combination with I-FFLs can enhance the robustness performance [22].

In addition to the susceptibility calculations, we carried out a stochastic analysis to study the robustness of the circuits against molecular noise [25–28]. We considered an intrinsic source of noise due to the low number of molecules together with a noisy input signal. We performed numerical simulations to calculate the noise level in the output gene at the state ON (Fig. 3.4) for different noise amplitudes in the input for the optimal circuits (I1-FFL-P, I2-FFL-P, I3-FFL-P, I4-FFL-P, and I4-FFL-A). Essentially, noise in gene expression can be decomposed into three terms, one intrinsic that is Poissonian for genes without self-regulation, another due to propagation effects, and a third extrinsic one accounting for sources common to all species [27]. In our case, we did not consider extrinsic noise, and the propagation term accounts for noise directly resulting from the input (N_u) and noise coming indirectly via the intermediary element (N_y). These terms are proportional to their susceptibilities. Then we can write the expression $\eta_z^2 = 1/z_0 + N_u + N_y$ for noise in the output. Circuits with similar transfer functions have similar susceptibilities, however noise tolerance is structure-dependent. Indeed, at the state ON, the concentration of the intermediary element is low for circuits I1-FFL-P and I2-FFL-P

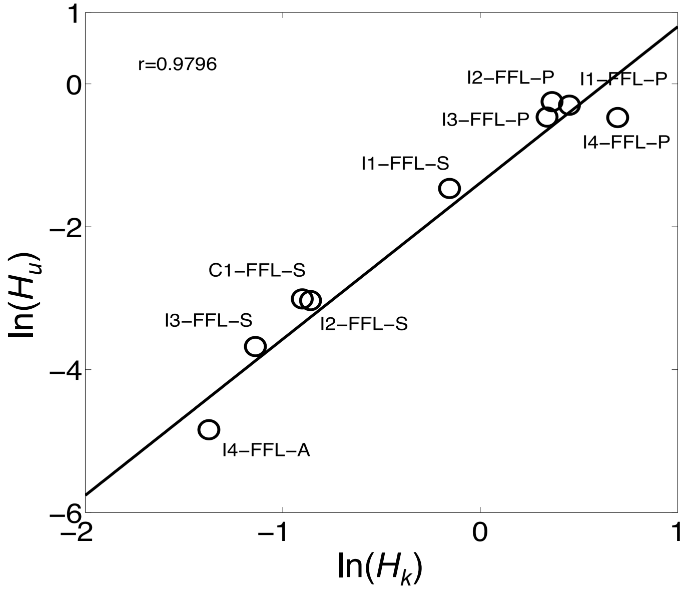


Figure 3.3: Adaptiveness versus parameter sensitivity. Correlation between the input and intrinsic susceptibilities (H_u and H_k respectively) in natural logarithmic scale, where each circle corresponds to one circuit. For this plot, to calculate H_k we considered the parameters θ_0, θ_1, θ_2, and ω. We represent the four circuits optimized for precision (I1-FFL-P, I2-FFL-P, I3-FFL-P and I4-FFL-P), the one optimized for adaptiveness (I4-FFL-A), and four more suboptimal circuits (I1-FFL-S, C1-FFL-S, I2-FFL-S and I3-FFL-S). The value of r corresponds to the linear correlation coefficient (solid line obtained by linear fit).

because this gene represses the output, whereas it is high for circuits I3-FFL-P and I4-FFL-P as in these cases it activates the output. This fact entails that the term N_y is higher for circuits I1-FFL-P and I2-FFL-P than for circuits I3-FFL-P and I4-FFL-P, since noise is inversely proportional to concentration. As we can observe, noise increases in circuits optimized for precision with randomly fluctuating input signals, whereas circuit I4-FFL-A is highly insensitive to such stochastic events, maintaining a constant Poissonian noise level ($\eta_z \simeq 1/z_0$). This can be rationalized knowing that $H_u \simeq H_y \simeq 0$ for this circuit. For high input fluctuations ($\nu = 4$), we have $N_u \gg N_y$ thereby precise circuits show similar noise levels.

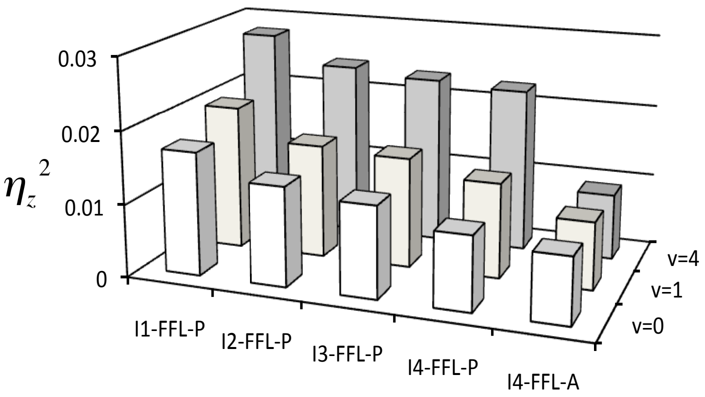


Figure 3.4: Noise tolerance for optimal designs. Noise in output expression (η_z^2) for different FFL circuits due to intrinsic effects and several noise levels at the input; ν represents the corresponding Fano factor.

3.5 Discussion

The knowledge of the dynamical properties of different fundamental regulatory networks is crucial to infer the selective pressures that the cell has suffered during its evolution. In fact, although the kinetic parameters are important to determine the dynamical behavior, a network topology by itself can determine or constrain the dynamics and provide structural sources of robustness [19–21] or noise tolerance [37]. Why a precise regulatory motif is prominent in Nature, whereas a functionally analogous circuit (same behavior but different topology) is less abundant or even not found, remains an intriguing question. Certainly, depending on the biological demands, a particular circuit will be more favorable for the cell. Regulatory networks based on I-FFLs can operate in a dosage-response manner to generate one-stripe spatial patterns. More complex (multiple-stripe) patterns can be obtained by interplaying several I-FFLs [9]. In fact, the segmentation network of *D. melanogaster* involves several cascades of genes that allow obtaining these banding patterns. For instance, while the gap genes form a one-stripe pattern, the downstream elements, such as the pair-rule or segmentation genes, give multiple-stripe patterns [38, 39]. Importantly, this supports the modular organization of the regulatory networks by which complex functions are reached by interconnecting small units. The four I-FFL architectures, with a proper parameterization, can operate with maximal precision having similar input and intrinsic susceptibilities. However, noise at the state ON is eventually higher for circuits I1-FFL-P and I2-FFL-P due to the monochromatic regulatory mode of the sensor, which leads to a repression exerted by the intermediary element. Remarkably, only the I4-FFL topology is able to provide adaptiveness at the state ON (while the four architectures can give an adaptive response at the state OFF).

In a recent work, Cotterell and Sharpe proposed different three-gene topologies, not necessarily FFLs, to produce one-stripe patterns [40]. Using a systematic design procedure, these authors found new structural elements for reading morphogen gradients and controlling developmental genetic units, some of which should still be discovered *in vivo*. Furthermore, the combination of these elements can enlarge the repertoire of circuit topologies and increase the level of robustness. However, unless bistable-like circuits, these topologies were essentially based on I-FFLs. Furthermore, some canonical functional topologies were mislaid, such as the I4-FFL, indeed because the search algorithm used by Cotterell and Sharpe did not account for synergistic actions (*e.g.*, promoters type AND). Herein, our design procedure has resulted more sensitive to study the transcriptional FFLs and has allowed us to refine such general approaches for a comprehensive study. Our model accounts for the intracellular circuit dynamics under certain level of an external signal and without tolerating the diffusion of proteins. In this sense, Cotterell and Sharpe illustrated that protein diffusion resulting in a cell-to-cell communication weakly affects noise tolerance but results into a mechanism that allows tuning the position and bandwidth of the stripe. Interestingly, diffusion affects the bandwidth differently depending on the circuit structure. Therefore, a logical further step concerning adaptiveness would be to study the addition of more regulations over single FFLs, the effect of diffusion and the signaling at the intermediary gene level to obtain a widespread analysis of the different genetic architectures that allow reading gradients and generate one-stripe patterns.

In addition, the I-FFL motif is also found in simple organisms that do not require the formation of spatial patterns (*e.g.*, bacteria or yeast). In this case, the filtering device normally operates at one state, and switch to the other state after environmental changes. According to the Savageau's demand principle [41], the mode of gene regulation should entail a maximization of the usage (binding to

DNA) of the transcription factors; otherwise, the regulators are lost during evolution. On the one hand, in circuits based on I1-FFL and I2-FFL topologies operating at the state ON only one regulator is functional, whereas in case of I3-FFL and I4-FFL topologies the state ON requires the function of the two regulators. This relates to the fact that in the I1-FFL and I2-FFL the sensor has a monochromatic regulatory mode, whereas for the I3-FFL and I4-FFL it acts as activator and repressor simultaneously. Hence, it would be expected that circuits operating at the state ON were preferentially based on I3-FFL and I4-FFL and were present within the regulatory map of highly demanded biological functions, such as central metabolism or transcription-translation machinery. On the other hand, only the I1-FFL entails the functionality of the two regulators at the state OFF but for low input levels, since we are considering that the input post-translationally inhibits the sensor. Then, circuits operating at the state OFF would be mostly based on the I1-FFL and would control genes of low demand (*e.g.*, secondary metabolism) or genes that need to be activated in specific situations such as stress responses or during development. Interestingly, I1-FFL architectures are the most abundant ones in bacteria and yeast [12], being reasonable that this abundance is a consequence of the specialization of the I-FFL to operate as time pulse generator and keep the expression of its target genes tightly suppressed in absence of external stimuli.

One open question that arises from our results is if given the properties of robustness associated to I4-FFLs, their abundance as regulatory module could be considered as an exaptation (an spandrel in S. J. Gould usage) that results from selection of larger and more complex network structures or, perhaps, as a direct consequence of selection for increased robustness [42]. In this second case, the consequent relevant question is how robustness mechanisms were selected for. If buffering mechanisms minimize the effect of every possible mutation, they will operate on the mutations created, thus making them invisible to natural selection and hence preventing their spread in the population. A possible solution to this paradox is that mutational robustness is a side effect of selection for mechanisms that buffer environmental perturbations [32, 33]. Our observation that when selection against adaptiveness was imposed, the optimal design I4-FFL-A was also robust against parameter perturbations (equivalent to mutational effects on catalytic/binding properties) gives further support to this possibility. Therefore, the recurrent inference of the design principles that confer adaptiveness to organisms would clarify our understanding of the causes of robustness to genetic perturbations and noise.

The following publication holds the contents presented in this chapter

- Rodrigo G, Elena SF (2011) Structural discrimination of robustness in transcriptional feedforward loops for pattern formation. *PLoS ONE*, 6: e16904.

Further reading

- Rodrigo G, Carrera J, Elena SF, Jaramillo A (2010) Robust dynamical pattern formation from a multifunctional minimal genetic circuit. *BMC Syst Biol*, 4: 48.

Bibliography

[1] Wolpert L (1969) Positional information and the spatial pattern of cellular differentiation. *J Theor Biol*, 25: 1-47.

[2] Ashe HL, Briscoe J (2006) The interpretation of morphogen gradients. *Development*, 133: 385-394.

[3] Reeves GT, Muratov CB, Schupbach T, Shvartsman SY (2006) Quantitative models of developmental pattern formation. *Dev Cell*, 11: 289-300.

[4] Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8: 450-461.

[5] Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R (2005) A synthetic multicellular system for programmed pattern formation. *Nature*, 434: 1130-1134.

[6] Entus R, Aufderheide B, Herbert M, Sauro MH (2007) Design and implementation of three incoherent feed-forward motif based biological concentration sensors. *Syst Synth Biol*, 1: 119-128.

[7] Sohka T, Heins RA, Phelan RM, Greisler JM, Townsend CA, *et al.* (2009) An externally tunable bacterial band-pass filter. *Proc Natl Acad Sci USA*, 106: 10135-10140.

[8] Mangan S, Alon U (2003) Structure and function of the feedforward loop network motif. *Proc Natl Acad Sci USA*, 100: 11980-11985.

[9] Ishihara S, Fujimoto K, Shibata T (2005) Cross talking of network motifs in gene regulation that generates temporal pulses and spatial stripes. *Genes Cells*, 10: 1025-1038.

[10] Kim D, Kwon YK, Cho KH (2008) The biphasic behavior of incoherent feed-forward loops in biomolecular regulatory networks. *Bioessays*, 30: 1204-1211.

[11] Kaplan S, Bren A, Dekel E, Alon U (2008) The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol Syst Biol*, 4: 203.

[12] Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feedforward loop accelerates the response-time of the gal system of *Escherichia coli*. *J Mol Biol*, 356: 1073-1081.

[13] Goentoro L, Shoval O, Kirschner MW, Alon U (2009) The incoherent feedforward loop can provide fold-change detection in gene regulation. *Mol Cell*, 36: 894-899.

[14] Shoval O, Goentoro L, Hart Y, Mayo A, Sontag E, *et al.* (2010) Fold-change detection and scalar symmetry of sensory input fields. *Proc Natl Acad Sci USA*, 107: 15995-16000.

[15] Basu S, Mehreja R, Thiberge S, Chen M, Weiss R (2004) Spatiotemporal control of gene expression with pulse-generating networks. *Proc Natl Acad Sci USA*, 101: 6355-6360.

[16] Macía J, Widder S, Solé R (2009) Specialized or flexible feed-forward loop motifs: a question of topology. *BMC Syst Biol*, 3: 84.

[17] Koshland DE Jr, Goldbeter A, Stock JB (1982) Amplification and adaptation in regulatory and sensory systems. *Science*, 217: 220-225.

[18] Sontag ED (2008) Remarks on feedforward circuits, adaptation, and pulse memory. *IET Syst Biol*, 4: 39-51.

[19] Barkai N, Leibler S (1997) Robustness in simple biochemical networks. *Nature*, 387: 913-917.

[20] Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. *Nature*, 397: 168-171.

[21] Levchenko A, Iglesias PA (2002) Models of eukaryotic gradient sensing: Application to chemotaxis of amoebae and neutrophils. *Biophys J*, 82: 50-63.

[22] Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. *Cell*, 138: 760-773.

[23] Bintu L, Buchler NE, Garcia H, Gerland U, Hwa T, *et al.* (2005) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15: 116-124.

[24] Tsang J, Zhu J, van Oudenaarden A (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell*, 26: 753-767.

[25] Thattai M, van Oudenaarden A (2002) Attenuation of noise in ultrasensitive signaling cascades. *Biophys J*, 82: 2943-2950.

[26] Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a

single gene. *Nat Genet*, 31: 69-73.

[27] Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. *Science*, 307: 1965-1969.

[28] Ghosh B, Karmakar R, Bose I (2005) Noise characteristics of feed forward loops. *Phys Biol*, 2: 36-45.

[29] Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW (2007) Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130: 141-152.

[30] Rodrigo G, Carrera J, Jaramillo A (2007) Genetdes: automatic design of transcriptional networks. *Bioinformatics*, 23: 1857-1858.

[31] Rodrigo G, Carrera J, Elena SF (2010) Network design meets in silico evolutionary biology. *Biochimie*, 92: 746-752.

[32] de Visser JAGM, Hermisson J, Wagner GP, Meyers LA, Bagheri-Chaichian H, *et al.* (2003) Evolution and detection of genetic robustness. *Evolution*, 57: 1959-1972.

[33] Wagner A (2005) Robustness and Evolvability in Living Systems. Princeton University Press, New Jersey.

[34] Papatsenko D, Levine M (2005) Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the *Drosophila* embryo. *Proc Natl Acad Sci USA*, 102: 4966-4971.

[35] Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. *Nature*, 405: 590-593.

[36] Yi TM, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci USA*, 97: 4649-4653.

[37] Cagatay T, Turcotte M, Elowitz MB, Garcia-Ojalvo J, Suel GM (2009) Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139: 1-11.

[38] Carroll SB (1990) Zebra patterns in fly embryos: Activation of stripes or repression of interstripes? *Cell*, 60: 9-16.

[39] Schroeder MD, Pearce M, Fak J, Fan HQ, Unnerstall U, *et al.* (2004) Transcriptional control in the segmentation gene network of *Drosophila. PLoS Biol*, 2: e271.

[40] Cotterell J, Sharpe J (2010) An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol Syst Biol*, 6: 425.

[41] Savageau MA (1977) Design of molecular control mechanisms and the demand for gene expression. *Proc Natl Acad Sci USA*, 74: 5647-5651.

[42] von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment

polarity network is a robust developmental module. *Nature*, 406: 188-192.

Chapter 4

Integral control networks: a natural design

Nature laughs at the difficulties
of integration.
– Pierre-Simon Laplace

Mathematical models of natural networks can help in our understanding on why genes interact and unveil the underlying control structures. In this chapter, we study plant gravitropism as an example of biological integral control, and analyze the integration of hormone signaling and gene regulation.

4.1 The case of plant gravitropism

Living organisms have the ability of sensing and processing many environmental signals to act accordingly. For this purpose, organisms have developed a potent sensory machinery that, coupled to the appropriate signaling circuits, can trigger specific cellular responses. The capacity of an organism to adapt to varying environmental conditions therefore depends on several intrinsic properties established by the topology of the networks involved in this response. Plants display a particularly good adaptive ability, and it has been proposed that this advantage may rely on the architecture of their signaling networks [1]. Among all external stimuli, gravity is invariant and plants use it as a reference for the orientation of the growth of their organs. For instance, plants placed in a horizontal position reorient growth of the aerial part in the direction opposite to the gravity vector. According to the early Cholodny-Went theory [2], the perception of a change of position with respect to the gravity vector triggers the formation of a gradient that determines differential growth rates on either side of the organ, thus causing the formation of a curvature and the reorientation of the whole organ. More recent work has established that this gradient is formed by differential distribution of the phytohormone auxin [3, 4]. In aerial tissues, auxin accumulation triggers a cascade of molecular events [5–7] that ultimately promote the expression of growth-related genes in one side of the organ subject to the gravitropic stimulus.

Although the auxin gradient is instrumental in the differential promotion of growth, the phytohormones gibberellins have been recently involved in the regulation of the response to gravity [8, 9]. Gibberellins are well-known growth-promoting hormones [10, 11] that sometimes act as a subsidiary signal for auxin [12]. However, in the case of the gravitropic response, they display a counterintuitive effect because they delay reorientation, and they do so by attenuating auxin signaling through the transcriptional regulation of an auxin signaling element [9]. According to these recent experiments, we propose the construction of a mathematical model to study the combined effect of the two hormones on plant gravitropism and make predictions of the expected behavior under different conditions.

Given the complex interactions that modulate the gravitropic response, we have attempted to elucidate the quantitative and dynamical properties of the signaling circuit by modeling the molecular interactions that subtend this response. We have paid particular attention to the type of control mechanism in the circuit, and to the capacity of the circuit to generate noise in gene expression. How cell fate is switched by environmental stimuli and how precise molecular interactions implement a control on plant physical behavior are intriguing questions herein we have addressed. Here we present a stochastic dynamical model to dissect the particular hormonal interplay, between auxins and gibberellins, which is key for plant behavior under gravitropic stimuli. The whole model consists of a

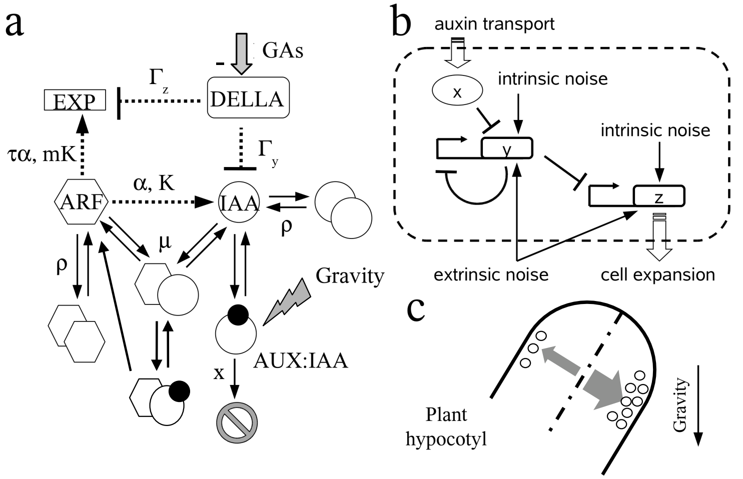


Figure 4.1: (a) Schematic representation of the genetic network that controls gravitropism in plants. Dotted lines denote transcription regulations. The network involves auxins (AUX) and gibberellins (GAs), two central hormones in plant signaling, auxin/indoleacetic acid-induced proteins (Aux/IAA, labeled as IAA), auxin-response transcription factors (ARF), DELLA proteins, and cell expansion proteins (EXP). AUX promote the degradation of IAA proteins and GAs control negatively the synthesis of DELLA proteins. Letters denote kinetic parameters of the model. (b) Simplified regulatory network, involving two genes (y and z) and one hormone (x). See the text for a complete explanation. (c) Planar representation of the plant organ. The gravity vector leads to a differential accumulation of auxins, represented by small circles.

molecular description of gene interactions, assumed in quasi-steady state, together with a physical model accounting for the reorientation of the plant. We have analytically developed the model to illustrate an integral control mechanism and to obtain a theoretical prediction of noise in gene expression.

4.2 Modeling at molecular and physiological levels

Although gravitropic reorientation affects a whole organ including multiple cell types, experimental observations lead to the assumption that gravity is perceived in the endodermis [9, 13] and that the molecular interactions that initially regulate the gravitropic response occur in these cells (Fig. 4.1a). Cell expansion is accomplished by expressing growth-related genes (*e.g.*, expansins and others) that drive the elongation of the plant. These genes are activated by auxin-response transcription factors (ARFs) [14, 15], a pivotal family of transcriptional regulators in plants, and repressed by the action of DELLA proteins [9], which are a family of putative regulators that inhibit the cell proliferation and expansion. ARFs also activate transcription of auxin/indoleacetic acid-induced (Aux/IAA) proteins [14, 15], which implement a post-translational negative feedback loop providing robustness to system [16, 17]. Although in some cases hormonal signals might slightly influence on the expression of ARFs [18, 19], here we assume that it does not depend on auxins or gibberellins and then the total amount of ARFs is taken constant [15]. In addition, ARFs and Aux/IAA proteins form homo- and heterodimers, although the kinetics for heterodimerization is much faster [20]. This regulatory loop is closed by the action of auxins, which trigger the degradation of Aux/IAA proteins by the proteasome through the formation of a complex between the hormone, the auxin receptor, and the target Aux/IAA protein [5, 21]. The hormonal crosstalk between gibberellins and auxins emerges due to gibberellins repress the synthesis of DELLA proteins, and, as recent investigations have shown, DELLA proteins down-regulate Aux/IAA proteins and moderate the response to the auxin gradient induced by gravity [9].

Based on these interactions, we construct a reduced molecular model (Fig. 4.1b) defined by differential equations involving the concentration of auxins (x, dimensionless variable), ARFs (u_{tot}, total amount), Aux/IAA proteins (y), and the generic gene (referred just as expansins for the following) activity executing cell expansion (z). In the model, auxins just promote the degradation of Aux/IAA proteins, and the effect of gibberellins by means of DELLA proteins is reduced to coefficients that modulate the protein synthesis rate. The deterministic dynamics are governed by

$$\begin{aligned} \frac{dy}{dt} &= \alpha\Gamma_y f_1(y) - xy - y, \\ \frac{dz}{dt} &= \alpha\Gamma_z f_m(y) - \frac{1}{\tau}z, \end{aligned} \tag{4.1}$$

where α is the maximal synthesis rate, Γ_y and Γ_z the repression coefficients of DELLA over Aux/IAA proteins and expansins respectively, and $f_1(y)$ and $f_m(y)$ the regulatory functions of ARFs. Notice that these functions also depend on u_{tot}. Variations in the levels of gibberellins are set by changing the values of Γ_y and Γ_z. For normal levels of gibberellins we just set $\Gamma_y = \Gamma_z = 1$, whereas for low levels of this hormone DELLA proteins are up-regulated and they can exert the repression resulting in $\Gamma_y \leq \Gamma_z < 1$ (we also assume that the repression over Aux/IAA proteins is stronger than over expansins). Time is conveniently rescaled by the degradation coefficient of Aux/IAA proteins, while τ accounts for the higher stability of expansins. We assume that auxins do not saturate the proteolytic degradation of Aux/IAA proteins and that the kinetics of this process is equivalent to that of thermodynamic degradation [21]. The values of the model parameters are shown in Table 4.1.

By exploiting the different time scales (binding reactions are much faster than protein synthesis) and assuming much faster kinetics for heterodimerization ($\mu \gg \rho$), we obtain the expression for the free amount of ARFs

$$u(y) = \frac{1}{4\rho}(\sqrt{1 + 8\rho(u_{tot} - y)} - 1), \tag{4.2}$$

for $y < u_{tot}$, and $u = 0$ elsewhere. In addition, the transcriptional activation function of ARFs (assumed of Hill-type) reads

$$f_h(y) = \frac{u(y)^n}{(hK)^n + u(y)^n}, \tag{4.3}$$

where n is the Hill coefficient, K is the protein-DNA binding coefficient, and the activation threshold of transcription can be modulated using different values of h. Herein, we consider $h = 1$ for Aux/IAA activation, whereas $h = m$ for expansin activation.

We simulate the model to study the sensitivity of the kinetic parameters in the stationary regime (Fig. 4.2). The stationary solution is given by $y_0(1 + x_0) = \alpha\Gamma_y f_1(y_0)$ and $z_0 = \tau\alpha\Gamma_z f_m(y_0)$, where y_0 and z_0 denote the steady state values. Certainly, the level of Aux/IAA proteins is fundamental to determine the functioning point, and this is controlled by three elements in the circuit: auxins, gibberellins (via DELLA proteins), and ARFs. The accumulation of auxins decreases the abundance of Aux/IAA proteins, whereas higher levels of DELLA proteins (modeled by $\Gamma_y \leq \Gamma_z < 1$) boost the expression of expansins and their differential expression (∇z) computed between the two sides of the elongating plant organ. In fact, the differential growth remarkably reaches a maximum at certain level of DELLA proteins, higher than in the wild-type case.

Table 4.1: Kinetic parameters (with typical values) used in this work.

Parameter	Description	Value
α	Protein synthesis amount	400 molec [a]
K	ARF-DNA binding coefficient over the Aux/IAA promoter	10 molec [b]
m	Relative ARF-DNA binding affinity over the expansin promoter	10 [c]
n	Hill coefficient (ARF multimerization degree)	2 [c]
ρ	Homodimerization equilibrium constant	0.01 molec^{-1} [d]
μ	Heterodimerization equilibrium constant	10 molec^{-1} [d]
u_{tot}	Total ARF amount in the nucleus	100 molec [e]
x_{tot}	Total auxin amount (normalized)	2 [f]
τ	Relative stability of expansins over Aux/IAA proteins	3 [f]
Γ_y	Repression coefficient of DELLA on Aux/IAA proteins	1 [g]
Γ_z	Repression coefficient of DELLA on expansins	1 [g]
λ	Relative elongation rate	1 mm/(molec h) [h]
D	Diameter of the elongating organ	1 mm [g]
ζ_0	Minimal proportion of auxins in the upper side	0.33 [i]

[a] Estimated from the amount of some nuclear proteins in yeast [22] given that Aux/IAAs are short-lived nuclear proteins [5]. [b] Estimated from the MAPK transcription factor Ste12 in yeast [23, 24]. [c] Assuming that ARF dimers preferentially bind to Aux/IAA promoters [15]. [d] Based on quantification of homo- and heterodimers in HeLa cells [20]. [e] Estimated from the amount of the MAPK transcription factor Ste12 in yeast [22]. [f] Based on the kinetics of degradation of Aux/IAA proteins [21]. [g] This work. [h] Based on hypocotyl elongation in soybean [25]. [i] Based on quantification of auxins under gravistimulation in *Brassica oleracea* [3] and in pea [4].

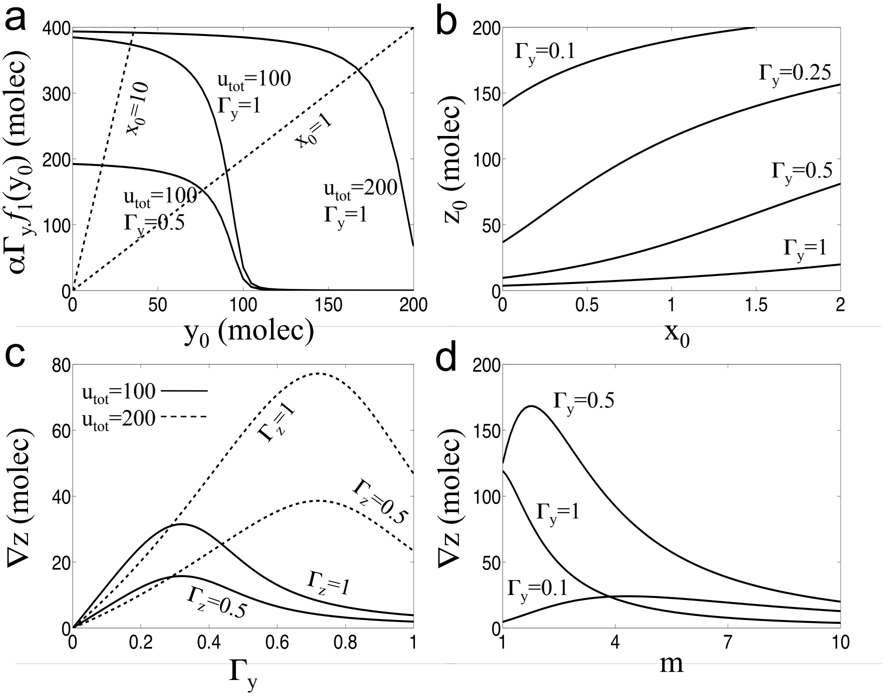


Figure 4.2: Simulation of the molecular model in the deterministic regime for a fixed amount of auxins (x_0). (a) Synthesis rate of Aux/IAA proteins for different levels of them (y_0) showing a repressive function. Dotted lines show the degradation of Aux/IAA. (b) Expression of cell expansion genes (z_0) versus the auxin amount. Differential expression of expansins (∇z) versus (c) the repression coefficient of DELLA on Aux/IAA proteins (Γ_y) and (d) the parameter that accounts for the relative ARF-DNA binding affinity (m). If not specified, kinetic parameters take values from Table 4.1.

Although DELLA proteins directly repress the expansion genes, their action over the self-repressed Aux/IAA proteins counteracts that effect. Accordingly, we corroborate that the accumulation of auxins stimulates elongation as a direct consequence of the up-regulation of cell expansion genes, which is also in tune with the experimental evidence [3].

As mentioned, the ability of plants to describe curve trajectories relies on a differential growth in both sides of the elongating organ (planar projection), caused by a differential accumulation of auxins induced by the gravity action (Fig. 4.1c) [26]. For simplicity, we consider a linear distribution of auxins along the transversal axis.

This distribution depends on the angle of the plant with respect to the vertical (θ). Certainly, when the plant is straight ($\theta = 0$) the auxin distribution is symmetric. At maximal bending ($\theta = \pi/2$) the ratio of auxins between the two sides is also maximal and has been experimentally estimated to be at most the double [3, 4]. By continuity, we assume that the total amount of auxins (x_{tot}) is constant, being $x_{down}(\theta) = \zeta(1, \theta)x_{tot}$ and $x_{up}(\theta) = \zeta(0, \theta)x_{tot}$. To achieve that $x_{down}(0) = x_{up}(0) = x_{tot}/2$ and that $x_{up}(\pi/2) = \zeta_0 x_{tot}$, the distribution ζ follows $\zeta(r, \theta) = r + (1-2r)(\frac{1}{2} - \frac{\theta}{\pi} + \frac{2\theta}{\pi}\zeta_0)$, where ζ_0 is the minimal proportion of auxins in the upper flank (to obtain this expression we have assumed a linear distribution). The maximal auxin ratio is given by $1/\zeta_0 - 1$. Hence, $\nabla z(\theta) = z(x_{down}(\theta)) - z(x_{up}(\theta))$.

As stated above, we consider that elongation is proportional to the level of expression of elongation genes at a given position (up or down) and orientation (θ). Since this expression is modulated by the levels of auxins and a change in the angle provokes a redistribution of auxins, the physiological response is time-coupled to the dynamics of the genetic circuit. Thus, by considering that the differential elongation provokes the curvature of the organ [2], the dynamics follows

$$\frac{d\theta}{dt} = \lambda \frac{\nabla z(\theta)}{D}, \tag{4.4}$$

where λ is the elongation rate relative to the expression of expansins, and D the organ diameter. We also define a physiological dimensionless time $T = t\lambda/D$. We assume that the physiological time scale is greater than the molecular one and, hence, the concentrations of the relevant molecules in the cell are considered to have reached their steady states. We use these values to compute the elongation and the corresponding degree of reorientation at each time step. Moreover, the diffusion of auxins is sufficiently rapid, as is the protein synthesis, to ensure the decoupling of the time scales, which supports the fact of assuming the molecular system in quasi-steady state. Indeed, the diffusion coefficient of auxins is $\sim 10^{-3}$ mm^2/s [27], then for a space of ~ 1 mm we have a diffusion time of ~ 20 min, which is of the order of the half-life of Aux/IAA proteins.

Having such a physical description, we couple this with the molecular model to simulate the gravitropic response. Given that the position of DELLA proteins in the regulatory circuit could contemplate the possibility of both a positive and a negative effect of gibberellins upon the gravitropic response, we investigated the dynamics of organ reorientation with our model under two hypothetical control strategies

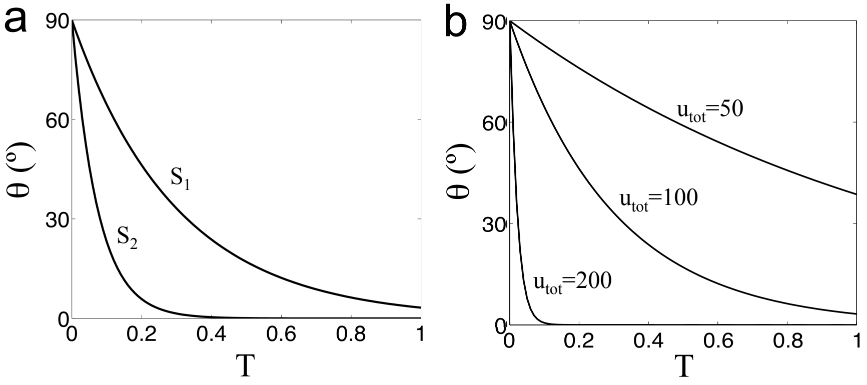


Figure 4.3: Dynamic plant response in a simulated experiment under gravistimulation (plant artificially rotated 90°), (a) for two control strategies modulated by gibberellins (S_1 for low levels and S_2 for high levels of DELLA proteins), and (b) for different amounts of total ARF ptoteins (u_{tot}). The rest of the parameter values are shown in Table 4.1.

involving gibberellins: one (S_1) for low levels of DELLA with $\Gamma_y = 1$ and $\Gamma_z = 1$ (wild-type scenario, with normal levels of gibberellins), and other (S_2) for high levels of DELLA with $\Gamma_y = 0.5$ and $\Gamma_z = 0.75$ (with low levels of gibberellins). Accordingly, we simulate the dynamics of the organ reorientation under gravistimulation, where the plant is artificially rotated 90° (Fig. 4.3). Interestingly, our model predicts that the speed of the response would be higher in S_2 than in S_1. In fact, this discrepancy could be higher since DELLA proteins enhance the gradient of auxins by means of the activation of efflux carriers [28]. While the repression over expansins by DELLA (Γ_z) gives a monotonic effect, the repression over Aux/IAA (Γ_y) entails an optimal point in the reorientation ability. Also, the higher stability of expansins (or the higher degradation of Aux/IAA proteins) would allow a more rapid tropic response. In addition, we investigated the effect of the total amount of ARFs (u_{tot}). Our model predicts that multiple knockouts in some genes of the ARF family [18] would also cause a decrease in the speed of the response. In that way, this genetic engineering could counteract deficient levels of gibberellins in the system. Remarkably, these predictions have been confirmed in parallel experimental work [9]. Therefore, our model finely predicts the plant gravitropic response based on simple molecular interactions and allows depicting the role of gibberellins (attenuation of the speed of reorientation) in such a response.

4.3 Integral control and stochasticity

Biological systems, like in mechanics or electronics, implement automatic control strategies to accommodate the developmental behavior to the environment or to be robust under perturbations. The automatic control allows a continuous sensing of the output (z) and acting over the input (x) to maintain the reference state ($\theta_0 = 0$) [29]. In control theory, a system is assumed in equilibrium and subjected to external perturbations that can alter the desired mode of operation. Control loops are designed to automatically correct such perturbations over the system. In that way, does the network of genetic interactions that governs the tropic plant response provide the expected robustness [16] in biological systems? This depends on the network topology and a proper parameterization ensuring the stability of the control system. At first sight, the network consists in a negative feedback loop, which has been demonstrated in other systems to be responsible of implementing an integral control (Fig. 4.4). This type of control uses the past trajectory to compute the deviation with respect to the reference value (steady state), and, in our case, perturbations in the auxin level could be counteracted [30].

To dissect the control structure and study its stability, we apply the Laplace transform ($\hat{\ }$ with domain variable s) on the system (Eqs. 4.1) linearized around the steady state, to have

$$\begin{aligned}(\phi + s)\Delta\hat{y} &= -y_0\Delta\hat{x},\\ (1/\tau + s)\Delta\hat{z} &= \alpha\Gamma_z f_m'(y_0)\Delta\hat{y}.\end{aligned} \tag{4.5}$$

where $\phi = 1 + x_0 - \alpha\Gamma_y f_1'(y_0)$ (f' denotes derivative). The basic scheme of control of a system consists of a sensor-controller that implements the negative feedback loop. In our case, the system consists of two subsystems (represented by the states of two proteins) and the sensory machinery is implemented by the plant through a spatial hormone gradient. Gravity modulates the level of auxins in both sides of the organ to lead to reorientation. Here, the system is of second-order, whereas the global system is of third-order due to the integral sensor-control. Thus, the stability condition (necessary and sufficient) is reduced to $3.2\cdot 10^9 m^2 K^6 > \tau\alpha^3 u_{tot}^4$ (by imposing negative roots of the system). In the particular case of choice of parameters shown in Table 4.1, the stability condition is satisfied.

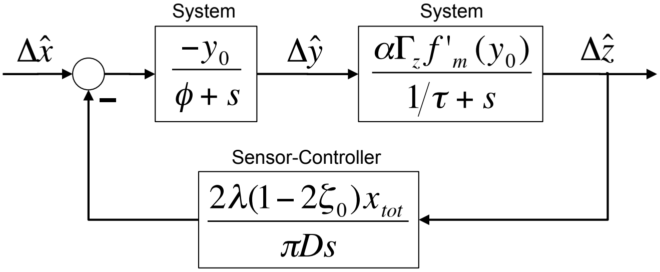


Figure 4.4: Control diagram (negative feedback loop) implemented in plants for gravitropic response (Eqs. 4.5). In this case, $\hat{\cdot}$ represents the Laplace transform, and s the corresponding domain variable.

Hence, perturbations in the level of auxins are corrected to guide the system to the reference state. In addition, gibberellins modulate the magnitude of the response, in such a way that the accumulation of DELLA proteins during the deficiency of gibberellins accelerates the corrective response. In fact, what gibberellins control is ultimately the transient time to reach a symmetric auxin distribution along the organ transversal. In that case (S_2), random fluctuations in auxin levels could induce a disproportional response. However, the strategy S_1 (the natural one) appears to provide a more flexible control (*i.e.*, slower corrective response) over the tropic response to overcome possible stochastic effects on hormonal signaling. In addition, if the redistribution of auxins is caused by light stimuli, a more flexible control would allow a higher bending during shade avoidance.

Our analysis of the model indicates that the main result of the gibberellin regulation in the circuit that regulates gravitropism is to modulate the sensitivity to auxin in the cells that perceive and respond to gravity. To investigate if the topology of the circuit provides additional regulatory features to the system, we examined the stochasticity of it and specially how this affects the expansin expression as the final output. The topology of the circuit suggests that gibberellins could modulate the sensitivity to auxins. Can different levels of gibberellins significantly influence noise tolerance? Hence, we further investigate the noise propagation in single cells from auxins to expansins via Aux/IAA proteins, and the effect that gibberellins exert in such a noise propagation.

For this purpose, we adopt a Langevin formulation to account for stochastic events [31]. Now, the system of differential equations accounting for molecular noise (intrinsic and extrinsic) reads

$$\begin{aligned} \frac{dy}{dt} &= \alpha\Gamma_y f_1(y) - xy - y + \xi_y(t) + \xi_g(t), \\ \frac{dz}{dt} &= \alpha\Gamma_z f_m(y) - \frac{1}{\tau}z + \xi_z(t) + \xi_g(t), \end{aligned} \tag{4.6}$$

where the stochastic processes $\xi_y(t)$ and $\xi_z(t)$ account for the intrinsic noise, whereas the common process $\xi_g(t)$ for the extrinsic noise. According to previous experimental results [32], the autocorrelation time for the intrinsic noise is very small and therefore we can assume that their statistics are $\langle\xi_i(t)\rangle = 0$ and $\langle\xi_i(t_0)\xi_i(t_0 + t)\rangle = q_i^2\delta(t)$ for $i = y, z$, where $\langle.\rangle$ represents the ensemble average. However, the autocorrelation time for the extrinsic noise is of the order of the protein half-lives [32], so we assume $\langle\xi_g(t)\rangle = 0$ and $\langle\xi_g(t_0)\xi_g(t_0+t)\rangle = q_g^2\frac{1}{2\tau}e^{-|t|/\tau}$. For auxins, we consider a distribution with $\langle x(t)\rangle = x_0$ and $\langle\Delta x(t_0)\Delta x(t_0 + t)\rangle = x_0 q_x^2\frac{1}{2\tau}e^{-|t|/\tau}$. Here, we take the approximation of mean field theory, assuming a perturbative regime, by which the dynamics is decomposed as $z(t) = \langle z(t)\rangle + \Delta z(t)$, where mean value is the deterministic solution ($\langle z(t)\rangle = z_0$), and the perturbative term only depends on the mean field. Hence, we have $q_y^2 = 2y_0(1 + x_0)$ and $q_z^2 = 2z_0/\tau$. Besides, q_x and q_g are free parameters that control the amplitude of the auxin and extrinsic (global) noise.

We define noise as $\eta_z^2 = \langle\Delta z^2\rangle/z_0^2$, which can be analytically calculated taking advantage of the previous considerations. Thereby, being $\phi \gg 1/\tau$ and introducing $A = -\alpha\Gamma_z f_m'(y_0)$, it turns out

$$\begin{aligned} \eta_y^2 &= \frac{1 + x_0}{\phi y_0} + \frac{x_0}{2\phi(\tau\phi + 1)}q_x^2 + \frac{1}{2\phi(\tau\phi + 1)y_0^2}q_g^2, \\ \eta_z^2 &= \frac{1}{z_0} + \frac{\tau^2 A^2 y_0^2}{2z_0^2}(\eta_y^2 - \frac{1 + x_0}{\phi y_0}) + (\frac{1}{2} - \frac{\tau A}{\tau\phi + 1})\frac{\tau}{2z_0^2}q_g^2. \end{aligned} \tag{4.7}$$

In essence, noise can be decomposed into three terms, one intrinsic to the gene (mostly Poisson-like), another due to propagation, and one last extrinsic due to global effects that are common to all species [33]. Certainly, the stochasticity arises from a low number of molecules, which induces fluctuations in gene expression. To study noise propagation, we plot the noise in protein concentrations for different amounts of auxins (Fig. 4.5). For negligible noise levels in auxins ($q_x = 0$), noise in expansins is mostly Poissonian in absence of extrinsic sources, indicating that propagation events from upstream proteins are not significant. In fact, in this case, the noise in expansins is basically inversely proportional to the noise in Aux/IAA proteins

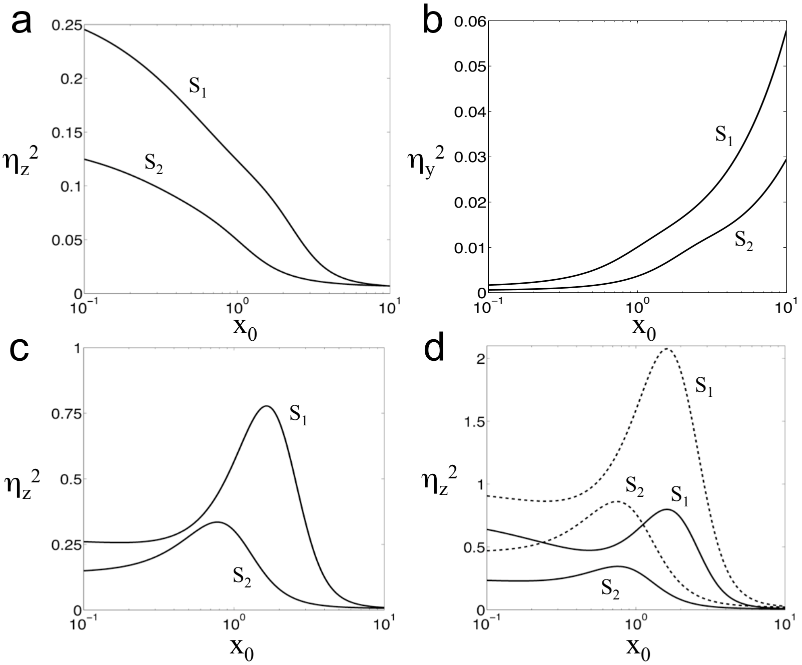


Figure 4.5: Noise in expansin expression (η_z^2) and Aux/IAA protein expression (η_y^2) versus the mean auxin amount (x_0), for two control strategies modulated by gibberellins (S_1 for low levels and S_2 for high levels of DELLA proteins). (a,b) No noise in auxins and no global noise ($q_x = q_g = 0$), (c) Poissonian noise in auxins ($q_x = 1$) and no global noise ($q_g = 0$), and (d) noise in auxins and global noise ($q_x = q_g = 1$), where dotted lines correspond to expansins with low stability ($\tau = 1$). The rest of the parameter values are shown in Table 4.1.

($1/\eta_z^2 \propto \eta_y^2 \propto x_0$). Since the level of auxins positively correlates with the expression of expansins, its noise will decrease with auxins, thus reducing the variability in the cells located in the lower side of the organ. However, for high noise levels in auxins ($q_x = 1$), there is a maximum in the noise in expansins at intermediate auxin amounts, due to the tradeoff between the intrinsic and propagation terms ($\eta_z^2 \propto p(x_0)/x_0$, being p a quadratic polynomial). This is interesting because small perturbations in the amount of auxins could lead to notable changes in noise in expansins. In addition, our model predicts that deficient levels of gibberellins (*i.e.*, high levels of DELLA proteins) would entail a reduction of the noise in protein expression, which could reduce the variability in the physiological response of a population of plants gravitropically stimulated. Very strikingly, this prediction on the variability in the physiological response has been confirmed in parallel experimental work [9].

4.4 Discussion

In this chapter, we have proposed a model, based on nonlinear dynamics and stochastic modeling, to show how plants have programmed an integral control by coupling transcription circuits with hormonal signaling. Importantly, our molecular model integrates a novel regulatory interaction, the repression of the expression of Aux/IAA proteins (encoding auxin signaling elements) by DELLA proteins (which are gibberellin signaling elements). This interaction has been shown to affect gravitropic reorientation in etiolated seedlings [9], and our model establishes that the mechanism relies in the generation of a negative feedback loop involving the two hormones that implements a system of integral control. Interestingly, as in bacterial chemotaxis, such a control strategy is generally responsible for perfect adaptation, by which the output of the system always reaches its operating point after a transient response when varying the input level [30]. Alternative modes to regulate the auxin level by other types of control, such as the proportional control, would not return the system to the reference state, being the output level in steady state dependent on the input signal. On the other hand, an integro-derivative control could give a finer strategy, since it can anticipate the future of the signal. However, such a control is not applicable to real systems that are subjected to random fluctuations, since for a noisy signal the derivative control stage would introduce an undesirable deviation.

Hormones are known to redundantly regulate gene expression during plant development. Crosstalk between hormones has been generally depicted as occurring at the level of signal transduction, although more recent molecular evidence points to multiple integration points including gene regulation [34, 35]. The circuit that we have modelled here represents a mechanism in which signal transduction and gene regulation are intertwined, and in fact transcriptional regulation becomes an integral part of the feedback regulatory module that provides plasticity to the output trait.

Recently, a deterministic mathematical model of the regulatory feedback loop of auxins was developed, but without coupling to a physical model of plant reorientation, to analyze the dynamical features of the system [17]. Our model simplifies the underlying complexity to capture the essential elements that play in the gravitropic response, and it allows us to predict the physiological response under molecular changes. In addition to the role of gibberellins as modulators of auxin sensitivity, the analysis of our model highlights a previously unsuspected feature of the hormonal circuit that regulates gravitropic respones: the positive effect of gibberellins upon noise propagation. This occurs in such a way that gibberellins are found to decrease the response to gravity, and also increase the variance of this response, and both phenomena have been confirmed *in vivo* [9]. Interestingly, the increase in noise propagation represents an intrinsic property of the regulatory circuit studied here, and it is caused by the incorporation of high levels of gibberellins into the circuit. From this perspective, our analysis suggests for the first time a molecular basis for noise generation in the biological response to gravity.

Finally, one question that becomes relevant from a biological point of view is why Nature has selected a molecular mechanism that attenuates the ability of plants to respond to gravity, which is an important environmental cue that determines growth orientation. In other words, what selective advantage is provided by this attenuation? In this particular case, one possibility is that the generation of variance in gravitropism allows the individuals to respond in a more precise way towards light cues, for instance when seedlings emerge from the soil or during shade avoidance [36]. In general, our results suggest that partially redundant signaling pathways might impinge on each other not just to regulate the magnitude of the response, but to maintain an elevated degree of plasticity from individual to individual [37].

The following publication holds the contents presented in this chapter

- Rodrigo G, Jaramillo A, Blázquez MA (2011) Integral control of plant gravitropism through the interplay of hormone signaling and gene regulation. *Biophys J*, 101: 757-763.

Bibliography

[1] Casal JJ, Fankhauser C, Coupland G, Blázquez MA (2004) Signalling for developmental plasticity. *Trends Plant Sci*, 9: 309-314.

[2] Went FW, Thimann KV (1937) Phytohormones. Macmillan, New York.

[3] Esmon CA, Tinsley AG, Ljung K, Sandberg G, Hearne LB, Liscum E (2006) A gradient of auxin and auxin-dependent transcription precedes tropic growth responses. *Proc Natl Acad Sci USA*, 103: 236-241.

[4] Haga K, Iino M (2006) Asymmetric distribution of auxin correlates with gravitropism and phototropism but not with autostraightening (autotropism) in pea epicotyls. *J Exp Botany*, 57: 837-847.

[5] Chapman EJ, Estelle M (2009) Mechanism of auxin-regulated gene expression in plants. *Annu Rev Genet*, 43: 265-285.

[6] Santner A, Calderon-Villalobos LIA, Estelle M (2009) Plant hormones are versatile chemical regulators of plant growth. *Nat Chem Biol*, 5: 301-307.

[7] Hagen G, Guilfoyle TJ, Gray WM (2010) Auxin signal transduction. *Plant Hormones*, D: 282-307.

[8] Tatematsu K, Kumagai S, Muto H, *et al.* (2004) MASSUGU2 encodes Aux/IAA19, an auxin-regulated protein that functions together with the transcriptional activator NPH4/ARF7 to regulate differential growth responses of hypocotyl and formation of lateral roots in *Arabidopsis thaliana. Plant Cell*, 16: 379-393.

[9] Gallego-Bartolomé J, Kami C, Fankhauser C, Alabadí D, Blázquez MA (2011) A hormonal regulatory module that provides flexibility to tropic responses. *Plant Physiol*, doi: 10.1104/pp.111.173971.

[10] Cowling RJ, Harberd NP (1999) Gibberellins control *Arabidopsis*

hypocotyl growth via regulation of cellular elongation. *J Exp Bot*, 50: 1351-1357.

[11] Ubeda-Tomás S, Swarup R, Coates J, *et al.* (2008) Root growth in *Arabidopsis* requires gibberellin/DELLA signalling in the endodermis. *Nat Cell Biol*, 10: 625-628.

[12] Frigerio M, Alabadí D, Pérez-Gómez J, Garcia-Carcel L, Phillips AL, Hedden P, Blázquez MA (2006) Transcriptional regulation of gibberellin metabolism genes by auxin signaling in *Arabidopsis. Plant Physiol*, 142: 553-563.

[13] Tasaka M, Kato T, Fukaki H (1999) The endodermis and shoot gravitropism. *Trends Plant Sci*, 4: 103-107.

[14] Ulmasov T, Hagen G, Guilfoyle TJ (1999) Activation and repression of transcription by auxin-response factors. *Proc Natl Acad Sci USA*, 96: 5844-5849.

[15] Ulmasov T, Hagen G, Guilfoyle TJ (1999) Dimerization and DNA binding of auxin response factors. *Plant J*, 19: 309-319.

[16] Kitano H (2004) Biological robustness. *Nat Rev Genet*, 5: 826-837.

[17] Middleton AM, King JR, Bennett MJ, Owen MR (2010) Mathematical modelling of the Aux/IAA negative feedback loop. *Bull Math Biol*, 72: 1383-1407.

[18] Guilfoyle TJ, Hagen G (2007) Auxin response factors. *Curr Opin Plant Biol*, 10: 453-460.

[19] Wilmoth JC, Wang S, Tiwari SB, *et al.* (2005) NPH4/ARF7 and ARF19 promote leaf expansion and auxin-induced lateral root formation. *Plant J*, 43: 118-130.

[20] Muto H, Nagao I, Demura T, Fukuda H, Kinjo M, Yamamoto KT (2006) Fluorescence cross-correlation analyses of the molecular interaction between an Aux/IAA protein, MSG2/IAA19, and protein-protein interaction domains of auxin response factors of *Arabidopsis* expressed in HeLa cells. *Plant Cell Physiol*, 47: 1095-1101.

[21] Zenser N, Ellsmore A, Leasure C, Callis J (2001) Auxin modulates the degradation rate of Aux/IAA proteins. *Proc Natl Acad Sci USA*, 98: 11795-11800.

[22] Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. *Nature*, 425: 737-741.

[23] Shao D, Zheng W, Qiu W, Ouyang Q, Tang C (2006) Dynamic studies of scaffold-dependent mating pathway in yeast. *Biophys*

J, 91: 3986-4001.

[24] Hu B, Rappel W-J, Levine H (2009) Mechanisms and constraints on yeast MAPK signaling specificity. *Biophys J*, 96: 4755-4763.

[25] Bensen RJ, Beall FD, Mullet JE, Morgan PW (1990) Detection of endogenous gibberellins and their relationship to hypocotyl elongation in soybean seedlings. *Plant J*, 94: 77-84.

[26] Masson PH, Tasaka M, Morita MT, Guan C, Chen R, Boonsirichai K (2002) *Arabidopsis thaliana*: a model for the study of root and shoot gravitropism. *The Arabidopsis Book*, 1: e0043.

[27] Goldsmith, M. H., T. H. Goldsmith, and M. H. Martin (1981) Mathematical analysis of the chemosmotic polar diffusion of auxin through plant tissues. *Proc. Natl. Acad. Sci. USA*, 78: 976-980.

[28] Rakusová H, Gallego-Bartolomé J, Vanstraelen M, Robert HS, Alabadí D, Blázquez MA, Benková E, Friml J (2011) Polarization of PIN3-dependent auxin transport for hypocotyl gravitropic response in *Arabidopsis thaliana. Plant J*, doi: 10.1111/j.1365-313X.2011.04636.x.

[29] Doyle J, Francis B, Tannenbaum A (1990) Feedback Control Theory. Macmillan, New York.

[30] Yi T-M, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci USA*, 97: 4649-4653.

[31] Wilkinson DJ (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet*, 10: 122-133.

[32] Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. *Science*, 307: 1962-1965.

[33] Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. *Science*, 307: 1965-1969.

[34] Liu J, Mehdi S, Topping J, Tarkowski P, Lindsey K (2010) Modelling and experimental analysis of hormonal crosstalk in *Arabidopsis. Mol Syst Biol*, 6: 373.

[35] Jaillais Y, Chory J (2010) Unraveling the paradoxes of plant hormone signaling integration. *Nat Struct Mol Biol*, 17: 642-645.

[36] Jiao Y, Lau OS, Deng XW (2007) Light-regulated transcriptional networks in higher plants. *Nat Rev Genet*, 8: 217-230.

[37] Wolters H, Jurgens G (2009) Survival of the flexible: hormonal growth control and adaptation in plant development. *Nat Rev Genet*, 10: 305-317.

Chapter 5

Balance of integral and derivative control strategies

Not to anticipate
is already to moan.
– Leonardo da Vinci

RNA silencing constitutes a control mechanism to eliminate undesired RNA molecules, in particular RNA viruses. In this chapter, we show how a mathematical model of this network, in addition to help in our understanding on the dynamics of viral infection, illustrates a balance between integral and derivative control in RNA silencing and provides further understanding of the evolved viral strategies to subvert such a control mechanism.

5.1 The case of RNA silencing

The underlying working principle of RNA silencing relies on the repressive action triggered by the intracellular presence of double-stranded RNAs (dsRNA) [1] (Fig. 5.1). In the case of single-stranded RNA viruses (ssRNA), dsRNAs are byproducts of genome replication mediated by virus-encoded RNA-dependent RNA polymerases (RdRp). During viral genome replication, the dsRNA intermediates become the target of the first component of the silencing pathway, DICER, a type-III RNase that degrades these dsRNA into units of 21 to 24 nucleotides called small interfering RNAs (siRNAs) [2]. Subsequently, the cellular RNA-induced silencing complex (RISC), that contains the argonaute (AGO) endonuclease [3], loads the antisense siRNAs resulting in an active form. Using the antisense siRNA as a guide, AGO cleaves the target viral ssRNA [4]. Furthermore, in a secondary cycle of amplification, the host RNA-dependent RNA polymerase VI (RDR6) uses siRNAs as primers, together with partially degraded ssRNAs, to produce long dsRNAs that serve as new substrates for DICER, a process known as transitivity [5]. Then, siRNAs systemically move from cell-to-cell immunizing new cells against infection [5, 6]. Given the properties of the RNA silencing pathway (specificity and amplification), it represents a sort of innate immune system for plants [7, 8].

Not surprisingly, viruses have evolved strategies to actively evade the RNA silencing surveillance while promoting their own replication [9]. Many viruses encode a suppressor protein (viral suppressor of RNA silencing or VSR) that interacts with elements of the silencing pathway blocking it [10–12]. The targets of these VSRs within the RNA silencing pathway are diverse: DICER, the dsRNA, the siRNA, RISC, or the systemic signal [8, 9, 13, 14] (Fig. 5.1). For example, the helper component-protease (HC-Pro) encoded by the *Potyvirus* works as suppressor by sequestering siRNAs [15–18]. This binding prevents the incorporation of siRNAs into RISC. Furthermore, by also binding plant endogenous micro-RNAs (miRNA) and controlling the expression of other genes, HC-Pro may interfere the expression of DICER proteins [19, 20], reducing the degradation of dsRNAs and, thus, favoring potyvirus replication. Similarly, the *Nodavirus* B2 suppressor also sequesters siRNAs [9]. The *Tombusvirus* P19 and *Cucumovirus* 2b suppressors interfere with the systemic spread of the 24 nucleotides siRNAs produced by DCL3 [21]. Some suppressors act on RISC, either avoiding the upload of siRNAs into AGO, like the *Closterovirus* P21 [22], by binding to AGO1 and avoiding its interaction with other proteins required to assemble the RISC, as the coat protein (CP) of *Tombusvirus* [23], by inhibiting the RISC

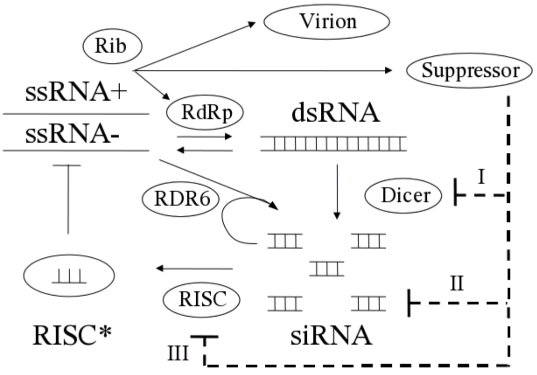


Figure 5.1: Schematic representation of the RNA silencing pathway and its interaction with viral replication. RNA viruses encode for replicase, suppressors of silencing (VSR) and coat proteins for virions. Three types of suppressors are considered in the scheme: suppressors of DICER (I), sequesters of siRNA (II), and suppressors of RISC (III).

activity after its maturation, like the *Begomovirus* AC4 [24], or by targeting AGO for degradation, as it is the case for *Polerovirus* P0 protein [25, 26]. It has also been recently shown that V2 suppressor of *Geminivirus* competes with SGS3, a key component of the secondary cycle of siRNAs amplification, in binding dsRNAs and thus interferes with transitivity [27]. Finally, the CP of some carmoviruses [28] and the P14 of *Aureusvirus* [29] can also bind long dsRNAs, resulting in the protection of the intermediaries of replication from DICER activity. Accordingly, VSRs have been divided into three families [14]: (*i*) those enhancing within cell virus accumulation, (*ii*) those essential for cell-to-cell movement but dispensable on virus accumulation in single cells, and (*iii*) those that facilitate virus long-distance movement and/or intensify disease symptoms but are not essential for viral replication and cell-to-cell movement.

The first mathematical models of the RNA silencing pathway focused on aberrant cellular mRNA as triggers of the silencing response [30, 31]. More recent models considered viral RNAs as triggers of the response and focused on virus spread in the plant [32]. However, on the one hand, these studies did not analyze in detail the possible effect that different viral suppressor strategies may have in the outcome of the interaction. On the other hand, although many several kinetic models of intracellular growth have been proposed for different viruses, none of them specifically incorporates the silencing response [33–37].

Herein, we perform a dynamical analysis of viral RNA accumulation under central parameter values of the infection, modeling the pathway shown in Fig 5.1 by differential equations. The system can reach two different stationary states (virus silencing or virus replication). In this case, we cannot apply perturbation theory to analyze the dynamics of the system, because a perturbation (one viral particle infects the cell) can lead the system to a different state (virus accumulation and spread).

5.2 Mathematical modeling

We constructed a simple mathematical model based on nonlinear differential equations involving the following species: ssRNA (S), dsRNA (D), siRNA (I), and activated RISC (R^*). The intial condition was $S_0 \geq 1$, $D_0 = 0$, $I_0 = 0$, and $R^*_0 \geq 0$. The model reads

$$\begin{aligned} \frac{dS}{dt} &= 2\beta D - \alpha S^2 - \nu R^* S - \kappa_0 S, \\ \frac{dD}{dt} &= \alpha S^2 - \delta D - \beta D, \\ \frac{dI}{dt} &= n\delta D - \rho I - \kappa_1 I, \\ \frac{dR^*}{dt} &= \rho I - \nu R^* S - \kappa_2 R^*, \end{aligned} \tag{5.1}$$

where greek letters denote kinetic parameters (see Table 5.1). This system has two stable states: either the silencing pathway is able to suppress all viral particles, or the virus bypasses the silencing response and replicates and accumulates in the cell.

To further analyze the dynamics of system, we also constructed a more detailed model accounting for positive and negative ssRNA (S^+ and S^- respectively), dsRNA (D), siRNA (I), viral proteins (P), virions (V), primed ssRNA (S^*), and secondary dsRNA (D^*). Three different viral proteins were considered: the nonstructural replicase, the VSR, and the structural coat protein. Their corresponding relative abundances were p, q and $1-p-q$. In addition, the model accounted for several cellular components: the ribosomes (Z), the RDR6 polymerase involved in transitivity (Y), DICER-like proteins (C), and inactivated and activated RISC (R and R^* respectively). We assumed that at the beginning of infection, a single viral ssRNA genome is present, which can be either sense (+) or antisense (-) depending on the nature of

Table 5.1: Kinetic parameters (with typical values) used in this work.

Parameter	Description	Value
α	Replication rate of ssRNA	10 h^{-1}
β	Dissociation rate of dsRNA	10 h^{-1}
ν	Cleavage rate by RISC	0.025 $(mol.h)^{-1}$
δ	Cleavage rate by DICER	10 h^{-1}
ρ	Rate of RISC activation	1 h^{-1}
κ_0	Degradation rate of ssRNA	0.1 h^{-1}
κ_1	Degradation rate of siRNA	1 h^{-1}
κ_2	Degradation rate of proteins	0.01 h^{-1}
n	Number of siRNAs per dsRNA	10

the virus. This model was constructed following a generalized enzyme kinetics scheme where both substrates and enzymes are limited in the medium [38], and there are competitions between different enzymes for the same substrate and different substrates for the same enzyme [39], resulting in a highly coupled formulation.

5.3 Strategies for bypassing RNA silencing

First, we considered the case of RNA viruses that do not encode VSRs. An efficient RNA silencing mechanism can prevent virus replication, although if the cleavage by DICER is not the virus may escape. In this case, after a latency period, viral proteins reach a critical concentration and promote further exponential replication. Analytically, the latency period could be estimated when RdRp reaches the corresponding binding affinity coefficient. In all these simulations the condition (+)ssRNA > (-)ssRNA holds, in excellent agreement with the observation of an excess of sense siRNAs for positive-sense viral genomes [40]. The transfer of siRNAs from infected to neighboring healthy cells, which allows the peripheral cells to activate RISC in the absence of viral infection, has the expected effect. In the absence of triggering siRNAs, infection progresses with the time delay already described above. However, if the cell has been already activated, the virus is not able of overcoming the cleavage by RISC and runs into extinction (Fig. 5.2). The multiplicity of infection (initial amount of single viral genomes per cell) has also a decisive role in preventing the extintion of the population. Indeed, the virus can bypass the silencing mechanism if the multiplicity of infection just

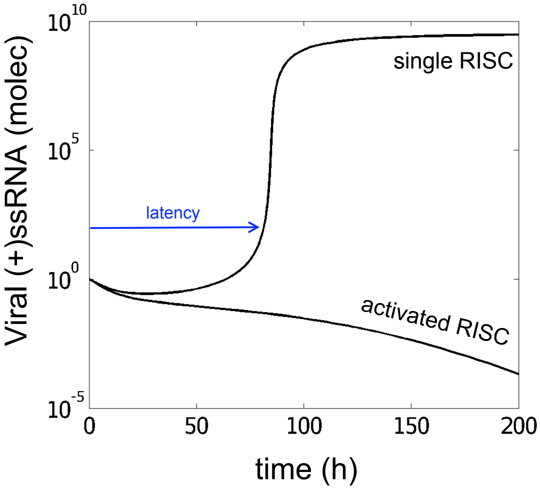


Figure 5.2: Viral RNA dynamics in a cell that has not been immunized by receiving siRNA from neighboring cells (single RISC) and in a cell that has received a small input of siRNA from an infected neighbor cell (then activating RISC).

increases. The effect of further increasing the multiplicity of infection is to reduce the latency period.

In addition, we performed several sensitivity analyses to study the regions in parameter space in which viral replication occurs or for which viral silencing takes place. We found that the higher the affinity for the negative strand, the wider is the parameter space for viral replication. In fact, this can be rationalized because viral RdRps compete with ribosomes and with the activated RISC for genomic strands, whereas they do not compete for antigenomic strands. In addition, high replication rates also allow the virus to escape from the silencing machinery and to minimize the effect of non-specific thermodynamic degradation. One question that arises here is whether a tradeoff between replication and translation exists. Upon uncoating and the strictly necessary first event of translation, a viral genome can either be directed to transcription, and thus increase the concentration of RNA, or to translation, and thus increase the concentration of viral proteins (in this case only replicase and coat). In Fig. 5.3 we analyzed such a tradeoff by considering the binding affinities to positive strands of replicase and ribosomes. We showed that in the absence of a VSR, silencing is the outcome favored when translation

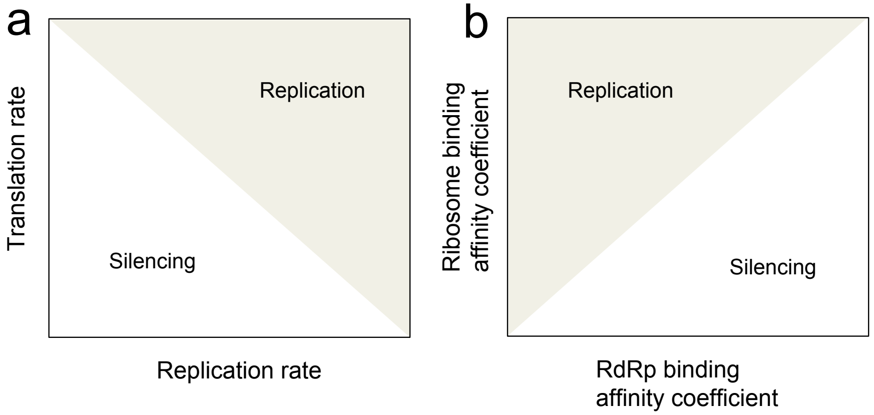


Figure 5.3: Phase diagrams identify different viral strategies. Diagram (a) shows the sensitivity of the replication and translation rates of the virus, whereas diagram (b) illustrates the relationship between the binding affinities of replication (RdRp) and translation (ribosome).

is more frequent than transcription. Accordingly, the best strategy for a virus to bypass the RNA silencing response in the absence of a VSR would be to increase the affinity of its RNA to the replicase rather than to optimize its binding affinity to the ribosomes. Likewise, by increasing its transcription efficiency, a virus will produce more copies of its genome up to the point in which the cleavage by DICER would no longer control the accumulation of viral genomes. We also show, as expected, that the higher are the catalytic constants for transcription and translation, the higher are the chances for a successful viral replication.

The fact that, in the presence of an active silencing response, it is in the benefit of the virus to invest into a transcriptional strategy rather than in translation may be somehow counterintuitive because one may expect more replication to generate more dsRNA and, therefore, to strength the silencing response and, likewise, more translation to produce more suppressor protein. It can be argued that, after the very initial burst of translation from the infecting genomic sequence resulting in a few viral proteins, the optimal strategy involves synthesizing antigenomic strands and use them as templates for producing a large excess of genomic strands (*i.e.*, using an stamping machine replication strategy) without diverting them into translation. If replication is fast enough, this replicative strategy works even in the absence of a suppressor protein: a positive feedback is established such as the replication overcomes the capacity of the available DICER molecules to keep virus replication under control. Once a significant amount of genomic strands has been produced, then translation may take place. If translation results in a VSR, then a synergistic effect between fast transcription and translation appears, resulting in a successful viral replication.

Among many possibilities, we have focused on four viral strategies. The first one, consisting in blocking DICER, turns out to be the most efficient promoting viral replication. This result is somehow logical from an optimal design perspective. By hitting the first bottleneck in the pathway the virus ensures its own replication. Hitting downstream steps would allow DICER to still exert partial control on virus replication. The other three strategies explored, sequestering siRNA, blocking RISC and disrupting the secondary amplification via RDR6, have resulted less efficient in promoting intracellular virus accumulation, although they may gain some benefit when looking at cell-to-cell movement. This finding is in good agreement with the observation that *Cucumovirus* 2b and *Tombusvirus* P19, which promote systemic and cell-to-cell movements, are not required for intracellular accumulation [9].

This has allowed us to propose an operational model of RNA silencing (Fig. 5.4). Whether a virus infects a cell, it engages its amplification machinery through the buffer dsRNA. Then, DICER attacks this unit for cleavage and produce siRNAs. If this process is efficient enough, it results in a sufficient condition to silence the virus at the single cell level. The subsequent cleavage by RISC would close the loop to provide a sort of nonlinear integral control. However, as we have shown, this process is inefficient and could be neglected in the case of viruses. However, the plant is able to transfer siRNAs from cell-to-cell, which incorporate into RISC resulting in a pre-arranged cell to eradicate *ab initio* the viral strand. According to our results, this cleavage by RISC is only instrumental in a non-infected cell. This results in a sort of nonlinear derivative control mechanism by which the sensor-controller of the system (siRNA) is able to anticipate the future of the input signal (virus infection) [41]. Because there is a delay between cell infections, such a derivative control would be of high order. In that way, VSRs that only target siRNAs try to neutralize the derivative control mechanism that allows anticipating the infection in a neighbor cell, and they are not much concerned about the viral

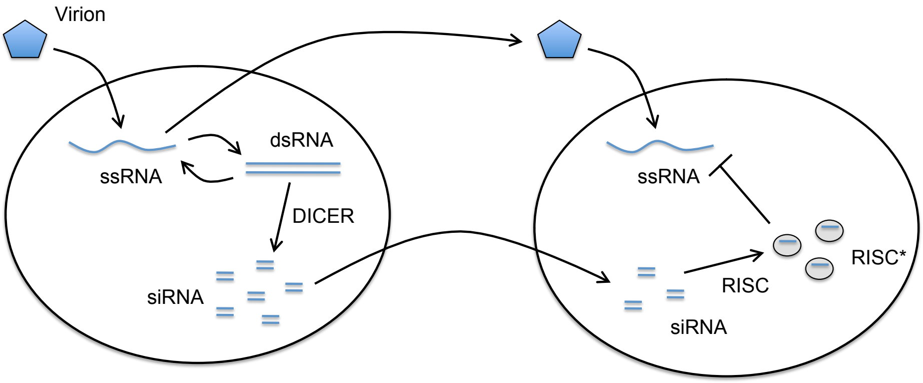


Figure 5.4: Operational model of RNA silencing.

elements in the current infected cell because the promotion of the integral control results in an inefficient task.

5.4 Discussion

In this chapter we have shown that, from a system design perspective, the best strategy that a virus may take to ensure its replication in presence of the antiviral response mediated by RNA silencing would be to (*i*) replicate fast and produce an excess of genomic strands, (*ii*) encode for a VSR that interacts with the DICER protein and (*iii*) exert some control on the multiplicity of infection, ensuring that multiple genomes infect each cell. Obviously evolution is not a perfect designer and viruses have acquired suppressor proteins that target at different steps of the silencing pathway. Understanding the exact mechanisms by which these VSRs operate will allow to develop better models and increase our ability to predict the outcome of the host-virus interaction. Furthermore, VSRs have clear biotechnological potential as they can be used to maximize the expression of transgenes [9].

Although mathematically convenient, it may be a biological oversimplification to assume that suppressors act at a single stage of the silencing pathway. Evidences exist showing that VSRs may well simultaneously operate at diverse stages of the pathway. For example, the potyviral HC-Pro sequesters siRNAs but also affects the expression of plant genes, including the *dcl*-like genes encoding for the different DICER proteins in *Arabidopsis thaliana* [19], or by reducing the 3' methylation of siRNAs making them sensitive to oligouridilation and subsequent degradation [42, 43]. Another example of multiple actions is the *Polerovirus* P0 that interferes with the silencing pathway at least at two levels: binding to siRNAs and also avoiding the formation of the activated AGO complex and labeling it for degradation [25, 26]. Also, a virus may carry more than one VSR, as it seems to be the case for some *Tombusviruses* (P19 and CP).

The following publication holds the contents presented in this chapter

- Rodrigo G, Carrera J, Jaramillo A, Elena SF (2011) Optimal viral strategies for bypassing RNA silencing. *J R Soc Interface*, 8: 257-268.

Bibliography

[1] Fire A, Xu S, Mongomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391: 806-811.

[2] Hamilton AJ, Baulcombe DC (1999) A species of small antisense RNA in post-transcriptional gene silencing in plants. *Science*, 286: 950-952.

[3] Bohmert K, Camus I, Bellini C, Bouchez D, Caboche M, Benning C (1998) AGO1 defines a novel locus of *Arabidopsis* controlling leaf development. *EMBO J*, 17: 170-180.

[4] Rand T, Petersen S, Du F, Wang X (2005) Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell*, 123: 621-629.

[5] Voinnet O, Vain P, Angell S, Baulcombe DC (1998) Systemic spread of sequence-specific transgene RNA degradation is initiated by localised introduction of ectopic promoterless DNA. *Cell*, 95: 177-187.

[6] Himber C, Dunoyer P, Moissiard G, Ritzenthaler C, Voinnet O (2003) Transitivity-dependent and -independent cell-to-cell movement of RNA silencing. *EMBO J*, 22: 4523-4533.

[7] Lecellier CH, Voinnet O (2004) RNA silencing: no mercy for viruses?. *Inmunol Rev*, 1998: 285-303.

[8] Ding S, Voinnet O (2007) Antiviral immunity directed by small RNAs. *Cell*, 130: 413-426.

[9] Li F, Ding SW (2006) Virus counterdefenses: diverse strategies for evading the RNA-silencing immunity. *Annu Rev Microbiol*, 60: 503-531.

[10] Brigneti G, Voinnet O, Li, WX, Ji LH, Ding SW, Baulcombe DC (1998) Viral pathogenicity determinants are suppressors of transgene silencing in *Nicotiana benthamiana*. *EMBO J*, 17: 6739-6746.

[11] Kasschau KD, Xie Z, Allen E, Llave C, Chapman EJ, Krizan KA, Carrington JC (2003) P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. *Dev Cell*, 4: 205-217.

[12] Baulcombe D (2004) RNA silencing in plants. *Nature*, 431: 356-363.

[13] Moissiard G, Voinnet O (2004) Viral suppression of RNA silencing in plants. *Mol Plant Pathol*, 5: 71-82.

[14] Díaz-Pendón JA, Ding SW (2008) Direct and indirect roles of viral suppressors of RNA silencing in pathogenesis. *Annu Rev Phytopathol*, 46: 303-326.

[15] Mallory AC, Reinhart BJ, Bartel D, Vance VB, Bowman LH (2002) A viral suppressor of RNA silencing differentially regulates the accumulation of short interfering RNAs and microRNAs in tobacco. *Proc Natl Acad Sci USA*, 99: 15228-15233.

[16] Chapman EJ, Prokhnevsky AI, Gopinath K, Dolja V, Carrington JC (2004) Viral RNA silencing suppressors inhibit the microRNA pathway at an intermediate step. *Genes Dev*, 18: 1179-1186.

[17] Dunoyer P, Lecellier CH, Parizotto EA, Himber C, Voinnet O (2004) Probing the microRNA and small interfering RNA pathways with virus-encoded suppressors of RNA silencing. *Plant Cell*, 16: 1235-1250.

[18] Lakatos L, Szittya G, Silhavy D, Burgyan J (2004) Molecular mechanism of RNA silencing suppression mediated by P19 protein of *tombusviruses. EMBO J*, 23: 876-884.

[19] Deleris A, Gallego-Bartolomé J, Bao J, Kasschau KD, Carrington JC, Voinnet O (2006) Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. *Science*, 313: 68-71.

[20] Mlotshwa S, Schauer SE, Smith TH, Mallory AC, *et al.* (2005) Ectopic DICER-LIKE1 expression in P1/HC-Pro *Arabidopsis* rescues phenotypic anomalies but not defects in microRNA and silencing pathways. *Plant Cell*, 17: 2873-2885.

[21] Qi Y, Zhong X, Itaya A, Ding B (2004) Dissecting RNA silencing in protoplasts uncovers novel effects of viral suppressors on the silencing pathway at the cellular level. *Nucl Acids Res*, 32: e179.

[22] Peremyslov VV, Hagiwara Y, Dolja VV (1999) HSP70 homolog functions in cell-to-cell movement of a plant virus. *Proc Natl Acad Sci USA*, 96: 14771-14776.

[23] Azevedo J, *et al.* (2010) Argonaute quenching and global changes in Dicer homeostasis caused by a pathogenencoded GW repeat

protein. *Genes Dev*, 24: 904-915.

[24] Vanitharani R, Chellappan P, Pita JS, Fauquet CM (2004) Differential roles of AC2 and AC4 of *Cassava Geminiviruses* in mediating synergism and suppression of post-transcriptional gene silencing. *J Virol*, 78: 9487-9498.

[25] Baumberger N, Tsai CH, Lie M, Havecker E, Baulcombe DC (2007) The polerovirus silencing suppressor P0 targets ARGONAUTE proteins for degradation. *Curr Biol*, 17: 1609-1614.

[26] Csorba T, Lozsa R, Hutvagner G, Burgyan J (2010) Polerovirus protein P0 prevents the assembly of small RNA containing RISC complexes and leads to degradation of ARGONAUTE1. *Plant J*, 62: 463-472.

[27] Fukunaga R, Doudna JA (2009) dsRNA with 5' overhangs contributes to endogenous and antiviral RNA silencing pathways in plants. *EMBO J*, 28: 545-555.

[28] Meng C, Chen J, Peng J, Wong SM (2006) Hostinduced avirulence of *Hibiscus chlorotic ringspot virus* mutant correlates with reduced gene-silencing suppression activity. *J Gen Virol*, 87: 451-459.

[29] Mérai Z, Kerencyi Z, Molnar A, Barta E, Bisztray G, Havelda Z, Burgyan J, Silhavy D (2005) *Aureusvirus* P14 is an efficient RNA silencing suppressor that binds double-stranded RNAs without size specificity. *J Virol*, 79: 7217-7226.

[30] Bergstrom CT, McKittrick E, Antia R (2003) Mathematical models of RNA silencing: Unidirectional amplification limits accidental self-directed reactions. *Proc Natl Acad Sci USA*, 100: 11511-11516.

[31] Groenenboom MAC, Maree AFM, Hogeweg P (2005) The RNA silencing pathway: the bits and pieces that matter. *PLoS Comp Biol*, 1: e21.

[32] Groenenboom M, Hogeweg P (2008) The dynamics and efficacy of antiviral RNA silencing: a model study. *BMC Syst Biol*, 2: 28.

[33] Reddy B, Yin J (1999) Quantitative intracellular kinetics of HIV type 1. *AIDS Res Hum Retrovir*, 15: 273-283.

[34] Sidorenko Y, Reichl U (2004) Structured model of influenza virus replication in MDCK cells. *Biotechnol Bioeng*, 88: 1-14.

[35] Lim K, Lang V, Lam T, Yin J (2006) Model-based design of growth-attenuated viruses. *PLoS Comput Biol*, 2: e116.

[36] Dahari H, Ribeiro RM, Rice CM, Perelson AS (2007)

Mathematical modeling of subgenomic *Hepatitis C virus* replication in Huh-7 cells. *J Virol*, 81: 750-760.

[37] Sardanyés J, Solé RV, Elena SF (2009) Replication mode and landscape topology differentially affect RNA virus mutational load and robustness. *J Virol*, 83: 12579-12589.

[38] DeAngelis DL, Goldstein RA, O'Neill RV (1975) A model for tropic interaction. *Ecology*, 56: 881-892.

[39] MacRae IJ, Zhou K, Doudna JA (2007) Structural determinants of RNA recognition and cleavage by Dicer. *Nat Struc Mol Biol*, 14: 934-940.

[40] Qi X, Bao FS, Xie Z (2010) Small RNA deep sequencing reveals role for *Arabidopsis thaliana* RNA-dependent RNA polymerases in viral siRNA biogenesis. *PLoS ONE*, 4: e4971.

[41] Doyle J, Francis B, Tannenbaum A (1990) Feedback Control Theory. Macmillan, New York.

[42] Ebhart HA, Thi EP, Wang MB, Unrau PJ (2005) Extensive 30 modification of plant small RNAs is modulated by helper component-proteinase expression. *Proc Natl Acad Sci USA*, 102: 13398-13403.

[43] Mérai Z, Kerenyi Z, Kerstesz S, Magna M, Lakatos L, Silhavy D (2006) Double-stranded RNA binding may be a general plant RNA viral strategy to suppress RNA silencing. *J Virol*, 80: 5747-5756.

Chapter 6

Design of riboregulatory networks

...engineers create the world
that never was.
– Theodore von Karman

RNA-mediated regulations have been shown to be crucial in many cellular functions. In this chapter, we follow an optimization scheme to design the sequences of nucleic acids that implement riboregulatory circuits. Our ability to engineer those circuits will serve to gain quantitative insights about the design principles of riboregulation. We also develop a statistical mechanics model to predict the activity of the designed riboregulators.

6.1 Riboregulation in bacteria

That RNA molecules are not only information buffers but they play a decisive role in the complex regulatory map of the cell has constituted a breakthrough in our understanding of the central dogma of Molecular Biology [1, 2]. Indeed, small non-coding RNAs can silence gene expression [3], immunize the host against alien nucleic acids [4], block or induce translation [5], or even catalyze biochemical reactions [6]. Remarkably, what is common to all these regulatory mechanisms is a precise secondary structure that allows establishing interactions with further nucleic acids or even proteins while preventing degradation. Promptly, rational design techniques have been applied to exploit those RNA-mediated mechanisms, both in prokaryotes and eukaryotes, for metabolic control [7], logic gene silencing [8, 9], protein sensing and signaling [10], conditional cell death [11], activation of protein expression [12, 13], or transcription attenuation [14–16]. We now report the successful application of a novel computational, fully automated methodology to design original regulatory RNAs that can be experimentally validated *in vivo*.

Significant advances for working *in vitro* have been applied to design DNA-based computation circuits [17–19] and allosteric ribozymes [20]. *In vivo*, antisense RNA is among a conserved mechanism in all organisms relying on conformational changes in the system species [21], and it still offers wide potential applications to engineer synthetic RNA-based circuits for cell reprogramming [22]. However, the design of such circuits poses serious challenges for the rational techniques, especially by the necessity of accounting simultaneously for structures, free energies, partition functions, and allosteric motion. Certainly, all this information is difficult to manage without demanding for computational methods. In that way, optimization methods can rapidly find the sequences for the RNA circuits that are too complex for rational design approaches or too large for experimental library screening, and they are in fact the combination *in silico* of both techniques. Hence, similar to the development of the field of computational protein design more than one decade ago [23], optimization-based approaches embrace the fully *de novo* design of sequences of nucleic acids with regulatory ability.

In this chapter, we describe a novel automated strategy aimed to the *de novo* design of regulatory devices based on antisense RNA. We have validated the methodology by implementing several designs in the bacterium *E. coli.* Remarkably, our sequence selection algorithm is completely automated, from reading the design specifications. The algorithm optimizes simultaneously all RNA sequences, and during the optimization they are not imposed constraints based on natural systems, such as stems with high GC-content or loops with YUNR motifs [21, 24], thus providing unbiased synthetic sequences. Consequently, our designs are just based on low-level physicochemical principles and not on additional fitting. Thereby, we demonstrate that RNA folding models [25], which have been traditionally applied to disentangle the features of natural systems, can be also used to design artificial systems by solving an inverse problem.

The algorithm exploits the allosteric motion to design riboregulators of protein expression (Fig. 6.1), and it consists in a Monte Carlo Simulated Annealing optimization scheme [26] that assesses thousands, even millions, of different sequences. Because the secondary structures involve much higher energies than tridimensional architectures, we just modeled small RNAs at the 2D level [27]. Starting from random sequences, a mutation operator together with a bi-objective function served to evolve them towards the specific structures and interactions (Fig. 6.2A). The bi-objective function accounts, on the one hand, for the structures of the single species and also the complexes (structural term), and, on the other hand, for the free energies of the hybridization reactions (kinetic term). These two terms are then weighted in terms of energies resulting in a scalar optimization problem.

6.2 Computational method for sequence design

We developed an evolutionary algorithm to solve the inverse problem of interacting small RNA (sRNA) design. Riboregulation is based on conformational changes in the secondary structures of RNA molecules that allow controlling protein expression. In that way, the proper function of an RNA-based circuit relies on the structures of all species, since a disruption of the precise fold may result in a non-functional RNA, then affecting the circuit behavior. The annealing mechanism between two sRNAs is guided by the nucleotides that are not paired to form an intermediate complex (*e.g.*, kissing loops). Then, the stems next to that binding site from both sRNAs are destabilized to form a complex with another structure and minimal energy. We account for the hybridizing kinetics assuming a fast self-folding process, since its time scale is of microseconds whereas hybridization takes seconds or even minutes [28].

In our computational approach, the structures of all single species are design specifications (Fig. 6.2A). To address the computational design, a double inverse problem is formulated. First, we have to

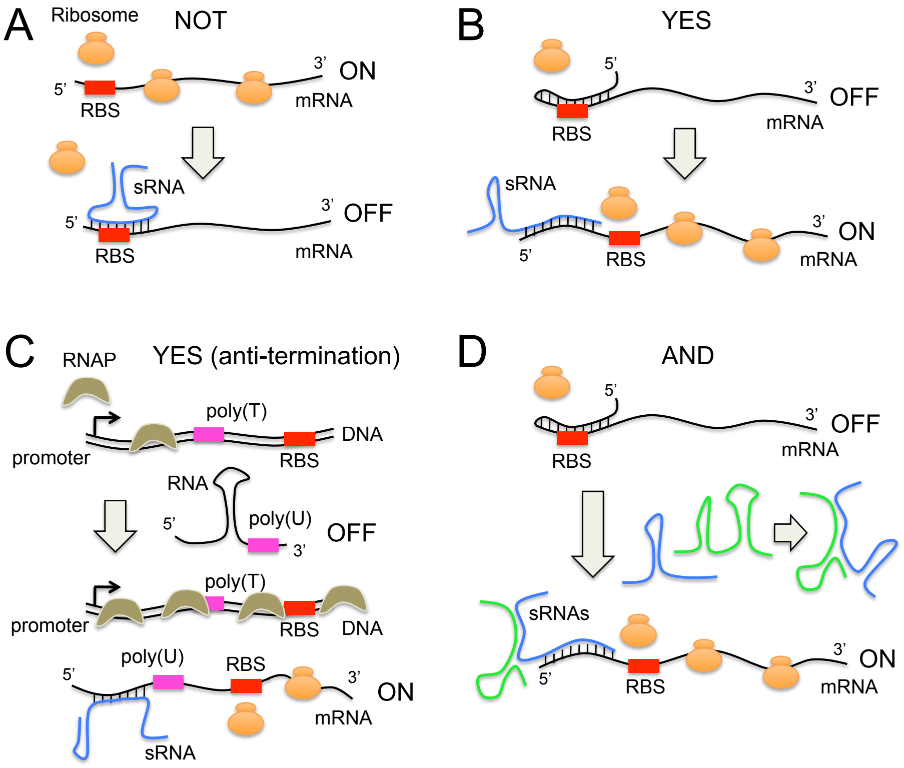


Figure 6.1: Riboregulation is based on conformational changes in the secondary structures of RNA molecules that allow controlling protein expression. The annealing mechanism between two sRNAs starts by the nucleotides that are not paired to form an intermediate complex and then follows to reach the structure of minimal energy. Herein, we illustrate different sRNA-based mechanisms to control protein expression. (A) One strategy to engineer a NOT gate consists in designing an sRNA able to bind to the RBS sequence to block translation. (B) On the contrary, to engineer a YES gate the sRNA is designed to release the RBS, which is trapped in a riboswitch. (C) Alternatively, a transcription terminator placed upstream the RBS prevents the formation of the mRNA. (D) In addition, two sRNAs can be designed to interact among them and form a complex that can release the RBS then implementing an AND gate.

find the sequences folding into the predefined structures and, second, find the sequences able to interact specifically among them to form the proper complexes to display the correct behavior. The structural constraints are exploited to considerably reduce the combinatorial space and accelerate the design of nucleic acid sequences. Our computational procedure optimizes at the same time all RNA sequences of the circuit. It consists in optimizing an objective function accounting for the stability and structure of the RNAs and the irreversibility of the reactions that lead to the target behavior. To compute the energy and folding of all species and complexes of a system, we have used the ViennaRNA [29] and MultiRNAFold [30] packages. The designed sequences were also analyzed with the NUPACK webserver [31].

The design specifications comprise the secondary structures of all single RNAs, critical subsequences of nucleotides (*e.g.*, ribosome binding site –RBS–), the reaction kinetics, and the structure of the output complex. The algorithm starts from random sequences satisfying the structural and subsequence constraints. If the subsequence constraints do not allow satisfying the structures, the algorithm stops. Eventually, it can be imposed a tolerance to account for species having similar structures of their targets. Consequently, an iterative process of mutation and selection is implemented. The mutation operator consists in either random or directed nucleotide replacements. We do not consider additions or deletions. A directed replacement is performed by taking a word (*i.e.*, set of consecutive nucleotides) from one sequence, making its reverse complementary, and randomly inserting it into another sequence. Initially, the length of this word is three, and it is reduced to one (*i.e.*, single point mutation) during the optimization process. In this work, we have considered much more chance for directed mutations because it speeds the in silico evolution. If a nucleotide that has to be mutated belongs to a stem, it is also mutated its pair in the stem with the corresponding nucleotide with the aim of preventing the disruption of the secondary structure and improving the convergence. We avoid sequences having consecutive repeats of four or more identical nucleotides. The objective function is a weighted sum of two terms to be minimized. The first term accounts for the kinetics of the system. For that, we compute the free energy release (ΔG) and the length of the toehold (α) of all possible reactions [18, 32], having

$$\Delta G_{kinetics} = \begin{cases} |\Delta G| + \alpha \Delta G_{eff}, \quad \text{if reaction OFF} \\ max(0, \Delta G_{max} - |\Delta G|) + max(0, \alpha_{max} - \alpha)\Delta G_{eff}, \\ \qquad \text{if reaction ON} \end{cases} \tag{6.1}$$

where $\Delta G_{max} = 15$ Kcal/mol and $\alpha_{max} = 6$ (saturation levels). $\Delta G_{eff} = 1.28$ Kcal/mol is an effective parameter to work in terms of free energies [32]. The specification of ON/OFF for the possible reactions between strands serves to define the behavior of the system. The second term accounts for the structural change of the output RNA. In the case of a YES gate, we impose the release of the RBS. For that, we use a Hamming distance (d) between the current and target structures

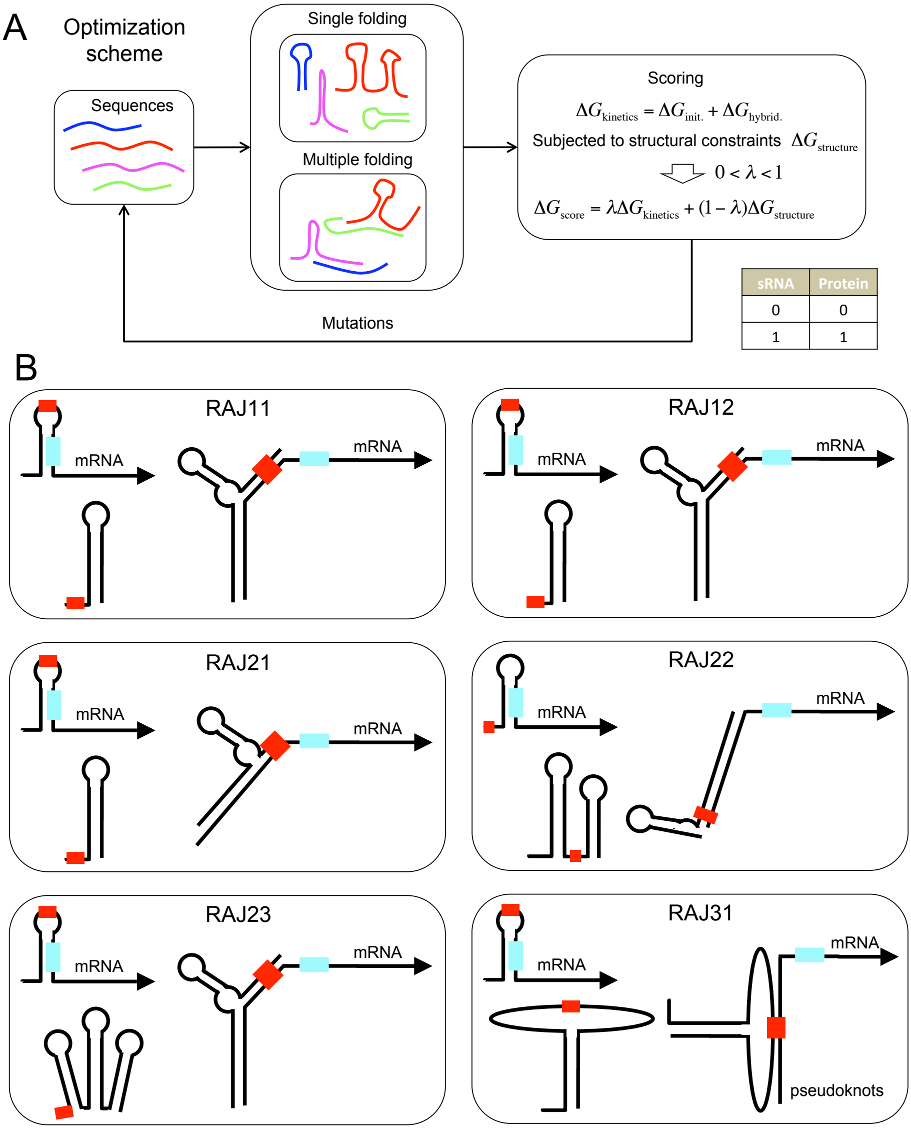
$$\Delta G_{structure} = d(S_{objective}, S_{complex})\Delta G_{eff}. \tag{6.2}$$

Thus, by selecting the λ factor between 0 and 1 (we usually select $\lambda = 0.5$), we can scalarize the problem resulting in

$$\Delta G_{score} = \lambda \Delta G_{kinetics} + (1 - \lambda)\Delta G_{structure}. \tag{6.3}$$

The algorithm converges rapidly ($\Delta G_{score} \rightarrow 0$) following an exponential scale and it can be launched in personal computers. It can be also launched in parallel in supercomputers to obtain multiple designs. The convergence scales with the number of species in the system and the complexity of their structures.

RNA molecules can fold into different structures (thermodynamic ensemble of structures). The free energy of each one (0 for the unfolded state) determines its probability of occurrence within the ensemble according to the Boltzmann factor. Our algorithm, instead of accounting for the whole ensemble, only accounts for the optimal solution, also called the minimal free energy (MFE) structure. To approach the whole population of structures by just the optimal structure, we have to assure that there is a given free energy gap between the MFE structure and any suboptimal one. Then, our idea relied on decreasing the free energy of the optimal point to guarantee that suboptimal structures are residual. This is of special interest for the formation of the complex. We have observed that reactions with moderated values of ΔG (about -5 Kcal/mol) do not ensure the major formation of the complex at the equilibrium. For that, lower values of ΔG are required (about -15 Kcal/mol).


A
Optimization scheme
Sequences
Single folding
Multiple folding
Scoring
$\Delta G_{\text{kinetics}} = \Delta G_{\text{init.}} + \Delta G_{\text{hybrid.}}$
Subjected to structural constraints $\Delta G_{\text{structure}}$
$0 < \lambda < 1$
$\Delta G_{\text{score}} = \lambda \Delta G_{\text{kinetics}} + (1 - \lambda) \Delta G_{\text{structure}}$
Mutations
B
RAJ11
mRNA
mRNA
RAJ12
mRNA
mRNA
RAJ21
mRNA
mRNA
RAJ22
mRNA
mRNA
RAJ23
mRNA
mRNA
RAJ31
mRNA
mRNA
pseudoknots

sRNA	Protein
0	0
1	1

Figure 6.2: Schemes of method and designs. (A) Optimization scheme followed to design the RNA devices. (B) Schematic representation of the six different RNA devices for riboregulation implementing a YES gate. Table 6.1 gives the thermodynamic properties of the systems.

6.3 Design of synthetic riboregulators

Our automated design method can be applied to generate a wide variety of RNA devices implementing logic computation circuits with high specificity (Fig. 6.1). To illustrate the efficacy of such an approach, we designed *cis*-repressing RBS elements in the 5'-untranslated region (5'-UTR) of a *gfp* gene. These elements are receptors of specific small RNAs working in *trans*, or riboregulators, with the ability of inducing a conformational change and then releasing the RBS. This mechanism implements a YES gate, where the GFP is expressed in presence of the riboregulator. In *E. coli*, the small RNA DsrA is responsible of activating in a similar fashion the expression of the sigma factor RpoS, which mediates the stress response [33]. In our designs, we used the RBS sequence 5'-AGGAGA, which is a small variant of the Shine-Dalgarno box (5'-AGGAGG), with a spacer of six nucleotides with the start codon. For a complete activation, we also imposed the release of the four nucleotides upstream of the RBS [34]. To highlight the versatility of our methodology, we specified different secondary structures for the riboregulators.

We designed six RNA devices implementing YES gates (Fig. 6.2B). By imposing the same design specifications, the algorithm could give singular sequences implementing different devices (RAJ11 and RAJ12). Moreover, the algorithm could accommodate different structural constraints for the riboregulator (RAJ21, RAJ22 and RAJ23). We selected naturally occurring ones from the bacterial small RNAs SokC, FinP, and DsrA [35]. Additionally, by taking advantage of a particular structure for the riboregulator [14], the algorithm could design devices based on pseudoknotted inter-molecular interactions (RAJ31). For illustrative purposes, Fig. 6.3 depicts the riboregulatory mechanism for the device RAJ11. A toehold of six nucleotides (5'-GGGAGG reading the riboregulator), complementary to the loop of the 5'-UTR structure, guides the hybridization, and the conformational change releases the twenty-one nucleotides upstream of the start codon, including the RBS. Table 6.1 shows the corresponding thermodynamic properties.

Given the library of riboregulatory modules, we investigated their orthogonality (*i.e.*, their mutual independence). For that, we tested the hybridization ability between the possible combinations of *cis*-repressing and *trans*-activating RNAs. Although the annealing by itself does not guarantee a given activation fold because the

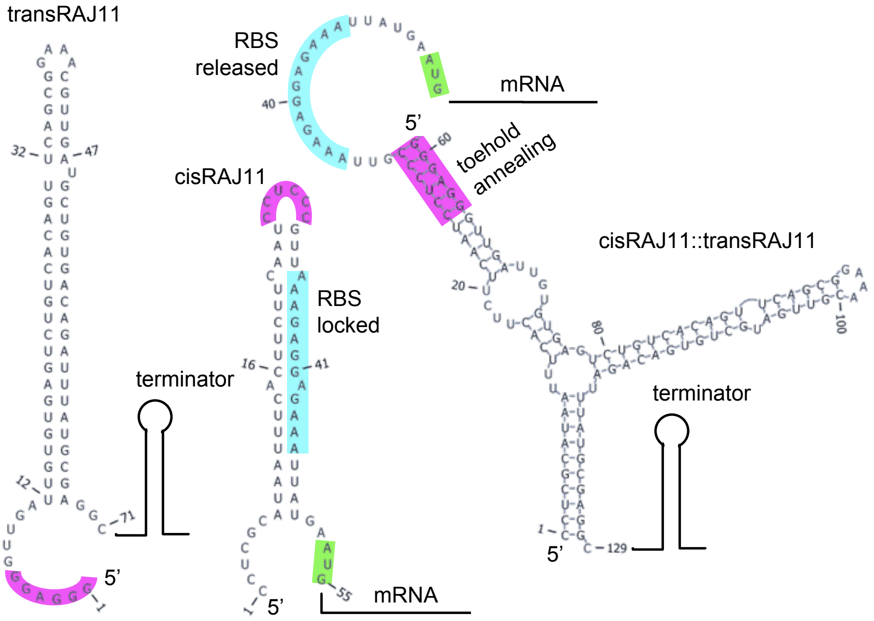


Figure 6.3: Illustration of the RNA device RAJ11. Sequences and secondary structures of the species of the system (cisRAJ11, transRAJ11, and the complex cisRAJ::transRAJ11). The RBS is colored in cyan, the toehold in magenta, and the start codon in green. Nucleotides are numbered relative to the predicted natural transcription start site; in the complex, transRAJ11 is numbered consecutive to cisRAJ11. The conformational change induced by the riboregulator release the RBS to activate translation.

resulting structure could still maintain the RBS locked, for an independent functioning of two RNA devices the species have to interact specifically. Computationally, we estimated the percentage of the complex at the equilibrium from the partition function, showing that our RNA devices, despite of the homologies found in the sequences and structures due to imposing a common RBS sequence, are highly orthogonal (Fig. 6.4). Therefore, by designing highly specific riboregulators, we are ensuring an elevated degree of orthogonality, which would allow implementing within the same cellular compartment complex systems by plugging the RNA devices as regulatory modules.

By harnessing the modularity of our designs, we engineered a genetic AND logic gate (Fig. 6.5). We placed the riboregulator

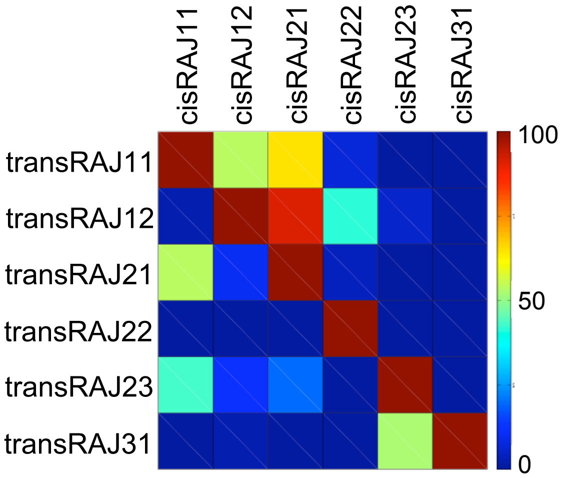


Figure 6.4: Computational prediction of specificity (probability of binding) between all designed RNA species. The results show high orthogonality between devices.

and the 5'-UTR-*gfp* under the control of tunable promoters. We used *tet* and *lac* promoters together with a strain that constitutively expressed the transcription repressors TetR and LacI [36]. These promoters respond to the inducers aTc and IPTG, respectively. To implement such a circuit, we took the device RAJ11. The resulting AND gate had a great overall activation fold and had low leakage. By setting high levels of IPTG, the concentration of aTc allows tuning the activation fold of the RNA device. Furthermore, a mathematical model allowed us to predict the surface response of the device. To construct this model, we used the derived equations, together with previously reported parameters for the *tet* and *lac* promoters and natural riboregulators. The engineering of this device demonstrates the integration of transcription and post-transcription control mechanisms to design synthetic gene circuits.

Apart of this library of YES gates, we have also applied the algorithm to design further regulatory modules. The minimal circuit would consist in one RNA species with repressive action working in *trans* (NOT function). This sRNA binds specifically to a segment of its target mRNA in order to inhibit translation. The most intuitive mechanism consists in blocking the RBS for preventing ribosome docking. For instance, in *E. coli* plasmid *F*, sRNA

Table 6.1: Thermodynamic properties of the designed RNA devices. The probability for specificty is estimated from the concentration of the complex in the equilibrium. The probability for RBS releasing is estimated from the ensemble of structures and the partition function.

Device	$-\Delta G$	$-G_{cis}$	$-G_{trans}$	Specificity	Releasing
	(Kcal/mol)	(Kcal/mol)	(Kcal/mol)	(%)	(%)
RAJ11	17.2	15.0	40.0	100	95
RAJ12	15.6	19.2	36.4	97	92
RAJ21	15.0	24.6	17.0	95	88
RAJ22	17.5	21.2	24.5	98	97
RAJ23	20.2	15.4	32.9	100	98
RAJ31	14.1	13.4	56.4	92	80

FinP directly binds to the RBS of protein TraJ [21]. Interestingly, the same sRNA can down-regulate many genes sharing a given RBS sequence. In Figs. 6.6A and B, we show two distinct computational designs of NOT gates. In addition, we have applied the algorithm to design riboregulatory activations based on the mechanism of anti-termination. Transcription terminators generally consist in simple hairpins of 10-15 base-pairs rich in GC-content followed by a poly(U) tail of 6-9 bases [15]. This structure entails the binding disruption of the RNA-polymerase. Hence, the idea behind this design consists in optimizing a *trans*-regulating RNA able to destabilize the structure of the terminator, which is the *cis*-regulating element, and form a new complex that allows the progression of the RNA-polymerase. In Fig. 6.6C, we show a computational design of a YES gate based on anti-termination. In the resulting structure of the complex, the terminator hairpin is purged and the poly(U) tail does not have any effect. Motivated by these previous results, we aimed at the design of combinatorial riboregulators. The regulatory function of multiple-sRNA complexes has not been reported in prokaryotes, which further fosters the exploration by means of computational methods. We illustrate the power of our approach by focusing on the design of synergistic activation (AND function), where two *trans*-regulating RNAs first interact among them to form a complex that will then activate a riboswitch. In Fig. 6.6D, we show a design of an AND gate. By themselves, the *trans*-regulating RNAs cannot release the RBS. However, the heterodimer they form has a distinct structure that allows interplaying with the *cis*-repressing element.

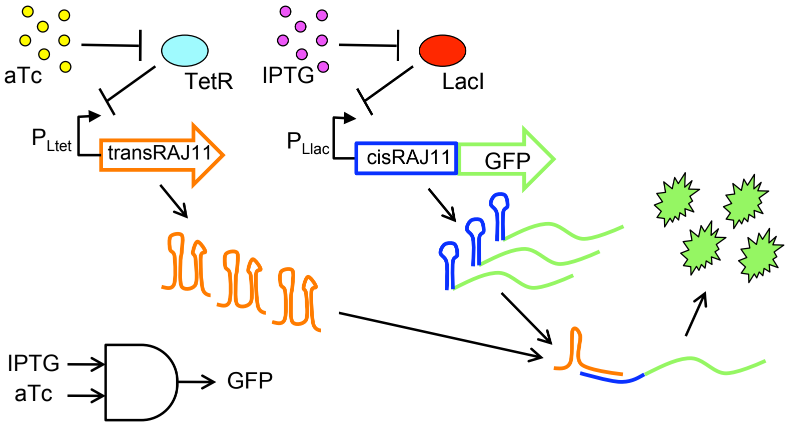


Figure 6.5: The RNA device can implement an AND gate. Scheme of a circuit coupling riboregulation with transcription control, where IPTG and aTc are the two inputs and GFP the output. To implement this circuit we used the RNA device RAJ11.

Importantly, given that the ribosome recruitment by heterologous systems is the primary cause of the growth rate reduction in bacteria [37, 38], the control of protein expression by riboregulators would be definitively less demanding for the cell. In that way, artificial systems expressing transcription factors will entail a much higher impact on the cell than those employing riboregulators. Being lower the impact, the heterologous system will reach a higher evolutionary stability.

6.4 Theoretical model for activity

We use the expression fold-change (f) as a metric of riboregulatory activity. For the case of a YES gate, this fold-change can be written as

$$f = \frac{(1 - P_{bind})r_0 + P_{bind}r_1}{r_0} = 1 + P_{bind}\left(\frac{r_1}{r_0} - 1\right), \qquad (6.4)$$

where r_0 is the mRNA translation rate (basal) and r_1 the sRNA::mRNA translation rate. P_{bind} is the probability of binding between the mRNA (*cis*) and sRNA (*trans*). We note that r_1 comes from the average over the thermodynamic ensemble of the complex.

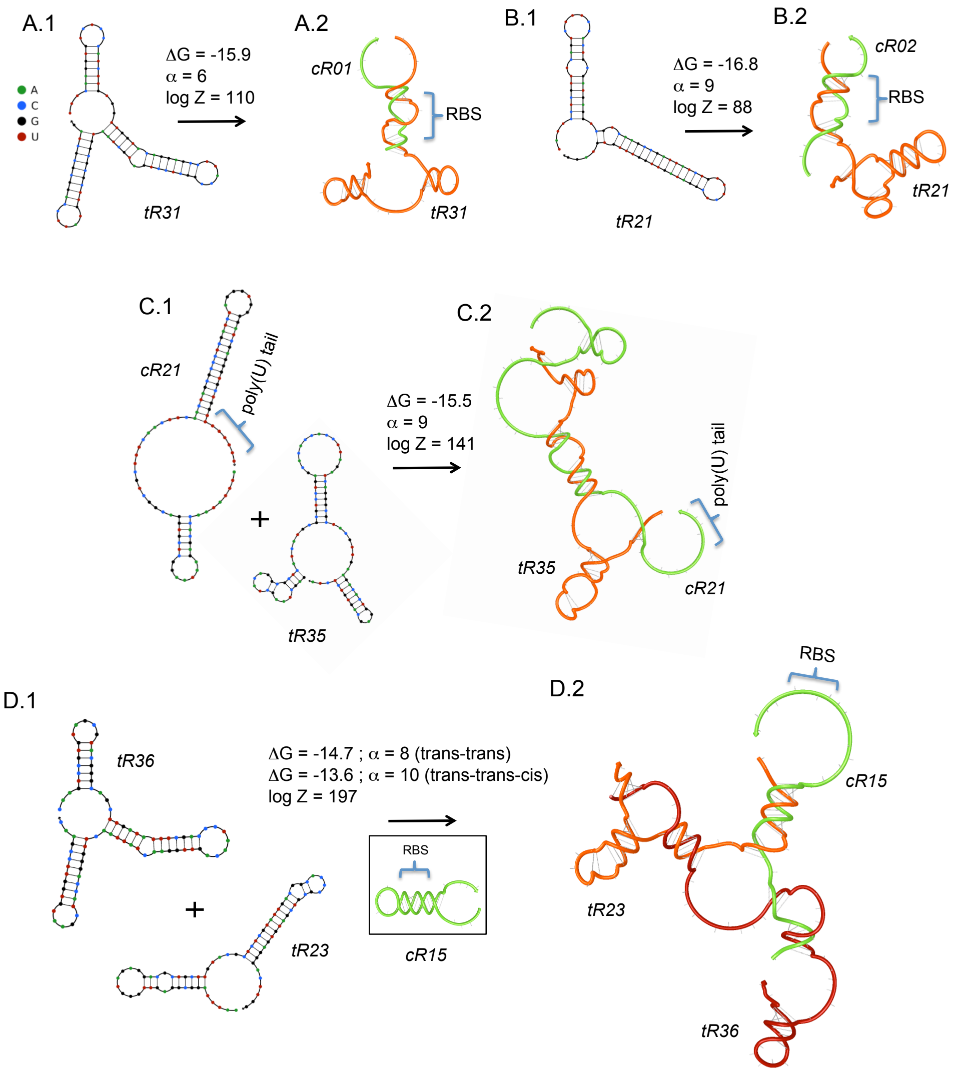


Figure 6.6: (A, B) Designs of NOT gates, using different naturally occurring secondary structures as scaffolds for the *trans*-repressing RNAs. Helical plots of the complex, where the RBS region is blocked, are shown. (C) Design of a YES gate based on anti-termination. Helical plot of the two-strand complex, where the hairpin before the poly(U) tail is destabilized, is shown. (D) Design of an AND gate. In the small inset we show the helical structure of the *cis*-regulating RNA (riboswitch). Helical plot of the three-strand complex, where the RBS region is released, is shown.

For simplicity we assume that

$$r_1 = \frac{1}{Z_{complex}} \sum_i r_1(G_i) e^{-G_i/kT} = P_{rbs} r_1' + (1 - P_{rbs}) r_1'', \tag{6.5}$$

where P_{rbs} is the probability of finding the RBS free in the complex; r_1' is the translation rate from a free RBS and r_1'' an effective rate from the other configurations of the ensemble. For a functional design, we then assume that $r_1' \gg r_1'' \gg r_0$ and that P_{rbs} is close to 1. P_{rbs} can be directly calculated by using the partition function $Z_{complex}$. Accordingly, we can write

$$f = 1 + P_{bind} P_{rbs} \frac{r_1'}{r_0}, \tag{6.6}$$

For simplicity we assume that r_1' is constant for all our RNA devices, because we use the same RBS sequence and because we imposed the same structural release on the complex. The basal rate is assumed to scale exponentially with $G_{cis} < 0$ (the free energy of the *cis*-repressing RNA) given by

$$r_0 \propto e^{B_0 G_{cis}}. \tag{6.7}$$

Accordingly, the lower is the free energy lower is the basal rate. Certainly P_{bind} depends on the relative concentration of sRNA/mRNA and the value of ΔG of the hybridization reaction. We introduce the dissociation constant (K_d), inverse of the equilibrium constant, and we denote by M_0 the total concentration of mRNA (*cis* element), by S_0 of sRNA (*trans* element) and by C of the formed complex. At the equilibrium we can write

$$\begin{aligned} M_0 &= M + C, \\ S_0 &= S + C, \\ MS &= K_d C. \end{aligned} \tag{6.8}$$

By solving this algebraic system, and denoting by σ the stoichiometry between sRNA and mRNA, we obtain the concentration of mRNA free given by

$$M = \frac{1}{2} \left(\sqrt{(K_d + (\sigma - 1)M_0)^2 + 4K_d M_0} - (K_d + (\sigma - 1)M_0) \right). \tag{6.9}$$

Thereby, we can write

$$P_{bind} = \frac{C}{M_0} = 1 - \frac{M}{M_0}. \tag{6.10}$$

The stoichiometry coefficient is assumed to scale exponentially with $G_{trans} < 0$ (the free energy of the *trans*-activating RNA) and it reads

$$\sigma = \frac{S_0}{M_0} = \sigma_0 e^{-B_1 G_{trans}}, \tag{6.11}$$

where σ_0 accounts for the relative promoter strength and gene dosage. In addition, assuming that the concentration of small RNA does not saturate the system and that, from statistical mechanics, we know that the equilibrium constant scales with $e^{-\Delta G/kT}$, we write

$$P_{bind} \propto \frac{\sigma}{1 + e^{(\Delta G + A)/kT}}, \tag{6.12}$$

where A is a parameter to be fitted (they may depend on different cellular factors). By producing computationally a large set of RNA devices, we found $A = 12.5$ Kcal/mol the threshold for binding. Therefore, we obtain a theoretical prediction of fold-change given by

$$f = 1 + kP_{rbs}\sigma_0 \frac{e^{-B_0 G_{cis} - B_1 G_{trans}}}{1 + e^{(\Delta G + A)/kT}}, \tag{6.13}$$

where k is a normalization constant. Because the sequences of our designs have been optimized to ensure a given energy gap, the use of the free energies of the optimal structures or the use of partition functions give similar results to calculate ΔG.

Despite of conceptual limitations and the lack of more accurate and comprehensive models, our theoretical model of riboregulation activity can address the causes of different performance in the designs (Fig. 6.7A). For instance, expression platforms where the riboregulator is in higher concentration than its target would lead to have more activity. In addition, it is expected that more stable riboregulators entailed a higher fold-change. Devices with lower ΔG would also entail better functionality. The model is adjusted to predict the experimentally reported activation fold of the six designs with high agreement (Fig. 6.7B). An interesting strategy to increase the fold-change of future designs could be the development of a more comprehensive model accounting for the interaction with the 16S rRNA [34] and using it to compute the objective function in the optimization algorithm. However, there are still many cellular factors that may influence the activation fold. In particular, bacterial RNase III is a potent and fast RNase that cleavages double strands from twelve base pairs [21], so bulges within the structure stems serve to prevent this degradation.

It is then expected certain dependence of the performance of the RNA devices on the expression of the RNase III. Another important issue is the impact of the RNA chaperone Hfq in the designs. It has been suggested that Hfq works as a helper for producing the hybridization between RNAs. Experimentally it has been shown that its expression is essential for the activity of natural riboregulators [39]. In addition, natural small RNAs integrate the transcription terminator within their structures. In fact, a poly(U) tail downstream of the last hairpin would act as such [15] and can be introduced as a design constraint. The addition *a posteriori* of a terminator over a designed sequence may alter the structure of the resulting strand and consequently the performance of the device. We can also rationalize that in absence of an efficient terminator the thermodynamic ensemble of riboregulators has much higher variability and then the set of functional configurations is reduced.

6.5 Discussion

The success of our design method relies on a tight coupling of theoretical principles, efficient numerical computation, and a decomposed empirical RNA interaction model. When combined, computational optimization methods can readily provide the sequences that implement riboregulatory circuits. The improvement in the accuracy of the physicochemical model together with the incorporation back into the design procedure of several cellular factors (*e.g.*, Hfq, RNase III, or 16S rRNA) would lead to the continuous development of our methodology. Given these first achievements, the prospect for pursuing even larger and comprehensive designs is outstanding. In addition, our method could be expanded to account for systems operating in eukaryotes, including mammalian cells. For instance, internal ribosome entry sites (IRES) [40] could play for the computational method the homolog role to the RBS in the bacterial designs. In that way, the design of future mammalian riboregulators would have important biomedical applications. A higher degree of sophistication would be therefore expected in the close future for the computationally designed RNA devices.

With the computational approach presented here and its experimental validation *in vivo*, we are opening new horizons for synthetic regulatory RNAs. In combination with the next-generation DNA synthesis techniques [41], this will permit the massive generation

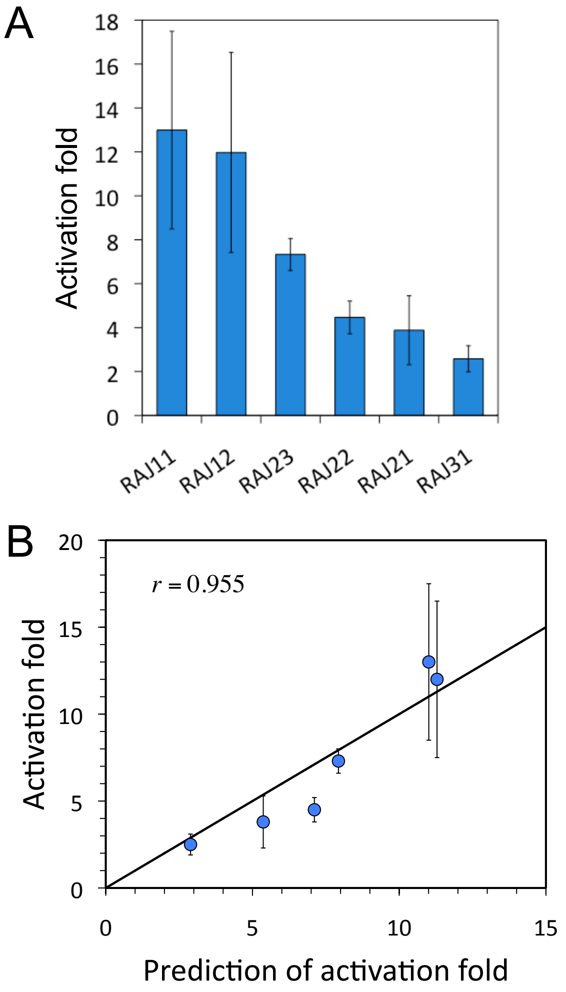


Figure 6.7: (A) Experimental characterization of the RNA devices given by the activation fold (f), measured as the ratio of GFP expressions. (B) Theoretical versus experimental activation fold (r gives the Pearson coefficient).

of RNA devices. Our devices could be advantageous for several biotechnological applications, where post-transcriptional control, specificity, and tunability are required features. Of relevance, the predictability of inter-/intra-molecular RNA interactions has allowed us to demonstrate for the first time that a fully automated design method can provide multiple reliable sequences of nucleic acids with regulatory ability that can be implemented *in vivo*. All in all, by designing genetic systems that never were, ultimately the goal of Synthetic Biology [42], we not only provide instrumental insights for cell reprogramming but we foresee the frontier of our knowledge about the life-driving principles.

The following publication holds the contents presented in this chapter

- Rodrigo G, Landrain TE, Jaramillo A (2011) *De novo* automated design of riboregulation. *Submitted.*

Bibliography

[1] Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2: 919-929.

[2] Ulveling D, Francastel C, Hubé F (2011) When one is better than two: RNA with dual functions. *Biochimie*, 93: 633-644.

[3] Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391: 806-811.

[4] Marraffini LS, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*, 11: 181-190.

[5] Itoh T, Tomizawa J (1980) Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proc Natl Acad Sci USA*, 77: 2450-2454.

[6] Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, 31: 147-157.

[7] Pfleger BF, Pitera DJ, Smolke CD, Keasling JD (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat Biotechnol*, 24: 1027-1032.

[8] Rinaudo K, Bleris L, Maddamsetti R, Subramanian S, Weiss R, Benenson Y (2007) A universal RNAi-based logic evaluator that operates in mammalian cells. *Nat Biotechnol*, 25: 795-801.

[9] Beisel CL, Bayer TS, Hoff KG, Smolke CD (2008) Model-guided design of ligand-regulated RNAi for programmable control of gene expression. *Mol Syst Biol*, 4: 224.

[10] Culler SJ, Hoff KG, Smolke CD (2010) Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science*, 330: 1251-1255.

[11] Venkataraman S, Dirks RM, Ueda CT, Pierce NA (2010) Selective cell death mediated by small conditional RNAs. *Proc Natl Acad Sci USA*, 107: 16777-16782.

[12] Isaacs FJ, Dwyer DJ, Ding C, Pervouchine DD, Cantor CR, Collins JJ (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nat Biotechnol*, 22: 841-847.

[13] Callura JM, Dwyer DJ, Isaacs FJ, Cantor CR, Collins JJ (2010) Tracking, tuning, and terminating microbial physiology using synthetic riboregulators. *Proc Natl Acad Sci USA*, 107: 15898-15903.

[14] Nakashima N, Tamura T (2009) Conditional gene silencing of multiple genes with antisense RNAs and generation of a mutator strain of *Escherichia coli*. *Nucl Acids Res*, 37: e103.

[15] Dawid A, Cayrol B, Isambert H (2009) RNA synthetic biology inspired from bacteria: construction of transcription attenuators under antisense regulation. *Phys Biol*, 6: 025007.

[16] Lucks JB, Qi L, Mutalik VK, Wang D, Arkin AP (2011) Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proc Natl Acad Sci USA*, 108: 8617-8622.

[17] Seelig G, Soloveichik D, Zhang DY, Winfree E (2006) Enzyme-free nucleic acid logic circuits. *Science*, 314: 1585-1588.

[18] Yin P, Choi HMT, Calvert CR, Pierce NA (2008) Programming biomolecular self-assembly pathways. *Nature*, 451: 318-322.

[19] Ran T, Kaplan S, Shapiro E (2009) Molecular implementation of simple logic programs. *Nat Nanotechnol*, 4: 642-648.

[20] Penchovsky R, Breaker RR (2005) Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nat Biotechnol*, 24: 545-554.

[21] Brantl S (2002) Antisense-RNA regulation and RNA interference. *Biochim Biophys Acta*, 1575: 15-25.

[22] Isaacs FJ, Dwyer DJ, Collins JJ (2006) RNA synthetic biology. *Nat Biotechnol*, 23: 1424-1433.

[23] Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science*, 278: 82-87.

[24] Franch T, Petersen M, Wagner EGH, Jacobsen JP, Gerdes K (1999) Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure. *J Mol Biol*, 294: 1115-1125.

[25] Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded

sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288: 911-940.

[26] Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science*, 220: 671-680.

[27] Flamm C, *et al.* (2000) RNA folding at elementary step resolution. *RNA*, 6: 325-338.

[28] Sosnick TR, Pan T (2003) RNA folding: models and perspectives. *Curr Opin Struct Biol*, 13: 309-316.

[29] Hofacker IL (2009) RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics*, 12: 12.2.

[30] Andronescu M, Zhang ZC, Condon A (2005) Secondary structure prediction of interacting RNA molecules. *J Mol Biol*, 345: 987-1001.

[31] Dirks RM, Bois JS, Schaeffer JM, Winfree E, Pierce NA (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev*, 49: 65-88.

[32] Yurke B, Mills AP Jr (2003) Using DNA to power nanostructures. *J Genet Prog Evol Mach*, 4: 111-122.

[33] Lease RA, Belfort M (2000) A trans-acting RNA as a control switch in *Escherichia coli*: DsrA modulates function by forming alternative structures. *Proc Natl Acad Sci USA*, 97: 9919-9924.

[34] Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*, 27: 946-950.

[35] Huang H-Y, Chang H-Y, Chou C-H, Tseng C-P, Ho S-Y, Yang C-D, Ju Y-W, Huang H-D (2009) sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucl Acids Res*, 37: D150-D154.

[36] Lutz R, Bujard H (1997) Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucl Acids Res*, 25: 1203-1210.

[37] Scott M, Gunderson CW, Mateescu WM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330: 1099-1102.

[38] Carrera J, Rodrigo G, Singh V, Kirov B, Jaramillo A (2011) Empirical model and in vivo characterization of the bacterial response to synthetic gene expression show that ribosome allocation limits growth rate. *Biotechnol J*, 6: 773-783.

[39] Sledjeski DD, Whitman C, Zhang A (2001) Hfq is necessary for regulation by the untranslated RNA DsrA. *J Bacteriol*, 183: 1997-2005.

[40] Hellen CU, Sarnow P (2001) Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes Dev*, 15: 1593-1612.

[41] Tian J, *et al.* (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, 432: 1050-1054.

[42] Lu TK, Khalil AS, Collins JJ (2009) Next-generation synthetic gene networks. *Nat Biotechnol*, 27: 1139-1150.

Chapter 7

Designability of metabolic pathways

There's plenty of room
at the bottom.
– Richard Feynman

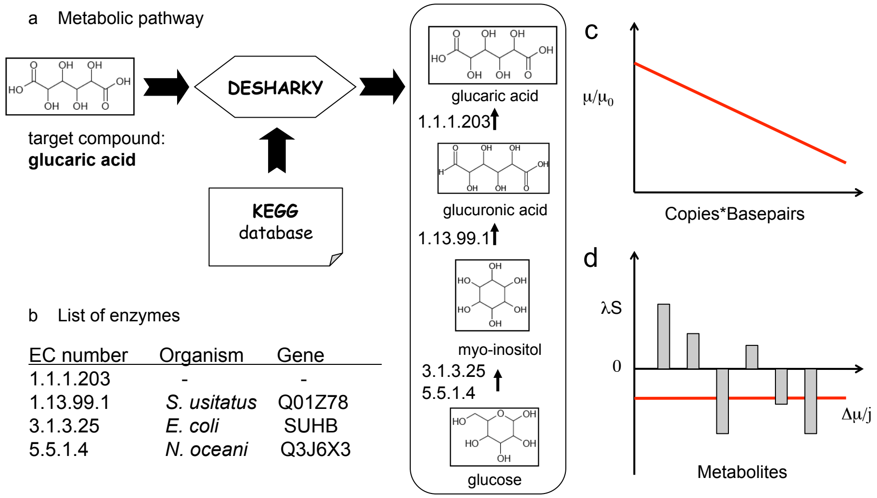
In this chapter, we extend our previous computational method to design and account for the designability of metabolic pathways, given a library of enzymatic reactions.

7.1 Metabolic engineering

Biotechnology process development is frequently equated with the production of biologics, such as proteins and viral vaccines [1]. Yet the use of biological systems for the production of small molecules goes back thousands of years and has been increasing since the discipline of metabolic engineering was defined fifteen years ago [2]. Initially, metabolic engineering efforts were primarily focused on improving the productivity of naturally occurring metabolites within an organism, such as for overexpressing glycolytic enzymes in yeast [3]. More recently, the field has expanded to encompass a number of examples of introducing new enzyme activities into a host cell in order to produce non-natural products [4, 5] or to engineer degradation of toxic compounds [6].

The use of automated techniques to design biological systems constitutes a breakthrough in biotechnology, and it has previously been applied to predict biodegradation pathways [7, 8]. Interestingly, functional approaches [7, 9, 10] could reveal novel pathways, but these are ultimately limited by the availability of naturally-occurring enzymes. In that sense, recent work shows how to construct biochemical pathways using atomic information [11, 12], and this approach could be used to enlarge our enzyme database by adding abstract reactions corresponding to functional enzymes. This would allow the design of metabolic pathways that incorporate enzymes not found in nature but which could be engineered by directed evolution or using computational design [13]. In this work we propose to go beyond by extending the design to biosynthesis and predicting the cell behavior when implementing a pathway in a given host using plasmids [14].

On the other hand, one of the major challenges in Synthetic Biology is engineering as far as possible orthogonal systems [15]. In that way, quantitative models provide fruitful insights. We propose the use of two different models to quantify the readjustment of fluxes [16] and the consumption of cellular resources [17] that results from the expression of heterologous pathways. We select the growth rate as the control parameter for the cellular behavior evaluation. From the transcriptional approach, we consider a dynamical model involving RNAs, RNA-polymerases, proteins and ribosomes [18, 19]. Accordingly, we compute the reduction in the growth rate due to the sequestration of RNA-polymerases and ribosomes. On the other hand, since the cell is metabolically altered, we use Flux Balance Analysis (FBA) to predict the new growth rate. These two strategies give different predictions about the cell behavior, but they constitute two scores to be considered when implementing a designed pathway. Further approaches will use more complex models by integrating the metabolic and transcriptomic systems, and also taking advantage of databases of Gibbs free energies for all enzymatic reactions [20]. Importantly, as the desired route could be not unique, we provide a methodology to rank different pathways according to their genetic/metabolic loads.



EC number	Organism	Gene
1.1.1.203	-	-
1.13.99.1	*S. usitatus*	Q01Z78
3.1.3.25	*E. coli*	SUHB
5.5.1.4	*N. oceani*	Q3J6X3

Figure 7.1: (a) Scheme of the design procedure. Application to glucaric acid biosynthesis. (b) List of enzymes involved the designed pathway. In (c) genetic load: the plasmid copy number times the length of all enzymes is assumed to be proportional to the fraction of ribosomes allocated for heterologous expression. In (d) metabolic load: list of the shadow prices for all cofactors required in that pathway and the source compound (D-Glucose-6-Phosphate). The growth rate variation comes from the stoichiometric sum of these shadow prices.

7.2 Computational method

We have developed a Monte Carlo algorithm with the aim of designing metabolic pathways. The purpose is to find a possible route connecting a given compound of interest with a metabolite from the considered hosting organism. These routes can be for biodegradation (reactant as source) or biosynthesis (product as source). For the source compound, we find the possible enzymatic reactions and select one among them with equitable probabilities. We repeat this process for the new source compound. Moreover, we consider with a given probability a move to go back, removing the previous reaction, to improve the convergence and to avoid long pathways. This probability is a function of the number of the already introduced steps, as the longer the pathway, the higher is the probability to go back, and here we have used a sigmoid function. We do not consider metabolic steps involving many compounds which are not specific to the hosting organism (here, one non-specific reactant and one product at most).

The microbial production or degradation of chemical compounds usually requires the expression of foreign enzymes. This expression consumes cellular resources such as RNA-polymerases and ribonucleotides for transcription, and ribosomes and amino acids for translation. Using previous knowledge on heterologous expression, we assume that RNA-polymerases ($RNAP_h$) and ribosomes (RIB_h) are the two critical pools [19]. Using the experimental measurements of these resources in *E. coli* [17], we have constructed an empirical model of the chassis (host resources for heterologous expression). Furthermore, we have modeled the total heterologous expression of RNA (R_h) by

$$\frac{d}{dt}R_h = \phi C - (\delta_r + \mu)R_h, \tag{7.1}$$

and enzymes (E_h) following

$$\frac{d}{dt}E_h = \psi R_h - (\delta_e + \mu)E_h, \tag{7.2}$$

where ϕ is the average transcription rate, C the number of copies of external DNA, ψ the average translation rate, μ the cell growth rate, and δ_r and δ_e the degradation rates of the RNA and enzymes, respectively. Hence, a first order approach is to compute the consumption of cellular resources by the heterologous system ($RNAP_h = \phi C t_r$ and $RIB_h = \psi R_h t_p$, where t_r is the transcription time and t_p the translation time) and then to recompute the growth rate using the phenomenological chassis model (Fig. 7.1).

We have adressed the metabolic burden with FBA [16]. This linear program, in which we maximize the cell growth rate (μ), can be written as

$$\begin{aligned} \max\ \mu &= cv \\ \text{subject to } Sv &= b, \end{aligned} \tag{7.3}$$

where v are the cell metabolic fluxes, c their contributions to the growth rate, S the stoichiometry matrix, and b the uptake fluxes. Then, we have constructed the corresponding dual problem [21], which is equivalent to its primal, given by

$$\begin{aligned} \min\ \mu &= \lambda b \\ \text{subject to } \lambda S &= c, \end{aligned} \tag{7.4}$$

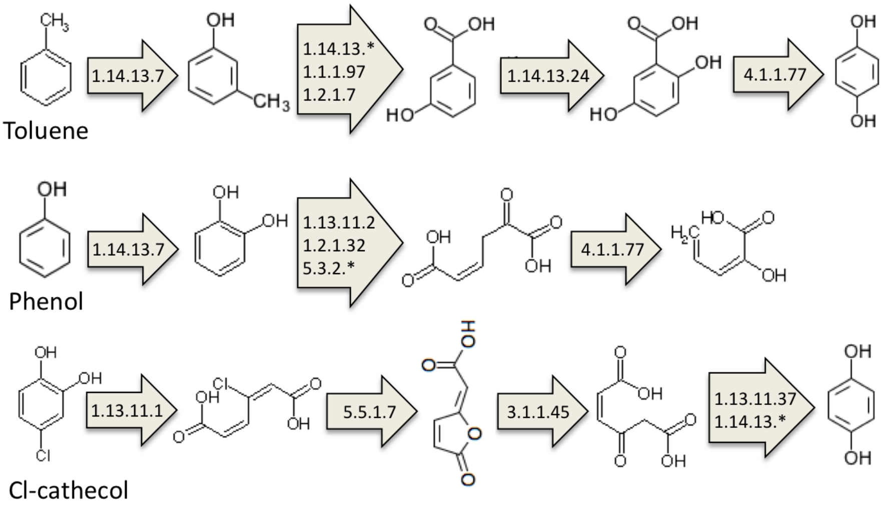


Figure 7.2: Examples of metabolic pathways designed with the algorithm for biodegradation. The final product is a metabolite present in the host organism.

where λ, usually called shadow prices, are the contributions to the growth rate when perturbing the uptake fluxes ($\Delta\mu = \lambda\Delta b$). Therefore, we can precompute λ since it is a property of the host organism. In that way, the fact of introducing a new metabolic route in the host can be treated in a perturbative way. Then, $\Delta b = S^{*}j$ where S^{*} is the stoichiometry matrix for this pathway and j its flux.

Here we have taken *E. coli* as the cell model. We have used an extended description of *E. coli* metabolism involving 1,039 compounds, including extracellular compounds, and 2,381 biochemical reactions [22]. We provide the KEGG [23] databases for chemical compounds and enzymatic reactions in a depured format. There are 14,965 chemical compounds, of which 826 are present in the host, 4,942 enzymes, of which 2,350 have available their sequence, and 7,400 enzymatic reactions from 650 organisms. Also we consider a set of compounds eventually in the medium that can be used as substrates by the cell. To enlarge the capabilities of the algorithm, we can assume reversible reactions. In addition, we can introduce reactions which are not found in KEGG. The input of our algorithm is the target compound. The output is the designed metabolic pathway together with the quatification of the transcription, translation and metabolic load. In addition, we provide the sequence of amino acids of the enzymes involved in the pathway. These sequences are the closest phylogenetically to *E. coli* according to the KEGG classification of organisms. Here we have assumed an initial growth rate of $\mu_0 = 2$ doublings/h, a transcription kinetics of $\phi = 0.1$ RNA-polymerases/s, a translation kinetics of $\psi = 0.4$ ribosomes/s, a number of DNA copies for the enzymes of $C = 100$, a transcription velocity of $1/t_r = 45$ nt/s, a translation velocity of $1/t_p = 16$ aa/s, and a metabolic pathway flux of $j = 1$ mmol/(gh) (arbitrary values).

7.3 Metabolic designability

We applied the algorithm to design several metabolic pathways including biodegradation of toluene or phenol and bioproduction of sorbitol and glucaric acid. For instance, the microbial production of glucaric acid is important for therapeutic purposes including cholesterol reduction and cancer chemotherapy, and for the synthesis of new nylons and hyperbranched polyesters. In Fig. 7.1 we show design of this pathway, including biochemical transformations and the list of genes encoding the corresponding enzymes. In addition, we compared the biodegradation pathways we found with those obtained from UM-BBD [7] showing alternative routes (Fig. 7.2).

Furthermore, we studied the capabilities of the KEGG database to design metabolic pathways of interest. Likewise, we randomly selected a set of 196 compounds of the 14,965 from KEGG. For those, we executed the algorithm. We found that only 37 of them (about the 20%) can be connected with the *E. coli* metabolism, considering all enzymatic reactions as reversible (Fig. 7.3). Subsequently, we executed the algorithm 600 times for those 37 compounds. Then, we counted the distinct pathways, removing the pathways with loops. The probability distribution shown in Fig. 7.4 has a mean of 18 and a median of 4. The probability to have a compound with more than 100 different pathways is close to 0. This reveals that the number of possible pathways, according to KEGG, for a given compound is reduced and then Monte Carlo techniques allow us to explore all the solution space.

In addition, we performed an exhaustive study to obtain all possible pathways for a set of compounds with special interest. For this study we considered the following compounds: C00116 (Glycerol), C00146 (Phenol), C00379 (Xylitol), C00794 (Sorbitol), C00818 (Glucaric-acid), C01407 (Benzene), C01455 (Toluene), and

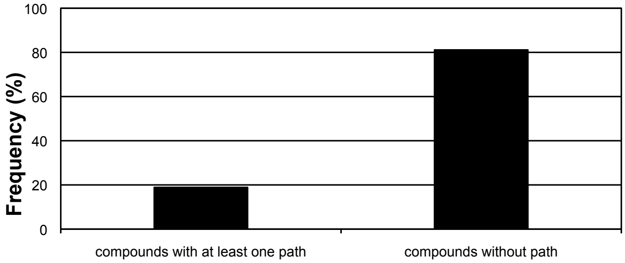


Figure 7.3: Distribution of possible pathways connecting a target compound with the *E. coli* metabolism according to the KEGG database.

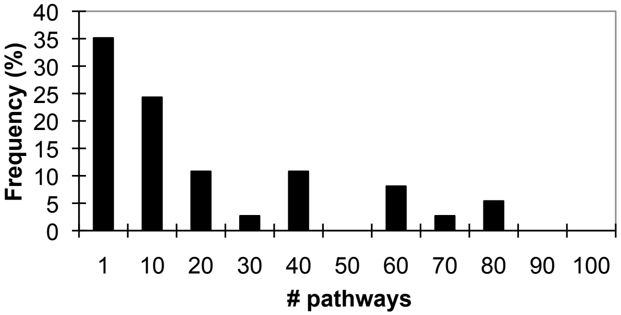


Figure 7.4: Histogram of the possible pathways between a given compound and the *E. coli* metabolism.

C02375 (4-Chloro-catechol). For that, we considered all reactions as reversible and a selected set of host compounds. As Monte Carlo techniques can efficiently sample the solutions space, we applied our method 300 times for each compound and stored all the resulting designs. Then, we counted the distinct pathways (Fig. 7.5). Pathways with metabolic loops were removed. In principle, those routes give the solution network to connect the specified compound with the metabolic chassis. The algorithm filters the designed pathways to avoid loops of 2 metabolites when using reversible reactions. Accordingly, our algorithm can be also used to find exhaustively all possible metabolic pathways.

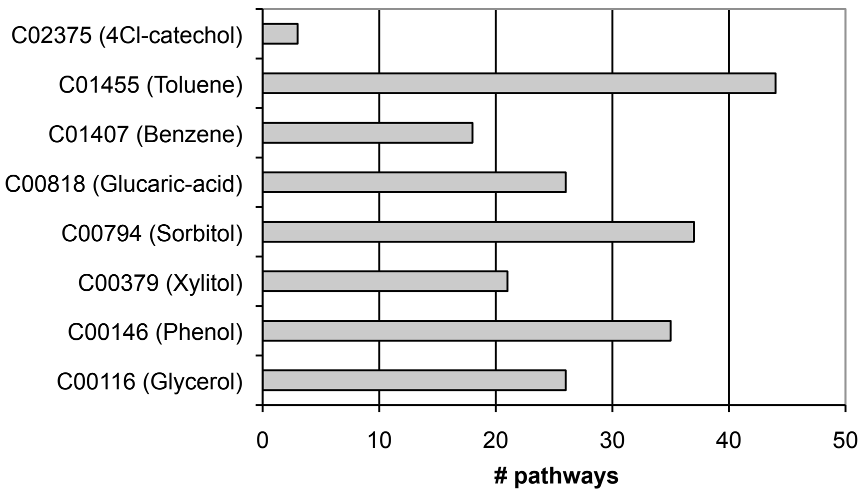


Figure 7.5: Number of possible pathways for the compounds C00116 (Glycerol), C00146 (Phenol), C00379 (Xylitol), C00794 (Sorbitol), C00818 (Glucaric-acid), C01407 (Benzene), C01455 (Toluene), and C02375 (4-Chloro-catechol).

7.4 Molecular hydrogen bioproduction

Cells use protons and electrons for synthesis of ATP and NAD(P)H, and ultimately for fixation of CO_2 to produce carbohydrates such as starch and glycogen. These organic compounds can be considered as stores of protons and electrons for further H_2 production, which has great potential as an environmentally clean energy fuel. Under anaerobic conditions, in which oxidative phosphorylation is blocked, many organisms have evolved a survival mechanism that extracts energy from these stores, coupled to creating gradients of protons to maintain the essential ATP level. This ATP synthesis is driven by the chemiosmotic gradient of protons and by phosphorylated intermediates produced during fermentation. Electroneutrality demands the release of both the protons and electrons, which in the absence of O_2 are ultimately recombined to produce H_2 instead of H_2O as in the aerobic case [24].

The reaction $2H^+ + 2e^- \rightarrow H_2$ is catalyzed in several microorganisms by metalloenzymes known as hydrogenases. In particular, cyanobacteria have evolved the biochemical machinery necessary to photoproduce molecular hydrogen from water. However, under normal conditions where oxygen is a byproduct of photosynthesis, sustained photoproduction is not possible since oxygen is a powerful inhibitor of most hydrogenases. Therefore, using a computational approach we found several oxygen consuming pathways to be engineered in the cyanobacterium with the aim of provoking an anaerobic environment within the cell. These devices are implemented by a set of enzymes that catalyze new metabolic routes aiming for the oxidation of natural compounds of the cyanobacterium. We designed all possible metabolic pathways of one or two reactions. In addition, we implemented a catalase (encoded in gene *katG* in many bacteria) when H_2O_2 is a product of the pathway. The catalase transforms the peroxide into water (*i.e.*, $2\ H_2O_2 \rightarrow O_2 + 2\ H_2O$). We also eliminated those pathways in which the enzymes did not have a known sequence. We imposed that the substrates for these reactions have to be present in the wild-type metabolism of *Synechocystis*. For that, we used the metabolic model presented by Fu and co-workers [25]. Moreover, we checked that the pathway stoichiometry allows the oxygen consumption. In Table 7.1 we summarize all obtained pathways.

The first strategy consists in oxidizing an organic compound such as glucose, glycerol, or galactose. For instance, a glucose oxidase from *A. niger* transforms β-D-glucose into D-glucono-1,5-lactone. However, the principal inconvenient of these pathways is the consumption of cellular resources, together with the production of a peroxide. Additionally, the enzyme H16-B1863 from *R. eutropha H16* can be engineered in the cyanobacterium to oxidize L-cysteine at the amino acid depletion expense. Another proposed reaction is the oxidation of S-dihydroorotate by overexpressing PyrD. Nevertheless, the enzyme PyrD belongs to the biosynthetic pathway of pyrimidines and its overexpression could induce a suboptimal state of the cell. Alternatively, the oxidation can be of inorganic compounds such as nitric oxide, which is transformed into nitrate using the enzyme Hmp from *E. coli*, although *Synechocystis* has a similar protein (NorB) and could produce an undesired interference, noting that nitric oxide is a toxic compound. Another option consists in the acidification of compounds such as pyridoxamine-5P and pyridoxine-5P (overexpressing PdxH), hypoxanthine and xanthine (AO090003001099 from *A. oryzae*), or sulfite (H16-B0860 from *R. eutropha H16*). The eventual issues of these pathways could be that

PdxH is involved in vitamin B6 synthesis that is toxic in excess, AO090003001099 produces, under some conditions, superoxides rather than peroxides, and H16-B0860 can interfere with the sulfur metabolism.

7.5 Discussion

Our tool uses a heuristic algorithm based on Monte Carlo to find a possible route connecting a specified target metabolite with the host metabolism, instead of using a pathway selection by enumeration of all possible metabolic routes [11, 26]. The algorithm finds a proper pathway and computes its associated genetic/metabolic load in a few seconds. In addition, our software can be used in distributed computing to sample most of the solution space. For illustration purposes, we have shown several biodegradation routes. Here, we have assumed non-weighted reactions for the heuristic procedure and we compute the genetic/metabolic load *a posteriori* using the transcription and metabolic models. Alternatively, a global optimization could be addressed by considering the load of each reaction during the heuristic procedure [27].

The design of new functional enzymes able to catalyze reactions for which natural catalysts are not available will open new avenues for the *de novo* design of metabolic networks. Protein design has been taking advantage of automated methods to engineer new enzymes [13] but the automatic design of enzyme networks is a different problem. The tremendous increase of known enzyme sequences from genomic and metagenomic projects [23], together with the appearance of combinatorial methods for gene synthesis [28], will facilitate the experimental test of novel synthetic metabolic pathways. Furthermore, metabolic models can be combined with transcription elements [29, 30] for suggesting different combinations of enzyme/regulator deletions/additions in order to optimize the production of specific biochemical compounds by restricting to a minimum cell growth rate. Likewise, the up/down-regulation of transcription factors serves as a tuning mechanism of the metabolic capabilities of the cell. In that way, the integration of gene regulations as metabolism-conditioning elements allows the prediction and, ultimately, the design of complex multilayer networks. Nevertheless, the automatic design of those networks is at last limited by the available information. Thereby, the *de novo* design approaches for metabolic engineering together with

Table 7.1: Enzymatic reactions designed to induce a local anaerobic environment within the cell by consuming molecules of oxygen. We used *Synechocystis* as bacterial chassis.

<table>
<tr><th>Enzymes</th><th>Organisms</th><th>Genes</th><th>Feature</th></tr>
<tr><td>1.1.3.4
1.11.1.6</td><td><i>A. niger</i>
<i>Synechocystis</i></td><td>gox
katG</td><td>Consumption of cellular resources.</td></tr>
<tr><td colspan="4">b-D-glucose + O_2 = D-glucono-1,5-lactone + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
<tr><td>1.14.12.17</td><td><i>E. coli</i></td><td>hmp</td><td><i>Synechocystis</i> has a similar protein (NorB). Also, nitric oxide is toxic.</td></tr>
<tr><td colspan="4">2 Nitric oxide + 2 O_2 + NAD(P)H = 2 Nitrate + $NAD(P)^+$ + H^+</td></tr>
<tr><td>1.13.11.20</td><td><i>R. eutropha H16</i></td><td>H16_B1863</td><td>Amino acid (cys) depletion.</td></tr>
<tr><td colspan="4">L-Cysteine + O_2 = 3-Sulfino-L-alanine</td></tr>
<tr><td>1.1.3.21
1.11.1.6</td><td><i>P. haloplanktis</i>
<i>Synechocystis</i></td><td>PSHAb0547
katG</td><td>Equivalent to glucose oxidase. Consumption of resources.</td></tr>
<tr><td colspan="4">sn-Glycerol + O_2 = Glycerone + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
<tr><td>1.3.3.1
1.11.1.6</td><td><i>Synechocystis</i></td><td>pyrD
katG</td><td>PyrD belongs to the biosynthetic pathway of pyrimidines.</td></tr>
<tr><td colspan="4">(S)-Dihydroorotate + O_2 = $Orotate^+$ + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
<tr><td>1.4.3.5
1.11.1.6</td><td><i>Synechocystis</i></td><td>pdxH
katG</td><td>PdxH is involved in vitamin B6 synthesis that is toxic in excess.</td></tr>
<tr><td colspan="4">Pyridoxamine 5P + H_2O + O_2 = Pyridoxal 5P + NH_3 + H_2O_2
Pyridoxine 5P + O_2 = Pyridoxal 5P + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
<tr><td>1.1.3.9
1.11.1.6</td><td><i>B. pseudomallei</i>
<i>Synechocystis</i></td><td>BURPS1710b_1758
katG</td><td>Equivalent to glucose oxidase. Consumption of resources.</td></tr>
<tr><td colspan="4">D-galactose + O_2 = D-galacto-hexodialdose + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
<tr><td>1.17.3.2
1.11.1.6</td><td><i>A. oryzae</i>
<i>Synechocystis</i></td><td>AO090003001099
katG</td><td>Under some conditions the product is mainly superoxide rather than peroxide.</td></tr>
<tr><td colspan="4">Hypoxanthine+ O_2+ H_2O = Xanthine + H_2O_2
Xanthine + H_2O + O_2 = Urate + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
<tr><td>1.8.3.1
1.11.1.6</td><td><i>R. eutropha H16</i>
<i>Synechocystis</i></td><td>H16_B0860
katG</td><td>Interference with the sulfur metabolism.</td></tr>
<tr><td colspan="4">Sulfite + O_2 + H_2O = Sulfate + H_2O_2
2 H_2O_2 = O_2 + 2 H_2O</td></tr>
</table>

the design of synthetic enzymes by directed evolution, computational design, or a combination of both could overcome the lack of specific reactions.

The following publication holds the contents presented in this chapter

- Rodrigo G, Carrera J, Prather KJ, Jaramillo A (2008) Desharky: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, 24: 2554-2556.

Further reading

- Landrain T, Carrera J, Kirov B, Rodrigo G, Jaramillo A (2009) Modular model-based design for heterologous bioproduction in bacteria. *Curr Opin Biotechnol*, 20: 272-279.

Bibliography

[1] Nielsen J (2001) Metabolic engineering. *Appl Microbiol Biotechnol*, 55: 263-283.

[2] Bailey JE (1991) Toward a science of metabolic engineering. *Science*, 252: 1668-1675.

[3] Schaaff I, Heinisch J, Zimmermann FK (1989) Overproduction of glycolytic enzymes in yeast. *Yeast*, 5: 285-290.

[4] Martin JJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003) Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat Biotechnol*, 21: 796-802.

[5] Ro DK, Paradise EM, Ouellet M, Fisher KJ, *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440: 940-943.

[6] Haro M-A, de Lorenzo V (2001) Metabolic engineering of bacteria for environmental applications: construction of *Pseudomonas* strains for biodegradation of 2-chlorotoluene. *J Biotechnol*, 85: 103-113.

[7] Hou BK, Wackett LP, Ellis LBM (2003) Microbial pathway prediction: a functional group approach. *J Chem Inf Comput Sci*, 43: 1051-1057.

[8] Pazos F, Guijas D, Valencia A, de Lorenzo V (2005) MetaRouter: bioinformatics for bioremediation. *Nucl Acids Res*, 33: D588-D592.

[9] Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem Eng Sci*, 59: 5051-5060.

[10] Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics*, 21: 1603-1609.

[11] Arita M (2003) In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism.

Genome Res, 13: 2455-2466.

[12] Arita M (2004) The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA*, 101: 1543-1547.

[13] Rothlisberger D, Khersonsky O, Wollacott AM, *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature*, 453: 190-195.

[14] Jones KL, Kim SW, Keasling JD (2000) Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. *Metabolic Eng*, 2: 328-338.

[15] Sprinzak D, Elowitz MB (2005) Reconstruction of genetic circuits. *Nature*, 438: 443-448.

[16] Varma A, Palsson BO (1994) Metabolic Flux Balancing: basic concepts, scientific and practical use. *Bio/Technology*, 12: 994-998.

[17] Bremer H, Dennis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In Ed. Neidhardt FC, *et al. Escherichia coli* and *Salmonella.* ASM Press, Washington DC.

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

[19] Scott M, Gunderson CW, Mateescu WM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330: 1099-1102.

[20] Mavrovouniotis ML (1991) Estimation of standard Gibbs energy changes of biotransformations. *J Biol Chem*, 266: 14440-14445.

[21] Schrijver A (1998) Theory of Linear and Integer Programming. John Wiley & Sons, New York.

[22] Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli. Mol Syst Biol*, 3: 119.

[23] Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl Acids Res*, 28: 27-30.

[24] Rupprecht J, Hankamer B, Mussgnug JH, Ananyev G, Dismukes C, Kruse O (2006) Perspectives and advances of biological H2 production in microorganisms. *Appl Microbiol Biotechnol*, 72: 442-449.

[25] Fu P (2009) Genome-scale modeling of *Synechocystis* sp. PCC 6803 and prediction of pathway insertion. *J Chem Technol Biotechnol*, 84: 473-483.

[26] Eppstein D (1998) Finding the k shortest paths. *SIAM J Comput*, 28: 652-673.

[27] Croes D, Couche F, Wodak SJ, van Helden J (2006) Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol*, 356: 222-236.

[28] Gibson DG, Young L, Chuang RY, Venter JC, Hutchison III CA, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth*, 6: 343-345.

[29] Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429: 92-96.

[30] Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: A computational framework for redesign of microbial production systems. *Genome Res*, 14: 2367-2376.

Chapter 8

Hierarchical regulatory networks: a natural design

I just wondered how things
were put together.
– Claude Shannon

So far we have studied networks at a small scale. However, cellular functions are orchestrated by hundreds of interlinked regulatory networks. In this chapter, we analyze how the cell canalizes the signals and responds to a biotic perturbation. This study would serve to elucidate design principles of genome-scale networks.

8.1 The case of plant viruses

For over decades, plant molecular virology has been primary focused on the pathogen itself, studying their individual genes and products, and their local effects on certain regulatory pathways related to antiviral responses. However, with the arrival of modern genomic tools allowing for high-throughput screenings, we can now tackle the problem of the plant host-virus interaction from a systemic perspective that would allow us reaching a deeper understanding on how host and virus genotypes, environmental effects and stochasticity interplay in determining the pathological outcome of an infection. Viral infections typically alter host physiology, notably by diverting cellular resources for the production of virus-specific components, and by actively suppressing host defenses [1, 2]. As a response to infection, hosts compensate by over- or under-expressing certain cellular pathways, and deploying specific antiviral measures. Collectively, these alterations determine the type and strength of symptoms displayed by infected organisms as well as the global virulence of the infection. Much effort has gone into identifying individual cellular traits that may change as a consequence of viral infection [3] and this has greatly benefited from the contemporary development of genome-wide investigation technologies and their successful application to plant diseases research [4, 5]. These technologies have further demonstrated great potential in providing insights into multidimensional networks of plant-virus interactions [6], notably by allowing combined analyses at the host transcriptome and proteome levels, as was recently shown for an animal virus such as HIV-1 [7].

Based on the above, it has been anticipated that a Systems Biology approach to infections should allow the identification of universal principles and features of host-virus interactions, as opposed to scrutinizing many specific aspects of any given viral infection [8–10]. Such generic principles may indeed prove more predictive of the outcome of viral diseases and therefore, more efficient in the prophylaxis, diagnosis, and even treatment of such diseases. In a network approach, viral pathogenesis can be viewed as the expression of new constraints imposed by the virus upon the cellular interactome: while the host initiates a reprogramming of its genetic profile to activate the immune system to counteract the infection effects, replication and suppression of host defenses by viruses entail the manipulation of molecular connections that ultimately result in the misregulation and/or silencing of genes that trigger defense functions, and eventually in the emergence of new topological properties of the host interactome. Thus, understanding the bases for such modifications is crucial to acquire a systemic view of the infection process [1, 11, 12]. One of the main goals to this end would be the identification of the virus effectors (*i.e.*, the targets of the viral proteins). Instead, herein we focus on the study of the mechanisms by which the host canalizes these virus effectors to trigger the global immune system. We propose a reverse-engineering approach by which

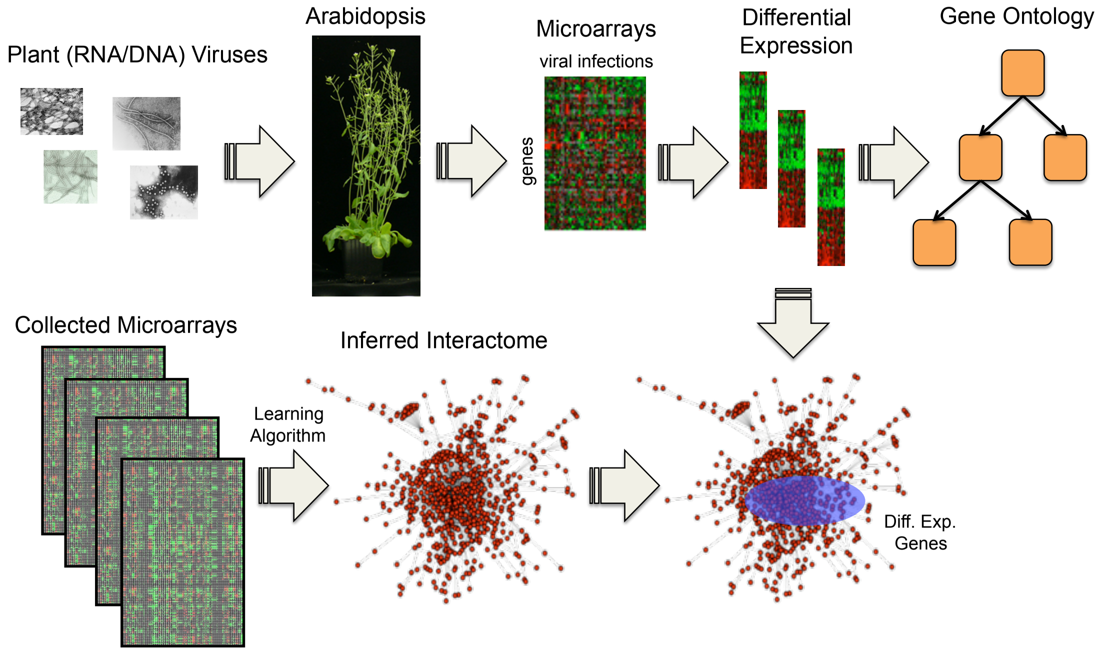


Figure 8.1: Overview of the Systems Biology approach we followed to study the viral infection in plants. We considered *A. thaliana* as model host. Microarray data from several infection experiments with viruses were collected to analyze the differentially expressed genes, and to perform functional analyses by harnessing GO annotations. In addition, by taking advantage of large databases of expression profiles derived from transcriptional perturbations, the global host interactome could be as a first approach unveiled by applying learning algorithms. The differential expression was then contextualized within the inferred network.

we analyze the genetic profile of the cell upon viral infection and contextualize this information onto the host interaction network.

Analyses of interaction networks have already uncovered global, dynamic features that relate directly to biological properties [13]. For example, proteins with a large number of interactions within a network, also referred to as *hubs*, have a higher impact on multiple phenotypic traits (pleiotropy) than loosely connected proteins, and moreover essential proteins for the survival are highly clustered [14]. Hub proteins can be further partitioned into those that function in a specific biological module and those that connect different modules. The existence of such hub proteins generates two interesting properties in networks. First, the network is scale-free, which implies that the probability distribution of the number of connections per node (*i.e.*, its connectivity degree) follows asymptotically a power-law. Second, the network presents the characteristic of small-worlds, in which the average number of intermediary nodes connecting any random pair is small [15]. These two properties confer robustness against random perturbations in the network, but at the cost of strong sensitivity to attacks directed against hubs [16]. Although RNA plant viruses usually encode for few proteins, the resulting genetic profile of the host upon viral infection presents hundreds even thousands of significant changes. A plausible explanation for this scenario is that virus effectors are highly connected proteins that spread the signal, and additionally interact in a short downstream pathway with the immune response genes. Of relevance, very recently it has been experimentally shown that bacterial effectors in *A. thaliana* are hubs and canalize the signal onto the regulators of the global immune system [17]. Interestingly, these results are in concordance with those of previous studies with *Epstein-Barr virus* [18], *Hepatitis C virus* [19], *Influenza A H1N1 virus* [20] and other viral and bacterial pathogens of mammals [8, 21]. These studies have shown that viral proteins preferentially target hub proteins in the human interactome. Herein, by assuming that virus effectors are hubs, we investigate whether this information is propagated following the same scale of the plant interactome.

8.2 Genetic profile targeted by plant viruses

Microarray-based functional genomics, which provides a global view of transcriptional changes in host cells, has been the most commonly used method to study global changes during plant-virus interactions [6, 22–29]. However, the comparison of results obtained in distinct experiments involving different viruses is both complex and challenging; it has not been attempted in a systematic manner. Here, we present the results of a meta-analysis (Fig. 8.1) of microarray data gathered from infections of the same host plant, *A. thaliana*, by seven plant RNA viruses belonging to four taxonomic families (*Tobacco etch potyvirus* –TEV–, *Turnip mosaic potyvirus* –TuMV–, *Plum pox potyvirus* –PPV–, *Tobacco mosaic tobamovirus* –TMV–, *Tobacco rattle tobravirus* –TRV–, *Turnip crinkle carmovirus* –TCV–, and a laboratory-evolved strain of *Tobacco etch potyvirus* –TEV-At17–) and one DNA geminivirus (*Cabbage leaf curl geminivirus* –CaLCuV–). Using transcriptomic data (steady-state RNA levels) extracted from these distinct virus infections on the model plant *A. thaliana*, we identified lists of genes with altered expression levels, referred herein to as virus-responsive genes (or VRGs).

TEV and TEV-At17 expression data (two-color raw data, NCBI GEO accession GSE11088) were obtained from ecotype Ler-0 plants 14 days post-inoculation (dpi) [26, 27]. TuMV data (Affymetrix raw data, ArrayExpress accession e-mexp-509) were obtained 5 dpi from ecotype Col-0 plants [25]. These three data sets were normalized using the RMA method [30] for background correction and quantiles for array scaling, and the list of differentially expressed genes was obtained by performing a Limma test [31] with a correction for multiple testing using the false discovery rate (FDR) procedure [32] (adjusted $P < 0.05$). PPV data (Affymetrix preprocessed data, NCBI GEO accession GSE11217) were obtained 17 dpi from Col-0 plants [29]. In this case, data normalization was done using the Affymetrix MAS 5.0 software package, and the differential expression using a one-way ANOVA test with a correction for multiple testing using the FDR procedure (adjusted $P < 0.05$), followed by a fold-change criterion of 1.5 in z-score over all genes (averaging replicates). TMV data (two-color raw data, deposited in www.bio.puc.cl/labs/arce/index.html) were obtained from ecotype Uk-4 plants 10 dpi [24], and normalized using the RMA method for background correction and quantiles for array scaling. The list of differentially expressed genes was obtained by performing a fold-change criterion of 1.96 in z-score over all genes (averaging replicates). TRV data (two-color raw data, NCBI GEO accession GSE15557) were measured 8 dpi from Col-0 leaves. TCV data (two-color raw data, NCBI GEO accession GSE29387) were quantified 10 dpi in Col-0 plants. These two data sets were normalized using the CATMA BGS procedure [33], and the list of differentially expressed genes was obtained by performing a Limma test with FDR correction (adjusted $P < 0.05$). In addition, for TCV data, a fold-change criterion of 1.96 in z-score over all genes (averaging replicates) was applied. Finally, CaLCuV data (Affymetrix raw data, ArrayExpress accession E-ATMX-34) were collected from Col-0 plants 12 dpi [28]. These data were normalized using subtraction for background correction and LOWESS [34] for array scaling, and the list of differentially expressed genes was obtained by performing a mixed ANOVA test with a correction for multiple testing using the FDR procedure (adjusted $P < 0.05$). To perform the data normalization and to obtain the differentially expressed genes, we used the GEPAS tool [35], which is implemented within the BABELOMICS webserver [36].

Those VRGs were then used to establish both general and specific genetic profiles associated to the pathogens of interest. We found that among the $>$ 22,000 genes inspected, a set of 5,296 VRGs (2,646 over- and 2,650 under-expressed) is altered by at least one of the eight viruses studied (summarized in Table 8.1). This VRG set may thus be used to reflect the global plant response to any viral infection. We found that the number of VRGs shared by more than one virus declines exponentially. Seven VRGs were found up-regulated in common by six viruses, of which, surprisingly, six play a role in cell migration (At3g57260, At5g10380, At3g14990, At3g28510, At5g52640, and At4g24690) and one (At1g75040) encodes a PR-5 thaumatin-like protein, factors known for their involvement in pathogens responses. While no single VRG was identified in common among the eight infections, one VRG was systematically up-regulated by seven viruses (all except PPV) and found to encode an aspartyl protease involved, again, in cell migration in the diencepahlon (At5g10760). Three VRGs were down-regulated by six viruses, two of which correspond to different subunits of the NADPH dehydrogenase complex (At1g18730 and At5g58260).

Not surprisingly, infections by the two different strains of TEV under study share the largest number of VRGs (197 over- and 282 under-expressed genes), although this may probably reflect, to some extent, homogeneity in experimental procedures. In the overlapping set, over-expressed genes principally have roles in response to stress (*e.g.*, fungal resistance TIR-NB-LRR protein At1g56510, transcription factor At1g22070, U-box-domain-containing E3 ubiquitin ligase

Table 8.1: Summary of the number of VRGs and VRFs (over- and under-expressed) from several viral infections in *A. thaliana*.

	VRGs		VRFs	
	Over	Under	Over	Under
TEV	356	322	35	41
TEV-At17	950	1441	32	90
TuMV	754	390	29	30
PPV	747	740	98	8
TMV	498	225	62	0
TRV	215	284	14	26
TCV	708	846	91	70
CaLCuV	454	732	66	107

At3g11840 that acts as a negative regulator of immune responses, or the aforementioned At1g75040), transport (*e.g.*, the mitochondrial inner membrane translocase At1g20350, the high-affinity ammonium transporter At2g38290, or the glycolipid transfer protein At4g39670), transcription (*e.g.*, the Myb-like transcription factor At1g25550, or the C2H2-type zinc finger At3g46080), and protein metabolism (*e.g.*, the chaperone DnaJ-domain At1g56300, or the eukaryotic aspartyl protease At5g10760). The overlapping set of under-expressed genes is mostly composed of factors involved in basic metabolic and cellular processes (*e.g.*, the member of the R2R3 factor At1g18710, the enzyme At1g03630 that is NADPH- and light-dependent, or the hydrolase At1g10740).

Interestingly, a set of 27 VRGs was significantly over-expressed upon infections by the three viruses that naturally infect hosts from the *Brassicaceae* family (TuMV, TCV and CaLCuV) and by the TEV laboratory strain, which has been experimentally adapted to *A. thaliana* (TEV-At17); hereafter, we will refer to this set of four viruses as *Brassica*-infecting viruses. A common feature of these VRGs is that all of them play roles in stress response, including, among others, the disulfide isomerase At1g21750 implicated in the regulation of apoptosis during endoplasmic reticulum stress as well as in osmotic stress. The set also includes the homolog of mammalian Bax inhibitor 1, At5g47120, which functions as an attenuator of biotic and abiotic stress-associated cell death, and the cytosolic heat shock protein At5g52640. The list further comprises several genes involved in signal transduction, such as the BAK1-interacting receptor-like kinase At5g48380 that regulates multiple signaling routes for plant resistance, or the ATP binding kinase At5g45800 involved in embryonic development. A set of 22 VRGs was also under-expressed in common, in plants infected by the *Brassica*-infecting viruses. This list includes, as in the afore-mentioned study of the two TEV strains, genes involved in central metabolic and cellular processes.

Next, we sought to establish an overall comparison of the lists of VRGs identified from any of the eight viruses included in the analysis. To do so, we computed similarity scores among all pairs of lists, and constructed a dendrogram to visualize which viruses showed more closely related lists. The eight viruses do not represent independent draws from a population; rather, some are phylogenetically related. It was therefore important to test whether the above overlap in VRGs reflected taxonomic correlations. In other words, do closely phylogenetically related viruses tend to share a higher number of VRGs, and does this overlap reduce as phylogenetic distance between viruses increases? To address this issue, we first used an alignment of the replicase genes from the eight viruses (the replicase-associated protein in the case of the geminivirus) to construct a maximum-likelihood phylogenetic tree (using the WAG + Γ model of amino acid susbstitutions and evaluating the significance of tree topology by 1,000 bootstrap replicates). Next we computed a congruency index [37] measuring the overlap between the tree topology obtained from the VRG similarity matrix, on the one hand, and the topology of the estimated phylogenetic tree, on the other. The congruence index ($I_{cong} = 1.4720$) was significantly larger than expected by mere chance ($P = 0.0052$), suggesting that the two topologies are indeed highly congruent. This result supports the hypothesis that the overlap between VRG lists reflects the taxonomic relationships among viruses: two closely related viruses (*e.g.*, the potyviruses TEV and TuMV) tend to alter the expression of a similar set of genes, whereas two non-related viruses (*e.g.*, TEV and TRV) tend to alter different subsets of genes. The various viruses included in the study have distinct replication, gene expression, movement, and RNA silencing-suppression strategies that should somehow impact transcriptomic profiles differently. It is highly likely, however, that these strategies may be more conserved between phylogenetically related viruses than among viruses with weak or no phylogenetic relationship and, hence, the above study most certainly already accounts for the differences and commonalities observed among virus

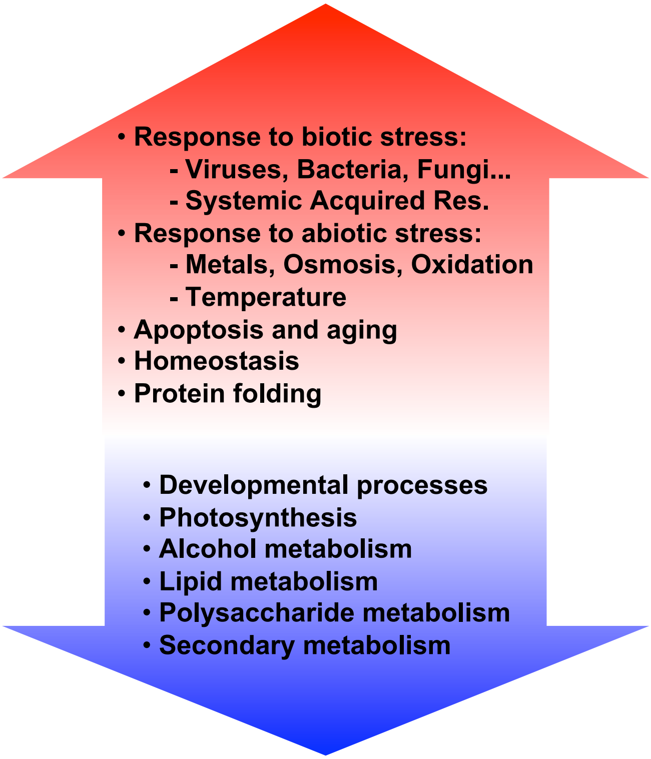

• Response to biotic stress:
- Viruses, Bacteria, Fungi...
- Systemic Acquired Res.
• Response to abiotic stress:
- Metals, Osmosis, Oxidation
- Temperature
• Apoptosis and aging
• Homeostasis
• Protein folding
• Developmental processes
• Photosynthesis
• Alcohol metabolism
• Lipid metabolism
• Polysaccharide metabolism
• Secondary metabolism

Figure 8.2: List of common biological functions up- and down-regulated by plant viruses in *A. thaliana*.

phyla. This being said, convergent evolution in phylogenetically unrelated viruses may contribute to increase the overlap of VRG lists. For instance, potyviruses and carmoviruses employ overlapping RNA silencing-suppression strategies affecting the global metabolism of miRNAs, which may lead to a set of related host responses.

A potential weakness of the above meta-analysis of gene lists is that the different experiments not only differed in the methodological details and plant ecotypes, but also in that different experiments took samples at different time points during the infection process and, in some cases, different tissues were also sampled. Ecotype-specific, time-dependent and tissue-specific responses to viral infection may turn ON/OFF different subsets of genes [25, 38] and thus may not receive a high enough score to be classified as VRG according to the stringent statistical criteria used in the study. To minimize as much as possible these potential problems, only data from leaves were included in the present analysis, although the possible effect of ecotype and sampling time may still exist. We performed several statistical analyses to assess this heterogeneity in the data, and we concluded that differences in ecotype or in sampling time would neither have a significant effect on the conclusions drawn from our meta-analysis.

8.3 Biological functions triggered by plant viruses

Subsequently, we performed a functional analysis to map changes in gene expression onto regulations effecting global biological functions, thus establishing lists of virus-responsive functions (or VRFs). For each list of VRGs, (over- or under-expressed), we looked for the significant over-represented biological processes (GO terms between levels 3 and 9) within that list. The statistical significance was performed by means of a Fisher's exact test for $2x2$ contingency tables with a correction for multiple testing using the FDR procedure (adjusted $P < 0.05$). To perform the functional analysis of the VRGs, we used the FatiGO tool [39], implemented in the BABELOMICS webserver [36].

Some generalities can be drawn from this study, highlighting the fact that different viruses alter common sets of VRFs (summarized in Fig. 8.2). On the one hand, approximately one-third of over-expressed VRGs are associated with cell rescue, defense, apoptosis and cell death and aging, including several defense- and stress-associated genes. Responses to biotic (viruses, bacteria, or fungi) and abiotic (metal ions, osmosis, oxidation, or temperature) stresses, including systemic acquired resistance [40] and the innate immune system, are upregulated by the plant to counteract viral infection. Such a defense response in *A. thaliana* to viruses is dependent on salicylic acid [38]. In addition, a variety of heat-shock proteins are also over-expressed after infection with any viruses. Although this might just be a generic nonspecific response by the plant to stress, we suggest that the virus directly triggers chaperones to assist in correct folding of its own proteins, since many of them could misfold (and thus aggregate) as a consequence of mutations produced during error-prone replication [41]. Ribosomal proteins and protein turnover genes are also up-regulated. Again, this could either reflect an increased demand on the host cells for protein synthesis or a response triggered by a virus to enhance its own production (or presumably both). On the other hand, several developmental functions, biosynthesis of lipids, alcohols and polysaccharides, and secondary metabolism constitute the principal down-regulated processes. For example, biosynthesis of lipids is pivotal for cell membrane construction and modification and carbohydrates biosynthesis is essential for building cell walls; therefore, because this expression is correlated to plant cell growth and expansion, reduced expression could well result in the stunting syndrome associated with some infections. Similarly, plastid genes and genes involved in chloroplast functioning are also preferentially under-expressed, resulting in chlorosis. The unspecific gene down-regulation response as a part of the viral reprogramming of metabolism is consistent with the previously proposed idea [42] that viruses impel the plant to redirect resources towards immune systems and, in particular, biotic stress responses, to the detriment of developmental processes.

As in the previous section, we tested if the dendrogram topology was congruent with the phylogenetic history of the viruses (from the similarity matrix computed from overlapping lists of VRFs obtained for the eight viruses). In this case, the congruence index ($I_{cong} = 1.2267$) did not significantly differ from what was expected by chance ($P = 0.1005$), thus suggesting that both topologies are not highly congruent; in other words, that the set of VRFs altered by two related viruses is similar to the one altered by two non-related viruses. At first, this result may be seen as contradicting the previous one, obtained by comparing lists of VRGs. Broad GO terms, however, encompass a multitude of genes and it may well be that different viruses affect the same VRF by modifying the expression of different target genes. Consistent with this idea, TEV and TRV infections were both associated with a significant over-representation of the GO term *stress response*. In both cases, the number of VRGs connected to this specific GO is similar (108 for TEV and 93 for TRV); yet only four genes are affected in common between the two viruses. Therefore, this analysis reveals that comparable trends in the global reprogramming of cellular functions might be achieved via highly dissimilar gene expression changes induced by distinct viruses.

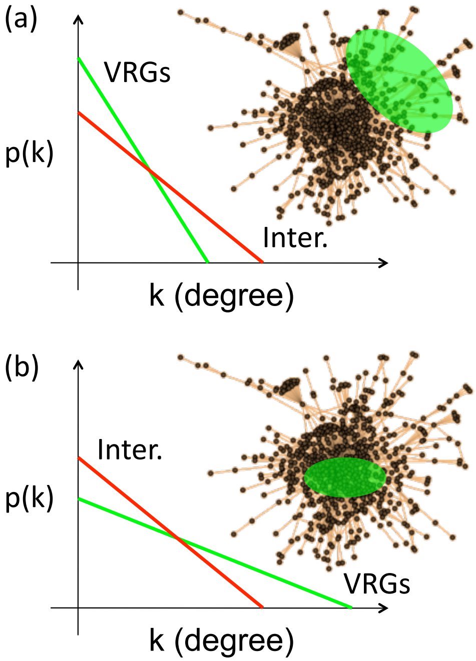


Figure 8.3: Connectivity distribution for the whole plant interactome (red line) and the distribution generated by the VRGs (green line) for two opposing modes of virus action. Panel (a) illustrates the case of VRGs being sparsely distributed in the network and poorly connected. This situation translates into connectivity distributions steeper than observed the whole interactome. Panel (b) exemplifies the opposite situation of VRGs being highly connected hubs. In this case the connectivity distribution is flatter.

8.4 Viruses preferentially alter highly connected genes

We then focused on the impact of viruses on two different predicted interactomes of *A. thaliana*: a transcriptional regulatory network (TRN) [43] and a protein-protein interaction network (PPIN) [44]. The TRN was inferred using a reverse-engineering procedure, based on mutual information with a local significance (z-score computation) as estimator of the likelihood, for capturing coexpression patterns between transcription factors and genes, and has optimal levels of confidence and coverage. This network contains 139,440 interactions and involves 19,108 genes. The PPIN consists of a set of

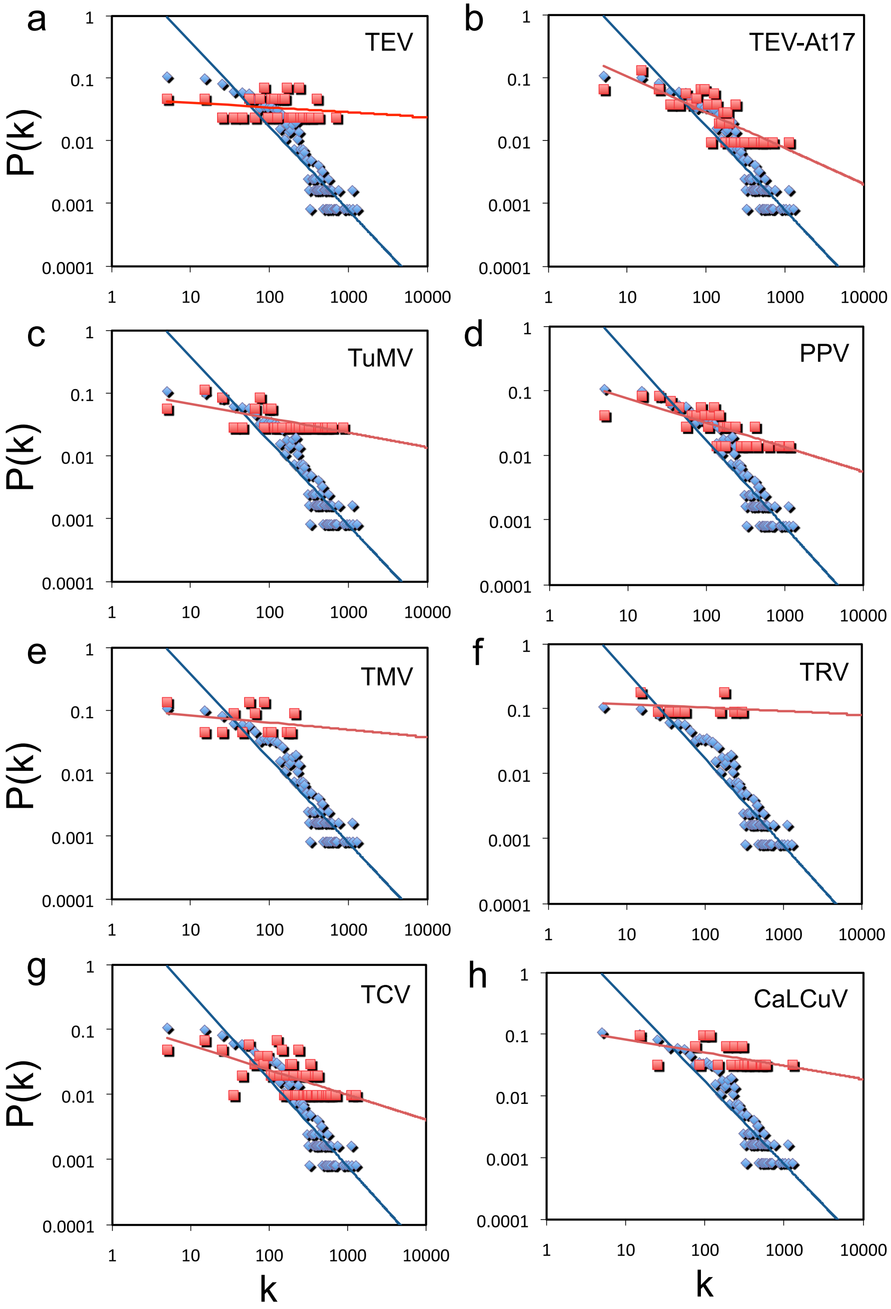


Figure 8.4: Outgoing connectivity distributions, contextualized in the TRN, for the VRGs (red) and the whole interactome (blue).

72,266 predicted interactions involving 7,177 proteins, based on the identification of orthologs of *A. thaliana* proteins in seven other species. Both TRN and PPIN have the properties of scale-free and small-worlds, the two major characteristic properties of real biological networks [15]. To analyze the impact of a viral infection in terms of genetic interactions, we studied the principal topological properties on the inferred networks: connectivity, clustering, connected components, shortest paths and modularity. For each VRG (up and down), we collected its connectivity degree and betweenness centrality, according to the global interactome. Differences in connectivity (k) and betweenness (b) among the VRGs and the total set of plant genes were analyzed by means of one-tailed Mann-Whitney U-tests ($P < 0.05$) considering the superior tails of the distributions (*i.e.*, the genes satisfying $k > \langle k \rangle$ or $b > \langle b \rangle$). Furthermore, we performed linear regressions in the log-log space to obtain the critical exponents, γ, of the power-law degree distribution $P(k) \sim k^{-\gamma}$ and assessed the statistical significance of the inferred values using Student t-tests ($P < 0.0001$).

First, we analyzed the connectivity degree distribution for the VRGs as compared to the global set of genes. Roughly, if those genes were located in the periphery of the network, their connectivity would be expected to be smaller than if they were central, since the interactome is scale-free (Fig. 8.3). As the TRN is a directed interactome, we focused on the outgoing connectivity, that is, the number of edges that leave from a given node to connect other nodes in the network. In Fig. 8.4 we show the connectivity distributions for all viral infections. Table 8.2 summarizes the value of the power-law exponent that better fits this particular distribution as well as the average connectivity. To statistically assess differences between the VRGs and the global set of genes, we used t-tests for differences in slopes and U-tests for differences in the location of the high-degree genes within the distributions. No significant differences were found between the incoming connectivity distributions of VRGs and the one characterizing the whole interactome. Fig. 8.5 shows the corresponding connectivity distributions using the PPIN (see also Table 8.2). We found that, in all cases (both in TRN and PPIN), the slope of the power-law distributions was significantly smaller than the slope estimated for the whole interactome. Apart of that *Brassica*-infecting viruses affect on average more genes, these viruses also manipulate more interactions, irrespective to the network model

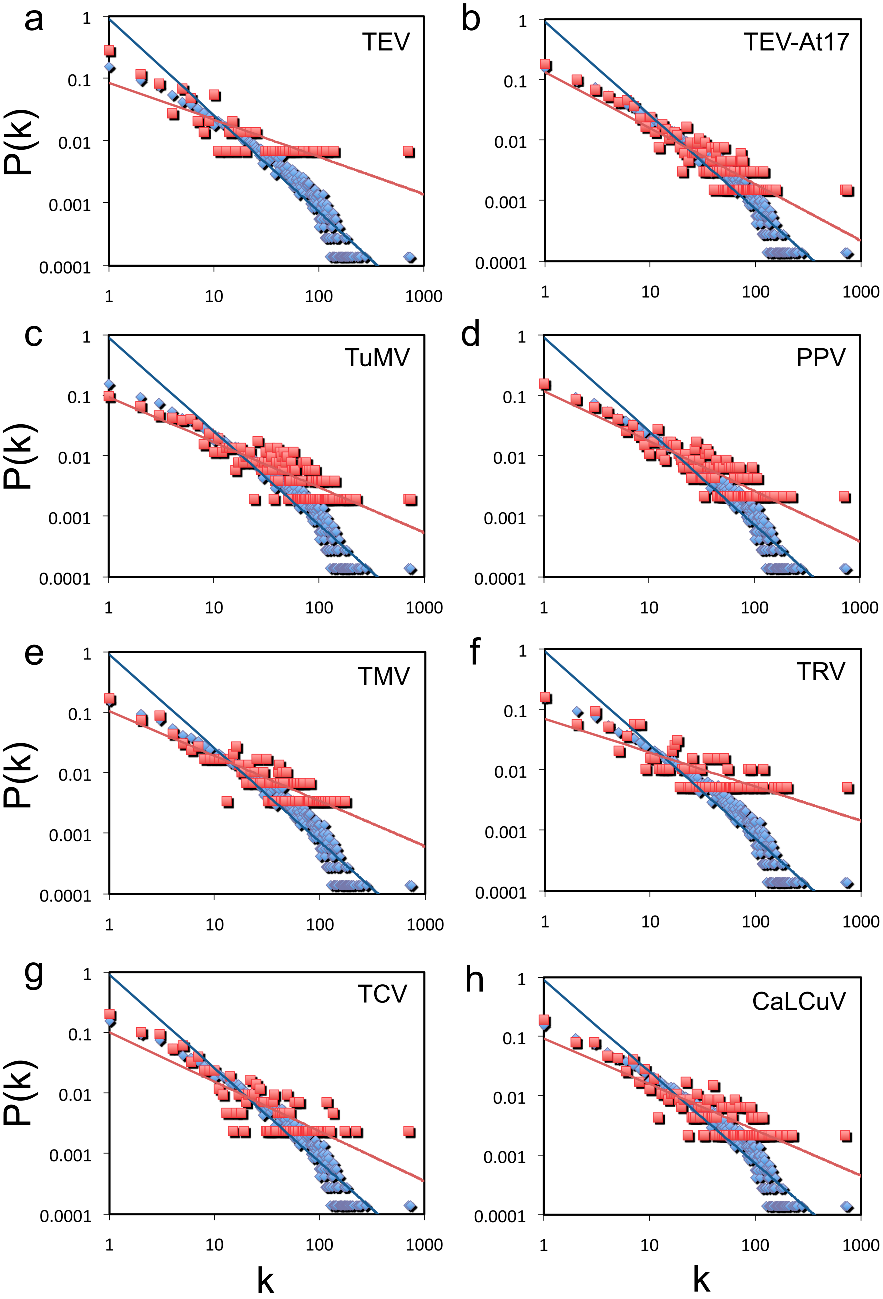


Figure 8.5: Connectivity distributions, contextualized in the PPIN, for the VRGs (red) and the whole interactome (blue).

(one-tailed t-test, $P < 0.05$). However, we did not find any significant difference among viruses in terms of the average connectivities and power-law distribution exponents. We conclude that a smaller slope of the power-law distribution is a general trend characterizing the VRGs, indicating that viral infection preferentially alters the expression of highly connected genes (hubs) rather than random genes within the whole network. This could reflect a cellular response to gain robustness against the manipulation of the host by viruses.

A more global scale analysis involved calculation of the betweenness centrality distribution, understood as the relative number of shortest paths traversing a given gene. Table 8.2 summarizes the values of the average betweenness for these sets of VRGs. Since PPIN represents the case of an undirected graph, we restricted the analysis to this interactome to evaluate the betweenness of the VRGs. We found that, as occurred with the connectivity at the local level, the VRGs were significantly central for seven out of eight viruses, with average betweenness centrality values significantly greater than observed for the whole interactome (U-test $P < 0.05$). TMV was the exception to this rule. We also found that betweenness and connectivity are significantly positively correlated (Spearman $\rho = 0.8885$ with $P < 0.0001$, releasing the isolated nodes), despite the high variability of betweenness at low connectivity values, a characteristic of hierarchical networks.

8.5 Discussion

The results of our meta-analysis combining transcriptomic data gathered for eight different viruses all infecting a common host, *A. thaliana*, confirm that host cells undergo significant reprogramming of their transcriptome during infection, which is possibly a central requirement for mounting the host defenses. Rather than focusing on the details of each virus infection, however, our study was designed to uncover generic features defining either the host response to, or the targets manipulated by, the various viruses tested. We found that the overlap in the lists of genes whose expression is altered upon infection (VRGs) decreases as the phylogenetic distance between the viruses increases, thus suggesting that related viruses may interact with similar host components, whereas non-related viruses may manipulate different targets. This association at the VRG level does not hold, however, at the level of altered, global biological functions (VRFs),

Table 8.2: Summary of topological properties of the differentially expressed VRGs from several viral infections contextualized in the *A. thaliana* interactomes. We show the number of interactions (edges) manipulated by the virus, the power-law distribution exponent for connectivity (γ), and the average connectivity ($\langle k \rangle$) for TRN and PPIN. The average betweenness ($\langle b \rangle$) is shown in case of PPIN. We also show the P-value for the tests comparing the shape and location of the VRGs distributions with respect to the corresponding whole interactome (^{a}Student t-test, ^{b}Mann-Whitney U-test).

	TRN			PPIN			
	Edges	γ (P^a)	$\langle k \rangle$ (P^b)	Edges	γ (P^a)	$\langle k \rangle$ (P^b)	$\langle b \rangle \cdot 10^{-4}$ (P^b)
TEV	1275	0.07 ($< 10^{-4}$)	162 (0.02)	64	0.59 ($< 10^{-4}$)	18 (0.18)	9.56 ($< 10^{-4}$)
TEV-At17	2850	0.57 ($< 10^{-4}$)	115 (0.23)	881	0.92 ($< 10^{-4}$)	22 (0.03)	5.72 ($< 10^{-4}$)
TuMV	1034	0.23 ($< 10^{-4}$)	172 ($< 10^{-3}$)	1665	0.74 ($< 10^{-4}$)	34 ($< 10^{-4}$)	8.63 ($< 10^{-4}$)
PPV	945	0.37 ($< 10^{-4}$)	153 ($< 10^{-2}$)	535	0.82 ($< 10^{-4}$)	24 (0.02)	6.46 ($< 10^{-4}$)
TMV	67	0.11 ($< 10^{-4}$)	76 (1)	214	0.74 ($< 10^{-4}$)	22 (0.15)	3.35 (0.40)
TRV	82	0.05 ($< 10^{-4}$)	111 (0.45)	154	0.56 ($< 10^{-4}$)	26 ($< 10^{-3}$)	8.20 ($< 10^{-4}$)
TCV	4328	0.38 ($< 10^{-4}$)	188 ($< 10^{-4}$)	364	0.81 ($< 10^{-4}$)	19 (0.04)	5.50 ($< 10^{-3}$)
CaLCuV	2117	0.21 ($< 10^{-4}$)	255 ($< 10^{-4}$)	664	0.77 ($< 10^{-4}$)	24 ($< 10^{-4}$)	6.14 ($< 10^{-4}$)
Interactome	139,440	1.33 (-)	114 (-)	72,266	1.54 (-)	20 (-)	3.43 (-)

thus suggesting that a common set of overall functional responses to infection may result from the manipulation of sometimes drastically different target genes. One caveat of the meta-analysis studies such as the one reported here, however, is that they are conservative in design. They will at best identify shared (sub) responses that are strong enough to be detected against the intrinsically high noise level as a consequence of the diversity of viral systems and microarray platforms used in the original studies that served as the basis for the present one. While reductionism through single-cell transcriptome analyses has been successfully employed in virus-infected mammalian cell cultures [9] and in plant protoplasts [29], studying *in vivo* host-virus interactions obviously adds many layers of complexity and variability, which are clearly reflected here. Nonetheless, our study shows that such complexity does not, *a priori*, constitute an insurmountable obstruction to the discovery of generic patterns associated to plant viral infections.

Our study points out that VRGs are, in general, more highly connected, central and modular than expected by chance. This result agrees with the fact that viral proteins preferentially interact with hub regulator genes [17–19, 21], although VRGs not necessarily entail virus effectors. Probably as a plant strategy, through hub genes the signal can be disseminated at large to change the whole genetic profile. Then, even a small number of viral proteins can affect a considerable number of host genes. In the case of *Potyviruses*, 11 mature proteins provoke significant changes in expression in about a thousand of host genes. That more hub genes (both from TRN and PPIN) than expected by chance were differentially expressed indeed reflects an effect of the virus over them, and also indicates that the information flow from virus effectors to immune response proteins is strengthened *ab initio* (lower slope in the power-law distribution). We therefore hypothesize that this over-stimulation of hubs is a mechanism that confers robustness to the plant to express the immune system. Whether a virus deactivated a recognition pathway, redundant hubs would emerge to counteract this viral action. We have confirmed this observation for all the plant viruses included in our meta-analysis, thus uncovering a possible universal pattern in host-virus interactions.

The following publications hold the contents presented in this chapter

- Rodrigo G, Carrera J, Ruiz-Ferrer V, del Toro FJ, Llave C, Voinnet O, Elena SF (2011) A meta-analysis reveals the commonalities and differences in *Arabidopsis thaliana* response to different viral pathogens. *Submitted.*

- Elena SF, Carrera J, Rodrigo G (2011) A systems biology approach to the evolution of plant-virus interactions. *Curr Opin Plant Biol*, 14: 372-377.

Bibliography

[1] Penga X, Chana EY, Lia Y, Diamonda DL, Kortha MJ, Katze MG (2009) Virus-host interactions: from systems biology to translational research. *Curr Opin Microbiol*, 12: 432-438.

[2] Dodds PN, Rathjen JP (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet*, 11: 539-548.

[3] Maule A, Leh V, Lederer C (2002) The dialogue between viruses and hosts in compatible interactions. *Curr Opin Plant Biol*, 5: 279-284.

[4] Whitham SA, Yang C, Goodin MM (2006) Global impact: elucidating plant responses to viral infection. *Mol Plant-Microb Interact*, 19: 1207-1215.

[5] Bailer SM, Haas J (2009) Connecting viral with cellular interactomes. *Curr Opin Microbiol*, 12: 453-459.

[6] Whitham SA, Quan S, Chang HS, Cooper B, Estes B, Zhu T, Wang W, Hou YM (2003) Diverse RNA viruses elicit the expression of common sets of genes in susceptible *Arabidopsis thaliana* plants. *Plant J*, 33: 271-283.

[7] MacPherson JI, Dickerson JE, Pinney JW, Robertson DL (2010) Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput Biol*, 6: e1000863.

[8] Jenner RG, Young RA (2005) Insights into host responses against pathogens from transcriptional profiling. *Nat Rev Microbiol*, 3: 281-294.

[9] Andeweg AC, Haagmans BL, Osterhaus ADME (2008) Virogenomics: the virus-host interaction revisited. *Curr Opin Microbiol*, 11: 461-466.

[10] Elena SF, Carrera J, Rodrigo G (2011) A systems biology approach to the evolution of plant-virus interactions. *Curr Opin Plant Biol*, 14: 372-377.

[11] Tan SL, Ganji G, Paeper B, Proll S, Katze MG (2007) Systems biology and the host response to viral infection. *Nat Biotechnol*, 25: 1383-1389.

[12] De la Fuente A (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet*, 26: 326-333.

[13] Albert R (2005) Scale-free networks in cell biology. *J Cell Sci*, 118: 4947-4957.

[14] Yu H, Braun P, *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, 322: 104-110.

[15] Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5: 101-113.

[16] Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature*, 406: 378-382.

[17] Mukhtar MS, *et al.* (2011) Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, 333: 596-601.

[18] Calderwood MA, Venkatesan K, Xing L, *et al.* (2007) *Epstein-Barr virus* and virus human protein interaction maps. *Proc Natl Acad Sci USA*, 104: 7606-7611.

[19] De Chassey B, Navratil V, Tafforeau L, Hietet MS, *et al.* (2008) *Hepatitis C virus* infection protein network. *Mol Syst Biol*, 4: 230.

[20] Shapira SD, Gat-Viks I, Shum BOV, *et al.* (2009) A physical and regulatory map of host-*influenza* interactions reveals pathways in H1N1 infection. *Cell*, 139: 1255-1267.

[21] Dyer MD, Murali TM, Sobral BW (2008) The landscape of human protein interacting with viruses and other pathogens. *PLoS Pathog*, 4: e32.

[22] Golem S, Culver JN (2003) *Tobacco mosaic virus* induced alterations in the gene expression profile of *Arabidopsis thaliana. Mol Plant-Microb Interact*, 16: 681-688.

[23] Ishihara T, Sakurai N, Sekine KT, Hase S, Ikegami M, Shibata D, Takahashi H (2004) Comparative analysis of expressed sequence tags in resistant and susceptible ecotypes of *Arabidopsis thaliana* infected with *Cucumber mosaic virus. Plant Cell Physiol*, 45: 470-480.

[24] Espinoza C, Medina C, Somerville S, Arce-Jonhson P (2007) Senescence-associated genes induced during compatible viral interactions with grapevine and *Arabidopsis. J Exp Bot*, 58:

3197-3212.

[25] Yang C, Guo R, Jie F, Nettleton D, Peng J, Carr T, Yeakley JM, Fan JB. Whitham SA (2007) Spatial analysis of *Arabidopsis thaliana* gene expression in response to *Turnip mosaic virus* infection. *Mol Plant-Microb Interact*, 20: 358-370.

[26] Agudelo-Romero P, Carbonell P, De la Iglesia F, Carrera J, Rodrigo G, Jaramillo A, Perez-Amador MA, Elena SF (2008) Changes in the gene expression profile of *Arabidopsis thaliana* after infection with *Tobacco etch virus. Virol J*, 5: 92.

[27] Agudelo-Romero P, Carbonell P, Perez-Amador MA, Elena SF (2008) Virus adaptation by manipulation of host's gene expression. *PLoS ONE*, 3: e2397.

[28] Ascencio-Ibánez J, Sozzani R, Lee TJ, Chu TM, Wolfinger RD, Cella R, Hanley-Bowdoin L (2008) Global analysis of *Arabidopsis* gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant Physiol*, 148: 436-454.

[29] Babu M, Griffiths JS, Huang TS, Wang A (2008) Altered gene expression changes in *Arabidopsis* leaf tissues and protoplasts in response to *Plum pox virus infection. BMC Genomics*, 9: 325.

[30] Irizarray RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4: 249-264.

[31] Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist Appl Genet Mol Biol*, 3: 3.

[32] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57: 289-300.

[33] Allemeersch J, Durinck S, Vanderhaeghen R, *et al.* (2005) Benchmarking the CATMA microarray. A novel tool for *Arabidopsis* transcriptome analysis. *Plant Physiol*, 137: 588-601.

[34] Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, 74: 829-836.

[35] Tárraga J, Medina I, Carbonell J, Huerta-Cepas J, *et al.* (2008) GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucl Acids Res*, 36: W308-W314.

[36] Al-Shahrour F, Mínguez P, Vaquerizas JM, Conde L, Dopazo J (2005) BABELOMICS: a suite of web tools for functional

annotation and analysis of groups of genes in high-throughput experiments. *Nucl Acids Res*, 33: W460-W464.

[37] De Vienne DM, Giraud T, Martin OC (2007) A congruence index for testing topological similarity between trees. *Bioinformatics*, 23: 3119-3124.

[38] Wise RP, Moscou MJ, Bogdanove AJ, Whitham SA (2007) Transcript profiling in host-pathogen interactions. *Annu Rev Phytopathol*, 45: 329-369.

[39] Al-Shahrour F, Mínguez P, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J (2007) FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucl Acids Res*, 35: W91-W96.

[40] Ryals JA, Neuenschwander UH, Willits MG, Molina A, Steiner H-Y, Hunt MD (1996) Systemic acquired resistance. *Plant Cell*, 8: 1809-1819.

[41] Jockusch H, Wiegand C, Mersch B, Rajes D (2001) Mutants of *tobacco mosaic virus* with temperature-sensitive coat proteins induce heat shock response in tobacco leaves. *Mol Plant-Microb Interact*, 14: 914-917.

[42] Pagán I, Alonso-Blanco C, García-Arenal F (2008) Host responses in life-history traits and tolerance to virus infection in *Arabidopsis thaliana. PLoS Pathog*, 4: e1000124.

[43] Carrera J, Rodrigo G, Jaramillo A, Elena SF (2009) Reverse-engineering the *Arabidopsis thaliana* transcriptional network under changing environmental conditions. *Genome Biol*, 10: R96.

[44] Geisler-Lee J, O'Toole N. Ammar R, Provart NJ, Millar AH, Geisler M (2007) A predicted interactome for *Arabidopsis thaliana. Plant Physiol*, 145: 317-329.

Chapter 9

General conclusion

Design is not just what it looks like...
Design is how it works.
– Steve Jobs

Cell fate is programmed through gene regulatory networks that perform several calculations to take the appropriate decision. In addition, signaling pathways are interconnected to regulatory networks for sensing the environment and expressing the appropriate genetic profile. Therefore, whether we could disentangle the design principles of these networks, we would gain fundamental insights about the way cells behave. Have natural networks a precise fine-tuning of kinetic parameters? Or, on the contrary, is it all about network architecture? One approach to understand natural regulatory networks consists in designing artificial systems and confronting them with the natural ones. This, indeed, challenges our knowledge about the mode of action of molecular systems and could reveal hidden traits on which natural selection operated. To this end, *in silico* evolutionary optimization mimics the way Nature has designed such genetic networks. In this thesis, we have focused on Monte Carlo techniques. We have discussed the basic principles of this heuristic approach and how can be applied to design different regulatory systems. To balance such an optimization framework, we also applied mathematical analysis techniques to rationalize the regulatory mechanisms found in natural genetic networks.

The engineering of synthetic gene circuits based on transcription has mostly relied on the assembly of few characterized regulatory elements using rational design principles. Then, it is of outmost importance to analyze the scalability and limits of such a design workflow. To analyze the design capabilities of libraries of regulatory elements, we developed an automated design approach that combined such elements to search the genotype space associated to a given phenotypic behavior. In this context, we calculated the designability of dynamical functions obtained from circuits assembled with a given genetic library. By designing circuits working as amplitude filters, pulse counters and oscillators, we could to infer new mechanisms for such behaviors. We also highlighted the hierarchical design and the optimization of the interface between devices. Finally, we dissected the functional diversity of a constrained library and we found that even such libraries can provide a rich variety of behaviors. We also found that intrinsic noise slightly reduces the designability of networks with digital behavior, but it increases the designability of oscillators.

Computational design can be also applied to discover design principles of specific network architectures. In particular, gradients of diffusing molecules (morphogens) determine cell fate at a given position, dictating development and spatial organization. Among, the FFL circuit is the simplest genetic architecture able to generate one-stripe patterns by operating as an amplitude detection device, where high output levels are achieved at intermediate input ones. Herein, we dissected the design space containing all possible topologies and parameter values of the FFL circuits. We explored the ability of being sensitive or adaptive to variations in the critical morphogen level where cell fate is switched. We found four different solutions for precision, corresponding to the four incoherent architectures, but remarkably only one mode for adaptiveness, the incoherent type 4. We further carried out a theoretical study to unveil the design principle for such structural discrimination, finding that the synergistic action and cooperative binding on the downstream promoter are instrumental to achieve absolute adaptive responses. Subsequently, we analyzed the robustness of these optimal circuits against perturbations in the kinetic parameters and molecular noise, which allowed us to depict a scenario where adaptiveness, parameter sensitivity and noise tolerance are different, correlated facets of the robustness of the FFL circuit. Strikingly, we showed a strong correlation between the input (environment-related) and the intrinsic (mutation-related) susceptibilities. Finally, we discussed the evolution of incoherent regulations in terms of multifunctionality and robustness.

The application of engineering methodologies in Molecular Biology can result very advantageous because, apart of providing a systematic way to redesign the control mechanisms that manage biological processes, they allow uncovering the design principles that were naturally selected to ensure the survival of the organism. If so, are these natural mechanisms similar to the ones we would engineer following theoretical principles? In this thesis, we focused on two responsive systems present in plants: gravitropism and RNA silencing. Certainly, plants have evolved mechanisms to sense gravity and orient themselves accordingly. We constructed a mathematical model that reproduced the plant gravitropic response. The model was based on known genetic interactions and hormone signaling coupled to a physical description of plant reorientation. Indeed, the interplay between hormone signaling and gene regulatory networks is instrumental in living organisms in order to promote their own development. The model allowed the analysis of the spatiotemporal dynamics of the system, triggered by a hormone gradient that induces differential growth of the plant with respect to the gravity vector. Our model predicted two important features with strong biological implications. First, the robustness of the regulatory circuit as a consequence of integral control. And second, the higher degree of plasticity generated by the molecular interplay between two classes of hormones.

Moreover, we studied the RNA silencing pathway and how it is applied to fight against an RNA virus. We found that this immune system performs a sort of derivative control for the suppression of the virus, by which siRNA is the central element. The cell-to-cell movement of siRNAs serves to anticipate the cleavage response by RISC. However, viruses have evolved strategies to escape from silencing surveillance while promoting their own replication. By examining the system at the single cell level, we found that an appropriated virus strategy would be to devote more time into transcription and target the first protein component of the pathway.

In addition, we tackled the problem of designing riboregulatory modules in bacteria. These modules could be used in combination with further regulatory elements to construct more complex systems. In this thesis, we have described the first fully automated design procedure and experimental validation in bacteria of synthetic RNA circuits. An optimization algorithm based on a physicochemical RNA model and exploiting allosteric motion was used to computationally screen iteratively generated sequences of nucleic acids. Our RNA devices implemented a mechanism of post-transcriptional control of protein expression with tunable performance, highly specific, and orthogonal. To obtain these devices, different structural constraints and inter-molecular interactions were specified. Such results demonstrate that computational methods can explore the large combinatorial space of sequences to *de novo* design genetic circuits, and that a quantitative, first principles-based, unbiased algorithm is useful in diverse riboregulation frameworks.

Furthermore, we applied the computational techniques to study the designability of metabolic pathways. The biological solution for synthesis or remediation of organic compounds using living organisms, particularly bacteria and yeast, has been promoted because of the cost reduction with respect to the non-living chemical approach. In that way, computational frameworks can profit from the previous knowledge stored in large databases of compounds, enzymes and reactions. In addition, the cell behavior can be studied by modeling the cellular context. We implemented a heuristic algorithm to find a metabolic pathway from a target compound by exploring a database of enzymatic reactions. We provided examples of designed metabolic pathways using bacteria as host organisms. We further analyzed the designability of certain metabolic pathways of special interest.

The cellular background of operation imposes global constraints on the elements and mechanisms of the networks. Still it is possible to analyze the genome-scale network and how the cell rearranges its genetic profile according to external perturbations. As a case of study, we focused on RNA viruses. Understanding the mechanisms by which the host cell mounts its defenses, and viruses overcome in order to proliferate, has been a challenging problem owing to the multiplicity of factors and complexity of interactions involved. Here, we identified and compared genes that are differentially expressed upon infection. Our results confirm that host cells undergo significant reprogramming of their transcriptome during infection, and that perturbations preferentially affect genes that are highly connected. This indicates redundancy in the information transmission from virus effectors to immune response genes and allowed us to suggest that this over-stimulation of hub proteins could be a mechanism to confer robustness for expression of host defenses.

All in all, our results demonstrate that computational methods are useful in several genetic frameworks. These methods can efficiently explore the fitness landscape to *de novo* design gene regulatory networks, and also they can be applied to analyze the designability of behaviors provided a limited number of interoperable elements. In addition, mathematical analyses revealed the natural design of control mechanisms and organization structures that can be used as rational resources to supplement computational methods. Ideally, these tools would input a set of specifications, in the form of human-readable programs, and would output a biological model together with its compilation into a reliable DNA sequence. The next steps in this field would be directed towards providing a common design platform going from single molecules with regulatory ability to complex biological systems integrating multilayer regulatory elements. However, one of the major challenges in the design of genetically engineered organisms with novel functions is their potential evolution after being deployed. In many instances, the traits of interest may be lost after subsequent cell replications. Accordingly, the quantification of the evolutionary reliability (how many generations the target phenotype is sustained) is pivotal for real biotechnological applications. Ultimately, the engineered cells with synthetic circuitries will result in micromachines working according to the design specifications.

Acknowledgements

I am especially grateful to my advisors, Santiago Elena and Alfonso Jaramillo, for introducing me to the exciting world of Systems and Synthetic Biology. I loved their original inspiration, the scientific discussions with, and remarkably the proximity in the interaction. During this time I have found this research field exciting enough to spend the next years of my work on it. I particularly admire the vision of my advisors and their ability to motivate an industrial engineer like me to become fascinated by Molecular Biology.

Special thanks to Javier Carrera, my office mate, colleague and friend, who has provided valuable feedback throughout my thesis. We both now laugh at the singular circumstances we lived abroad. Also thanks to the past and current members from Elena's group, Nicolas, Guillaume, Susana, Patricia, Clara, Puri, Jasna, Àngels, Stéphanie, Josep, and Mark, and from Jaramillo's group, Pablo, María, Josselin, Filipe, Daniel, Boris, Vijai, and Thomas. They all have fostered camaraderie as well as non-traditional and imaginative thinking.

Special thanks to Javier Forment (bioinformatic service at IBMCP) for providing support with the computational resources. Also thanks to Francisca de la Iglesia for instructive insights in the laboratory when dealing with bacteria. Also mention Miguel Blázquez and Mario Fares, who have engaged me in many interesting discussions about Molecular Biology.

I am also indebted to my advisor at MIT, Kris Prather, when I was in her laboratory. She created a unique and wonderful environment. I want to recall the friends I met there, Barry amongst.

Thanks to all my close friends, David amongst, always an interesting guy to talk about many different subjects.

I would like to also thank Drs. Díaz and Baquero, together with their respective teams, for their special attention. And I do not want to forget the nursery team of La Fe for the extraordinary care and

attention.

Finally, I would like to special thank my parents for all their love and care they have instilled in me since long time, early when starting my engineering studies, following by my Parisian adventure, and finally in my doctorate stage. Also thank all members of my family that supported me, and appreciate the amazing chess games with my brother Miguel. Most of all, I would like to thank, from the deepest of my heart, the support, care, and love I received from my girlfriend Maribel. She has helped me throughout this entire period of my life and indeed deserves much credit for this thesis.