AUTOMATIC COMPUTATIONAL DESIGN OF SYNTHETIC GENOMES

by

Javier CARRERA

Submitted to the Department of Biotechnology
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
Universidad Politécnica de Valencia

Thesis advisors:
Prof. Santiago F. ELENA FITO
Prof. Alfonso JARAMILLO ROSALES

July, 2012

To my parents,
my sister
and Rai.

Abstract

The development of the technology to synthesize new genomes and to introduce them into hosts with inactivated wild-type chromosome opens the door to new horizons in synthetic biology. Here it is of outmost importance to harness the ability of using computational design to predict and optimize a synthetic genome before attempting its synthesis. The aim of this thesis is to help enable the engineering of synthetic genomes of one prokaryotic and two eukaryotic cells by using quantitative genome-scale models. Here, I develop a novel methodology to the automatic design of synthetic genomes which is based on an optimization that computationally mimics genome evolution. First, I address the design of the genomic transcriptional network of *Escherichia coli* with adaptation to varying environments. Applying reverse-engineering methods to the large amount of transcriptomic and signalling data available for the bacterium, I seek to understand the design principles determining the regulation of its transcriptome. I find that *E. coli* genome could be reengineered in such a way that it has a simpler transcriptional regulatory structure while still maintaining the global physiological response to fluctuating environments. These genomes are more sensitive and show a more robust response to challenging environments. Second, I address how virus reprogram the cellular chassis of their host assuming that exist mechanisms by which virus are able to overcome the defenses exposed by the host and modify its gene expression on its own benefit. I develop a novel genome-scale quantitative model of transcriptional regulation of *Arabidopsis thaliana* for exploring the landscape of possible re-engineered genomes. I find a core set of host genes whose knockout or overexpression resulted in predicted transcriptional profiles that minimally deviate from the observed in plants infected. I perform this search for a set of eight viruses for which transcriptomic data are available and compared the results among them. Third, I extend the computational methodology for genome redesign to address the fine-tuning of the tomato fruit agronomic properties. I apply reverse engineering computational methods to transcriptomic, metabolomic

and phenomic data obtained from a collection of tomato recombinant inbreed lines to formulate a kinetic and constrain-based model that efficiently describes the cellular metabolism from the expression of a minimal core of genes. Based on the predicted metabolic profiles, a close association with agronomic and organoleptic properties of the ripe fruit could be revealed with high statistical confidence. The model was used for exploring the landscape of all possible local transcriptional changes with the aim of engineering tomato fruits with fine-tuned biotechnological properties. In sum, our results demonstrate that automated computational methods can efficiently explore the fitness landscape of re-engineered genomes with desired specifications.

Resumen

El desarrollo de tecnologías para sintetizar nuevos genomas e introducirlos dentro de hospedadores con sus respectivos cromosomas naturales inactivados abre las puertas a nuevos horizontes en biología sintética. Es de suma importancia aprovechar la habilidad de usar métodos computacionales para predecir y optimizar genomas sintéticos antes de llevar a cabo su síntesis. El objetivo de esta tesis es propulsar la ingeniería de genomas sintéticos de una célula procariota y otras dos eucariotas usando modelos cuantitativos a escala genómica. As pues, he desarrollado una nueva metodología para el dise no automatizado de genomas sintéticos que se basa en una optimización que computacionalmente imita la evolucin natural de genomas. Primero, he abordado el diseño de la red genómica transcripcional de *Escherichia coli* con adaptación a entornos cambiantes. Aplicando métodos de ingeniería reversa a los datos masivos disponibles de la transcripción y señalización para la bacteria, tratamos entender los principios de diseño que determinan la regulación de su transcriptoma. Encontramos que el genoma de *E. coli* puede ser rediseñado de tal forma que tenga una estructura reguladora transcripcional más simple manteniendo aún la respuesta fisiológica global ante ambientes fluctuantes. Estos genomas son más sensibles y muestran una respuesta más robusta ante ambientes agresivos. Segundo, he estudiado cómo los virus reprogramanlos chasis celulares de sus hospedadores asumiendo que existen mecanismos por los cuales los virus son capaces de superar con éxito las defensas expuestas por los hospedadores y modificar la expresión de sus genes en su propio beneficio. He desarrollado un nuevo modelo cuantitativo de la regulación transcripcional a nivel genómico de *Arabidopsis thaliana* para explorar el paisaje de posibles genomas rediseñados. Encontré un conjunto básico de genes del hospedador cuyos silenciamiento o sobre-expresión dieron pie a la predicción de perfiles de transcripción que se desvían mínimamente de los observados en plantas infectadas. Esta exploración la he realizado para un conjunto de ocho virus para los que se disponían datos transcriptómicos y consecuentemente, los resultados fueron comparados. Finalmente, he ex-

tendido la metodología computacional presentada para rediseño de genomas con el objetivo de abordar la optimización de propiedades agronómicas del fruto del tomate. Apliqué métodos computacionales de ingeniería reversa a datos transcriptómicos, metabolómicos y fenotípicos obtenidos de una colección de linajes recombinantes para formular un modelo cinético basado en restricciones que eficientemente describiese el metabolismo celular usando la expresión de un conjunto mínimo de genes. Basándose en los perfiles metabólicos predichos, relaciones entre las propiedades agronómicas y organolépticas del fruto maduro pudieron ser reveladas con alta confianza estadística. El modelo fue usado para explorar el paisaje de todos los posibles cambios transcripcionales locales con el fin de obtener frutos de tomate con propiedades biotecnológicas de especial interés. Resumiendo, los resultados presentados en esta tesis demuestran que métodos automáticos computacionales pueden explorar eficientemente las puntos óptimos de un paisaje de genomas rediseñados con especificaciones deseadas.

Resum

El desenvolupament de tecnologies per a sintetitzar nous genomes i introduir-los en cèl.lules amb els seus respectius cromosomes naturals inactivats obri les portes a nous horitzons en biologia sintètica. És de la màxima importància aprofitar l'habilitat d'usar metodologies computacionals per a predir i optimitzar genomes sintètics abans de dur a terme la seua síntesi. L'objectiu d'esta tesi és propulsar la inginyeria de genomes sintètics d'una cèl.lula procariota i altres dos eucariotes usant models quantitatius a escala genòmica. Així, he desenvolupat una nova metodologia per al diseny automatitzat de genomes sintètics que es basa en una optimització que computacionalment imita l'evolució natural de genomes. Primer, he abordat el diseny de la xarxa genómica transcripcional d'*Escherichia coli* amb adaptació a entorns canviants. Aplicant mètodes d'inginyeria reversa a les dades massives disponibles de la transcripció i senyalització per al bacteri, tractem d'entendre els principis de diseny que determinen la regulació del seu transcriptoma. Trobem que el genoma d'*E. coli* pot ser redisenyat de tal forma que tinga una estructura reguladora transcripcional més simple mantenint la resposta fisiològica global davant d'ambients fluctuants. Estos genomes són més sensibles i mostren una resposta més robusta davant d'ambients agressius. A continuació, he estudiat com els virus reprogramen els xassissos cel.lulars dels seus hospedadors assumint que hi ha mecanismes pels quals són capaços de superar amb èxit les defenses exposades pels hospedadors i modificar l'expressió dels seus gens en el seu propi benefici. He desenrotllat un nou model quantitatiu de la regulació transcripcional a nivell genòmic d'*Arabidopsis thaliana* per a explorar el paisatge de possibles genomes redisenyats. Vaig encontrar un conjunt bàsic de gens de l'hospedador que variant adequadament la seua expressió van donar peu a la predicció de perfils de transcripció que es desvien mínimament dels observats en plantes infectades. Esta exploració l'he realitzat en un conjunt de huit virus del quals disponiem dades del seu transcriptoma infectat per a comparar els resultats derivats. Finalment, he estés

la metodologia computacional presentada per a rediseny de genomes amb l'objectiu d'abordar l'optimització de propietats agronòmiques del fruit de la tomaca. Vaig aplicar mètodes computacionals d'ingenieria reversa a dades transcriptòmiques, metabolòmiques i fenotípiques obtingudes d'una colecció de llinatges recombinants per a formular un model cinètic basat en restriccions que eficientment descriguera el metabolisme cel.lular usant l'expressió d'un conjunt mínim de gens. Basant-se en els perfils metabòlics predits, relacions entre les propietats agronòmiques i organolèptiques del fruit madur van poder ser revelades amb alta confiança estadística. El model va ser usat per a explorar el paisatge de tots els possibles canvis transcripcionals locals a fi d'obtindre fruits de tomaca amb propietats biotecnológiques d'especial interés. Resumint, els resultats presentats en esta tesi demostren que mètodes automàtics computacionals poden explorar eficientment els punts òptims d'un paisatge de genomes redisenyats amb especificacions desitjades.

Contents

Objectives **1**

1 Introduction **3**
References 5

I Design of a Synthetic Bacterial Genome in Dynamic Environments **9**
Introduction 11
References 13

2 Characterization of Bacterial Response to Synthetic Gene Expression **15**
2.1 Construction of a Predictive Model of the Cell Growth Rate . 17
2.2 Tuning Synthetic Gene Expression 21
2.3 Biological Implications for the Design 25
Appendix A Estimation of Plasmids Concentration and Characterization 26
Appendix B Phenomenological Cellular Chassis Model 27
Appendix C Heterologous Gene Expression Model 27
References 29

3 Design-Guided Models of global Transcription Regulation **31**
3.1 Genome-Wide Quantitative Model of Transcription Regulation of *E. coli* 32
3.2 Design of Artificial Genomes and Validation of their Transcription Profiles 33
3.3 Prediction of Wild-Type *E. coli* Trancriptome 36
3.4 Genome-Wide Model of *E. coli* Integrating Signal Transduction Data 36

3.5 Model Validation: Predicting Growth Rate of Perturbed Transcriptional Networks of *E. coli* 39
3.5.1 Model Validation 1: Prediction of Expression Profiles Upon Genetic and Environmental Changes 39
3.5.2 Model Validation 2: Predicting the Results of *E. coli* Experimental Evolution 41
3.5.3 Model Validation 3: Predicting the Growth Rate of Rewired Transcriptional Networks of *E. coli* 42
3.6 Discussion 43
Appendix A Mathematical Model of Transcription Regulation . . 45
Appendix B Using Network Inference to Obtain a Kinetic Model . 46
Appendix C Construction of a Transcriptional Regulatory Network That Integrates Signal Transduction 49
Appendix D Structure of the Wild-Type Global Transcriptional Model 51
Appendix E Prediction of Transcriptomic Profiles 52
Appendix F Designing Genomes and Expression Data 52
Appendix G *In Silico* Genome Evolution by Adaptive Mutation . 54
Appendix H Prediction of Rewired Transcriptional Network of *E. coli* 55
References 56

4 Automatic Design of a Genome by Gene Refactorization 63
4.1 Design by Gene Refactorization in Dynamic Environments . . 63
4.1.1 Refactored Genomes with a Reduced Number of Operons 63
4.1.2 Cellular Environments Selectively Correlate with Genome Regulatory Complexity in the Refactored Genomes . . 67
4.1.3 Analysis of Biochemical Adaptation in Refactored Genomes 68
4.2 Prediction of a Refactored *E. coli* Genome Sequence with Wild-type Behavior in Changing Environments 69
4.3 Conclusions 71
4.3.1 Biological Consequences of Computational Genome Refactorization 71
4.3.2 Experimentally Testable Predictions 74
Appendix A Automatic Genome Design 75
Appendix B Genome-Wide Optimization Procedure 76
Appendix C Objective Functions for Design 78
Appendix D Genome Optimality Degree in Changing Environments 79
Appendix E Functional Analysis of Genomes 79

Appendix F Topological Properties of Refactored Genomes 79
Appendix G Analysis of Biochemical Adaptation to Varying Environments of the Refactored Genome Sequences 80
Appendix H Selecting Challenging Environments to Generate Different Degrees of Optimality in the Wild-Type Genome . . . 81
References 81

II Viruses Reprogram the Cellular Chassis of their Host 83
Introduction 85
References 87

5 Reverse-Engineering of the *Arabidopsis thaliana* Transcription under Changing Environments 91
5.1 Genome-Wide Transcriptional Control in *A. thaliana* 92
5.2 Transcriptomic Profile Prediction 96
5.3 Selection of Optimality in Changing Environmental Conditions 101
5.4 Conclusions 103
Appendix A Microarray Data 105
Appendix B Inference Procedure 105
Appendix C Model Validation 106
Appendix D Motif Detection and Analysis 106
References 107

6 Reprogramming Plant Cellular Chassis to Mimic Viral Infection 113
6.1 Re-engineered TRNs that Mimic the Transcriptomic Response Characteristic of Different Viral Infections 116
6.2 A Minimal Set of Transcriptional Factors Guarantees TRN Re-designs that Mimic Viral Infections 117
6.3 The Number of Proposed TFs to be Perturbed Correlates with the Magnitude of the Alterations in Gene Expression Observed upon Viral Infection 120
6.4 A Crucial Set of TFs is Pervasively Proposed in the Re-designed TRNs 122
6.5 Proposed TFs Which Are Common for Different Viruses . . . 124
6.6 Discussion 125
Appendix A Plant Viruses and Transcriptomic Data 125
Appendix B Genome-Wide Multiple-Optimization 126
References 127

III Fine-Tuning of the Tomato Fruit Agronomic Properties by Computationa Design 131

7 Computational Optimization of the Tomato Fruit Agronomic Traits 133

Introduction 133

7.1 Dissecting Tomato Genome: Kinectics-Based Models of Transcription, Metabolism and Phenotype 135

7.1.1 A Genome-Wide Transcriptional Model Allows the Integration of Tomato Fruit Metabolism 135

7.1.2 Genome-Wide Transcriptional Model Integrating Metabolism 137

7.2 Computational Optimization of the Agronomic Properties by Lean Manufacturing 138

7.2.1 Genome Redesign by Using Single and Multiple Genetic Perturbations 138

7.3 Experimental Validation Via Prections of Volatile Compounds Correlations 143

7.4 Conclusions 144

Appendix A Plant Material, Transcriptomic, Metabolomic and Phenomic Data 146

Appendix B Mathematical Model 146

Appendix C Construction of an Effective Transcriptional Regulatory Network Connected with Metabolism to Explain Agronomic Properties 147

Appendix D Genome-Wide Multiple-Optimization 148

Appendix E Experimental and Computational Metabolite Correlation 149

References 150

8 General Discussion 155

References 160

Acknowledgements 165

Objectives

The objectives of this thesis are:

1. Development of a computational framework to design genomic transcriptional networks with adaptation to varying environments. Application of optimization methods to explore the combinatorial space.

2. Characterization of the designability of bacterial re-engineered genomes. Quantitative study of synthetic transcriptional regulatory networks. Application of optimization methods to unravel design principles.

3. Empirical model and *in vivo* characterization of the bacterial response to synthetic gene expression. Determination of the type of cellular resource allocation that limits growth rate in the genome redesign.

4. Reverse-engineering of the *A. thaliana* transcriptional network under changing environmental conditions. Genomic network-based dissection of the cell response under viral infection. Reprogramming cell to mimic the transcriptome of *A. thaliana* upon viral infection by using computational genome design.

5. Apply reverse-engineering computational methods to transcriptomic, metabolomic and phenomic data obtained from a collection of tomato recombinant inbreed lines to formulate a kinetic and constrain-based model that describes the cellular metabolism from the expression of a minimal core of genes and the tomato phenotype from a critical set of metabolites. Explore the landscape of local transcriptional changes with the aim of engineering tomato fruits with fine-tuned biotechnological properties.

Chapter 1

Introduction

Making biology easier to engineer.
– Drew Endy

The availability of technology to synthesize new genomes and transplant them in cells urges us to develop methodologies to design such genomes. Computational design provides a way to overcome the complexity of the system. In this review we will focus on a few essential aspects that would have to be solved to obtain a computational genome design methodology. The forward engineering of a genome requires a good understanding of the cell at the molecular level. This understanding could be incorporated as inferred models from functional genomics data.

Over the last ten years, there has been a continued effort in the systems biology community for developing genome-scale models [1, 2]. They consist of physico-chemical descriptions of intracellular processes implementing bottom-up strategies using biological databases and the scientific literature [3, 4]. Although such models are very useful to analyze small systems, they pose some difficulties at the large-scale level due to the complexity of the cell, the lack of kinetic parameters and evolutionary and population effects [5]. Contrarily, reverse-engineering methods are applied to experimental data to obtain models [6, 7]. In that way, the technological developments to produce high-throughput-omics data (such as genomics, transcriptomics, proteomics, metabolomics or phenomics) have been pivotal towards that

end. Such data are used to train learning algorithms that output mathematical models able to predict the molecular behavior of the cell under different conditions. Some of such algorithms are also able to generalize and make predictions for alternative cell rewirings.

The goal of this thesis is to show how some of the current methods for predicting the behavior at a genomic scale could be used in a global methodology to design computationally new genomes. They could be later synthesized from chemical components to obtain cells with targeted properties, such as the bio-production of combustibles or for medical applications. The large number of components of a cell will probably always require the use of automated methods able to make predictions with a large number of equations. The improvement of computational methods will allow creating experimentally a full synthetic genome implementing the re-engineering of cells with highly reduced genomic complexity.

A computational genome design methodology requires using a biochemical model of the cell able to predict a phenotypic fitness function such as cell growth. The usual way to do this is by reconstructing metabolic networks from sequenced organisms, which allows analyzing their metabolic capabilities and engineering synthetic strains [8, 9]. This process is based on generating a stoichiometric model by using genome annotation and specific data about the enzymatic reactions carrying out in the organism. On the other hand, the reconstruction of a global regulatory network is facilitated by using data from microarray experiments. We could highlight several approaches, such as clustering, information-theory or Bayesian methods [2]. The use of prior knowledge on regulatory targets and biological databases, such as RegulonDB [10] and EcoCyc [11] for *E. coli*, can improve the accuracy of the models. Particularly, metabolic and transcription networks of the bacterium *E. coli* [12, 13] or the yeast *S. cerevisiae* [14, 15] have been widely studied.

Now, with the development of DNA technology, the full synthesis of a genome is conceivable [16, 17, 18]. The computational approach allows us to integrate large-scale mathematical models in a common platform to facilitate genome design. There has been some initial work on the automated design of genomes by rewriting the codons to change gene expression or to avoid specific restriction sites [19]. Furthermore, the use of molecular models will allow the design of synthetic genomes based on functional DNA cassettes working as independent modules. Here, we describe the current state-of-the-art on computational methods that could be applied to the design of genomes. One possible starting point to computationally design or

redesign a genome would be starting by the design of the metabolism. Here we would also have to make sure we include the essential genes, those to be required in a minimal cell [20]. Later, we could incorporate transcription regulation, coupled to signal transduction, to better optimize the cell response to alternative environments. Finally, we would have to provide the nucleotide sequence, for which we would have to deal with the genome organization of genes. Of course, this is a crude simplification of the real steps needed to design a genome, but it is a good starting point given the primitive status of the modeling in this area. Here we apply optimization methods to design synthetic genomes in one prokaryotic (*E. coli*) and two eukaryotic (*A. thaliana* and tomato fruit) cells with a desired behavior. Although we abstract each problem, it is expected the use of computational techniques to design cells with highly reduced genomic complexity involving several regulatory mechanisms, as certainly it occurs in natural systems. We approach the design of genomes as the inverse problem of finding the right regulatory biological components at play and their precise position in the circuit, by taking advantage of libraries of inferred-characterized promoters and open reading frames of the wild-type genomes.

References

[1] Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L., Palsson, B. O. (2009). Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129-143.

[2] Bonneau, R. (2008). Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.* 4, 658-664.

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

[4] Leduc, M., Tikhomiroff, C., Cloutier, M., Perrier M., Jolicoeur, M. (2006). Development of a kinetic metabolic model: application to Catharanthus roseus hairy root. *Bioprocess Biosyst. Eng.* 28 295-313.

[5] Palsson, B.O. (2002) *In silico* biology through omics. *Nat. Biotechnol.* 20, 649-650.

[6] Yeung, M. K. S., Tegner J., Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6163-6168.

[7] Tegner, J., Yeung, M. K. S., Hasty, J., Collins, J.J. (2003) Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5944-5949.

[8] Stephanopoulos, G. (1994). Metabolic engineering. *Curr. Opin. Biotechnol.* 5, 196-200.

[9] Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryaninc, I. I., Selkov E., Palsson, B. O. (2001) Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.* 26, 179-186.

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

[11] Karp, P. D., Keseler, I. M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S. M., Paulsen, I., Collado-Vides, J., Gama-Castro, S., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Bonavides-Martinez C., Ingraham, J. (2007). Multidimensional annotation of the *Escherichia coli* K-12 genome, *Nucleic Acids Res.* 35, 7577-7590.

[12] Edwards, J. S., Palsson, B. O. (1999) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition characteristics and capabilities. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5528-5533.

[13] Thieffry, D., Huerta, A. M., Perez-Rueda E., Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *BioEssays* 20, 433-440.

[14] Forster, J., Famili, I., Fu, P., Palsson B. O., Nielsen, J. (2003). Genomescale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 13, 244-253.

[15] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R. J., Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281-285.

[16] Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., Stockwell, T. B., Brownley, A.,

Thomas, D. W., Algire, M. A., Merryman, C., Young, L., Noskov, V. N., Glass, J. I., Venter, J. C., Hutchison III, C. A., Smith, H. O. (2008). Complete chemical synthesis, assembly and cloning of a *Mycoplasma genitalium* genome. *Science* 319, 1215-1220.

[17] Dymond, J. S., Richardson, S. M., Coombes, C. E., Babatz, T., Muller, H., Annaluru, N., Blake, W., J., Schwerzmann, J. W., Dai, J., Lindstrom, D. L., Boeke, A. C., Gottschling, D. E., Chandrasegaran, S., Bader, J., Boeke, J. D. (2011). Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature*, 477, 471-6.

[18] Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest C. R., Church, G. M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894-898.

[19] Richardson, S. M., Wheelan, S. J., Yarrington R. M., Boeke, J. D. (2006). GeneDesign: Rapid, automated design of multikilobase synthetic genes. *Genome Res.* 16, 550-556.

[20] Forster A. C., Church, G. M. (2006) Towards synthesis of a minimal cell. *Mol. Syst. Biol.* 2, 45.

Part I

Design of a Synthetic Bacterial Genome in Dynamic Environments

Introduction

The doors to new horizons in genome-scale synthetic biology have been opened by the recent and rapid development of technologies allowing the synthesis of novel genomes and their introduction into hosts with inactivated or deleted wild-type chromosomes [1, 2, 3]. The *de novo* design of cells with synthetic genomes that are viable in a well-defined environment might require only the constitutive expression of the minimal set of genes required for life [4]. This engineering approach, however, has several problems, including the absence of all necessary blocks (*e.g.*, genes, signaling cascades, etc.), the absence of a good definition of the minimal set of genes required, and a poor understanding of the pleiotropic negative effects that these genes may have when put together. In contrast, the re-engineering of an existing genome to change its regulation network would not require adding new genes to the genome but only their rearrangement with respect to promoter sequences. Previous work has considered the rearrangement of genomic sequences. For example, Chan et al. successfully modified the T7 genome to remove overlapping translational frames [5]. This approach was inspired by the engineering practice called refactoring, in which the internal structure of an already existing system is rearranged while its external function is maintained. Based on the same refactoring principle, and considering cell behavior as the external function, we have created a system for the design of a novel genome sequence with a refactored transcriptional regulatory network (TRN) that maintains its original behavior.

In the context of synthetic biology [6], the design of an organism that can respond in a directed way to variations in its environment has been a particularly interesting and challenging problem. This design would require the reengineering of suitable signal transduction and regulation systems [7, 8, 9, 10]. Because transcriptional regulation is the most well-studied regulatory system in bacteria, it may be a good starting point for those interested in the design of such systems. In fact, the recent experimental evolution of *E. coli* under changing environments has provided evidence of regulatory network rearrangements that allow anticipatory behavior [11]. However, the *de novo* design of a genome that can adapt to changing environments may be very challenging. A simpler alternative might be to alter a pre-existing genome by reshuffling its genes in such a way that its behavior is maintained. In particular, this problem can be treated computationally if restricted to the re-design of the global transcriptional network for an organism for which sufficient transcriptomic information is available.

To evolve new genomes *in silico*, a necessary first condition is to de-

fine a biologically meaningful fitness function that allows changes that are introduced during the evolution process to be evaluated. How can we define such fitness function? Interestingly, it has recently been shown that the transcriptomic expression profile is a good predictor of instantaneous cell growth in *S. cerevisiae* [12]. Assuming that this relationship is true for other organisms, it can be hypothesized that the expression profile of a given system determines cell growth. This can also be rationalized by arguing that natural selection results in nearly optimal biomass production by favoring regulation pathways that confer optimal levels of gene expression in a given environment. In this line, Tagkopoulos et al. used a Pearson correlation between the abundance of cell resources and the response of gene expression as a fitness to computationally evolve the biochemical network of *E. coli* in variable environments.

We can evaluate the validity of this hypothesis by analyzing the effect of mutations on the growth of a wild-type strain. Notably, this evaluation still requires the accurate prediction of a genome-scale expression profile. More modifications to the genome will lead to less growth and more differences in the expression profile. Therefore, we have developed an automated methodology for designing a genome based on an *in silico* evolution process; the methodology uses similarity to a wild-type transcriptional profile as its fitness function, which provides the variation of cell growth. We have used experimental short-term evolution results to validate the hypothesis that the distance between predicted and wild-type expression profiles is correlated with the difference between cell growth rates. Furthermore, it is possible to construct regulatory network models that accurately predict the global transcriptional profile for some organisms [13, 14]. These regulatory network models can be used to predict the growth of cells with modified transcriptional networks, thereby providing the fitness function required to evaluate their performance under various environmental conditions.

Recent experiments investigating the evolvability of bacterial TRNs have shown that adding new links to the network does not significantly alter cell growth. Isalan et al. added transcriptional fusions of *E. coli* promoters with different *E. coli* master transcription regulators [15] and showed that the bacteria tolerated almost all rewired networks; however, their growth was perturbed by as much as 5% [3]. This inherent predisposition of bacterial networks to dampen extreme changes in their circuitry enables the possibility of conducting genome-wide rewiring [17] in the *E. coli* TRN.

In this Part I, we describe a methodology for designing genomes that produce cells with targeted physiological responses to a set of environments.

This requires the integration of current known phenomic (Chapter 2), transcriptomic and signaling (Chapter 3) data into a global model consisting of differential equations, allowing the assignment of parameters to promoter and transcription factor (TF) coding sequences. We validate the TRN using experimental expression profile data. After a suitable model was generated, we validated the fitness function to be used in our *in silico* genome design procedure. We used experimental results from a laboratory evolution experiment to show that measured growth rate differences correlate with variations in fitness. This allowed us to perform an *in silico* genome evolution simulation with the aim of refactoring the *E. coli* genome to simplify its internal structure by reducing the number of operons and indirectly minimizing the interactions necessary for the TRN (Chapter 4). We found that we could dramatically reduce the number of operons while maintaining the organisms response to fluctuating environments. We also analyzed other properties of the synthetic TRN, such as its topology and adaptation to varying environments. Finally, we examine some design principles that can be inferred from our results, the tests of our experimental predictions and future experimental applications of this work.

References

[1] Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., Stockwell, T. B., Brownley, A., Thomas, D. W., Algire, M. A., Merryman, C., Young, L., Noskov, V. N., Glass, J. I., Venter, J. C., Hutchison III, C. A., Smith, H. O. (2008). Complete chemical synthesis, assembly and cloning of a *Mycoplasma genitalium* genome. *Science* 319, 1215-1220.

[2] Lartigue, C., Glass, J. I., Alperovich, N., Pieper, R., Parmar, P. P., Hutchison III, C. A., Smith, H. O., Venter, J. C. (2007). Genome transplantation in bacteria: changing one species to another. *Science* 317, 632638.

[3] Dymond, J. S., Richardson, S. M., Coombes, C. E., Babatz, T., Muller, H., Annaluru, N., Blake, W., J., Schwerzmann, J. W., Dai, J., Lindstrom, D. L., Boeke, A. C., Gottschling, D. E., Chandrasegaran, S., Bader, J., Boeke, J. D. (2011). Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature*, 477, 471-6.

[4] Forster A. C., Church G. M. (2006) Towards synthesis of a minimal cell. *Mol. Syst. Biol.* 2, 45.

[5] Chan, L.Y., Kosuri, S., Endy, D. (2005). Refactoring bacteriophage T7. *Mol. Syst. Biol.* 1, 2005.0018.

[6] Khalil, A. S., Collins, J. J. (2010). Synthetic biology: applications come of age. *Nat. Rev. Genet.* 11, 367-379.

[7] Carrera, J., Rodrigo G., Jaramillo, A. (2009). Towards the automated engineering of a synthetic genome. *Mol. Biosyst.* 5, 733-743.

[8] Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92-96.

[9] Bonneau, R. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131, 1354-1365.

[10] Ulrich, L. E., Zhulin, I. B. (2009). The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res.* 38, D401-7.

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

[12] Airoldi, E. M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A. A., Dunham, M. J., Broach, J. R., Botstein, D., Troyanskaya, O. G. (2009). Predicting cellular growth from gene expression signatures. *PLoS Comp. Biol.* 5, e1000257.

[13] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

[14] De Smet, R., Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8, 717-729.

[15] Bhardwaj, N., Kim, P. M., Gerstein, M. B. (2010). Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci. Signal.* 3, 146.

[16] Isalan, M., Lemerle, C., Michalodimitrakis, K., Beltrao, P., Horn, C., Raineri, E., Garriga-Canut, M., Serrano L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840845.

[17] Bashor, C.J., Horwitz, A.A., Peisajovich, S.G., Lim, W.A. (2010). Rewiring cells: Synthetic Biology as a tool to interrogate the organizational principles of living sytems. *Annu. Rev. Biophys.* 39, 515-537.

Chapter 2

Characterization of Bacterial Response to Synthetic Gene Expression

Evolution has optimized organisms at the genetic level to maximize fitness in their living environment. This entails a proper expression of the transcriptome, proteome and metabolome. In addition, the cell expresses the required machinery for replication and for maintaining appropriate intracellular metabolic patterns to profit from the available external substrates. Plasmids use the same machinery to express their encoded genes and to replicate. Therefore, the expression of heterologous systems encoded by plasmids would require a fraction of resources [1, 2]. In addition, the external elements could interact with the host ones, resulting in an eventual disruption of the precise intracellular organization and regulation [3].

Herein, we present a phenomenological model for calculating cell growth rate when expressing a heterologous system. For simplicity, we focused on prokaryotic hosts and heterologous devices implemented in plasmids that apparently do not disrupt the interactome of the host. We developed a model to estimate the consumption by a foreign genetic device of resources from the host cell, such as DNA polymerases for replication, RNA polymerases for transcription, or ribosomes for translation. The extent of this resource usage is rarely quantitatively known or reported. Furthermore, the level of resources that can be drained from a host cell by a foreign system without perturbing the behavior of the cell is poorly understood. In addition, perturbations in cellular behavior, in turn, affect the heterologous expression of the device. Together, this hampers the prediction of the device function and results into many ad hoc redesigns (often based on an inadequate knowledge

of resource usage) required to avoid malfunctions in such combination.

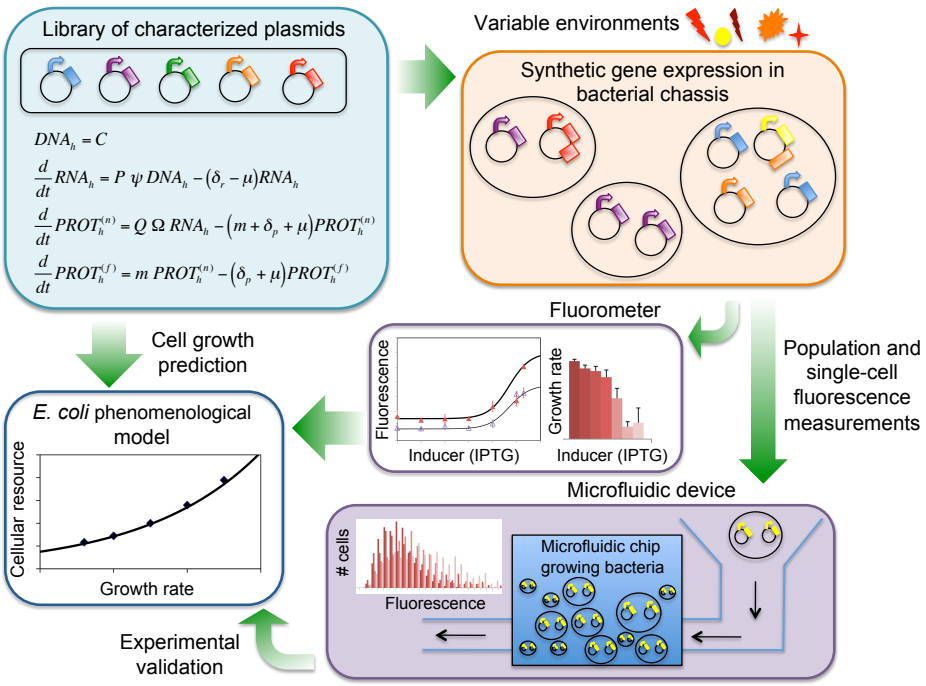

Figure 2.1: Scheme of the phenomenological model of a bacterial chassis under synthetic gene expression and its validation using population and single-cell fluorescence measurements.

Over the last decade, there have been multiple examples of engineer systems performing useful functions [4]. Generally, such devices have been expressed in plasmids. The choice of system components is often based on considering appropriate promoter strength, protein stability, or plasmid copy number. In principle, focusing on the internal determinants of the device would ideally be sufficient to predict its behavior within the cellular background and to allow for the reliable construction of systems that function as expected. In practice, however, the behavior of the engineered devices is partially determined by external factors to the devices themselves, such as unpredictable interactions between the system and the host cell. Still, even in the theoretical case of absence of interactions (*i.e.*, orthogonal systems), the dynamics of the system depends on the cell growth rate [5], which is determined by the level of heterologous expression [1, 2]. Therefore, it should be of extremely utility for an accurate prediction of the behavior of cells expressing such heterologous devices to provide a simple model of an interface that couples device dynamics and cell growth rate.

In this chapter, we experimentally construct and characterize a simple

genetic device to study the cost that the cell pays to express it. A simple mathematical model based on ordinary differential equations (ODEs) was used to calculate the amount of foreign mRNAs and proteins and hence to estimate the resources drawn from the cell by the heterologous device [6]. We took advantage of a detailed description of the chemical composition of the host cell, comprising of the set of molecular species that directly interact with the device, as a function of growth rate [7]. Such a host cell model could be used, in turn, to predict the reduction in growth rate caused by the consumption of cellular resources. With these very simple complementary models we could begin to tackle how the host cell fitness changes in response to a depletion of individual or multiple cellular resources caused by the heterologous expression and what is the limiting resource that entails the reduction in the host fitness (Figure 2.1).

2.1 Construction of a Predictive Model of the Cell Growth Rate

We developed a model based on ODEs to quantify the different levels in steady state of DNA, RNA and protein molecules of a plasmid. It allowed for computing the number of molecules of consumed resources, which the plasmid required. Hence, we predicted the cellular fitness as the cell growth rate in presence of plasmids, which is a function of their genetic load. In the light of previous experimental results, we investigated the origins that cause the fitness reduction in the host organism when expressing a heterologous device. The estimation of the new growth rate comes from solving our model by using the amount of resources that were consumed by the plasmid (see Appendix B). Hence, we obtained

$$\begin{aligned} \mu &= \frac{1}{0.923} \ln \frac{\mathrm{DNAP} - \xi_d \mathrm{DNAP}_h}{155} \\ \mu &= \frac{1}{1.529} \ln \frac{\mathrm{RNAP} - \xi_r \mathrm{RNAP}_h}{910} \\ \mu &= \frac{1}{1.775} \ln \frac{\mathrm{RIB} - \xi_p \mathrm{RIB}_h}{3690} \end{aligned} \tag{2.1}$$

in which the factors ξ_i account for the variation in the demand of resources by the host cell with the growth rate, increasing the expression of the cell genetic profile. Analogously to equation (2.4C) for the plasmid demand of cellular resources, they scale as $\xi_d \propto D\mu/c_d$, $\xi_r \propto RD/c_r$ and $\xi_p \propto AR\mu/c_p$. Afterwards, the fraction of heterologous RNA and protein is

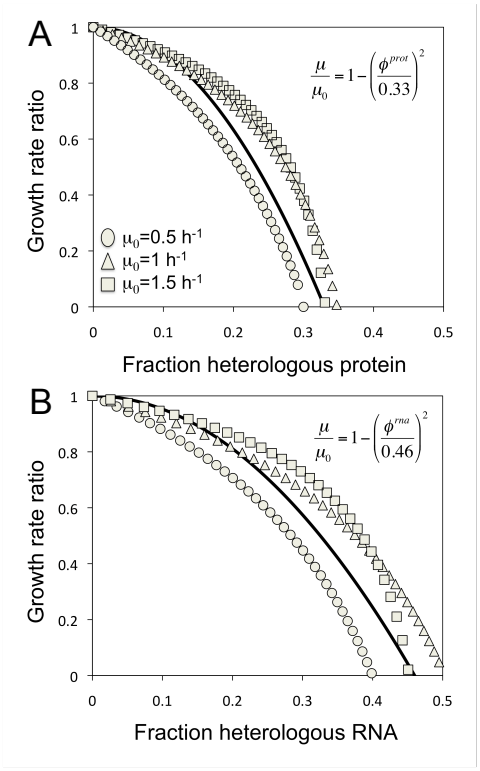

Figure 2.2: Estimation of the growth rate reduction expressing a heterologous device. The growth rate ratio (μ/μ_0) depends on the (a) fraction of heterologous protein (ϕ_{prot}) fitted with $R^2 = 0.89$ and (b) heterologous RNA (ϕ_{rna}) fitted with $R^2 = 0.77$. We applied the equations 2.3 with $\mu_0 = 0.5h^{-1}$ (circles), $\mu_0 = 1h^{-1}$ (triangles) and $\mu_0 = 1.5h^{-1}$ (squares), and we then inferred the model (equation 2.3) as an empirical quadratic relation (solid line). We used $\psi P_0 = 1000$ RNA-molecules/h, $\Omega Q_0 = 600$ protein-molecules/h, $\xi_{RNAP} = 1 + 0.3e^{0.923\mu}\mu^{-0.238}$, and $\xi_{RIB} = 1 + 0.15e^{1.125\mu}\mu^{0.613}$ in the simulations, and varied the plasmid copy number (C).

given by

$$\phi^{rna} = \frac{\mathrm{RNA}_h}{R/l + \mathrm{RNA}_h} \tag{2.2}$$
$$\phi^{prot} = \frac{\mathrm{PROT}_h^{(total)}}{3A/l + \mathrm{PROT}_h^{(total)}}$$

We observed that the amount of DNA polymerases available is very high compared to the eventual requirements for plasmid replication, so their effect on the growth rate can be neglected. Thereby, we could apply our model, based on physical principles and a modulation of the chemical composition of the cell with the growth rate, to obtain an empirical formulation for the dependence between the growth rate reduction and the load imposed by the heterologous device on the cell (Figure 2.2). For a same conceptual system with different genetic loads (specifically, we varied the plasmid copy number), we calculated the corresponding amounts of mRNA and protein (equations (2.1C), (2.2C) and, (2.3C)). With such, we then calculated the factors ϕ^{rna} and ϕ^{prot} (equations (2.2)), and the amount of cell resources required for heterologous expression (equations (2.4C)), which allows for the estimation of the new growth rates (equations 2.3). By representing the growth rate as a function of the factors ϕ_{rna} and ϕ_{prot} (Figure 2.2), we indentified a quadratic dependence following

$$\frac{\mu}{\mu_0} = min\left(1 - \left(\frac{\phi^{rna}}{0.46}\right)^2, 1 - \left(\frac{\phi^{prot}}{0.33}\right)^2\right) \tag{2.3}$$

Our model predicts that the maximal cell capacity that could be allocated for heterologous expression is 46% of the total RNA and 33% of the total protein. These levels are derived by imposing $\mu = 0$ in equation 2.3. This result indicates that the cell cannot dedicate unlimited amounts of resources for the expression of synthetic genes, because this affects the expression of the host transcriptome and proteome. Hence, the cell stops growing when the metabolism is not capable of meeting the internal demand for essential components. The latter is a consequence of the reduction in expression of certain key wild-type genes. Usually, ribosomes are the limiting resources, since the amount of cellular protein is considerably higher than the one of mRNA. Nevertheless, for heterologous systems that are expressing solely non-coding RNAs, RNA polymerases will be the resources to take into account when analyzing the bacterial response. Importantly, these maximal capacities, together with the quadratic relationship, were es-

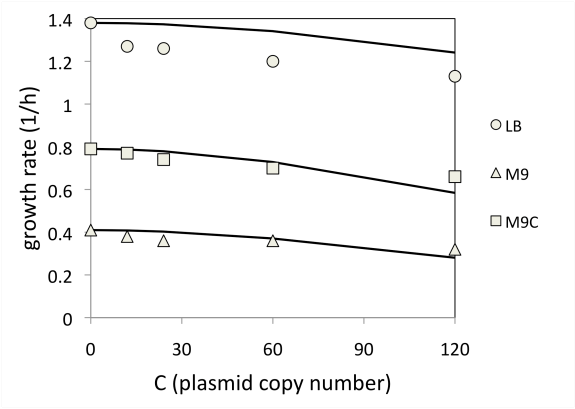

Figure 2.3: Prediction of cell growth rate using our empirical model. Initial specific growth rates $\mu_0 = 1.38h^{-1}$, $0.41h^{-1}$ and $0.79h^{-1}$ corresponding to LB, M9 and M9C media measured experimentally were represented by circles (9.8% of average relative error), triangles (8.7% of average relative error) and squares (5.8% of average relative error), respectively. The model predictions were given by the continuous lines corresponding to ribosomes as limiting resources. Values of the input model variables are $l = 2500$ base pairs, $\psi P_0 = 1000$ RNA-molecules/h, and $\Omega Q_0 = 500$ protein-molecules/h.

timated from basic physical principles and a simple phenomenological model, and remarkably they are in tune with previous experimental work [1].

Consequently, we applied our model (equation 2.3) to predict cellular growth rates in several experimental conditions including differential physical loads as the ones published by Bailey [8]. In the cited work, three distinct media were utilized to grow cultures of *E. coli*: LB, M9, and M9 with casamino acids (M9C), all of them at 37C. In these media, bacteria grew at different rates, $\hat{\mu}_0 = 1.99$, 0.59 and 1.14 doublings/h respectively. Different copy numbers were studied by using plasmids derivative of RSF1050 [9]. For that the analysis of the system, we considered the following physical properties: number of promoters $\psi = 2$, number of ribosome-binding sites $\Omega = 2$, $L = 8000$ and $l = 2500$ base pairs, $C = 12$, 24, 60 and 120 plasmid copies per cell, $P_0 = 500$ RNA-molecules/h and $Q_0 = 250$ protein-molecules/h [10]. Figure 2.3 shows the prediction of growth rate for each copy number in the three different media (LB, M9 and M9C). Not surprisingly, cells grew faster for richer culture media and we used the growth rate in absence of plasmid as initial condition to apply our model. Interestingly, the limiting cellular resources in all cases were ribosomes, being RNA polymerases subsidiary elements in the model. Even though, due to the low genetic load imposed by the plasmid used in [8], our model predicted small changes in the growth rate of the bacterium. In addition, our model shows that there should be a larger

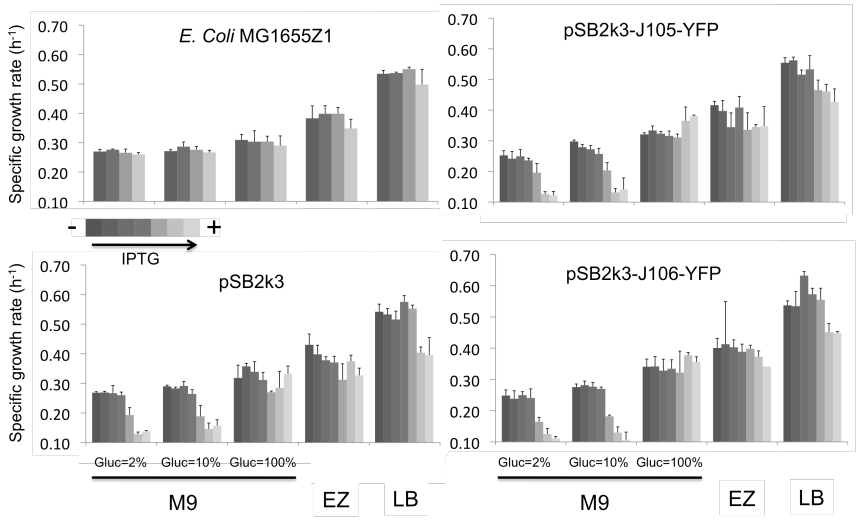

Figure 2.4: Specific growth rate for four different bacterial strains (*E. coli* MG1655Z1 without plamid and with a plasmid, pSB2k3 incorporating a insert which containg two different heterologous devices) growing in five different culture media (three M9 with different levels of the carbon source, EZ and LB) induced by seven concentrations of IPTG (0, 1, 10, 100, 1000, 10000, and 20000 μM).

dependence of the growth rate with copy number for richer mediums due to the fact that such growth rates depends linearly with the medium richness (defined as the growth rate under a given condition). This dependence was also observed in the experimental points of Figure 2.3.

2.2 Tuning Synthetic Gene Expression

To study the cost paid for the heterologous expression calculated in loss of host organism fitness, we used a plasmid as external genetic device and the bacterium *E. coli* as cellular chassis. The plasmid carried a yellow fluorescence protein (EYFP) under the control of a constitutive promoter. Moreover, the plasmid replication was under the control of a LacI repressible promoter. We used the strain MGZ1, which over-expressed the repressor LacI, and thus the plasmid copy number could be tuned according to the concentration of the external inducer IPTG. We studied two promoters with medium (J23106) and weak (J23105) strengths, in different culture media (M9 with three different concentrations of glucose, EZ and LB; see Appendix A). We would expect that rich media entailed a higher growth rate and then a higher support to express the heterologous device. Figure 3.6 shows the quantification of reduction in growth rate for each strain expressing the het-

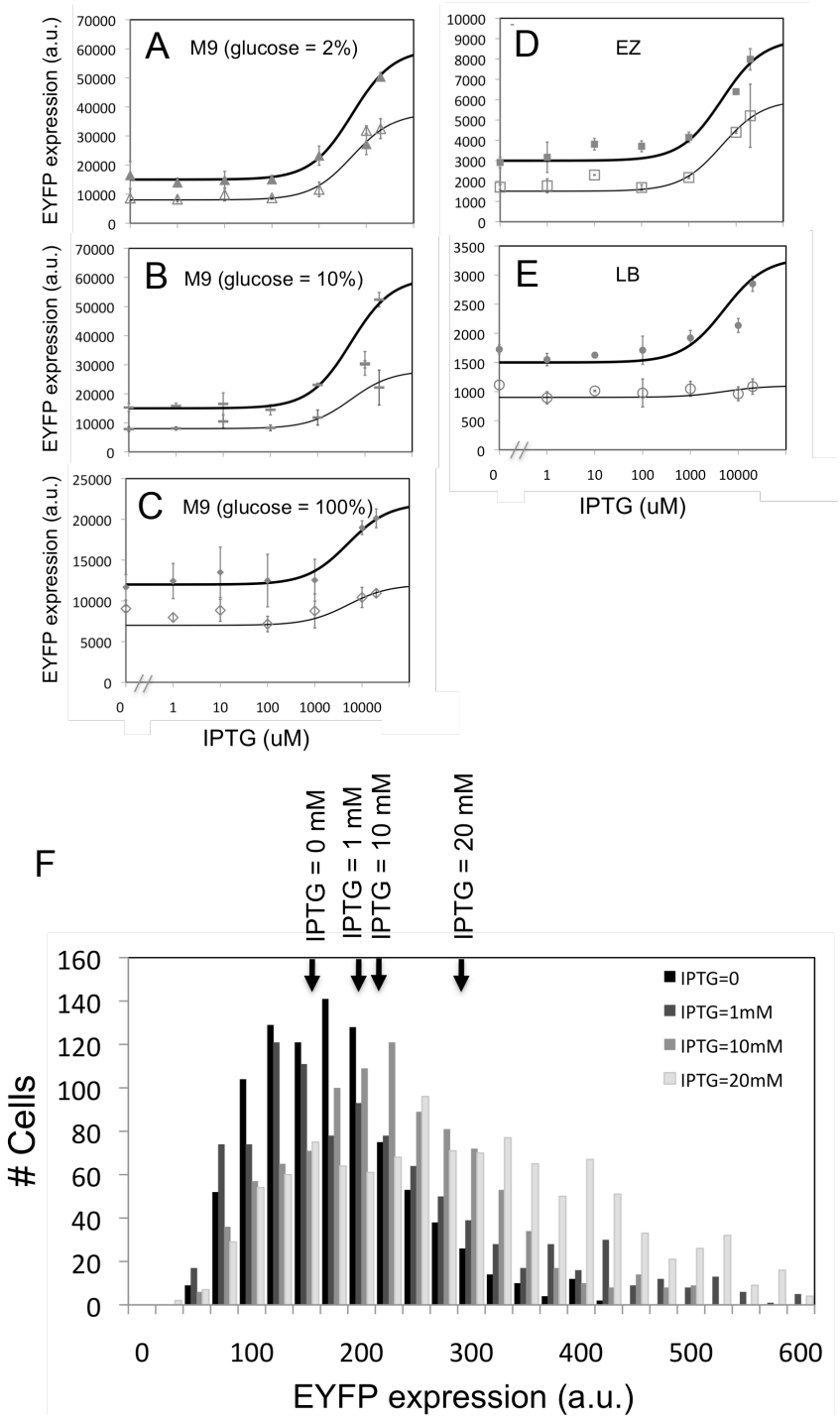

Figure 2.5: (A-E) Apparent EYFP expression from fluorescence data in different culture media for the two devices J23106+EYFP (thick line and filled symbols) and J23105+EYFP (thin line and open symbols). Data were fitted to the model $B_0\left(1+\frac{1}{1+(5000/IPTG)}\right)$, in which B_0 was the fitting parameter (a.u. indicates arbitrary units of protein expression). Error bars show fluorescence measurements of three replicates. (F) Apparent EYFP expression controlled by J23106 promoter inserted in pSB2k3 and measured by microfluidics platform in LB medium. Arrows show average values of EYFP expression for different fluorescence distributions.

erologous devices under different environmental conditions. Interestingly, we captured a slight decrease on the cell growth rate with or without heterologous expression when we added IPTG for the five culture media tested.

We then studied the tunning of the heterologous expression inducible by IPTG. In Figure 2.5 (A-E) we show the transfer functions for the apparent EYFP expression (*i.e.*, the experimental value of $PROT^{(f)}$ in steady state). Fluorescent protein expression was almost duplicated comparing the levels of both promoters. In the regime of IPTG studied, we inferred that the protein expression was proportional to $1 + \frac{1}{1+(5000/\mathrm{IPTG})}$. To further verify the increase of protein expression with IPTG, we performed an independent analysis using microfluidic techniques, testing the system pSB2k3-J23106. We also detected a significant increase in the EYFP expression (Figure 2.5), which was indicative of an induction caused by IPTG. Subsequently, we estimated the plasmid copy number by quantifying the plasmid DNA concentration. The number of plasmid copies displayed a similar dependence on IPTG concentration to the fluorescent protein amount.

At this point, we used the model to estimate the cost of expressing of the systems pSB2k3-J23106 and pSB2k3-J23105. From Figure 2.5(A-E) we inferred a ratio of 1.5 in terms of expression for these systems. Hence, we calculated the growth rate using our empirical model (equation 2.3), being ribosomes the limiting resources. Then, the genetic load of the system was quantified as

$$\phi^{prot} = A_0 \frac{e^{-1.125\mu_0}}{\mu_0 + 0.5} \left(1 + \frac{1}{1 + 5000/\mathrm{IPTG}}\right) \tag{2.4}$$

where A_0 was a parameter fitted to $A_0 = 0.04$ for pSB2k3-J23105 and $A_0 = 0.06$ for pSB2k3-J23105, resulted of increasing 1.5 times the value corresponding to the system with the weak promoter. The multiplicative factor depending on μ_0 derived from the transcription rate (equation (2.1C)) and the dilution term (equations (2.2C) and (2.3C)) to compute the amount of heterologous protein and the variation of the total amount of amino acids usage by the cell (equation (2.4B)). Hence, Figure 2.6 shows the predictions of our model observing a big agreement with the experimental data. However, we also found some discrepancies, certainly because the complexity in the accurate measurement of the bacterial growth rate. All in all, this evidences that the heterologous expression causes a cost for the host cell and that this effect can be explained by the sequestration of ribosomes by the mRNA expressed from the plasmid [11]. Very importantly, our model could be used in further genetic engineering projects of the bacterium *E. coli*, or eventually the approach could be reproduced to infer a model for

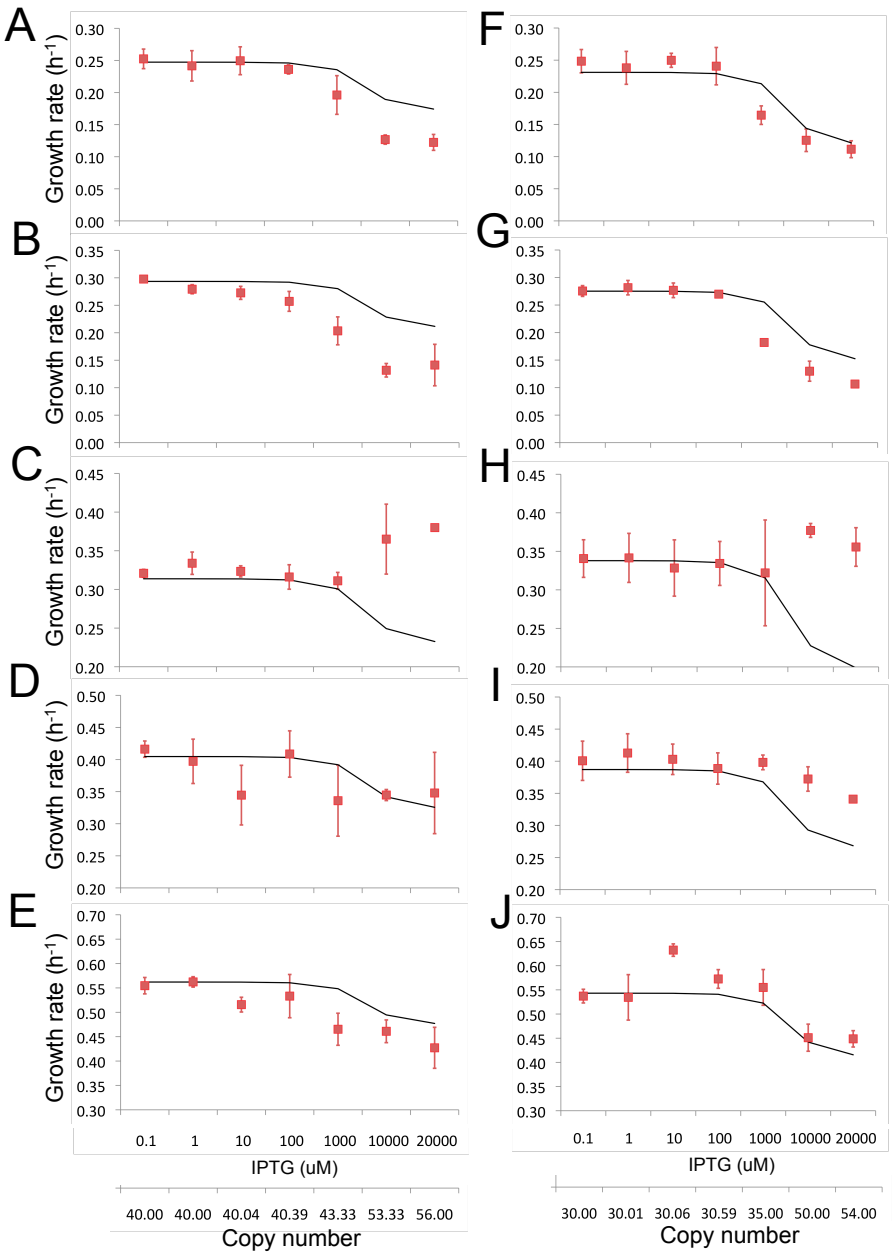

Figure 2.6: Specific growth rate predictions (solid line) by the model (equation 2.4) compared with respect to the experimental measurements (squares) for the strain *E. coli* MG1655Z1 with pSB2k3 containing an insert with the weak and strong promoter (left and right columns, respectively) (see appendices) growing in M9 with 2 (A, F), 10 (B, G) and 100% (C, H) of glucosce, EZ (D, I) and LB media (E, J). We estimated the initial cellular resources by the growth rate measured with pSB2k3 without insert. Notice that error bars in the experimental measurements represented three replicates.

other organisms.

2.3 Biological Implications for the Design

In this chapter, we have studied the impact that causes a heterologous system into a cellular background (here, we restricted ourselves to prokaryotes). To analyze the cellular growth rate we engineered a vector with a tunable origin of replication carrying a reporter device. We constructed a phenomenological model for the heterologous system by considering RNA and protein expression, which cause genetic load on the cell by requiring the allocation of certain resources from the host machinery. We relied on previous characterization of the growth rate dependence on cellular resources (in this case, DNA and RNA polymerases, and ribosomes), and we inferred an analytic model. To estimate the genetic load (fitness reduction at the expense of heterologous system expression), we calculated the amount of sequestered resources to come back to the model and obtain the new growth rate. Strikingly, we found that the cell supports at most a 46% more of heterologous RNA and at most a 33% more of heterologous protein, following a quadratic expression. To validate this model, we studied different systems constitutively expressing a fluorescent protein (caused by variation in plasmid copy number and promoter strength), which were grown in different culture media. Our model allowed us to predict the corresponding genetic loads and, among the considered resources, ribosomes appeared as the limiting resource that caused the growth reduction. Thereby, since the quantification of the cell growth rate is essential to better calculate the dynamical and stationary concentration levels of the engineered system [5], having a model able to predict such a growth rate before the implementation into a given chassis and environment would improve the design process. Our model will be useful to decide among alternative plasmid replicons or synthetic constructions on the basis of their effect in cell growth, or applied to subtract the effects of the chassis while characterizing the transcription or translation rate of a given regulatory element.

Biotechnological and biomedical applications would entail certain interplay between the heterologous system, the wild-type cell and the living environment [12]. In particular, expression of transcription factors that regulate wild-type genes or enzymes that reroute the host metabolism would change the transcriptome or metabolome and finally the fitness of the cell. In that way, the particular environment of application should play a decisive role to counteract the genetic load that would be produced and maximize the evolutionary stability of the synthetic strain. Thus, the heterologous sys-

tem should be designed so that all synthetic elements required for function deployment are beneficial for the cell. Otherwise, mutations affecting those functional elements would be selected, thus compromising the reliability of such synthetic strains [13, 14]. Furthermore, genome integration allows avoiding the addition of a selector (usually an antibiotic resistance gene) to keep the vector after cell replication. This will indeed reduce the genetic load and, hence, will increase the evolutionary stability. Further work will develop more complete bacterial cell models including regulatory interactions [15] and metabolic routes [16] in order to capture and model the majority of the internal biological processes and to be able to predict the cellular fitness as a balance between metabolic and genetic loads [17]. In addition, this type of models could be applied to other bacterial chassis, or even higher organisms upon viral infection.

Appendix A Estimation of Plasmids Concentration and Characterization

Single colonies of all strains were inoculated into 5 mL Luria Bertani (LB) broth medium containing kanamycin (50 μg/mL) and were incubated at 37C with 200 rpm overnight. We diluted overnight cultures in fresh LB medium with different concentrations of IPTG (0, 1, 10, 100, 1000, 10000, and 20000 μM) for induction of plasmid copy numbers. The final values of OD_{600} were between 0.3 and 0.4 for all cultures. We extracted plasmids from all constructs including the plasmid containing the ColE1 origin of replication (copy number 10-30) as a control. The amounts of plasmid DNA were quantified in triplicates by using NanoDrop 2000.

A single colony of *E. coli* MG1655Z1 was grown overnight in 5 mL LB medium at 37C with orbital shaking at 200 rpm. The overnight grown cultures were diluted 1:100 times in supplemented M9 media (M9 salts, 0.05% (w/v) casamino acids, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$ and 1.5 mM thiamine) in a 96 well microplate with final volume of 200 μL per well. Briefly described, we used flourometer (Tecan Infinite F500) for our experiments with a run of about 20 hours. The parameters used for these experiments were as following: temperature 37C, orbiter shaking, wait time 15 minutes, absorbance measurements at 600 nm, fluorescence measurements (510 nm excitation filter and 545 nm emission filter) with manual gain 40. Again we used different levels of IPTG concentrations (see above) with three levels of glucose concentrations (2, 10, 100 mg/L), LB and EZ media. We have used EYFP as reporter fluorescent protein. We have performed three days

independent experiments in triplicates that were averaged.

Appendix B Phenomenological Cellular Chassis Model

We considered an empirical model for the cellular chassis of the bacterium *E. coli* based on measurements of the intracellular chemical composition at different growth rates [7]. We derived from those experimental data the phenomenological equations describing the dependence of the levels of DNA polymerases, RNA polymerases, and ribosomes on the specific growth rate $\mu(h^{-1})$, giving

$$\begin{aligned} \text{DNAP} &= 155e^{0.923\mu} \\ \text{RNAP} &= 910^{1.529\mu} \\ \text{RIB} &= 3690^{1.775\mu} \end{aligned} \tag{2.1B}$$

In addition, we fitted the velocities of DNA (c_d in bp/h), mRNA (c_r in nt/h) and protein (c_p in aa/h) elongation as function of the specific growth rate resulting in

$$\begin{aligned} c_d &= 2.96 \times 10^6 \mu^{0.321} \\ c_r &= 1.75 \times 10^5 \mu^{0.238} \\ c_p &= 0.63 \times 10^5 \mu^{0.387} \end{aligned} \tag{2.2B}$$

In addition, we obtained a measure of the amount of DNA (D in base-pairs of the equivalent genomes), RNA (R in nucleotides usage), and proteins (A in amino acids usage), following

$$\begin{aligned} D &= 5.50 \times 10^6 \mu^{0.321} \\ R &= 1.75 \times 10^5 \mu^{0.238} \\ A &= 0.63 \times 10^5 \mu^{0.387} \end{aligned} \tag{2.3B}$$

Appendix C Heterologous Gene Expression Model

We considered a model based on ordinary differential equations to describe the device behavior in the cellular context, accounting for the physical properties of a particular plasmid, such as copy number (C), backbone length

(L) and length of the insert (l). Our device consists in an EYFP under the control of a constitutive promoter. The promoter strength was a function of the growth rate ($P = \frac{P_0}{1+0.5/\mu}$ where P_0 is the maximum transcription rate [5]), whereas the ribosome binding site affinity was almost constant due to saturation of the ribosomes ($Q = \frac{Q_0}{1+0.5/\mu}$ where Q_0 is the maximum translation rate [5]). The number of molecules of plasmid DNA in steady state was directly C. In addition, the dynamics for the total amount of heterologous mRNA (RNA_h) was given by

$$\frac{d}{dt}\mathrm{RNA}_h = P\psi\mathrm{DNA}_h - (\delta_r + \mu)\mathrm{RNA}_h \tag{2.1C}$$

where ψ is the number of promoters per plasmid, DNA_h is directly the copy number, and $delta_r$ is the RNA degradation rate. Moreover, the dynamics for the total amount of reporter proteins was modeled considering the non-fluorescent ($PROT_h^{(n)}$) and, fluorescent ($PROT_h^{(f)}$) content

$$\frac{d}{dt}\mathrm{PROT}_h^{(n)} = Q\Omega\mathrm{RNA}_h - (m + \delta_p + \mu)\mathrm{PROT}_h^{(n)} \tag{2.2C}$$

where Ω is the average number of ribosome binding sites per molecule of RNA_h and δ_p is the protein degradation rate [18]. By considering the maturation rate of the protein (m), we can obtain the fraction of fluorescent protein by solving its equation at steady state, which is ultimately what was experimentally measured, given by $\frac{1}{1+(\delta_p+\mu)/m}$, obtained from considering

$$\frac{d}{dt}\mathrm{PROT}_h^{(f)} = m\mathrm{PROT}_h^{(n)} - (\delta_p + \mu)\mathrm{PROT}_h^{(f)} \tag{2.3C}$$

Notice that the explicit term of μ (equations 2.1C, 2.2C and 2.3C) models the dilution of mRNA, and non-fluorescent and fluorescent protein, respectively, because cells are growing at rate μ. In that way, we computed the consumption of cell resources as functions of the DNA and RNA of the plasmid in steady state ($\mathrm{DNA}_h^{ss} = C$ and $\mathrm{RNA}_h^{ss} = P\psi\mathrm{DNA}_h^{ss}/(\delta_p + \mu)$, respectively). Specifically, consumption for replication ($DNAP_h$) was computed as a ratio between the production of DNA at growth rate given, μ, and DNA chain elongation. Analogously, consumptions for transcription ($RNAP_h$) and translation (RIB_h) were given by a ratio between the corresponding synthesis rates (equations 2.1C and 2.2C) and chain elongation (equation 2.3B) of RNA and protein, respectively:

$$\mathrm{DNAP}_h = \frac{\mu(L+l)\mathrm{DNA}_h^{ss}}{c_d} \tag{2.4C}$$

$$\mathrm{RNAP}_h = \frac{P\psi L\mathrm{DNA}_h^{ss}}{c_r}$$

$$\mathrm{RIB}_h = \frac{Q\Omega l/3\mathrm{RNA}_h^{ss}}{c_p}$$

References

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

[2] Shachrai, I., Zaslaver, A., Alon U., Dekel. E. (2010). Cost of unneeded proteins in *E. coli* is reduced after several generations in exponential growth. *Mol. Cell* 38, 758-767.

[3] Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M., Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840-845.

[4] Lu, T. K., Khalil, A. S., Collins, J. J. (2009). Next-generation synthetic gene networks. *Nat. Biotech.* 27, 1139-1150.

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

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

[7] Bremer, H., Dennis, P. P., Modulation of chemical composition and other parameters of the cell by growth rate. in: F. C. Neidhardt, R. I. Curtiss, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikow, M. Riley, M. Schaechter, H. E. Umbarge (Ed.), *Escherichia coli* and Salmonella, ASM Press, Washington, D.C. 1996, pp. 1553-1569.

[8] Peretti, S. W., Bailey, J. E. (1987). Simulations of host-plasmid interactions in *Escherichia coli*: copy number, promoter strength, and ribosome binding site strength effects on metabolic activity and plasmid gene expression. *Biotechnol. Bioeng.* 29, 316-328.

[9] Heffron, F., So, M., McCarthy, B. J. (1978). In vitro mutagenesis of a circular DNA molecule by using synthetic restriction sites. *Proc. Natl. Acad. Sci. USA* 75, 6012-6016.

[10] Seo, J. H., Bailey, J. E. (1985). Effects of recombinant plasmid content on growth properties and cloned gene. *Biotech. Bioeng.* 28, 1668-1674.

[11] Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpe, Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141, 344-354.

[12] de la Cueva-Mndez, G., Pimentel, B. (2007) Gene and cell survival: lessons from prokaryotic plasmid R1. *EMBO Rep.* 8, 458-464.

[13] Sleight, S.C., Bartley, B.A., Lieviant, J.A., and Sauro, H.M. (2010). Designing and engineering evolutionary robust genetic circuits. *J. Biol. Eng.* 4, 12.

[14] Canton, B., Labno, A., Endy, D. (2008). Refinement and standardization of synthetic biological parts and devices. *Nature* 6, 787-793.

[15] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

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

[17] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Towards the automated engineering of a synthetic genome. *Mol. BioSyst.* 5, 733-743.

[18] Leveau, J. H. J., Lindow, S. E. (2001). Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J. Bacteriol.* 183, 6752-6762.

Chapter 3

Design-Guided Models of global Transcription Regulation

Molecular regulations govern the cell response under environmental (extracellular) or genetic (intracellular) perturbations. The elucidation of these regulations with computational techniques will allow analyzing the cell behavior [1], since modeling in biology has boosted the understanding of the cell mechanisms by means of systemic approaches [2]. On the other hand, the design of new transcriptional networks requires a quantitative description of the transcription regulation. Thanks to new developments in the inference from transcriptomic data, now it is possible to reconstruct the regulatory network with enough accuracy to predict gene expression profile in presence of heterologous networks. We propose a procedure that, by extending a recent methodology, could be used to redesign transcriptional networks.

The continuous developments on genome sequencing and annotation allow us to identify the genes and transcription factors (TFs) of an organism. The development of the microarray technology has provided high-throughput genomic measurements, where cells are subjected to several conditions or stresses to measure their gene expression profiles [3]. Large-scale cell models, such as metabolic, transcription or protein networks, are distilled from high-throughput genomic data, which poses one of the most challenging problems in biology. The construction of a deterministic model would allow the prediction of the cell response under different stimuli [4].

To redesign the transcriptional regulation network, we need a quantitative model able to predict the gene dynamics. We propose to characterise

such model by using microarray data with a known transcriptional network inference method. We first infer the network topology and we later estimate the corresponding kinetic parameters. For the last decade, there has been an enormous effort in the improvement of techniques aimed at the inference of the connectivity of the transcription network. Clustering approaches [5, 6, 7, 8, 9] have been used to obtain information of regulatory networks but with low accuracy [10]. Information-theoretic inference provides more accurate networks [11, 12, 13, 14, 15] even from reduced expression datasets. A local significance calculation has been very fruitful to capture the network topology [14]. On the other hand, Bayesian methods [16, 17, 18, 19] give networks with high precision but low proportion of true recovered interactions (they introduce few regulations with high confidence). Moreover, such methods have a higher computational cost. Herein, we propose the construction of predictable genome models in a standard format from a regulatory scaffold captured by using probabilistic methods. Other approaches, instead, optimized directly the corresponding kinetic parameters for a linear regulatory model [20, 21]. In addition, recent algorithms [22, 23] applied sparse logistic regression [24] for gene selection in order to avoid overfitting.

3.1 Genome-Wide Quantitative Model of Transcription Regulation of *E. coli*

In the present chapter, we have applied inference methodologies recently used to obtain models suitable for genome redesign. We have considered the *E. coli* genome, which contains 4345 non-redundant genes, of which 328 are putative TFs. The genome is organized into 3333 operons, 2447 containing single genes and 886 polycistronic units. The reference regulatory set has been constructed according to RegulonDB [14]. For the inference procedure, we have used public microarray data [27] from Affymetrix normalized using RMA [28]. This is a microarray compendium containing 189 experiments. From this data set, 20 experiments were excluded in order to later predict expression profiles from unbiased data. The inferred network contains 525 regulatory interactions (z$>$ 6.92) and 566 combinatorial influences (z$>$ 12). *InferGene* predicts 3982 genes to be controlled by constitutive promoters.

To analyze those results in a biological context, we have used the EcoCyc [29] classification to group genes by biological functions and to rank those groups according to their level of prediction. We have scored each biological function as $\Delta_{bf} = \frac{1}{n}\frac{1}{m}\sum_{g\in bf}\sum_{c\in set}|\widehat{y}_{gc} - y_{gc}|$, where n is number of genes involved in the biological function , m the number of the new conditions

of the set ($m = 20$), $\widehat{y}_{gc}$ the predicted expression, and y_{gc} the measured expression of the gene g in the condition c. The best predicted functions are involved in the metabolism such as biosynthesis of lipoprotein, carnitine, glycolate and glycoprotein, or functions related with information transfer such as rRNA and stable RNA, ATP binding, and DNA degradation. In addition, we have observed two significant correlations between the number of constitutively expressed genes and the error in expression (Δ_{bf}). These genes are from biological functions involved in the location of gene products and the cell processes.

3.2 Design of Artificial Genomes and Validation of their Transcription Profiles

We have constructed several genomes *in silico* using GAG and we have compared the predefined regulations in our models with the regulations inferred by *InferGene*. We have constructed three types of transcription networks according to the mode of regulation of its constituent operons: i) networks with promoters regulated by at most one TF, ii) networks with promoters that can be regulated by more than one TF, and iii) networks with promoters that can be combinatorially regulated including synergistic effects. We have computed the precision rate and sensitivity (see section Methods) to quantify the efficiency of *InferGene*.

In Figure 3.1 we show the evaluation of the inference for different types of genome networks. *InferGene*, which at this stage relies on CLR, predicts the 85.4% (sensitivity) of the possible interactions although only the 15.7% (precision rate) of them are correct for a genome of 500 genes using 100 conditions (Figure 3.1A). However, if the number of conditions increases to 250, the precision rate reaches values around the 90% (see Figure 3.1B). The same trend occurs with larger genomes as we can see from Figs. 3.1C-D, where we have worked with genomes of 5000 genes with 300 and 600 conditions, respectively. Thus, we improve 6-fold the precision rate, maintaining a given level of sensitivity, when increasing the number of conditions 2.5 fold. Therefore, the efficiency of algorithm has a nonlinear behavior regarding the number of conditions used for training. We have also extended the inference capabilities of CLR to cooperative interactions. Our results show that we need a minimum set of microarray experiments to infer a transcriptional regulatory network with high precision rate for a given sensitivity. Furthermore, genomes with only promoters regulated by at most one TF reached higher values of precision rate and sensitivity.

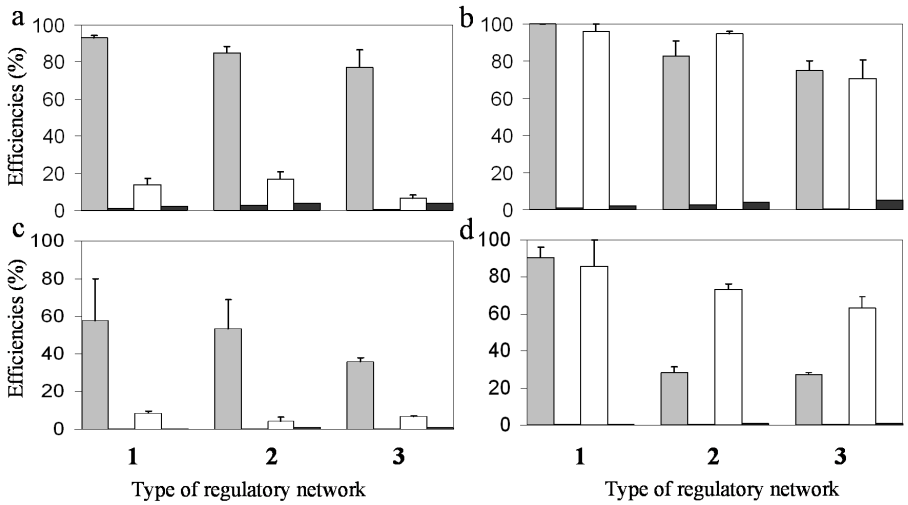

Figure 3.1: *InferGene* performance. Evaluation of sensitivity (gray) and precision rate (white) together with a random inference (black) of the transcriptional regulatory network. We used several types of synthetic genomes with different topological and parametrical properties generated by GAG. We constructed three types of genomes: (1) all promoters are regulated by at most one TF, (2) the promoters that can be regulated by more than one TF, and (3) promoters with combinatorial regulations including synergistic effects. Genomes for (a,b) had 500 genes and 50 TFs, and for (c,d) 5000 genes and 200 TFs. The number of conditions was in (A) 100, in (B) 250, in (C) 300, and in (D) 600. Deviations in precision rates and sensitivities were calculated using three different genomes for each type. The z-score threshold used was in (A) 0.5, in (B) 1, in (C) 3, and in (D) 7.

We have analyzed the predictive power of *InferGene* by calculating a score based on the error made on predicting the expression levels (Δ_{op}), and other score based on the error made on the prediction of the model parameters (Γ). We define $\Delta_{op} = \frac{1}{n}\frac{1}{m}\sum_{g\in op}\sum_{c\in set}|\widehat{y}_{gc} - y_{gc}|$, where $\widehat{y}_{gc}$ is the predicted expression profile, y_{gc} is the experimental value, n is the number of operons that are correctly inferred according to RegulonDB, and m is the number of conditions that were not used in the training set ($m = 20$). We also define $\Gamma = \frac{1}{n}\frac{1}{n_p}\sum_{g\in o}\sum_{p\in P}\left|\widehat{\beta}_{gp} - \beta_{gp}\right|$, where n_p is number of parameters we use to model the kinetics of the operon expression, $\widehat{\beta}_{gp}$ are the estimated model parameters, and β_{gp} are the model parameters from GAG. To perform such analysis, we have generated a network using the GAG algorithm with 500 genes across 250 conditions. The median for Δ_{op} was 0.009, and for Γ was around 0.01. Moreover, we have validated the estimated parameters by performing linear regressions with the predefined kinetic models and obtaining Pearson correlation coefficients above 0.90.

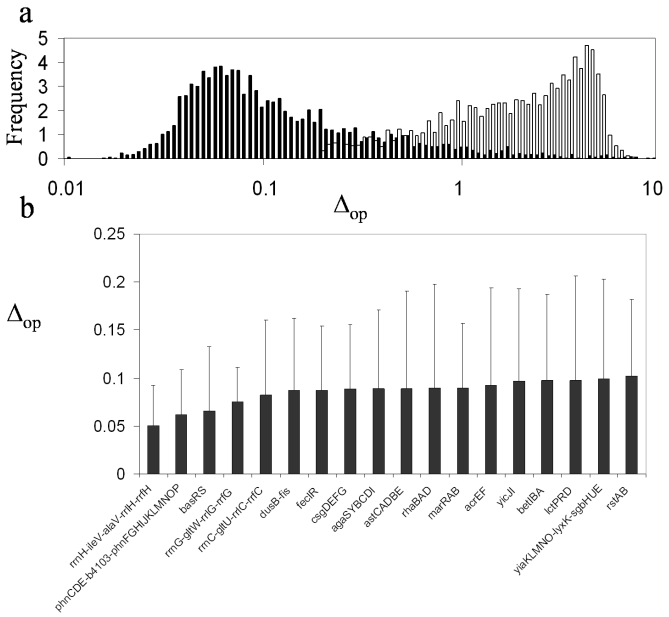

Figure 3.2: (A) Histogram of the expression error on the transcriptomic profile for each operon (Δ_{op}). In black, model with parameters from linear regression; in white, model with random parameters (for a fixed inferred topology). (B) We show the mean of Δ_{op} with the corresponding standard deviations for the best predicted operons. We measured the predictive power under the 20 conditions of the testing set.

3.3 Prediction of Wild-Type *E. coli* Trancriptome

Before proceeding to change the regulation of *E. coli*, we have calculated the ability of the inferred model to predict the steady state expression levels of the *E. coli* genes. For that, we have used the model together with the expression levels of all the TFs for each experimental condition to compute the global expression profile. Afterwards, we have compared the predicted expression values with the corresponding measurements, obtaining Δ_{op}. We have also determined the predictive power of the inferred model on the 20 experimental conditions excluded from training data set. The distribution of Δ_{op} for the 3333 operons of *E. coli* is shown in Figure 3.2a (black columns). The mean of this distribution is 0.048. White columns represent a model with random parameters for the inferred topology. In Figure 3.2b, we show the prediction for the best inferred operons. It is interesting to note that the genes from these operons are involved in functions related with information transfer (RNA related such as transcription related, tRNA, rRNA or stable RNA; and protein related such as translation), regulation, location of gene products (cytoplasm and *ompR*) and cell processes (adaptation and defense-survivial).

In Figure 3.3 we plot the predicted profiles with lowest Δ_{op} against the experimental profiles across all conditions (189 experiments, 169 conditions from the training set and 20 new conditions for prediction). We also perform a K-fold cross-validation to ensure that our results do not depend on the selection of the testing set.

3.4 Genome-Wide Model of *E. coli* Integrating Signal Transduction Data

To develop a methodology able to automatically design a genome for fast growing cells in changing environments, we have assumed the hypothesis that a cells growth is determined by its transcriptional profile. Therefore, we used the genome-wide model inferred of *E. coli* gene transcription in response to selected external signals to predict changes in cell growth after genome modification. Such a model allows the assignment of mathematical parameters to promoters and TF sequences, which we have assumed to be independent of genomic context. We extended our TRN to sense environmental changes at the molecular level. Recent studies have collected data describing thousands of interactions between environmental factors (EFs) and TFs that are involved in sensing environmental perturbations. These interactions were coupled to the TRN model (see Appendix C) such that

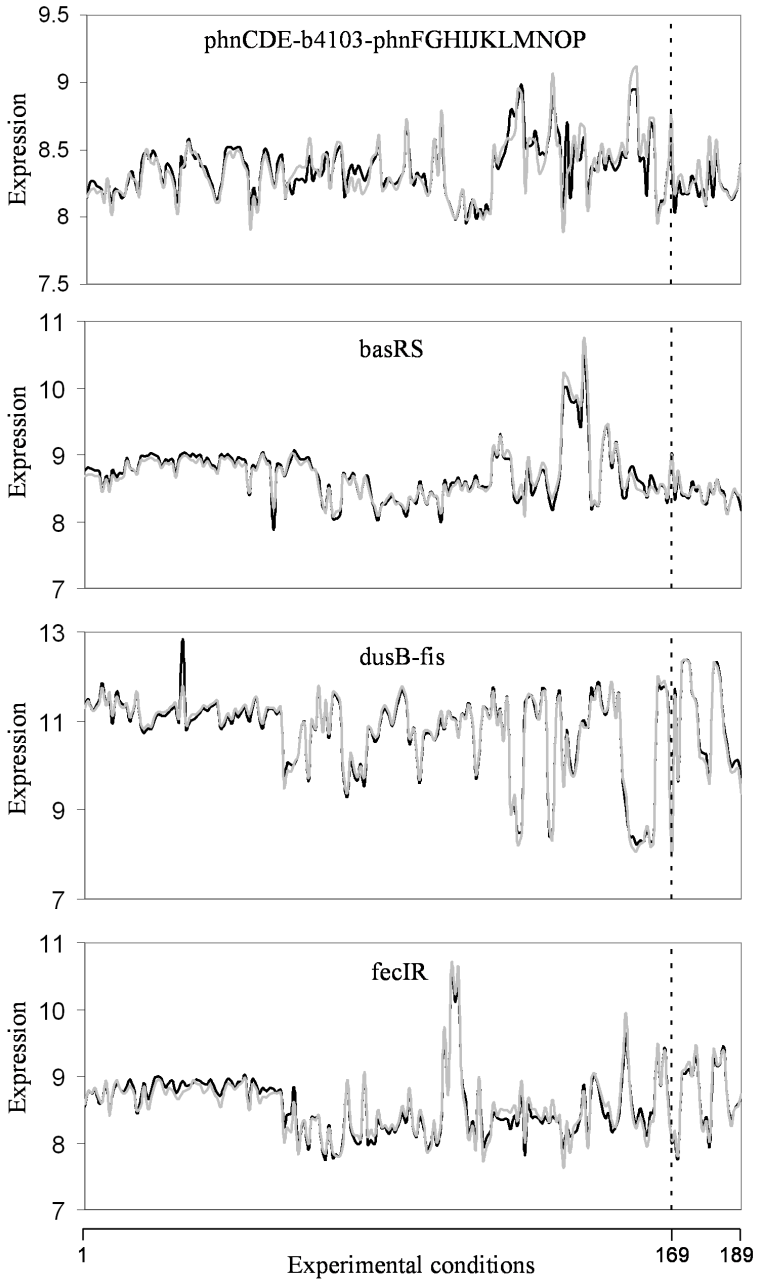

Figure 3.3: Prediction of expression profiles in *E. coli*. Each plot shows the experimental profile (gray line), and the profile predicted by our model (black line). The last 20 experiments, separated by a dashed line, correspond to conditions that were not included in the training data set with which we inferred the kinetic model.

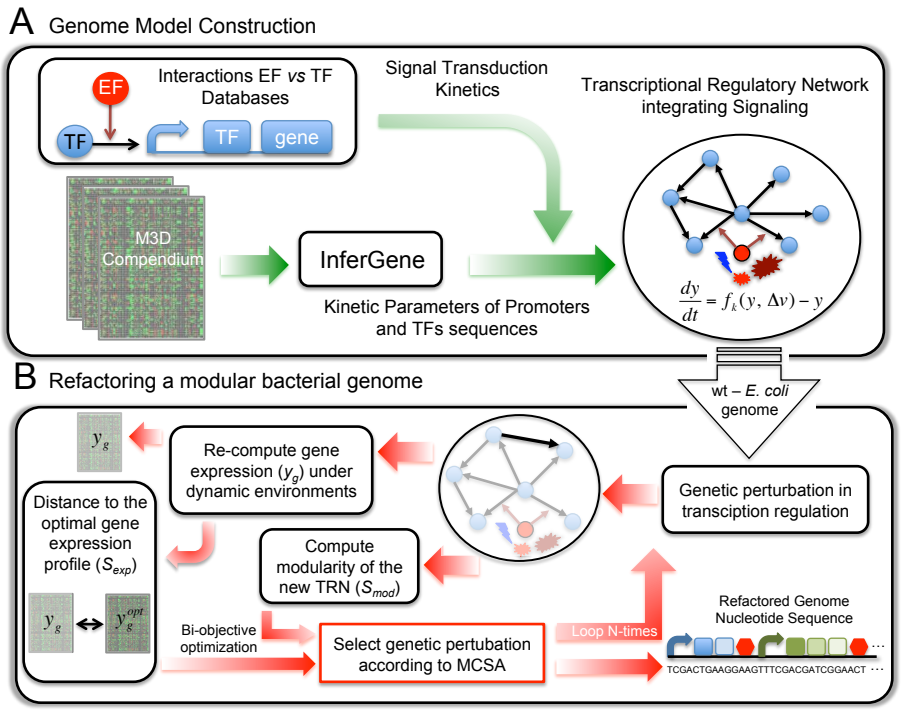

Figure 3.4: (A) Steps designed to construct the regulatory network of *E. coli* required to sense environmental changes. (B) A scheme of the algorithm used to re-design the *E. coli* TRN. The wild-type genome was used as the starting point for an optimization process based on Monte Carlo Simulated Annealing. During the *in silico* evolution, we modified gene regulation and computed the resulting genome fitness as a function combining the genome modularity and the distance between the gene expression levels of the re-engineered and wild-type genomes.

uptake factors modify the predicted expression of several TFs. We therefore quantified how the expression of a given TF changes upon the perturbation of a specific uptake factor(s) (Figure 3.4A).

Next, we investigated how the model system responds to environmental changes. We evaluated a distance, S_{exp}, between the optimal expression profile (defined as the expression profile measured for *E. coli* growing at the maximum rate for a given environmental condition) and the expression profile of the model in each environment. As it is not clear which genes will be most relevant to cell growth during genome evolution, we explored six sets of genes to define S_{exp} (physiological adaptation genes, defense pathway genes, a combination of genes related to these two functions, genes that protect against abiotic stresses, genes encoding central metabolism enzymes and all genes). Figure 3.5A shows the optimality degree, defined as the relative growth that *E. coli* exhibits in environments that are optimal except in the concentration of a single component, such as oxygen or glucose [16]. Figure 3.5B shows calculations of S_{exp} based on our model from the expression profiles predicted under 100 different environmental conditions. Here, each environment is defined using random values for external oxygen flux and carbon and nitrogen availability that range from minimal to saturating values, thereby simulating extreme environments. As expected, the largest variations of the expression score and optimality degree were obtained when selecting a gene set related to defense functions, and the smallest variation was obtained after considering genes related to enzymatic activity. This difference is expected, because the defense responses are highly inducible and specific to given environmental stimuli whereas metabolism is able to buffer external stimulus through a critical set of metabolic pathways.

3.5 Model Validation: Predicting Growth Rate of Perturbed Transcriptional Networks of *E. coli*

3.5.1 Model Validation 1: Prediction of Expression Profiles Upon Genetic and Environmental Changes

We sought to determine whether a genome model able to assign parameters to promoters and TF sequences would be able to predict the transcriptome of *E. coli* under different environmental conditions and/or after genetic modifications. We evaluated the performance of our model network in predicting responses to environmental stresses and genetic changes. For illustrative purposes, Figure 3.5C shows the predicted versus experimental profiles for two examples of master regulator knockouts (*fnr* and *soxS*) under aero-

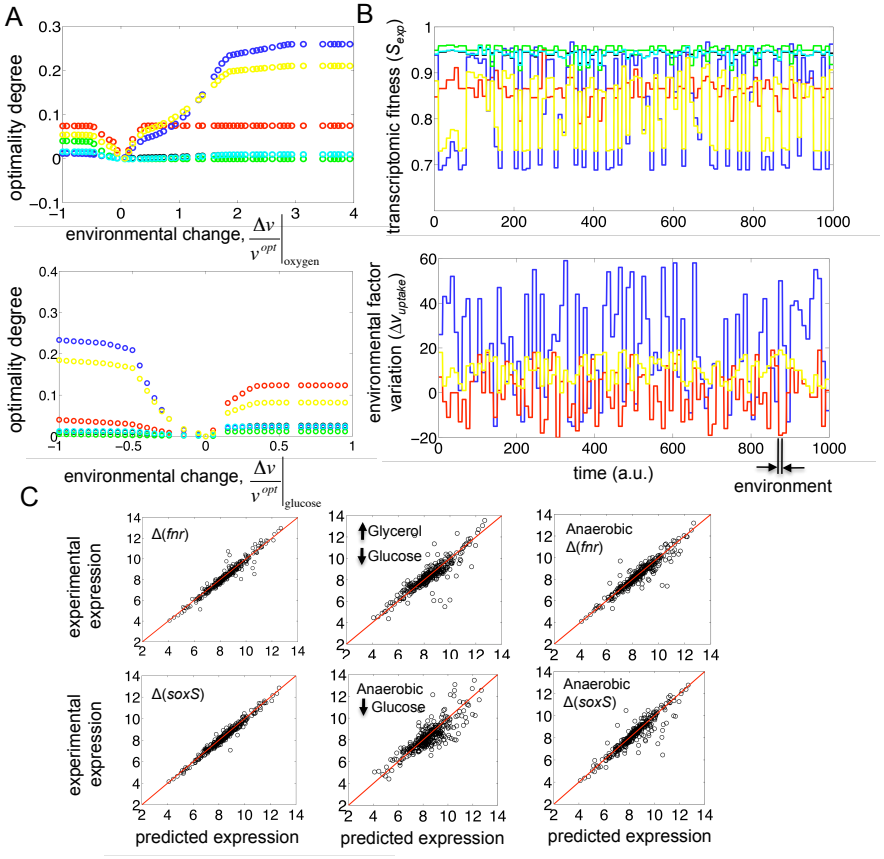

Figure 3.5: (A) Optimality of the wild-type genome under environments perturbed in a single direction (oxygen and glucose). Notice that each simulation considered different sets of genes related to adaptation functions (red line), defense pathways (blue line), a combination of genes related to adaptation and defense (yellow line), protection (green line), central metabolism (cyan), and all genes (black line). (B) Quantitative simulations of transcriptomic fitness, S_{exp}, under a variety of environmental conditions, , simultaneously (colors in panel A were maintained to indicate the different sets of genes selected to evaluate S_{exp} in the top of panel B). At the bottom of panel B, the external fluxes of glucose (red line), NO_3^- (yellow line) and oxygen (blue line) represent changes in carbon and nitrogen availability and the cellular uptake of oxygen, respectively. (C) Prediction of the expression profiles of *E. coli* upon genetic perturbation (knockout of *fnr* and *soxS*), environmental perturbation (modification of oxygen and carbon availability), or both (*fnr* and *soxS* knockout under anaerobic conditions). The red line represents the exact prediction.

bic and anaerobic conditions and for two environmental perturbations in which glucose, oxygen and glycerol sources were changed. Each dot in the scatter plots represents a value obtained from a different hybridization experiment plotted against the algorithm prediction. In a more general way, we compared the predicted expression levels of all the TFs ($\hat{y}_{gc}$) for each experimental condition, c, with respect to the corresponding empirical measurement, y_{gc}, using the normalized Euclidean distance (e_c) and the Pearson correlation coefficient (ρ_c) for all microarray experiments. Specifically, we used our model to predict expression profiles for TF knockouts by removing the corresponding transcription regulation in our model. We obtained values of $\rho_c > 0.87$ and $e_c < 3.2\%$, with the exception of the *recA* knockout ($\rho_c = 0.74$) made by Faith et al. Moreover, we applied this procedure to the modification of the global TRN by changing the environmental conditions. This was experimentally achieved under several different conditions included in the *M3D* compendium [27]. When oxygen and carbon sources were perturbed, we estimated $\rho_c > 0.74$ and $e_c < 7.3\%$. The model was able to capture whole transcriptome expression, obtaining values of $\rho_c > 0.85$ and $e_c < 5.4\%$.

3.5.2 Model Validation 2: Predicting the Results of *E. coli* Experimental Evolution

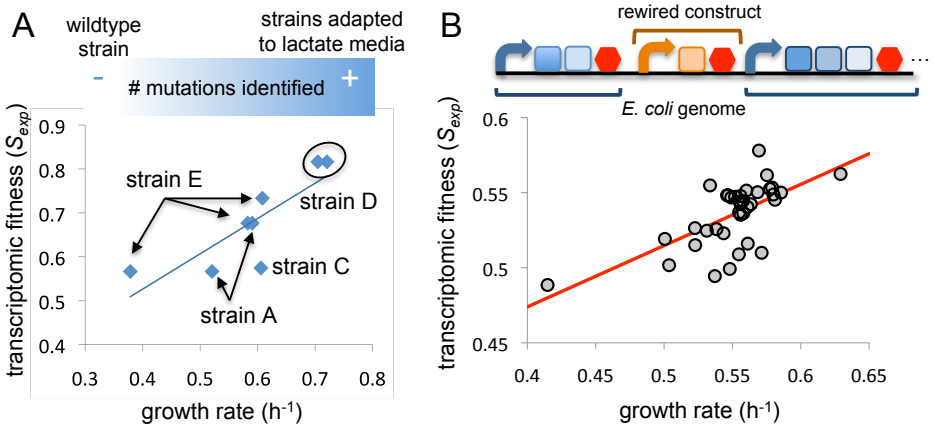

Figure 3.6: (A) Correlation between predicted fitness considering only TFs and the growth rates of four strains and their intermediaries evolved in the laboratory under minimal lactate media. (B) Correlation between predicted fitness considering only TFs and the growth rate of 37 strains with a rewired TRN.

After generating this predictive model for the TRN, we attempted to automatically design genomes by implementing an *in silico* evolution al-

gorithm in which a fitness function is used to select for beneficial genome modifications during the evolution process. To validate S_{exp}, we compared our fitness predictions to data obtained experimentally during *E. coli* evolution. Recently, Conrad et al. characterized all acquired adaptive mutations of *E. coli* strains from a short-term laboratory evolution in minimal lactate medium [31]. They measured growth rates and identified adaptive mutations using whole-genome sequencing for all evolved strains at specific time points. Interestingly, several mutations were identified in highly connected TFs in the TRN (*crp*, *ydcI* and *hfq*) in a gene related to transcription termination (*rho*) and in a gene responsible for recycling RNA polymerases (*hepA*). We predicted the transcriptome for each of these strains by modifying our *E. coli* network model to introduce a different gene expression value for each mutated gene. We then determined the S_{exp} fitness function of the predicted expression profile by predicting the transcriptome of a strain with the mutated genes set at optimal transcription levels and then calculating the distance between the mutant strain and the optimal strain evolved for adaptation in lactate culture (see Appendix G). Figure 3.6A demonstrates a significant linear Pearson correlation ($r = 0.82$; $p < 0.05$) between observed and predicted fitness when considering only the contributions of the TFs to S_{exp}, validating our fitness function. Similar correlations were observed when considering the contributions of central metabolism enzymes, genes related to stress or the full genome. Overall, we showed that growth rates predicted using genome design reached high correlations ($r > 0.72$, $p < 0.05$).

3.5.3 Model Validation 3: Predicting the Growth Rate of Rewired Transcriptional Networks of *E. coli*

We attempted to predict the phenotypic response of *E. coli* after adding new regulations in its TRN. A recent study of Isalan et al. systematically explored such problem by expressing endogenous promoters controlling different TFs or σ-factor genes and measuring the growth rate of each strain hosting a rewired TRN [32]. We only selected the subset of the 37 strains in which the rewired constructs (promoter region-TF fusions) were stably integrated in the *E. coli* chromosome. We then determined the S_{exp} fitness function of the predicted expression profile (see Appendix H). Figure 3.6B show a significant linear Pearson correlation ($r = 0.65$; $p < 0.0001$) between observed and predicted fitness when considering only the contributions of the TFs to S_{exp}, corroborating that our fitness function is able to capture large changes in the TRN.

3.6 Discussion

We have discussed a methodology to create quantitative models for transcription regulation aimed to future genome redesign projects. We have shown how we could use recent methodologies to infer the global topology of transcription regulation to produce the kinetic model able for genome redesign. We have successfully applied the inferred model to predict the transcriptomic response of *E. coli* under experimental conditions not included in the training set. The prediction has in average an error of 1-5% relative to the experimental value (average computed across all conditions). Furthermore, we have predicted the gene expression under knockouts of TFs and genetic rewirings [32] by solving a perturbed model, showing the predictive power of the inference procedure. Such perturbations change the regulatory map of the cell, but more complex redesigns, even a whole transcription refactorization, could be *in silico* explored by using our model. Our algorithm provides a global deterministic kinetic model of genetic regulations using microarray data. We show how to use this kinetic model to make predictions [23]. Thus, our approach constitutes an important step towards the large-scale design of cell behaviors by providing models which are validated using *in silico* genomes and experimetal transcription data. In this direction, we have accounted for simple transcription rewirings [32] by obtaining the gene expressions using computational methods. Such models can be used in the future to rewire the regulation of organisms without affecting their physiological behavior.

The algorithm reaches high efficiencies at the topology and kinetic level, based on the CLR algorithm [14] to infer the network together with an extension to include cooperations in combinatorial promoters. However, it could use other approaches such as Bayesian methods [19]. In addition, the generation of synthetic data from specified genome models has been essential to analyze the performance and limitations of *InferGene*. Indeed, we have shown how the precision rate is drastically improved, from 10-20% to 80-90%, by just doubling the number of perturbations in artificial genomes. Moreover, the error in the prediction of the expression value for correctly predicted regulations is of the order of magnitude of the standard errors on measured expression data, and the estimated parameters highly correlate with the predefined ones (correlation coefficient higher than 0.9). The inaccuracies in our prediction could be rationalized by the lack of modelling of many dynamic variables of the cell (*e.g.*, proteins or metabolites) or non-transcriptional regulations (*e.g.*, protein-protein or RNAi), since these variables are not experimentally measured using microarrays. Furthermore,

future works could consider confidence intervals on the model parameters to analyze the stochasticity in expression data.

Our approach can take advantage from additional sources of information. For instance, it can incorporate in the inferred model experimentally-validated interactions (*e.g.*, from functional genomics measurements or sequence analysis) as a regulatory background. In addition, the knowledge on the genome sequence can help in the inference procedure, by providing information about operon structure, identification of TFs and their regulations [34, 14, 35]. The prior knowledge about regulation provides a topology that can be added into the model and can be used to predict new interactions with high fidelity [36].

The identification of regulations is a high time-consuming activity. The running time scales with the number of genes and the square of the number of conditions. Nonetheless, the parameter estimation is a quick process (relative to the previous). For instance, in *E. coli* there are 4345 genes (strain K-12) clustered in 3333 operons, and 328 TFs and 53628 pairs of TFs [14]. The whole inference process took 6 hours accomplished on a computer Pentium M 2.00 GHz and 1 GB RAM (time resources for parameter estimation are neglected as they are around 2 minutes). However, all simulations can be run in parallel allowing the reduction of the execution time (less than 5 minutes on a simple cluster). In this way, distributed computing provides the necessary resources to apply our methodology to infer the regulations of much larger genomes. Our methodology provides a simple and fast way to obtain a quantitative global model of transcriptional regulation even for large networks. The incorporation of sparse Bayesian regression methods [19] provides a promising extension for further works. Such methods would provide better inference but increasing the computational cost.

The construction of genome-scale models is clearly a valuable step towards the understanding of the cellular behavior [4], but it is also of interest for the emerging field of synthetic biology, where functional genetic circuits are engineered into cells dealing to minimize the impact on the host [37]. Hence, *InferGene* provides an accurate model to predict the changes in the biological processes when perturbing the cell. In addition, this model can be applied to discover molecular targets of heterologous compounds [20, 21].

Appendix A Mathematical Model of Transcription Regulation

We aim to the development of a methodology able to *in silico* evolve a genome for having a predefined transcriptional profile. For this, we require to construct a predictive genome model of transcription, based on ordinary differential equations (ODEs), to account for global redesigns of the cellular regulatory map. Using such models we could study the evolution of gene regulations as a consequence of the environmental stimuli. To construct this we have to use as input microarray data properly normalized. In general, transcription involves protein-DNA interactions, but microarray data gives the genetic expression by quantifying the amount of mRNA. Thus, inferring just from transcriptomic profiles could introduce some inaccuracies due to, for instance, protein-protein interactions of TFs [38, 39]. Furthermore, some environmental stresses (*e.g.* heat shock) can alter globally protein expression. However, in this work we neglect these effects for simplicity, assuming that the mRNA amount is proportional to the protein expression and that it is function of the TFs only. In addition, as the precise kinetic model of transcription regulation is not known for any organism, we have generated *in silico* genomes having random regulatory maps with scale-free topology [40]. We have applied our methodology against synthetic transcriptomic profiles. We will only assume a previous knowledge of the list of all genes and TFs obtained from genome annotation (*e.g.*, RegulonDB [14] for *E. coli*). Eventually we can consider the genomic organization in operons (especially in case of bacteria). Such operons can be known *a priori* or inferred from the same microarray data. Our approach consists of two nested steps. Firstly, we obtain the topology of the network (*i.e.*, which TF regulates which gene or operon) by using an information-theory-based approach. We store in a matrix the likelihood of the mutual information (MI) among all the TFs and operons [41, 42, 43], computed as the z-scores from the distribution of MI using the transcriptomic expressions for all the perturbing conditions [14]. Then, using a suitable threshold, we infer the TFs regulating a given operon. Subsequently, for each operon we perform a multiple linear regression against the corresponding TFs to recover the model kinetic parameters [44]. To infer cooperative regulations, we create a set of artificial TFs whose expression profiles are obtained in a combinatorial way as the product of two TF profiles (with the aim of conserving linearity in the formalism). This model is subsequently exported into a SBML file [33], which could be visualised using Cytoscape [25]. We have measured the performance of our algorithm by using synthetic transcriptomic data from

artificially generated networks.

We describe the genetic regulations using a linear model for the mRNA dynamics. Here, we use as input data mRNA expression profiles in steady state derived from transcriptional perturbations. As transcriptomic data is normalized and usually represented in logarithmic scale, we have considered $\log_s(mRNA)$ as variables (where s can be 2 or 10). Therefore, the mRNA dynamics from gene y_i is given by

$$\frac{d}{dt}y_i = a_i + \sum_{j \in TF} b_{ij} y_j + \sum_{j \in TF} \sum_{k \in TF} b_{ijk} y_j y_k - \delta_i y_i, \tag{3.1A}$$

where a_i is the basal synthesis rate, b_{ij} the transcription regulatory coefficient of TF j, b_{ijk} the cooperative transcription regulatory coefficient of TFs j and k acting on the promoter controlling the gene i, and δ_i the degradation rate. We set $b_{ij} = 0$ and $b_{ijk} = 0$ when j and j, k are not TFs regulating the gene i. We assume that all the genes of an operon have the same expression value. We also consider that two regulators could act in a cooperative way (*i.e.*, synergistic inductions and cooperative repressions). We do not consider cooperation between more than two TFs.

Here, we use expression values in steady state. Nevertheless, it could be also possible to extend our approach to the use of time series to enrich the experimental input [45]. Hence, in the steady state we can write

$$y_i = \alpha_i + \sum_{j \in TF} \beta_{ij} y_j + \sum_{j \in TF} \sum_{k \in TF} \beta_{ijk} y_j y_k, \tag{3.2A}$$

where we have defined $\alpha_i = a_i/\delta_i$, $\beta_{ij} = b_{ij}/\delta_i$, and $\beta_{ijk} = b_{ijk}/\delta_i$. Notice that the resulting parameters are referred to the intensity scale of the microarray technology. We use a time scale such that the mRNA degradation constant is $\delta = 1$. To use a realistic mRNA degradation constant, it would require translating the Affymetrix [46] data to concentration units.

Appendix B Using Network Inference to Obtain a Kinetic Model

To obtain a kinetic model suitable for redesign, we take advantage of recent methods aimed to infer the topology of the global regulatory map. In particular, we have chosen one of the best performing methods, the CLR [14], although other methodologies providing a transcriptional map, such as sparse Bayesian methods [19] could also be used. Our approach consists of using multiple regressions to fit the kinetic parameters of a continuous

model of the transcription regulation. The approach for large-scale transcription inference is based on measuring the influence between the expression levels of TFs and operons across a large set of conditions. Here, we use MI to estimate the correlation between a TF t and an operon p by using $MI(y_t, y_p) = H(y_t)+H(y_p)-H(y_t, y_p)$, where H is the entropy of a variable. It is defined as $H(y_i) = -\sum_c p(y_{ic})\log(p(y_{ic}))$, where y_{ic} is the expression value of gene i in the condition c, and $p(y_{ic})$ the probability to reach that value. The MI is always a positive magnitude. Joint normal distributions are generated with independent variables MI_i and MI_j (values for gene i and TF j, in row i and column j). Thus, the MI matrix is converted into Z matrix where $Z_{ij} = \sqrt{Z_i^2 + Z_j^2}$ and Z_i and Z_j are the z-scores of MI_{ij} from the marginal distributions. According to this matrix we obtain the genomic interactions.

For completeness, we have developed an algorithm (*InferOpe*) to infer operons from microarray data. Since two genes from one operon share the same mRNA molecule, we would expect that their transcriptomic profiles would be similar. Our operon prediction is based on the use of co-expression patterns [47], assuming that two genes, i and j, belong to the same operon if they are highly correlated. We evaluate this by using the Pearson correlation coefficient (we assume correlation if $\rho_{ij} > \rho_0 = 0.5$). Moreover, we impose that the angle (θ_{ij}) of such correlation should be around 45° (*i.e.*, $tan(\theta_{ij}) \simeq 1$), where the relationship with ρ_{ij} is given by $tan(\theta_{ij}) = \rho_{ij}\frac{\sigma_j}{\sigma_i}$.

For each operon we compute the kinetic parameters for the TFs regulating its promoter. The experimental value of one operon is computed as the average of the expressions of all genes belonging to that operon (*i.e.*, $y_{op} = \frac{1}{n}\sum_{g\in op} y_g$, where n is the number of genes of the corresponding operon). To estimate the model parameters α_i, β_{ij} and β_{ijk} we use multiple linear regression [44], which is the result of a minimization problem (least squares) defined by

$$(\widehat{\alpha}_i, \widehat{\beta}_{ij}, \widehat{\beta}_{ijk}) = \arg\min\{(y_i - \alpha_i - \sum_{j\in TF} \beta_{ij}y_j - \sum_{j\in TF}\sum_{k\in TF} \beta_{ijk}y_j y_k)^2\}. \quad (3.1B)$$

We assume that the variability in the experimental conditions and the complexity of the natural regulation is high enough to prevent linear correlations between TFs, which would produce identifiability problems in the regression parameters. Even in such a case, our model is a valid solution although there could be alternative models. We have used the LINPACK libraries [48] to calculate the solution.

Our procedures are implemented in C++, and they run on any UNIX

environment. The *InferGene* software, a tutorial, the corresponding files and some examples are available upon request. The software consists of different functional modules to compute firstly the network topology and then the corresponding kinetic parameters. Below we present the procedure implemented in *InferGene*:

1. Represent the microarray data organized in matrix form, for instance, genes in rows and conditions in columns.

2. Obtain the list of TFs for the given organism.

3. Ensure that the microarray matrix contains the expression profiles for all TFs.

4. Add new rows corresponding to the combinations of two TFs obtained as the product of them (*i.e.*, $y_{TF_i}.y_{TF_j}$ are the new TF profiles).

5. In case of bacteria, have a file containing the list of operons with the corresponding genes. Otherwise, run *InferOpe*, our algorithm to infer clustered genes based on co-expression patterns. To maintain the same scheme in all cellular contexts, we can dispose one gene per operon in case of eukaryotes.

6. Compute the MI among all the TFs and operons by using the CLR algorithm [14].

7. Compute the z-score among all the TFs and operons from the MI distributions by using the CLR algorithm.

8. Infer the TFs regulating a given operon, single and cooperative interactions, according to a given threshold depending on the desired precision. The threshold for cooperative regulations is taken higher than for single ones (2-fold for the reported calculations, although it can be modified straightforwardly) to avoid overfitting in the computation of the combinatorial interactions.

9. For each operon, estimate the kinetic parameters for its regulating TFs by using multiple linear regressions (obtaining single and synergistic interactions). Eventually, remove regulations with low strength.

10. Construct a SBML file containing the ODE-based model using the inferred topology and the estimated kinetic parameters.

Appendix C Construction of a Transcriptional Regulatory Network That Integrates Signal Transduction

We constructed a TRN of the wild-type genome that was able to predict gene regulation at the transcriptional and environmental levels. For this, we adopted a linear model based on differential equations describing the time dynamics of each mRNA in order to infer real kinetic parameters for promoter and TF sequences. Thus, the mRNA dynamics from the i_{th} gene, y_i, is given by $\frac{dy_i}{dt} = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i$, where α_i represents its constitutive transcription rate, β_{ij} represents the regulatory effect that gene j has on gene i, γ_{ik} represents the effect that environmental factor (EF), *i.e.* the metabolic uptake factor k, has on the expression of gene i, $\Delta v_k = (v_k - v_k^{opt})$ is the difference between the uptake factor measured under a given environmental condition, v_k, and the uptake factor measured in the optimal environmental condition, v_k^{opt}, and δ_i represents the degradation and dilution rate constant.

Time was conveniently scaled such that $\delta_i = 1$ and the model was assumed to be in steady-state $y_i = \hat{\alpha}_i + \sum_j \beta_{ij} y_j$, where $\hat{\alpha}_i = \alpha_i + \epsilon_i + \sum_k \gamma_{ik} \Delta v_k$, because fitting the appropriate mRNA degradation constant would require time series data [45]. To calibrate TF expression, the newly redefined constitutive transcription rate included a perturbative term (ϵ_i) that fit only the TF expression profile (y_{opt}) for the defined optimal condition $\epsilon_i = \sum_j (1 - \beta_{ij}) y_j^{opt} - \alpha_i$. Each TF expression is bounded $\varphi y_i^{min} \leq y_i \leq \varphi^{-1} y_i^{max}$ by a range interval defined by the minimum (y_{min}) and maximum (y_{max}) value of all experimental measurements for that TF in the microarray compendium (M3D). $\varphi \geq 1$ is a tunable parameter that decreases the gene expression range to improve the predictive capacity of the presented model under environmental and genetic perturbations.

To construct the TRN model, we used steady-state mRNA expression profiles derived from transcriptional perturbations collected in *M3D* version 4.5 online (*M3D*). We identified 330 TFs by searching for the keyphrase transcription factor in the functionally annotated *E. coli* genome from RegulonDB (version 5). The dataset contains pre-processed expression data from 380 hybridization experiments using 4,289 probe sets spotted on an Affymetrix GeneChip. Data were normalized using the robust multi-array average method [28] and represented on the log_2 scale. The inference procedure consisted of three nested steps. In the first step, global network connectivity was inferred using the *InferGene* algorithm [52]. This method

uses mutual information (MI) with local significance (z-score computation) to compute the number of transcriptional regulations in the genome [14]. Hence, each potential interaction between a regulator and a gene receives a z-score, which provides an estimate of MI. This approach eliminates some false correlations and indirect influences [14]. Subsequently, we selected a z-score threshold for cut-off. We included transcriptional regulations that were experimentally compiled in RegulonDB [14], but not those inferred by our procedure. Then, multiple regressions based on ODEs were performed to estimate the kinetic parameters of the regulatory model.

The wild-type transcriptional network contains 2,987 inferred regulatory interactions with z-scores over the selected threshold of 5. The network also contains 3,388 interactions from the reference regulatory set constructed based on RegulonDB [14]; 179 of these experimental interactions also belonged to the inferred test. The performance of the inferred TRN model topology was evaluated using a reference network defined by genes with known transcriptional regulation. Only interactions among genes included in this reference set were considered. The fraction of interactions that were correctly predicted by the model (the precision, P) and the fraction of all known interactions that were discovered by the model (the sensitivity, S) were used to compute a global performance statistic defined as $F = 2PS/(P+S)$ [15]. This TRN has a global performance of $F = 11.8\%$ (35.1% precision and 7.1% sensitivity) in predicting the regulations identified in RegulonDB. While this provides far from complete understanding of the regulation network of *E. coli*, the model constructed demonstrates sufficient predictive power to be used as starting point for our design.

Biological systems optimize their regulation by monitoring changes in their environment. Gene expression is largely controlled at the level of transcription by TFs. In addition to a DNA-binding domain, TFs often have structural domains that can bind specific metabolites. Thus, we increased the TRN complexity by including 299 external metabolic fluxes [53] as environmental factors (EFs). These EFs are direct links from the environment to the genetic network, affecting the expression of several TFs, and are common signals for endogenous and exogenous changes in cell state.

To link the environment to the regulatory network of the genome, we used two sets of experimentally obtained EF-TF interaction data published by Martinez-Antonio et al. [54] and Wall et al. [55]. However, only regulations in which the EF represents one of the 299 external metabolic fluxes defined in the work of Feist et al. were considered, reducing the set to 65 interactions (EF-TF) involving 50 EFs and 53 TFs [53]. The transcriptional

sensing system that was added to the TRN incorporated three types of sensors: (i) 14 transported metabolites (E-TM) that are sensed externally, (ii) 4 TFs that sense metabolites that are generated internally (I-SM) and (iii) 37 TFs that sense metabolites that are both transported and generated in the cytoplasm, *i.e.*, a hybrid system (H) [54]. Hence, we focused our study on one-component signal transduction pathways, because these are more widely over-represented in bacteria and display a greater diversity of domains than do two-component systems [56].

We computed γ_{ik} as a perturbation of the expression in the optimal condition of the gene i due to an environmental change that also perturbs the optimal state of the metabolic flux k, $\gamma_{ik} = \frac{\Theta y_i^{opt}}{v_k - v_k^{opt}}$, where Θ is a parameter that represents the normalized variation of the optimal expression. This parameter is optimized to fit the experimental gene expression under genetic and environmental perturbations. If j or k have no effect on the expression of i, then $\beta_i j = 0$ and $\gamma_{ik} = 0$; in fact, only regulatory effects of EFs on TFs are considered. We have not incorporated the effects of cooperation in transcription regulation. We have used public microarray hybridization data (*M3D*) from an Affymetrix chip normalized using RMA [56]. This microarray compendium contains data from 380 experiments.

Two parameters were optimized: $\varphi = 0.9$ defines the model gene expression range, and $\Theta = 0.5$ characterizes the variation in the wild-type expression of a given TF due to the influence of a specified external metabolic flux. These parameters were optimized to fit several predicted gene expression profiles from 31 experiments (contained in the *M3D* compendium) corresponding to transcriptional and environmental perturbations. Specifically, we used data from 16 knockouts of transcriptional master regulators (*appY*, *arcA*, *fnr*, *soxR*, *soxS*, *recA*, *fis*, *yncC*, and *rpoS*), 8 environmental perturbations of oxygen and carbon sources (glucose, acetate, glycerol, and proline), and 7 conditions combining both types of perturbations.

Appendix D Structure of the Wild-Type Global Transcriptional Model

The proposed transcriptional network contains 1,684 genes controlled by constitutive promoters; thus, more than half of the genes are non-regulated or are controlled by promoters with only one TF. Also, from a purely topological perspective, this transcriptional network of *E. coli* has a high density (0.154%) in comparison to values reported in similar studies on the same [51] and other organisms [1]. The characteristic path length (the distance

between two genes for which a path exists) follows a Gaussian distribution and ranges from 1 to 9 edges with an average value of 2.897 edges. Moreover, both the out- and in-coming connectivity distributions of the TRN belong to the class of scale-free small-world networks. The average clustering coefficient of the network was 0.103 and is log-log linearly related with the number of connections per gene in the range from 1 to 10 with a slope of -0.98, suggesting that the *E. coli* wild-type TRN is hierarchically organized [57, 58].

Appendix E Prediction of Transcriptomic Profiles

To compute the performance of our algorithm, we defined a reference network taking those genes with known transcriptional regulation. In addition, the TFs that were present in our reference set regulating genes outside the reference set were also removed when determining the performance of the algorithm. Then, only the interactions among the genes present in that reference set were evaluated to compute the algorithm efficiency. All known interactions catalogued in RegulonDB version 4 [14] were used to construct the reference network in *E. coli*. However, we are still far from a complete understanding of the transcriptional regulation network of *E. coli*. Therefore, we designed *in silico* genomes with predefined regulations to validate the performance of our algorithm. For that, we did not consider: *i*) operons with self-regulations, *ii*) operons with constitutive promoters, nor *iii*) operons containing only TFs.

We calculated two types of efficiencies (precision rate and sensitivity) to compare the inferred network with the reference network. We defined precision rate as the fraction of predicted interactions that are correct ($TP/(TP+FP)$), and sensitivity as the fraction of all known interactions that are discovered by the algorithm ($TP/(TP+FN)$), where TP is the number of true positives, FN the number of false negatives and FP the number of false positives [49, 50].

Appendix F Designing Genomes and Expression Data

In order to evaluate the suitability of our procedure to redesign the transcription regulation, we will analyse our ability to infer the kinetic parameters. Since they are not known for any organism, this lead us to the development

of a Generator of Artificial Genomes (GAG) to *in silico* create expression profiles (Figure 3.7). To construct such genomes, we specify the number of genes and TFs (this last is usually taken one order of magnitude less than the number of genes), and eventually the ratio between inducers and repressors (we have used 2/3). We can also specify the degree of connectivity to obtain scale-free networks (we have considered a probability distribution $P(k) \propto k^{-2}$ where k is the number of regulators of an operon), and the law for clustering distribution (we have assumed $P(n) \propto 2^{-n}$ where n is the number of genes per operon).

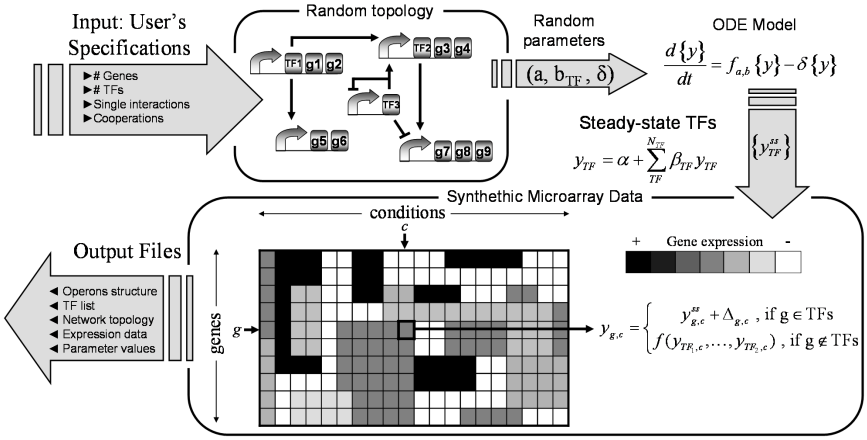

Figure 3.7: Generation of an artificial genome model to get synthetic microarray data. We have developed a computational algorithm (GAG) to construct such model, where the user inputs the total number of genes and TFs as well as the percentage of single regulations and cooperations. GAG generates a random network following those specifications with the corresponding model parameters: the constitutive transcription rate a, the regulatory coefficients b_{TF}, and the mRNA degradation coefficient δ. In the presented network, consisting of 5 operons, 3 TFs and 9 non-regulatory genes (g_i), arrows mean activation and blunt lines repression. The regulatory function (f) is assumed linear and the expression is calculated in the steady state (ss), where $\alpha = a/\delta$ and $\beta_{TF} = b_{TF}/\delta$. Later, GAG gives the *in silico* microarray data. We select a TF or a subset of TFs and we perturb the expression in the steady state ($\Delta_{g,c}$). Then we recomputed the whole expression profile using the model. Therefore, GAG outputs the list of operons and TFs, the regulatory network (adjacency matrix) with the corresponding model parameters.

To generate synthetic microarray data, we firstly obtain the steady state of the system ($y = f(y)$, since $dy/dt = f(y) - y$ with an arbitrary degradation rate of 1) without taking into account cooperations between different regulators (*i.e.*, $\beta_{ijk} = 0, \forall i, j, k$) as an approximate solution of the system (Eq. 3.2A). In fact, as the gene expressions (y) are only functions of the TFs (y_{TF}), we can write the system as $y = f(y_{TF})$. Subsequently, we generate a

new condition by randomly choosing a set of TFs with given size optimized for the inference and perturbing their steady state values, while maintaining constant the other TF expressions.

The perturbations over/under-express the TFs to a 50%, relative to their steady states. Hence, this perturbed value (y^*_{TF}) is used to recalculate the gene expressions by applying the model $y^* = f(y^*_{TF})$. Although this could be extended to more complicated conditions, where different gene categories are altered, the conditions based on TF perturbations are more revealing. Furthermore, to generate more realistic data we have added random fluctuations (which would simulate noisy data) in the expression values. We have studied the efficiency (precision rate and sensitivity) of our algorithm for different noise levels.

Appendix G *In Silico* Genome Evolution by Adaptive Mutation

With slight alterations, our methodology was able to predict the behavior of intermediary *E. coli* strains generated from laboratory evolution by local adaptive mutations in minimal lactate media [31]. We chose strains that were more adapted to the new media as the optimal model, in contrast to the automatic design by gene refactorization in which the wild-type strain was the optimal model. Consequently, this altered the expression profile required to maintain optimal cell behavior. Hence, by introducing all adaptive mutations to the wild-type genome-scale model, we were able to simulate the optimal gene expression profile, $y_g^{adapted}$. We replaced the corresponding ODEs of the mutated genes ($\hat{g}$), $\frac{dy_{\hat{g}}}{dt} = \alpha_{\hat{g}} + \sum_j \beta_{\hat{g}j} y_j + \sum_k \gamma_{\hat{g}k} \Delta v_k - \delta_{\hat{g}} y_{\hat{g}}$, with constant expression values to simulate the new steady state of that mutated gene, $y_{\hat{g}} \in [\varphi y_{\hat{g}}^{min}, \varphi^{-1} y_{\hat{g}}^{max}]$. Note that to simulate the appropriated minimal media, we imposed $\Delta v_k = 0$ for all metabolic uptake factors excepting the lactate ($\delta v_{lactate} = 20$) and glucose ($\Delta v_{glucose} = -10$) uptakes. Solving the new system of ODEs that incorporates the adaptive mutations, $\frac{dy_i}{dt} = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i,\ y_{\hat{g}} = Y_{\hat{g}}, Y \in [\varphi y_{\hat{g}}^{min}, \varphi^{-1} y_{\hat{g}}^{max}]$, we simulated different gene expression profiles, $y_g = y_g(Y_{\hat{g}})$, as functions of the steady state expression of the mutated genes, $Y_{\hat{g}}$, for intermediary adaptive *E. coli* strains. We computed transcriptomic fitness as the distance measured by the Pearson correlation coefficient, ρ, from the gene expression profile of the strain most adapted to the new environment with lactate ($y_g^{adapted}$) to the predicted profile incorporating adaptive mutations ($y_g = y_g(Y_{\hat{g}})$), $S_{exp}(Y_{\hat{g}}) = \rho\left(y_g^{adapted}, y_g(Y_{\hat{g}})\right)$.

Note that in Figure 3.5D, we selected $S_{exp}(Y_{\hat{g}})$ to optimize the correlation between growth rate and S_{exp} for the different intermediary steps of each strain evolved. Interestingly, we found that the gene mutations that caused maximal $S_{exp}(Y_{\hat{g}})$ also guaranteed maximal correlations.

Appendix H Prediction of Rewired Transcriptional Network of *E. coli*

Our methodology was able to capture the behavior of *E. coli* strains with TRNs rewired by adding on a wild-type genetic background new links from different recombinations of promoters with TFs. A recent study by Isalan et al. systematically explored such problem by expressing endogenous promoters controlling different TFs or σ-dependent genes and measuring the growth rate of each rewired strain [32]. In our study, we did not consider promoter region-open reading frame fusions that were constructed on high copy number plasmids because our model is limited to predict gene expression from the bacterial genome. Therefore, we selected 38 strains from Isalan et al. collection in which the rewired construct was stably integrated in the *E. coli* chromosome. For these strains, we computed their growth rate as the maximum value of $\Delta(lnOD_{600})/\Delta t$ (with Δt = 1 hour), achieving values between $0.39h^{-1}$ and $0.63h^{-1}$. The strain with the construct of the TF, *rpoE*, controlled by the promoter appY was not included in the dataset because it showed the largest lag phase and slowest growth rate compared to the rest strains, indicating that the levels of gene expression are not necessarily in steady state. Therefore, this strain violated our assumption of steady state gene expression as a proxy to fitness, S_{exp}.

Hence, by introducing the modification imposed by the rewired construct to the wild-type genome-scale model, we were able to simulate the gene expression profile of the rewired TRN, and consequently, predict fitness. We modified the corresponding set of ODEs that models the expression of the TF (TF_c) encoded in the construct, c, and controlled by the promoter, p. Specifically, we added the basal rate and that determines the gene expression of the genes controlled by p to the ODE models TF_c. Solving the new system of ODEs, $\frac{dy_i}{dt} = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i,\ i\ /TF_c\ \frac{dy_i}{dt} = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i + \alpha_p + \sum_j \beta_{pj} y_j + \sum_k \gamma_{pk} \Delta v_k - \delta_p y_p,\ i = TF_c$ we simulated the gene expression profile of the rewired TRN in order to compute the fitness S_{exp}. Note that to simulate the appropriated medium, we imposed $\Delta v_k = 0$ for all metabolic uptake factors excepting the glucose uptake ($\Delta v_g lucose = 50$) to provide an excess of carbon source.

References

[1] Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., Young, R. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804.

[2] de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comp. Biol.* 9, 67-103.

[3] Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126.

[4] Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., Palsson, B.O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92-96.

[5] Eisen, M., Spellman, P., Brown, P., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863-14868.

[6] Ben-Dor, A., Shamir, R., Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.* 6, 281-297.

[7] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. U. S. A.* 96, 6745-6750.

[8] Dhaeseleer, P., Liang, S., Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707-726.

[9] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31, 370-377.

[10] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., diBernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78.

[11] Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropymeasurements. *Pac. Symp. Biocomp.* 5, 415-426.

[12] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382-390.

[13] Margollin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., dellaFavera, R., Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7, S7.

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

[15] Meyer, P.E., Kontos, K., Lafitte, F., Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinf. Syst. Biol.* 2007, 79879.

[16] Yu, J., Smith, V., Wang, P., Hartemink, A., Jarvis, E. (2004). Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 3594-3603.

[17] Husmeier,D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* 19, 2271-2282.

[18] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C., Ferreira, C.E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Syst. Biol.* 1, 39.

[19] Steinke, F., Seeger, M., Tsuda, K. (2007). Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. *BMC Syst. Biol.* 1, 51.

[20] Gardner, T., di Bernardo, D., Lorenz, D., Collins, J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiles. *Science* 301, 102-105.

[21] di Bernardo, D., Thompson, M., Gardner, T., Chobot, S., Eastwood, E., Wojtovich, A., Elliott, S., Schaus, S., Collins, J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 3, 377-383.

[22] Shevade, S., Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* 19, 2246-2253.

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

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B.* 58, 267-288.

[25] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498-2504.

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

[27] Faith, J., Driscoll, M., Fusaro, V., Cosgrove, E., Hayete, B., Juhn, F., Schneider, S., Gardner, T. (2008). Many microbe microarrays database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* 36, D866-D870.

[28] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249-264.

[29] Karp, P., Riley, M., Saier, M., Paulsen, I., Collado-Vides, J., Paley, S., Pellegrini-Toole, A., Bonavides, C., Gama-Castro, S. (2002). The EcoCyc DataBase. *Nucleic Acids Res.* 30, 56-58.

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

[31] Conrad, T. M., Joyce, A. R., Applebee, M. K., Barrett, C. L., Xie, B., Gao, Y., Palsson B. O. (2009). Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol.* 9, R118.

[32] Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M., Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840-845.

[33] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H. and the rest of the SBML Forum. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524-531.

[34] Price, M., Huang, K., Alm, E., Arkin, A. (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* 33, 880-892.

[35] Reiss, D., Baliga, N., Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* 7, 280.

[36] Mordelet, F., Vert, J.P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics* 24, i76-i82.

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

[38] Behrens, J., vonKries, J., Khl, M., Bruhn, L., Wedlich, D., Grosschedl, R., Birchmeier, W. (1996). Functional interaction of bold β-catenin with the transcription factor LEF-1. *Nature* 328, 638-642.

[39] Stewart, V., Bledsoe, P. (2005). Fnr-, NarP- and Narl-dependent regulation of transcription initiation from the *Haemophilus influenzae* Rd napF (Periplasmic Nitrate Reductase) promoter in *Escherichia coli* K-12. *J. Bacteriol.* 187, 6928-6935.

[40] Long, J., Roth, M. (2008). Synthetic microarray data generation with RANGE and NEMO. *Bioinformatics* 24, 132-134.

[41] Gray, R. (1990) Entropy and Information Theory. Springer-Verlag, New York, NY, USA.

[42] Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18, S231-S240.

[43] Daub, C., Steuer, R., Selbig, J., Kloska, S. (2004). Estimating mutual information using B-spline functions an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* 5, 118.

[44] Cohen, J.P.C., West, S., Aiken, L. (2003). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ, U. S. A.

[45] Bar-Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics* 20, 2493-2503.

[46] Affymetrix (1999). Affymetrix Microarray Suite User Guide, version 4. Affymetrix, Santa Clara, CA, USA.

[47] Sabatti, C., Rohlin, L., Oh, M., Liao, J. (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* 30, 2886-2893.

[48] Dongarra, J., Bunch, J., Moler, C., Stewart, P. (1979). LINPACK Users Guide. SIAM, Philadelphia, PA, USA.

[49] Altman, D., Bland, J. (1994). Statistics notes: diagnostic tests 1: sensitivity and specificity. *Br. Med. J.* 308, 1552.

[50] Altman, D., Bland, J. (1994). Statistics notes: diagnostic tests 2: predictive values. *Br. Med. J.* 309, 102.

[51] Carrera, J., Rodrigo, G., Jaramillo, A., Elena, S.F. (2009). Reverse-engineering the *Arabidopsis thaliana* transcriptional network under changing environmental conditions. *Gen. Biol.* 10, R96.

[52] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

[53] Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., Palsson, B.O. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3, 121.

[54] Martnez-Antonio, A., Janga, S.C., Salgado, H., Collado-Vides, J. (2006). Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol.* 14, 22-7.

[55] Wall, M.E., Hlavacek, W.S., Savageau, M.A. (2004). Design of gene circuits: lessons from bacteria. *Nat. Rev. Genet.* 5, 34-42.

[56] Ulrich, L.E., Koonin, E.V., Zhulin, I.B. (2005). One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.* 13, 52-6.

[57] Barabasi, A.L., Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101-113.

[58] Ravasz, E., Barabasi, A.L. (2003). Hierarchical organization of complex networks. *Phys. Rev. E* 67, 026112.

Chapter 4

Automatic Design of a Genome by Gene Refactorization

4.1 Design by Gene Refactorization in Dynamic Environments

4.1.1 Refactored Genomes with a Reduced Number of Operons

In the previous chapter, we have demonstrated that we can predict experimental growth rates by assigning transcriptional parameters to genome regulatory sequences. Such assignment allows us to predict the TRN model after reshuffling genetic elements. Instead of trying to solve the challenging problem of evolving a genome for better growth, which would require a greater degree of accuracy for our fitness function, we attempted to reorganize the genome of *E. coli* while maintaining its functionality. We rearranged the nucleotide sequence to simplify the TRN in terms of regulatory complexity and modularity. We used our automatic design methodology to perform the genome refactoring, which entailed rearrangement of the operon structure while maintaining the organisms original behavior.

This problem can be solved because of our current understanding of the relationship between sequence and transcription regulation. For example, the operator sites of many promoter sequences are known, and mutations in such sequences would presumably eliminate the regulation of the promoter. This genetic modification could be implemented in our evolutionary algorithm in such a way that, once the optimal TRN is found, the genome engineer would need to further engineer such mutated promoter sequences

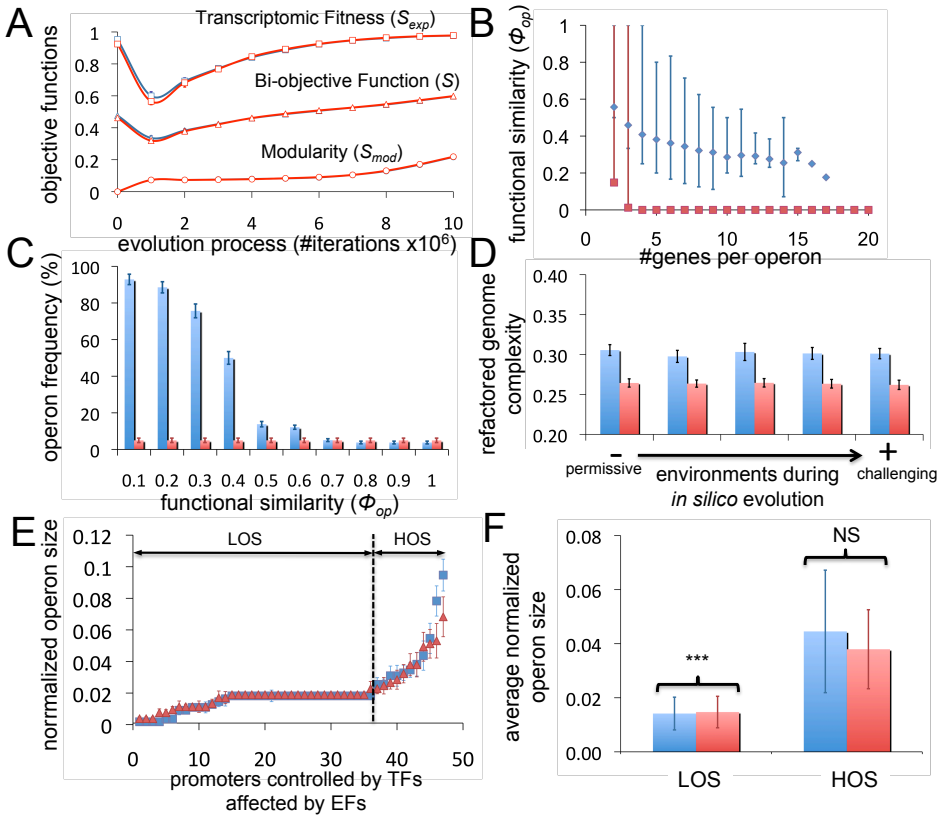

Figure 4.1: (A) Evolution during the optimization process, expression and modularity score, and bi-objective function for genome design. Random optimization produced significantly lower bi-objective function values than those of the wild-type genome. (B) Functional similarity, depending on operon size, of refactored genomes that have evolved in environment 3. (C) Histogram of functional similarity of the operons of genomes refactored under permissive environments (type 1). (D) Complexity reduction (the number of regulators and operons are represented by blue and red bars, respectively) of refactored genomes with respect to the wild-type genome under different levels of environmental extremity with regard to each altered factor (these range from type 1 (most permissive) to type 5 (most challenging)). Operons with a size equal to one (null-similarity) were not plotted in panels C and D. (E) Sizes of operons containing one or more genes controlled by promoters that interact with EFs in genomes refactored under permissive and challenging environments (blue squares and red triangles, respectively). Green diamonds indicate the gene distribution in operons for the wild-type genome. The vertical dashed line indicates the mean operon size. Low operon size (LOS) and high operon size (HOS) classes were defined using the average operon size as a cutoff. (F) Mean operon size of genomes refactored under weak and strong environmental conditions (blue and red bars, respectively) for LOS and HOS classes. The U-test significance is shown (***$p < 0.01$; NS: not significant). Error bars represent the standard deviations of scores obtained from 10 simulations. Selective pressure in the evolution process was only applied to genes relating to central metabolism.

(Figure 3.4B) to obtain the final genome sequence. In addition to the fitness related to growth, S_{exp}, we needed another objective function that is related to the expected genome arrangement (see Appendix A). Figure 4.1A illustrates the trajectories of the S_{exp} and S_{mod} functions and their weighted sum, which defines the fitness function to be used during the refactorization. This is carried out for different environments by maintaining the optimal gene expression levels of all enzymes. The fitness function achieved similar values during the last steps of the evolution process for all simulated replicates of the refactored genomes.

We then investigated whether genes with high functional similarity were grouped into the same operons or network modules, *i.e.*, we computed the functional similarity of all operons containing more than one gene in the refactored and random operon-organization genomes. Figure 4.1B shows the highly statistically significant functional similarity of genes refactored into the same operon with respect to random refactorizations (Kolmogorov-Smirnov test, $p < 0.001$; and U-test, $p < 0.001$; see Appendix D). Analogously, Figure 4.1C illustrates the distribution of functional similarities for all operons with non-zero values. It is especially interesting that the refactored genomes were characterized by operons containing genes with similar function, a property that was not imposed during the evolutionary process. Specifically, the number of refactored operons with degrees of functional similarity < 0.8 considerably exceeded the number of those with random organization. Figure 4.2 (B, C) illustrates the same functional operon-organization using an expression score based on genes related to adaptive and defensive processes; Figure 4.2 (H, I) also shows the analogous operon-structure for the entire genome.

We observed a significant reduction in the complexity of the refactored TRN with respect to the wild-type genome. Figure 4.1D illustrates the ratio between the number of regulatory interactions ($\Xi < 0.31$; $p < 0.001$) and the number of operons ($\Theta < 0.27$; $p < 0.001$) for the refactored and wild-type genomes, which do not appear to depend on the environment. These genomes were optimized under conditions requiring only central metabolism enzyme expression to remain close to the optimal level. How does the reduction of genome complexity depend on the selection of critical genes involved in the fitness function? To address this question, we also explored the possibility that limiting the expression of only those genes that relate to defense and adaptation would allow larger reductions in complexity (Figure 4.2D; $\Xi < 0.25$, $p < 0.001$; $\Theta < 0.23$; $p < 0.001$); the smallest reductions in complexity were obtained when the entire genome was restricted (Figure

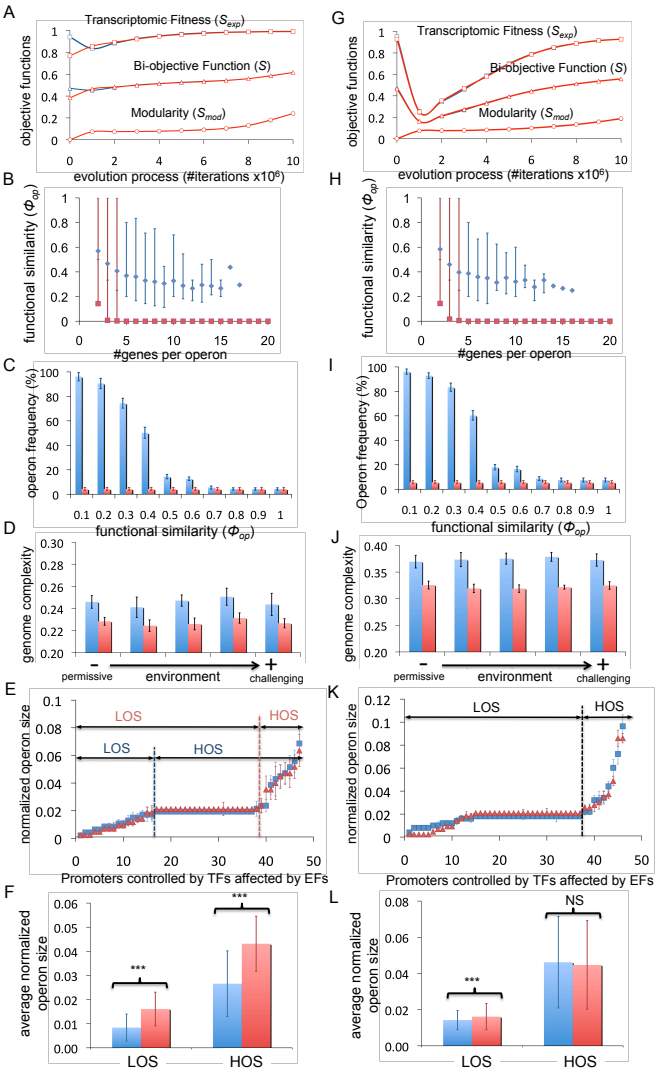

Figure 4.2: Genomes refactored under selective pressure based on the expression of stress-related genes involved in adaptation and defense (A-F) or the entire genome (G-L). (A and G) Evolution during the process of optimizing the expression and modularity score, and the bi-objective function for genome design. (B and H) Functional similarity, depending on operon size, for the refactored genomes that evolved in environment 3. (C and I) Histogram showing the functional similarity scores of the operons in genomes refactored under permissive environments (type I). (D and J) Complexity reduction (the number of regulators and operons are represented using blue and red bars, respectively) of the refactored genomes with respect to wild-type genome under different levels of environmental extremity (these range from type 1 (most permissive) to type 5 (most constrained)). (E and K) Sizes of operons containing one or more genes controlled by promoters that interact with EFs in the refactored genomes under permissive and harsh environments (blue squares and red triangles, respectively). The vertical dashed line indicates the mean operon size value. Low operon size (LOS) and high operon size (HOS) classes were defined using the average operon size as a cutoff. (F and L) Mean operon size in genomes refactored under weak and strong environments (blue and red bars, respectively) for LOS and HOS classes. The U-test significance is shown ($^{***}p < 0.01$; NS: not significant). Error bars represent the standard deviations of scores obtained from 10 simulations.

4.2J; $\Xi < 0.38$, $p < 0.001$; $\Theta < 0.33$; $p < 0.001$). Thus, high reductions in genome complexity were obtained independently of the set of genes selected as critical predictors of transcriptomic fitness.

4.1.2 Cellular Environments Selectively Correlate with Genome Regulatory Complexity in the Refactored Genomes

We compared the internal structures of networks evolved under permissive and challenging environments to determine whether the environmental conditions imposed in the refactorization confer any specific characteristics to the TRN of the refactored genomes. The clustering coefficients (CCs) of the refactored TRN were highly reduced with respect to the wild-type TRN, illustrating that the refactored genomes are composed of large modules that induce additional coregulation. Interestingly, genomes refactored in challenging environments show higher CCs than those evolved in more permissive environments, a difference supported by the positive Pearson correlation observed between the CCs and the gradient of environmental stress ($r = 0.63$; $p < 0.01$) when S_{exp} was computed considering stress genes only. Analyses of other topological properties of the TRNs of the evolved genomes are presented in the Appendix E.

We then analyzed the relationship between the reduction in genome complexity and the environment in which the genome evolved. We found no significant correlation between increased environmental challenge (measured as the variation in cell fitness) and the complexity of the refactored TRN (measured either as the number of operons or as the number of regulatory interactions). Surprisingly, the positive correlation observed between CCs and environmental stress did not contribute to a significant relationship in terms of the complexity of the refactored TRN.

Next, we focused only on the refactored operons that were regulated by promoters whose TFs interacted with EFs. Figure 2.1CE depicts the size of those 47 operons that were re-organized by forcing only central metabolism enzymes to achieve the optimal condition expression profile. The average normalized operon size was 0.0213 or 0.0204, depending on whether the environment in which the evolutionary process occurred was permissive or challenging, respectively. Surprisingly, we found a significant difference (U-test $p < 0.01$) between the average size of the refactored operons in permissive and challenging environments for low operon size (LOS) (Figure 4.1F). Analogously, we also found significant changes when the selection pressure forced optimization of all gene expression (Figure 4.2 (K, L), U-test $p < 0.001$) or the optimization of stress-related gene expression (Figure 4.2 (E, F), U-test

$p < 0.0001$). In fact, for stress-related gene expression, we also observed significant changes in the average size of the re-engineered operons for high operon size (HOS). This is direct evidence that environments in which cells perform poorly (*i.e.*, with very poor fitness) favored the emergence of operons containing a large number of genes co-expressed under promoters whose TFs respond to this environment.

4.1.3 Biochemical Adaptation in Refactored Genomes

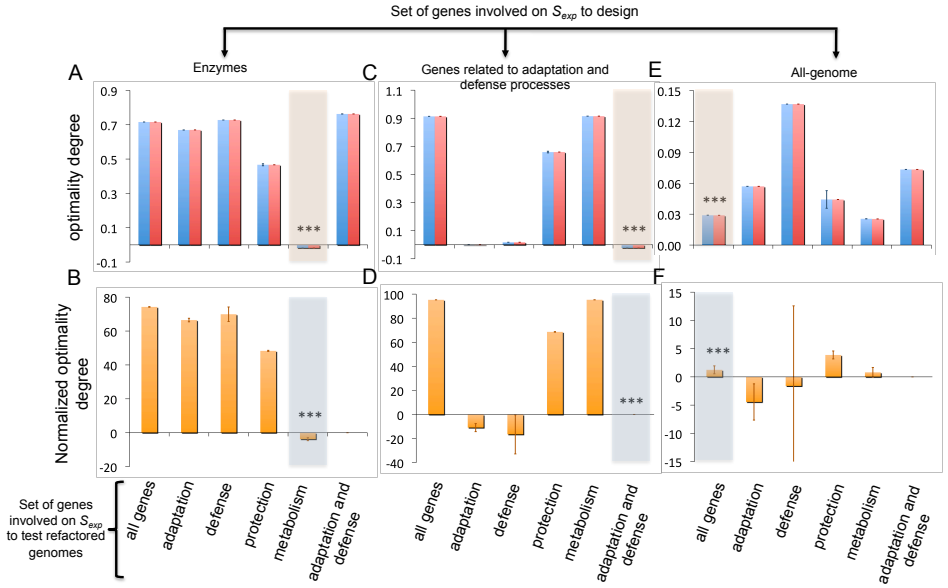

Figure 4.3: Adaptative behavior of the refactored *E. coli* genomes that evolved under selective pressure relating to genes coding for enzymatic activity (A and B), genes related to adaptation and defense functions (C and D) or to the entire genome (E and F). The behavior of the designed genomes was measured by applying single environmental perturbations (A, C and E) that modified external fluxes of oxygen or carbon sources (blue and red bars, respectively) or multiple, simultaneous changes in the oxygen, carbon and nitrogen sources (B, D and F). Adaptation was predicted for six cellular fitness constraints depending on the critical genes selected: I, all genes; II, III, IV and V, genes related to adaptation, defense, protection and enzymatic functions, respectively; and VI, genes associated with adaptation and defense processes. Genomes that were refactored by applying selection to the same category that was scored are highlighted in the plot (***). Error bars represent the standard deviations of scores obtained from 10 simulations.

Many signaling systems can adapt their expression programs in response to novel stimuli. Figure 3.5A shows that a single, strong environmental perturbation induced wild-type TRN to reduce cell fitness to a minimal, but

stable, value. This motivated us to investigate whether refactored systems acquired the ability to adapt to environmental changes more quickly than wild-type systems. To that end, we explored single environmental perturbations by simulating two sets of environments. We then used the optimality degree to test the adaptation of refactored TRN to the environments, considering three types of selection pressure in the expression score: selecting only genes coding for enzymes involved in central metabolism (Figure 4.3A), stress-related genes (Figure 4.3C) or the entire genome (Figure 4.3E). Interestingly, using the first two design criteria, the average of the optimality degrees $\left\langle \hat{\xi} \right\rangle$ around the set of environmental perturbations was negative (*i.e.*, cellular fitness exceeded the optimal value for all re-engineered genomes ($\left\langle \hat{\xi} \right\rangle$ = -0.018 or -0.023, respectively). On the contrary, genomes refactored based on the third criterion achieved significantly positive optimality degrees ($\left\langle \hat{\xi} \right\rangle$ = 0.029). Defining the fragility of a genome as its optimality degree in different environments, refactored genomes were more fragile; anticipatory behavior disappeared ($\left\langle \hat{\xi} \right\rangle > 0.467$, 0 or 0.025 for the three design criteria mentioned, respectively) when cell fitness was computed using an expression score from a set of critical genes different from that used during the design phase. It should be noted that the optimality degree under single perturbations did not significantly depend on alterations in metabolic uptake factors.

Next, we studied systems that were re-engineered under simultaneous multiple perturbations (Figure 4.3 (B, D, F)). As above, we tested genome optimality by altering oxygen and carbon source uptake factors in the same range defined by single perturbations, and we added a third sensing component related to the nitrogen source by adding NO_3^- to the environment. As before, refactored genomes achieved negative or zero degrees of normalized optimality with the two first design criteria ($\left\langle \hat{\xi} \right\rangle$ = -3.81%), respectively, but for the third criterion, the average normalized optimality ($\left\langle \hat{\xi} \right\rangle = 1.25\%$) indicated that new systems retained the fitness of the optimal system.

4.2 Prediction of a Refactored *E. coli* Genome Sequence with Wild-type Behavior in Changing Environments

Thus far, the experimental implementation of refactored genomes has required significant promoter engineering to obtain the desired synthetic promoters by adding operators that do not exist in wild-type promoters. Therefore, we implemented a second evolutionary process to design refactored

genomes containing only genetic building blocks that exist within the wild-type *E. coli* genome. The transcriptional regulation landscape that we explored contained all possible genome reconfigurations that could result from regrouping a set of genes under the control of a wild-type promoter. Similar to our previous designs, we observed a large reduction in the complexity of the refactored TRN with respect to the wild-type genome in terms of the number of regulations ($\Xi < 0.14$; $p < 0.001$) and operons ($\Theta < 0.14$; $p < 0.001$) using a design function based on scoring the expression of stress genes (Figure 4.4A). Analogously, we found that limiting only the expression of genes coding for enzymes or genes related to defense and adaptation in the design produced larger reductions in complexity ($\Xi < 0.18$, $\Theta < 0.19$ and $\Xi < 0.23$, $\Theta < 0.23$, respectively; $p < 0.001$ in all cases).

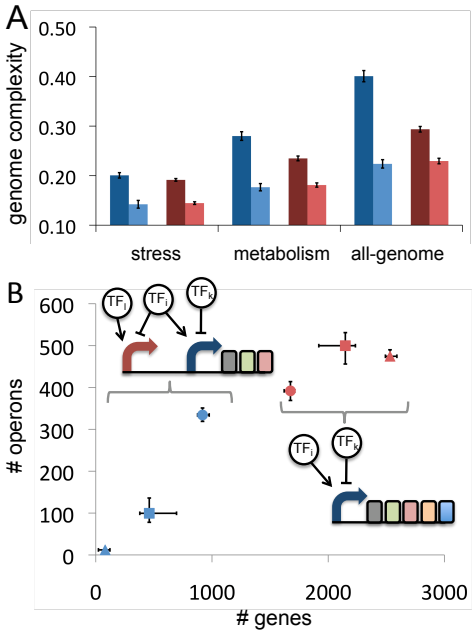

Figure 4.4: (A) Complexity reduction (the number of regulatory interactions and operons are represented using blue and red bars, respectively) of the refactored genomes with respect to the wild-type genome under permissive environments using an evolutionary process including tandem promoters (dark bars) or genetic re-organization of *E. coli* transcriptional units only (light bars). (B) Number of genes or operons regulated by a given wild-type promoter (blue points) or a tandem promoter added to the synthetic transcription unit (red points), as proposed for the refactored genomes evolved using selective pressure based only on genes coding for enzymatic activity (squares), genes related to adaptation and defense functions (triangles) or the entire genome (circles). Error bars represent the standard deviations of scores obtained from 10 simulations.

To enlarge the genome design landscape, we allowed the addition of a

maximum of three promoters in tandem to modify the regulation of a given operon (see the second evolutionary process described in Appendix A). We determined a set of *E. coli* promoters that were potential candidates to operate in tandem (sometimes using a suitable spacer sequence to isolate them). We selected the entire promoter library (27 promoters) used by Isalan et al. to exhaustively explore the effect of multiple genome rewirings on growth rate [1]. We also included all *E. coli* promoters that are regulated by fewer than two master regulator TFs, as defined by Isalan et al. Consequently, we considered 272 promoters susceptible to tandem incorporation. Figure 4.4A shows that the largest reductions in complexity were achieved using designs that consider stress genes in the objective function ($\Xi < 0.20$, $p < 0.001$; $\Theta < 0.19$, $p < 0.001$). Surprisingly, as shown in Figure 4.4B, few operons from the refactored genomes needed a promoter to be added in tandem to modify the gene expression provided by their wild-type promoter. Only 15 operons within the refactored genomes required the addition of two tandem promoters to guarantee that gene expression could adapt to changes in the environment. Such refactored genomes were characterized by operons that captured genes with similar functionality (Figure S4.5 A-F). We also tested adaptation in genomes that were refactored by considering each of the three types of selective pressures previously mentioned in the expression score (see Appendix F and Figure 4.5 G, H).

4.3 Conclusions

4.3.1 Biological Consequences of Computational Genome Refactorization

Genome Organization Can Be Simplified Without Disrupting the Response of the Genome to Environmental Changes

In this Chapter, we have developed a computational framework for the design of bacterial genomes that are able to respond to changes in environmental conditions. We used transcriptomic data to infer a continuous model for the transcription of all *E. coli* genes [2], which we then used to assign appropriate parameters to promoter and TF coding sequences. By assuming that these parameters do not depend on genomic context in most cases, we proposed our first methodology for the automatic design of genome rearrangements under changing environments. Our results demonstrate that it is possible to refactorize the genome of *E. coli*, achieving a 69% reduction in the number of regulatory interactions and a 73% reduction in the number of operons, while maintaining the ability to physiologically adapt

to environmental changes. We found that the refactored genomes contain operons that encode several genes with similar functionality. This is an important result, given that the fitness function imposed to evaluate genome performance did not consider gene function. This agrees with the experimental observation that genes within an operon have similar functions [3]. Moreover, these genomes acquired the ability to adapt more rapidly to environmental changes, probably as a direct consequence of the reduced number of regulatory elements.

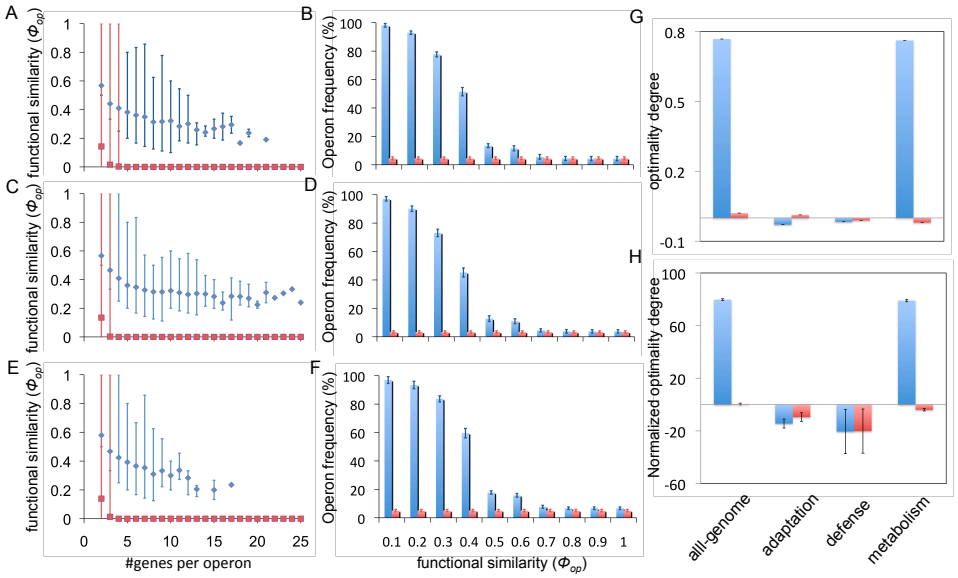

Figure 4.5: Functional similarity, depending on operon size, and a histogram showing the functional similarity of genes in the operons of genomes refactored under selective pressure on the expression of genes coding for enzymes (A and B), stress genes (C and D) and the whole genome (E and F). A second set of rules was used for the evolutionary process (see Figure S1) including a limit on the maximum number of tandem promoters allowed in each operon. (G and H) Adaptive behavior of the refactored genomes evolved under different selective pressures in environments with single or multiple perturbations, respectively, with tandem promoters allowed (red bars) or not (blue bars), using the second set of rules for the evolutionary process. Error bars represent the standard deviations of the scores obtained from 10 simulations.

The refactored genomes satisfied the main design specification, which was to maintain the global physiological response under both optimal and changing environments. In addition, we found that there was an increase in the complexity of the internal structure related to the signal transduction for all refactored genomes. More specifically, genomes that evolved under the most extreme environments required a greater re-organization of

critical genes under promoters that could sense greater numbers of environmental interactions. Interestingly, genomes that were refactored under stressful environments showed higher clustering coefficients than those that evolved under more permissive environments. An intuitive explanation for this observation relies on the differences in the selective pressures imposed by both types of environments. Survival and replication in a stressful environment represents stringent selection, requiring the coordinated expression of all genes involved in survival. By contrast, replicating in a permissive environment may be equated to soft selection and therefore does not require the coordination of expression because the cells remain able to exploit some components of their environment.

Design Principles of Genomic Adaptation to Environmental Changes

One important application of our results is the ability to infer some principles of genome design. In particular, we studied the refactored genomes that had achieved over-optimality or lost optimality. Our refactored genomes were more susceptible to environmental perturbations when we tested optimality using transcriptomic fitness based on a different set of genes than that selected for the refactorization. Recent work has shown that biochemical networks have evolved to capture the multidimensional structure of diverse environments and thus form internal representations (through regulatory networks) that allow the prediction of environmental changes. For example, Tagkopoulos et al. provided evidence of anticipatory behavior of *E. coli* to changes in temperature and oxygen levels that occurred over evolutionary time scales [4, 5, 6]. We tested the anticipatory ability of our refactored genomes by computing their optimality using transcriptomic fitness with the same set of genes used in the refactorization process. Interestingly, we found that refactored genomes achieved greater optimality degrees than those of wild-type genomes for both single and multiple environmental perturbations. This suggests that natural selection may be shortsighted (*i.e.*, it does not anticipate large changes over the long-term) and that actual genomes have thus evolved for optimal responses to regimes of small fluctuations.

Extension of this Methodology to Other Organisms

The methodology presented here could be extended to other organisms for which quantitative TRN and signal transduction models can be inferred. The models should be able to predict genomic transcriptional profiles under several external conditions in order to construct a transcriptomic fitness function. The computational refactorization of the genome of a given organ-

ism requires the following information: (i) genome annotation, (ii) a microarray compendium capturing genetic and environmental diversity, (iii and iv) datasets of transcriptional interactions (gene *vs.* TF) and two-component signal transduction pathways (TF *vs.* EF) that have been experimentally verified. Further extensions would have to consider the influence of other factors such as 3D localization [7], post-transcriptional regulation [8] and post-translational regulation in prokaryotes [9], or chromatin regulation [10] in eukaryotes.

4.3.2 Experimentally Testable Predictions

Strains from Short-Term Laboratory Evolution Show Growth Rates Similar to Those Predicted by Genome Design

Our methodology takes advantage of our ability to predict variations in cell growth based on changes in the transcriptome. The linear relationships found in Figure 3.6 guarantee that optimizing transcriptomic fitness also optimizes the growth rate of the cell. It should be noted that we have fixed the mutated genes to steady-state expression levels under the given lactate environment and that the fixed level is found by optimization. The experimental characterization of such transcription values could change this linear relationship and therefore the correlation between predicted fitness and growth rate. However, this would not change our results, as any monotonous relationship would be sufficient for optimization. Therefore, we have provided evidence that our fitness function, which is based on transcriptomic profile distances, is a suitable selection pressure for our in silico evolution procedure.

Proposition of a Testable Refactored *E. coli* Genome Sequence

This work also provides a generic procedure to generate a gene sequence for a synthetic *E. coli* genome with a targeted transcriptomic response, which we exemplify by proposing a genome that could be engineered by assembling known elements. For this quantitative prediction of a designed genome, we propose combining known promoter regions and transcriptional regulators such that the transcriptional profile could reasonably be predicted. To create more complex promoters, we propose taking advantage of their modularity and fusing some of them in tandem, and choosing a set of promoters for which transcriptional interference would be minimized [11]. Notice that the *E. coli* genome contains 166 non-overlapping tandem promoter pairs [11]. As sequence repetition could create ectopic recombination events, some care

will have to be taken in experimental testing. In addition, wild-type transcription terminators are not completely efficient and are sometimes even absent; therefore, some terminators may have to be replaced by stronger ones (probably synthetic, to avoid repetitive sequences). The neglect of non-transcriptional regulation may produce unexpected behavior in certain environments, but this could be remedied by selecting alternative conditions. Other undesired behaviors could be alleviated by suitable randomization of the nucleotide sequence with the restriction of maintaining the desired functionality (*e.g.*, the ribosome-binding site or protein coding sequence).

Application to Experimental *in vivo* Genome Engineering

Our computational procedure could also be adapted to the particular needs of experimentalists who are willing to create major gene rearrangements in genomes using *in vivo* techniques. Several affordable cloning-free techniques allow the generation of large insertions, deletions or inversions in the genome. Usually, random mutagenesis followed by screening is used, but this methodology is tedious when a specific locus is to be targeted, and only a small number of successive modifications (N) would be practical. Therefore, it would be particularly useful for the genome engineer to know in advance the most suitable sequence of experiments to introduce genome modifications. In *E. coli*, this could be readily done by appropriately adapting the mutational moves used by our *in silico* evolution methodology. Such moves should be restricted to their genome rearrangement technologies available at the laboratory. Then, one could computationally explore all possible evolutionary paths of N moves that would give the highest fitness under specific dynamic environments. The experimentalist could then engineer the genome by implementing the N consecutive experiments suggested by the algorithm. This computational procedure could also incorporate the constraint that each intermediate genome should be viable.

Appendix A Automatic Genome Design

The main variables required for automatic genome design are the same as those required for any evolutionary algorithm: (i) An initial genome, (ii) evolutionary steps represented by changes in the genome and (iii) a fitness function to evaluate the performance of each mutant genome (see Appendix B). For the first step, we used the genome of the model bacterium *E. coli*. The second step was achieved by dissecting the bacterial genome into elementary modules, to which evolutionary rules were applied [12].

One design approach that we used involved the *in silico* refactorization of the nucleotide sequence of the *E. coli* genome, a process where we pursued two goals simultaneously: (i) simplifying the internal structure of *E. coli* and (ii) maintaining the external system function. To maximize the modularity of the system and thus simplify the TRN, we defined a measure based on the entropy of the genome. We also aimed to maximize the similarity of the expression profiles of the wild-type and refactored genomes for a set of extreme environments and for a set of critical genes that guarantee the functionality of the refactored system. We used the TRN model integrated with signal transduction to measure that similarity. Considering these two aims, we developed an optimization algorithm (Figure 3.4B) based on the mutation rules described in the Appendix B to refactorize the wild-type *E. coli* genome. Genes that are controlled by constitutive promoters were not involved in the design. These genes could always be refactored in a straightforward way by assuming that they could be collapsed into large operons regulated by a gradient of different expression levels (produced by a library of several constitutive promoters or using tuned ribosome-binding sites).

Appendix B Genome-Wide Optimization Procedure

Our algorithm searches possible reconfigurations of the global transcriptional regulation of *E. coli* such that the resulting modular genome contains all genes in a minimal set of operons, thus decreasing the number of transcriptional regulatory elements, and with the constraint that the overall gene expression of the refactored genome shall be as close to the wild-type as possible. We used Monte Carlo Simulated Annealing [13] to perform the optimization in the space of all possible refactored transcriptional networks. The size of this combinatorial space is governed by the previously characterized variability in the *E. coli* natural promoters, and the diversity of synthetic promoters was obtained during the optimization process. As the starting condition, we assumed that the expression of each gene was controlled only by its natural promoter. Based on the transcriptional regulation landscape size, we defined two sets of optimization processes. In the first set, we introduced small transcriptional modifications in the genome at each step of the optimization by either changing the regulation of a gene (moving it downstream of another promoter) or eliminating regulation by natural or synthetic promoters according to the following rules:

1. Move gene *g* belonging to operon *op* and regulated by a non-constitutive promoter *P(op)* to another operon *op* regulated by a different non-constitutive promoter *P(op)*. When *g* moves to *op*, we add all regulatory operators of its natural promoter to *P*. However, the fact that *g* leaves *P* implies that if the gene is regulated by a promoter different from its natural promoter, then *P* will lose all inserted operators due to the regulatory effect of *P* on *g*. Co-expression of all genes expressed from a given operon was imposed.

2. Remove an operator from a synthetic promoter. Only operators associated with TFs are likely to be removed. Unlike transcriptional regulations, interactions of TFs associated with the binding with EFs remain linked to their corresponding genes throughout the optimization process.

To simplify the genome network structure and improve algorithm convergence, the probability of removing a regulation was made much larger than the probability of changing a genes promoter (*e.g.*, 10-fold). We also defined a second set of transcriptional perturbations that were more restrictive from a refactored genome diversity standpoint:

1. Move gene *g* belonging to operon *op* and regulated by non-constitutive promoter *P(op)* to another operon *op* regulated by a different non-constitutive promoter *P(op)* without adding regulatory operators to *P*.

2. Move gene *g* belonging to operon *op* and regulated by non-constitutive promoter *P(op)* to another operon *op* regulated by a different non-constitutive promoter *P(op)* and add a wild-type promoter in tandem position with *P*. In this case, gene expression is controlled by the addition of tandem promoters upstream of *op*. For simplicity and to avoid looping between promoter sequences, we limited the downstream addition of promoters in tandem to two.

3. Remove or permute promoters in tandem positions. Promoters added in tandem to a given transcription unit could be removed or replaced by other promoters. The probability of removing a promoter in tandem was set to be much larger than the probability of replacing one promoter in tandem with another promoter (*e.g.*, 10-fold).

Both sets of evolutionary processes share a rule to simulate the expression behavior of the newly created genome and compute its new objective

function (S_{new}), which depends on the full transcriptome predicted under a set of environments and the new modular organization of the operons. If the suggested mutation improves S ($S_{new} \geq S$), then it is accepted. Otherwise, it is accepted with probability $e^{(S-S_{new})/T}$, where T is a Boltzmann temperature parameter that decreases exponentially with the number of iterations. Hereafter, we loop back and introduce a new transcriptional modification.

Appendix C Objective Functions for Design

We aimed to rearrange genes (refactorization) within the genome of *E. coli* such that the information content of the distribution of genes in operons could be increased. We hypothesized that this would produce a genome with fewer operons but retaining the entire original set of genes. Therefore, we considered a measure based on Shannon entropy [15] as the first objective function. This measure is computed from the distribution of genes in the operons as $S_{mod} = 1 - \sum_{op}^{N_{op}} k_{op} log_{N_g} k_{op}^{-1}$, where $k_{op} = N_g^{op}/N_g$. N_g^{op} represents the number of genes in the operon *op*, N_g is the number of non-constitutive genes in the wild-type genome, and N_{op} is the updated number of operons contained in the designed genome. Genes initially controlled by constitutive promoters were not involved in the optimization because we assumed that unregulated genes with similar basal expression levels could be grouped into operons controlled by constitutive promoters that provide similar expression levels regardless of the environment. By defining the logarithm base as N_g, we ensured that S_{mod} ranges from 0 to 1, thereby obtaining null modularity for the wild-type genome. We assumed in our model that the sizes of all operons in the wild-type genome are equal to one because genes that are known to be controlled in the same operon did not share the same experimental interactions with TFs collected in RegulonDB [14] or inferred by the *InferGene* algorithm [2]. Thus, precision and recall in the inference of the TRN were maximized. The second objective function was defined as the distance from the wild-type gene expression profile to the predicted profile under various environmental conditions. This similarity was measured as the Pearson correlation coefficient (ρ) obtained when the predicted expression profiles for a set of extreme environments (N_{env}) were compared to the wild-type expression, $S_{exp} = \left[\prod_{env} \rho\left(y_g^{opt}, y_g^{env}\right)\right]^{\frac{1}{N_{env}}}$, where g denotes genes included in a set of critical genes that guarantee the optimal growth of the cell (*e.g.*, genes encoding enzymatic activity). We defined three sets of critical genes: (i) genes coding for enzymatic activity, (ii) genes related to the stress response, and (iii) all genes. Ultimately, we defined a bi-objective function

based on the weighted sum of both objectives, $S(\lambda) = \lambda S_{exp} + (1 - \lambda)S_{mod}$; thus, selecting a given weighting factor, $\lambda \in [0, 1]$, the bi-objective problem relies on maximizing S by the Monte Carlo Simulated Annealing optimization protocol. We used $\lambda = 0.5$ for the simulations.

Appendix D Genome Optimality Degree in Changing Environments

We assumed that cell fitness could be estimated in terms of the S_{exp} objective function. This allowed the study of genome adaptation under changing environments in one ($\Delta v_{k=i} \neq 0$ and $\Delta v_{k\neq i} = 0$) or multiple ($\Delta v_k \neq 0$) directions [16]. To do this, we defined the optimality degree, $\xi_{\Delta v_k}$, in a target environment characterized by Δv_k^* and different from the optimal environment as the difference between S_{exp} evaluated in an environment containing $\Delta v_k = 0$ (*i.e.*, fitness in the optimal condition) and that evaluated in the target environment containing Δv_k^*. Hence, we distinguished between positive and negative error adaptation corresponding to environmental states where cell fitness achieved sub- or over-optimal growth, respectively.

Appendix E Functional Analysis of Genomes

Genes contained in the operons of all refactored genomes were functionally identified using 184 biological functions in GO [17]. We defined the degree of functional similarity, Φ_{op}, of a given operon, *op*, as the ratio between the maximum number of genes with the same functionality and the operon size. We imposed $\Phi_{op} = 0$ for those operons containing only one gene because more than one gene was needed to assess functional similarity; all operons in the wild-type genome therefore received a score of 0.

Appendix F Topological Properties of Refactored Genomes

The evolved configuration based on interconnected building blocks provided a significant increase in the diameter and characteristic path length of the rewired networks. Similarly, refactored networks tend to lose the hierarchical scale-free system characteristics of the wild-type TRN. Whereas the slope of the log-log regression for the average clustering coefficient with the number of genes with k-connections is close to one for the TRN, it was significantly < 1 for the refactored genomes. Furthermore, the power-law that fits the

incoming ($\gamma_{incoming}$) and outgoing ($\gamma_{outcoming}$) connectivity distributions of the refactored genomes are both smaller than those observed for the wild-type TRN, corroborating the observation that in re-engineered TRN a large number of TFs are responsible for activating different biological modules that emerged spontaneously.

Next, we analyzed the changes in promoter type across the entire genome after refactorization. The number of genes controlled by promoters that interact with only one TF was significantly smaller for the rewired genomes than for the wild-type TRN, and the number of genes controlled by two or more TFs significantly increased. The minimum percentage of operons controlled by a synthetic promoter in the refactored genomes was 17%-20%, depending on the fitness definition (*i.e.*, whether fitness considered only genes coding for enzymes involved in central metabolism or only genes related to stress responses, respectively). Consequently, the minimum percentage of synthetic regulations added was greater than 9.5%.

Appendix G Analysis of Biochemical Adaptation to Varying Environments of the Refactored Genome Sequences

Two sets of environments were simulated to explore single environmental perturbations: (1) a set of 100 random perturbations that varied oxygen availability from a fully anaerobic environment to an environment with a rate that was 4-fold greater than the optimal flux value (75 mmol $g^{-1}h^{-1}$) and (2) a set of 100 perturbations that changed the availability of glucose as the carbon source, ranging from the negative value of the optimal uptake flux to the positive value (*i.e.*, -20 mmol $g^{-1}h^{-1}$ to 20 mmol $g^{-1}h^{-1}$).

We also tested adaptation in refactored genomes by considering the previous three types of selective pressure in the expression score. Interestingly, genomes that incorporated tandem promoters achieved low adaptation errors under single environmental perturbations ($\langle\xi\rangle < 0.021$) and over-optimality was even achieved by genomes designed with selection pressure based on genes with enzymatic activity ($\langle\xi\rangle <$ -0.019) or related to defense processes ($\langle\xi\rangle <$ -0.010). By contrast, genomes that were refactored without the design specification of tandem promoter addition had high error adaptation ($\langle\xi\rangle > 0.762$), except for those refactored considering stress genes. Furthermore, we tested the adaptation of genomes designed under multiple perturbations and concluded that evolved genomes that included tandem promoters exhibited over-optimality independent of the objective

function imposed in the design. By contrast, genomes refactored by only re-organizing wild-type genes had adaptation errors as large as $\left\langle \hat{\xi} \right\rangle = 79.8\%$.

Appendix H Selecting Challenging Environments to Generate Different Degrees of Optimality in the Wild-Type Genome

Five sets of six environments defined by external oxygen flux, carbon source (external glucose flux) and nitrogen source (external NO_3^-) were selected based on the decrease that each caused in the expression score. Specifically, we included environments in each set based on five levels of decreases in S_{exp} that range from 0 to 10%. All sets include the environment associated with the optimal condition, creating different ranges of environmental variability for each set.

References

[1] Isalan, M., Lemerle, C., Michalodimitrakis, K., Beltrao, P., Horn, C., Raineri, E., Garriga-Canut, M., Serrano L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840845.

[2] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

[3] Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A. G., Mackie, A., Paulsen, I., Gunsalus, R. P., and Karp, P. D. (2011). EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.* 39, D583-590.

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

[5] Perkins, T. J., Swain, P. S. (2009). Strategies for cellular decision-making. *Mol. Syst. Biol.* 5, 326.

[6] Koide, T., Pang, W. L., Baliga, N.S. (2009). The role of predictive modeling in rationally re-engineering biological systems. *Nat. Rev. Microbiol.* 7, 297-305.

[7] Carrera, J., Rodrigo G., Jaramillo, A. (2009). Towards the automated engineering of a synthetic genome. *Mol. Biosyst.* 5, 733-743.

[8] Isaacs, F. J., Dwyer, D. J., Collins, J. J. (2006). RNA synthetic biology. *Nat Biotechnol.* 24, 545-54.

[9] Kiel, C., Yus, E., Serrano, L. (2010). Engineering signal transduction pathways. *Cell* 140, 33-47.

[10] Wu, J. I., Lessard, J., Crabtree, J. R. (2009). Understanding the words of chromatin regulation. *Cell* 136, 200-206.

[11] Shearwin, K. E., Callen, B. P., Egan, J. B. (2005). Transcriptional interference-a crash course. *Trends Genet.* 21, 339.

[12] Kashtan, N., Alon, U. (2005). Spontaneous evolution of modularity and networks motifs. *Proc. Natl. Acad. Sci.* 102, 13773-13778.

[13] Kirkpatrick, S., Gelatt, C.D., Vechhi, M.P. (1983). Optimization by simulated annealing. *Science* 220, 671-680.

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

[15] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* 106, 620-630.

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

[17] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25-29.

Part II

Viruses Reprogram the Cellular Chassis of their Host

Introduction

Genomic tools have allowed assessment of gene expression at a genome-wide scale, providing unprecedented views of the host-virus interaction. To make use of all of the information contained in these large data sets, however, it is necessary to use computational and mathematical tools to disentangle the interactions between the molecular components of both biological entities and to identify how these interactions determine the outcome of the infection [1, 2], which is known as the field of genomic systems biology (GSB). GSB is a top-down approach that takes advantage of the recent development of high-throughput experimental techniques for obtaining omic data, and constitutes the antithesis of the reductionist paradigm (with a bottom-up perspective) that has been dominating molecular biology. The GSB approach consists of cycling between the generation of experimental data and modeling by means of reverse-engineering techniques to propose testable hypotheses about biological systems, experimental validation of these hypotheses and quantification of the relevant model parameters, and then using the newly acquired quantitative description to refine the computational model and finally make predictions of the system behavior [3, 4].

To complete its infectious cycle, a few components of a virus, including its nucleic acids and encoded proteins, must establish multiple and complex interactions not only among themselves [5, 6, 7, 8] but also with a myriad of components of the host cell [9, 7, 11]. The outcome of all these interactions is that the plant controls the spread of viral infection or, alternatively, the virus overcomes the host defenses and establishes a productive infection that may or may not be associated with the development of symptoms. Although the GSB approach is being extensively used in the analysis of animal virus interactions (*e.g.* hepatitis C, human immunodeficiency, yellow fever, influenza A and herpesviruses), plant virology has not yet benefited to the same extent, and the most relevant studies in the field generally apply some transcriptomic techniques to produce lists of genes with altered mRNA abundance in infected plants relative to controls. However, these types of studies still produce very useful data and serve to highlight some interesting features. Indeed, recent application of the GSB approach to the analysis of animal-virus interactions has revealed new and interesting insights. For example, thoughtful statistical analyses of expression data have facilitated identification of virusregulated genes (VRGs), some of which encode cellular factors required for completion of the infection cycle, while others are direct targets that the virus manipulates to deactivate the cell defense mechanisms [9, 12, 13, 20]. It has also been observed that VRGs are

preferentially highly connected elements in the host regulatory network [15, 16, 17]. Furthermore, it has been observed that the topological properties of the intraviral interaction network change as a consequence of its integration within the host network [6, 18].

Although some studies have analyzed changes in global profiling resulting from virus infection of natural hosts, such as infection of cassava by African cassava mosaic virus [19] and infection of rice by rice yellow mottle virus [20], *A. thaliana* has been the main model host used in combination with viruses belonging to different taxonomic families. These studies involved cauliflower mosaic caulimovirus (CaMV) [21]; turnip vein clearing (TVCV) [22], oilseed rape mosaic (OMRV) [22] and tobacco mosaic (TMV) tobamoviruses [23, 24]; potato X potexvirus (PVX) [22]; cucumber mosaic cucumovirus (CMV) [22, 25, 26]; turnip mosaic (TuMV) [22, 27], plum pox (PPV) [28] and tobacco etch (TEV) potyviruses [29]; and mung bean yellow mosaic (MYMV) [30] and cabbage leaf curl (CaLCuV) geminiviruses [31]. However, even using the same host species, direct comparisons across experiments are not straightforward because differences in profiling techniques and platforms, plant ecotypes, sampling schemes, inoculation conditions and dosages, and growth environmental variables may all exert unpredictable effects on the expression pattern of multiple genes. Furthermore, differences in statistical normalization methods and analyses also contribute to making comparisons difficult. Whitham et al. [22] carried out the most comprehensive of such studies for five viruses (CMV, ORMV, PVX, TVCV, and TuMV) while keeping constant all other experimental variables and techniques. Some generalities can be drawn from this study that can be extended to most of the other studies cited above, highlighting the fact that different viruses alter common sets of genes or biological functions. On the one hand, approximately one-third of overexpressed VRGs are associated with cell rescue, defense, apoptosis and cell death and aging, including several defense-associated and stress-associated genes. Responses to biotic (viruses, bacteria, or fungi) and abiotic (metal ions, osmosis, oxidation, or temperature) stresses, including systemic acquired resistance 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 [8]. In addition, a variety of heat-shock proteins are also overexpressed 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 muta-

tions produced during error-prone replication [33]. Rebosamos proteins and protein turnover genes are also upregulated. 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 downregulated processes. For example, biosynthesis of lipids is pivotal for cell membrane construction and modification and carbohydrate 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 underexpressed, resulting in chlorosis.

References

[1] Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662-1664.

[2] Barabasi, A.L., Oltvai, Z.N. (2004). Network biology: understanding the cells functional organization. *Nat. Rev. Genet.* 5, 101-113.

[3] Palsson, B.O. (2002). *In silico* biology through omics. Nat. Biotechnol. 20, 649-650.

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

[5] Guo, D., Rajamaki, M.L., Saarma, M., Valkonen, J.P.T. (2001). Towards a protein interaction map of potyviruses: protein interaction matrixes of two potyviruses based on the yeast two-hybrid system. *J. Gen. Virol.* 82, 935-939.

[6] Uetz, P., Dong, Y.A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S.V., Roupelieva, M., Rose, D., Fossum, E., Haas, J. (2006). Herpesviral protein networks and their interaction with the human proteome. *Science* 311, 239-242.

[7] Fossum, E., Friedel, C.C., Rajagopala, S.V., Titz, B., Baiker, A., Schmidt, T., Kraus, T., Stellberger, T., Rutenberg, C., Suthram, S.

Bandyopadhyay, S., Rose, D., von Brunn, A., Uhlmann, M., Zeretzke, C., Dong, Y.A., Boulet, H., Koegl, M., Bailer, S.M., Koszinowski, U., Ideker, T., Uetz, P., Zimmer, R., Haas, J. (2009). Evolutionarily conserved herpesviral protein interaction networks. *PLoS Pathog.* 5:e1000570.

[8] Lin, L., Shi, Y., Luo, Z., Lu, Y., Zheng, H., Yan, F., Chen, J., Chen, J., Adams, M.J., Wu, Y. (2009). Proteinprotein interactions in two potyviruses using the yeast two-hybrid system. *Virus Res.* 142, 36-40.

[9] Whitham, S.A., Wang, Y. (2004). Roles for host factors in plant viral pathogenicity. *Curr. Opin. Plant Biol.* 7, 365-371.

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

[11] Bailer, S.M., Haas, J. (2009). Connecting viral with cellular interactomes. *Curr. Opin. Microbiol.* 12, 453-459.

[12] Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., Elledge, S.J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921-926.

[13] Krishnan, M.N., Ng, A., Sukumaran, B., Gilfoy, F.D., Uchil, P.D., Sultana, H., Brass, A.L., Adametz, R., Tsui, M., Qian, F. Montgomery, R.R., Lev, S., Mason, P.W., Koski, R.A., Elledge, S.J., Xavier, R.J., Agaisse, H., Fikrig, E. (2008) RNA interference screen for human genes associated with West Nile virus infection. *Nature* 455, 242-245.

[14] Bushman, F.D., Malani, N., Fernandes, J., DOrso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., Konig, R., Bandyopadhyay, S., Ideker, T., Goff, S.P., Krogan, N.J., Frankel, A.D., Young, J.A.T., Chanda, S.K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 5:e1000437.

[15] Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., Ewence, A.E., Li, N., Hirozane-Kishikawa, T., Hill, DE., Vidal, M., Kieff, E., Johannsen, E. (2007). EpsteinBarr virus and virus human protein interaction maps. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7606-7611.

[16] Dyer, M.D., Murali, T.M., Sobral, B.W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.* 4:e32.

[17] De Chassey, B., Navratil, V., Tafforeau, L., Hiet, M.S., Aublin-Gex, A., Agaugue, S., Meiffren, G., Pradezynski, F., Faria, B.F., Chantier, T., Le Breton, M., Pellet, J., Davoust, N., Mangeot, P.E., Chaboud, A., Penin, F., Jacob, Y., Vidalain, P.O., Vidal, M., Andre, P., Rabourdin-Combe, C., Lotteau, V. (2008). Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230.

[18] MacPherson, J.I., Dikerson, J.E., Pinney, J.W., Robertson, D.L. (2010). Patterns of HIV-1 protein interaction identify perturbed hostcellular subsystems. *PLoS Comp. Biol.* 6:e1000863.

[19] Fregene, M., Matsumura, H., Akano, A., Dixon, A., Terauchi, R. (2004). Serial analysis of gene expression (SAGE) of hostplant resistance to the cassava mosaic disease (CMD). *Plant Mol. Biol.* 56, 563-571.

[20] Ventelon-Debout, M., Delalande, F., Brizard, J.P., Diemer, H., Van Dorsselaer, A., Brugidou, C. (2004). Proteome analysis of cultivar-specific deregulations of *Oryza sativa* indica and *O. sativa* japonica cellular suspensions undergoing rice yellow mottle virus infection. *Proteomics* 4, 216-225.

[21] Geri, C., Cecchini, E., Giannakou, M.E., Covey, S.N., Milner, J.J. (1999). Altered patterns of gene expression in *Arabidopsis* elicited by cauliflower osaic virus (CaMV) infection and by CaMV gene VI transgene. *Mol. Plant-Microbe Interact.* 12, 377-384.

[22] Whitham, S.A., Quan, S., Chang, H.S., Cooper, B., Estes, B., Zhu, T., Wang, X., Hou, Y.M. (2003). Diverse RNA viruses elicit the expression of common sets of genes in susceptible *Arabidopsis thaliana* plants. *Plant J.* 33, 271-283.

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

[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] Ishihara, T., Sakurai, N., Sekine, K.T., 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.

[26] Marathe, R., Guan, Z., Anandalakshmi, R., Zhao, H., Dinesh-Kumar, S.P. (2004). Study of *Arabidopsis thaliana* resistome in response to cucumber mosaic virus infection using whole genome microarray. *Plant Mol. Biol.* 55, 501-520.

[27] Yang, C., Guo, R., Jie, F., Nettleton, D., Peng, J., Carr, T., Yeakely, J.M., Fan, J.B., Whitham, S.A. (2007). Spatial analysis of *Arabidopsis thaliana* gene expression in response to turnip mosaic virus infection. *Mol. Plant-Microb Interact.* 20, 358-370.

[28] Babu, M., Griffiths, J.S., Huang, T.S., Wang, A. (2008). Altered gene expression changes in *Arabidopsis* leaf tissues and protoplasts in response to plum pox virus infection. *BMC Genomics* 9, 325.

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

[30] Trinks, D., Rajeswaran, R., Shivaprasad, P.V., Akbergenov, R., Oakeley, E.J., Veluthambi, K., Hohn, T., Pooggin, M.M. (2005). Suppression of RNA silencing by a geminivirus nuclear protein, AC2, correlates with transactivation of host genes. *J. Virol.* 79, 2517-2527.

[31] Ascencio-Ibanez, J.T., Sozzani, R., Lee, T.J., Chu, T.M., Wolfinger, R.D., 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.

[32] Wise, R.P., Moscou, M.J., Bogdanove, A.J., Whitham, S.A. (2007). Transcript profiling in hostpathogen interactions. *Annu. Rev. Phytopathol.* 45, 329-369.

[33] 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.

Chapter 5

Reverse-Engineering of the *Arabidopsis thaliana* Transcription under Changing Environments

Living organisms have evolved molecular circuitries with the aim of promoting their own development under dynamically changing environments. In particular, plants are not able to evade those changes and have had to evolve robust methods to cope with environmental stress and recovery mechanisms. Genomic sequences specify the context-dependent gene expression programs to render cells, tissues, organs and, finally, organisms. Then, at any moment during cell cycle and at each stage of an organisms development, and in response to environmental conditions, each cell is the product of specific and well defined programs involving the coordinated transcription of thousands of genes. Thus, the elucidation of such programs by means of the regulatory interactions is pivotal for the understanding of how organisms have evolved and what environments may have conditioned evolutionary trajectories the most. However, understanding how this highly tuned process is achieved is still beyond our knowledge for most organisms, and the surface of the problem is only being scratched for a handful of model organisms such as the bacterium *E. coli* [1], the yeast *S. cerevisiae* [2], the nematode *Caenorhabditis elegans* [3], the plant *A. thaliana* [4, 5], or to a lesser extent for humans [6].

Meta-analyses of microarray data collections may now be used to construct biological networks that systematically categorize all molecules and describe their functions and interactions. Networks can integrate biologi-

cal functions of cells, organs, and organisms. During recent years, there has been a tremendous effort in the development and improvement of techniques to infer gene connectivity. Clustering approaches [7, 8, 9, 10, 11] and information theory methods [12, 13, 14, 15, 16] have been used to infer regulatory networks. Bayesian methods [17, 18, 19, 20] can give accurate networks with low coverage but at a high computational cost.

The analysis of the expression of *A. thaliana* transcriptome offers the potential to identify prevailing cellular processes, to associate genes with particular biological functions, and to assign otherwise unknown genes to biological responses to which they are correlated. Previous attempts to model *A. thaliana* gene network used methods such as fuzzy k-means clustering [21], graphical Gaussian models [4], and Markov chain graph clustering [5, 15]. The inconvenience of the first approach is that clustering describes genes based on a characteristic property common to all genes but it is difficult to deduce a pathway structure from this property alone, because pathways would have to be concerned with co-expression features that transcend such cluster structure. The second approach assumes that the number of microarray slides should be much larger than the number of genes analyzed or approximations must be taken (*e.g.* empirical Bayes with bootstrap resampling or shrinkage approaches). The last approach is still based on Persons correlations and therefore, strongly sensitive to outliers and to violations to the implicit assumption of linear relationships among genes. In this article, we present a predictable genome model from a regulatory scaffold inferred by using probabilistic methods [15] and estimate the corresponding kinetic parameters using linear regression [22, 23, 24, 28]. We analyze the topological properties and predictive power of the inferred regulatory model. We evaluate the performance of the network by predicting already known transcriptional regulations and assess the functional relevance and reproducibility of the co-expression patterns detected. Finally, we discuss the evolutionary implications of the transcriptional control in plants [26].

5.1 Genome-Wide Transcriptional Control in *A. thaliana*

In the present Chapter, we have applied the previous developed inference methodology presented in the Chapter3, *InferGene* [28], to obtain a gene regulatory model, suitable for analyzing optimality and allowing studying the transcriptional control response under changing environments in *A. thaliana*. For that, we have considered the Affymetrixs chip for the *A.*

thaliana genome, from which we selected 22,094 non-redundant genes, of which about 1187 are putative transcription factors (TFs) (see Appendix A). The data used for the inference procedure were a compendium of 1436 Affymetrixs microarray hybridization experiments publicly available at the TAIR website and that were normalized using RMA [27]. Here we used the whole expression set (1436 experiments) to construct the model.

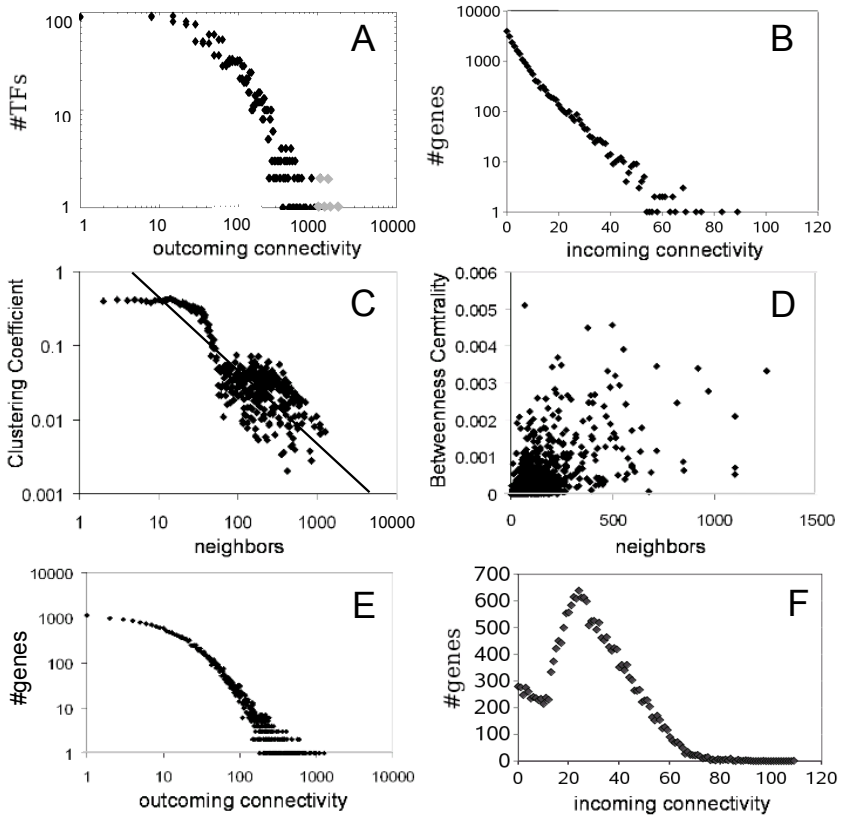

Figure 5.1: Analyses of the regulatory network of *A. thaliana*. Distributions for the transcriptional network of (A) outcoming connectivity showing the master regulators in a different color, (B) incoming connectivity, (C) clustering coefficient, and (D) betweenness centrality. Distributions for the non-transcriptional network of (E) outcoming connectivity and (F) incoming connectivity.

Three types of efficiencies, precision (P), sensitivity (S) and absolute efficiency (F), have been computed to assess the ability of the above inferred network to predict the 448 experimentally validated transcriptional regulations collected in the AtRegNet database. P is the fraction of predicted interactions that are correct $P = TP/(TP + FP)$ and S the fraction of all known interactions that are discovered by the model $S = TP/(TP + FN)$, where TP is the number of true positives, FN the number of false nega-

tives and FP the number of false positives. F thus represents the absolute efficiency and it is computed as $F = 2PS/(P + S)$ which is the harmonic mean of precision and sensitivity. Indeed, precision and sensitivity are necessarily negatively correlated performance statistics, and these two values were set up so they maximize global performance (F) by selecting values $>$ 5 for the z-score used as threshold to predict the transcriptional regulations. P, S and F were computed as a function of the z threshold. Sensitivity is maximized $S = 100\%$ for $z = 0$ (*i.e.*, high number of regulations but very low confidence) while precision is maximized $P = 100\%$ for $z = 11$ (*i.e.*, high confidence but very low number of regulations). The optimum value is reached for $z = 5$, a value for which $F = 26\%$ ($P = 40\%$ and $S = 20\%$). In a recent study, a smaller network topology has been proposed for *A. thaliana* [4]. This network contains 18.625 regulations and an $F = 3.7\%$ ($P = 88\%$ but $S = 1.8\%$), relative to the AtRegNet reference dataset.

InferGene predicts that more than half of the genes are controlled by constitutive promoters (17.89%) or by promoters regulated by less than three TFs. Also, from a purely topological perspective, the inferred transcriptional network of *A. thaliana* is weakly connected directed, containing 18.169 genes connected, while the size of the largest strongly connected component only contains 730 nodes, all of which are TFs. In addition, it has a high density (0.078%), understanding this parameter as the normalized average connectivity of a gene in the network, in comparison to values reported in similar studies done for other organisms. For example, Lee et al. [2] suggested a network density of 0.0027% for *S. cerevisiae*, while we previously reported a value of 0.036% for the network inferred for *E. coli* [28]. The characteristic path length [28] of the network follows a Gaussian distribution with an average value of 5.065 edges and, specifically, the distance between two genes for which a path exists ranges from 1 to 13 edges. In a previous study, we estimated that the characteristic path length for *E. coli* network was 1 [28], much smaller than for the case of *A. thaliana*. Furthermore, the *E. coli* inferred network, did not contain any strongly connected components and its largest weakly directed subnetwork only contained 4 TFs. Other relevant statistical properties of networks are the stress distribution, *i.e.*, the number of paths in which a gene is involved, and the betweenness centrality distribution (Figure 5.1D), *i.e.*, the number of shortest pathways in which a particular gene is involved. Both distributions are highly asymmetrical, with many nodes having a low betweenness centrality and a few cases with high values (Figure 5.1D) and with the number of shortest paths per gene smoothly increasing until reaching a maximum of approximately 105 short paths per

gene and then followed by a drastic drop, with very few genes (around 5) having 107 short paths. Ten genes (*At1g32330, At4g26930, At1g24110, At4g24490, At2g36590, At1g01030, At1g76900, At2g19050, At2g03840, and At3g19870*) are connected among them but remain isolated from the rest of the main network, the number of shortest paths for these genes ranges from 1 to 3. All these genes but the last one are involved in several and apparently loosely related GO functional categories that include regulation of transcription, transportation and signal transduction, and development and senescence.

Next, we sought to explore whether the inferred regulatory network has scale-free properties. It has been suggested that the distribution of outcoming connections should belong to the class of scale-free small-world networks, representing the potential of transcription factors to regulate multiple target genes whereas the distribution of incoming connectivities would be more exponential-like because the regulation by multiple TFs should be less common that the regulation of several targets by a given TF [28]. Figure 5.1A shows the distribution of outgoing connectivities per TF, whereas Figure 5.1B shows the same distribution but only for incoming connectivities per gene. As expected, the outcoming connectivity is best fitted by a truncated power-law (*i.e.*, the Weibull distribution) with exponent $\gamma = 0.902$ and cut-off $k_c = 99.093$ ($R^2 = 0.949$; Akaikes weight over a set of 10 competing models $> 99.99\%$). This distribution indicates that outcoming connectivities has a scale-free behavior in the range $1k < k_c$ but deviates from this for connectivities over the cut off. According to Barabasi and Oltvai [31], scale free properties arise when hub genes are related in a hierarchical way, with the hub receiving most links being connected to a small fraction of all nodes. In the case of incoming connectivities, the model that better describes the data is a restricted exponential, the half-Normal distribution ($R^2 = 0.983$; Akaikes weight $> 99.99\%$). Taken together, these two observations suggest that *A. thaliana* transcriptional network contains a few highly connected regulators that play a central role in mediating interactions among a large number of less connected genes. Notice that there are 88.4% TFs regulating more than 10 genes, 36.3% regulating more than 100 genes and just 2.6% that control over 500 genes. For the sake of comparison, it is worth mentioning that in the case of *S. cerevisiae* the critical exponents estimated for the outcoming connectivity distribution ($\gamma = 0.96$ [2, 31]) is quite similar to the one here reported. However, the estimate obtained for *E. coli* was smaller ($\gamma = 0.87$), a result that suggests that hubs are more important in bacteria than in the two eukaryotes [30]. We have validated the set of predicted

targets for the 25% most highly connected TFs using AtRegNet, recovering 80% of known interactions for the regulatory model and up to 85% for the effective model (*i.e.*, the one containing both gene-to-gene and gene-to-TF interactions). Figure 5.1C shows that the scaling of the average clustering coefficient with the number of genes with k-connections is approximately lineal in a log-log scale in the range (1-10000) of neighbors with slope -1.05 ($R^2 = 0.850$). Barabsi and Oltvai [30] and Ravasz and Barabsi [32] have suggested that whenever clustering scales with the number of nodes with slope -1, as it is our case, it has to be taken as a strong indication of hierarchical modularity, *i.e.* genes cluster in higher-order units of different modularity, a finding that has been suggested as general for system-level cellular organization in plants [33]. Similarly, when the effective model is analyzed, it shows similar results than for the regulatory model. The outcoming connectivities per gene follows a truncated power law with scale-free behavior up to $k_c = 21.341$ connections per gene and with an exponent $\gamma = 0.765$ ($R^2 = 0.998$, Akaikes weight $> 99.99\%$) (Figure 5.1E). Figure 5.1F shows that the incoming connectivity per gene does not present scale-free properties as it fits to a Normal distribution ($R^2 = 0.998$, Akaikes weight $> 99.99\%$).

The environment significantly influences the dynamic expression and assembly of all components encoded in the *A. thaliana* genome into functional biological subnetworks. We have computed the clustering coefficient for all subnetworks with the largest normalized index of connectivity between genes involved in the subnetwork. Interestingly, four of these highly connected subnetworks are involved in response to external influences as, for example response to pathogens and other processes related with abiotic stresses (heat, salinity, light, redox). For the sake of illustration, Figure 5.2 shows the inferred subnetworks for three abiotic and three biotic responses. Particularly, we have made a comprehensive analysis for the subnetwork of the systemic acquired resistance (Figure 5.2D) and found that the fraction of predicted interactions is $P = 33\%$. Not surprisingly, all genes involved in that subnetwork appear associated with GO categories related to response to stress, like defense to pathogens, response to other organisms such as fungus, bacterium and insects, and response to cold.

5.2 Transcriptomic Profile Prediction

The basic premise of our approach was to use transcriptomic data from multiple perturbation experiments (either genetic or environmental) and quantitatively measure steady-state RNA concentrations to assimilate these

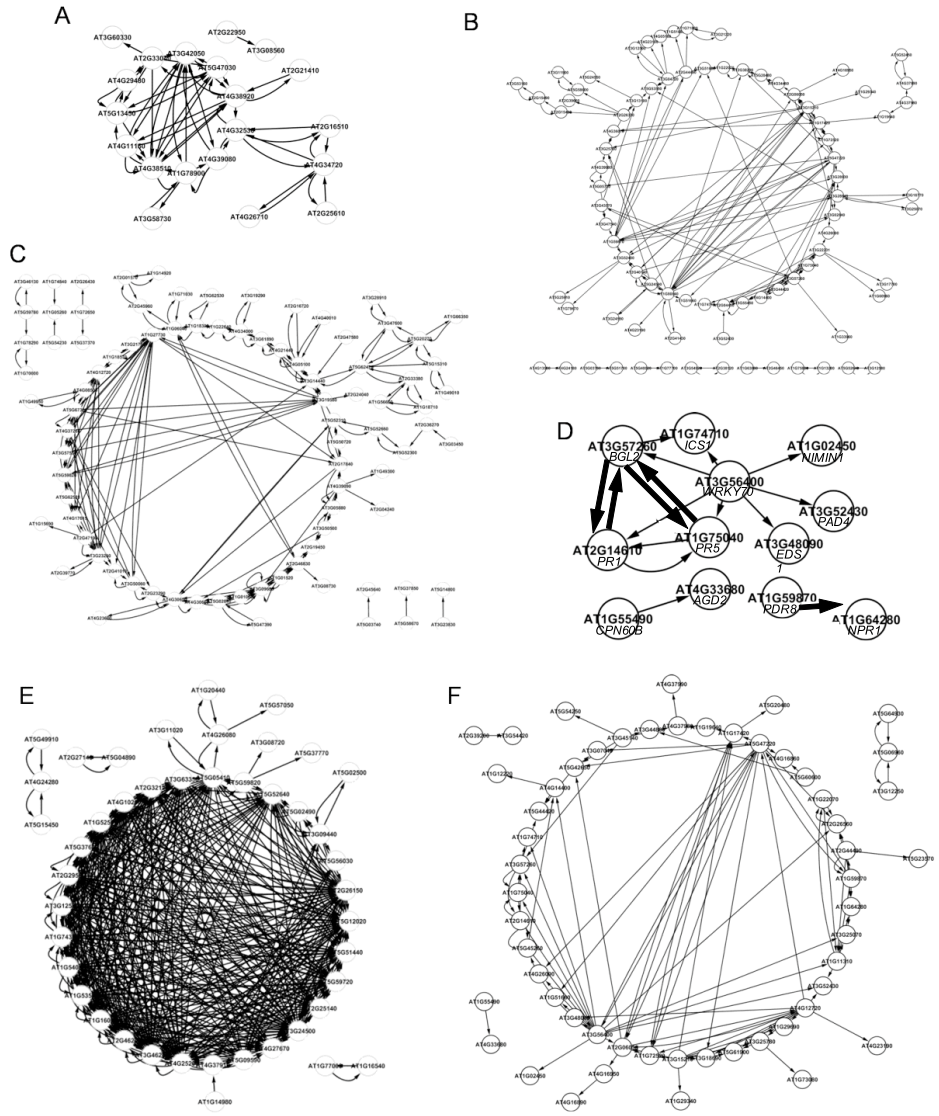

Figure 5.2: Transcriptional subnetworks with high clustering coefficients corresponding to the following GO pathways: (A) auxin metabolic process, (B) response to other organism, (C) response to heat, (D) systemic acquired resistance (experimentally verified regulations are represented with thick edges), (E) response to salt stress, and (F) immune response.

expression profiles into a network model that can recapitulate all observations. Now, we develop a second model (test model) excluding the 10% of experiments to quantify the prediction power. The data set was randomly split into two subsets. The first larger subset contained 1292 experiments and was used as training set for inferring a transcription network containing 128,422 regulatory interactions. The second, smaller, subset contained 144 array experiments and was used for validation purposes.

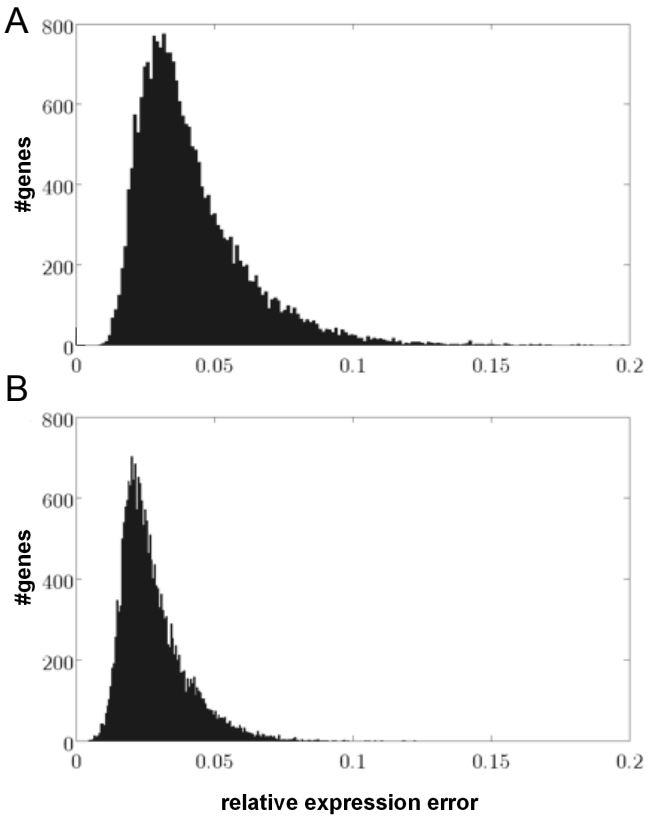

Figure 5.3: Histogram of the relative gene expression error in (A) the transcriptional test model (with an average error of 0.0402) and in (B) the effective model (with an average error of 0.0280). Errors were obtained from the comparison of the predicted model obtained from the training dataset and the experimental determinations contained in the random tester dataset.

As a first measure of the performance of our test model network in predicting responses to stresses, we have used it along with the expression levels of all the TFs for each experimental condition, c, to predict global expression proles. Then, the predicted expression values for each of the 22,094 individual genes included in the Affymetrix array, were compared with the

corresponding empirical measurements, y_{gc}, using the deviation statistic, $\Delta_g = \frac{1}{N_g}\sum_c^{N_c}\frac{y_{gc}-\hat{y}_{gc}}{y_{gc}}$, where $N_c = 144$ is the number of microarray experiments included in the random tester dataset. Figure 5.3A shows the distribution of Δ_g for all genes included in the predicted *A. thaliana* transcriptional network. The distribution of errors has a median value of 3.66% and is significantly asymmetrical (skewness 1.709 ± 0.017, $p < 0.0001$), with most genes having a relatively low error but with some genes whose expression is estimated with errors $> 10\%$ and even in a few instances $> 16\%$. How does this predictive performance compare to that obtained for other organisms, as for example *E. coli*? In a previous study, we constructed a transcriptional network containing 4345 genes and 328 TFs from *E. coli* [25] using a dataset containing 189 experimental conditions. For this network, the average error over the training set was similar (3.68%) to the values reported above but with the error distribution being even more asymmetrical (skewness 2.314 ± 0.017, $p < 0.0001$). The average error over the *E. coli* test set (4.80%) was larger. Figure 5.3B shows the distribution of Δ_g for gene-to-gene and gene-TFs interactions which is also significantly asymmetrical (skewness 1.455 ± 0.017, $p < 0.001$), although in this case the median error is reduced to 2.71% and in all cases the error was $< 9\%$. Both distributions significantly differ in shape (Kolmogorov-Smirnov test $p < 0.001$) and location (Mann-Whitney test $p < 0.001$), with the latter being narrower and centered around a lower expression error. One may ask whether the predictability of our model was driven by TFs and not by non-TF genes. To test this possibility we proceeded as follows. First, we selected a random set of 1187 non-TFs genes and used them to construct the corresponding pseudo-transcriptional network. Then we evaluated its performance as described above. The level of precision reached was undistinguishable from the previous one, with the distribution of relative expression error obtained fully overlapping with the one shown in Figure 5.3b (data not shown). Therefore, we conclude from this analysis that TFs do not have stronger predictive power than the rest of genes. This could be rationalized because, in terms of mathematical equations, genes that are coexpressed with the TFs have a priori equal chances to work as regulatory elements. On the other hand, we have also constructed an effective model excluding the TFs from the set of predictors and observed that the relative expression error decreased proportionally to the number of excluded TFs.

For illustrative purposes, Figure 5.4 shows the expression predicted for five best cases for the transcriptional network, each dot in the scatter plots representing a value obtained on a different hybridization experiment. The

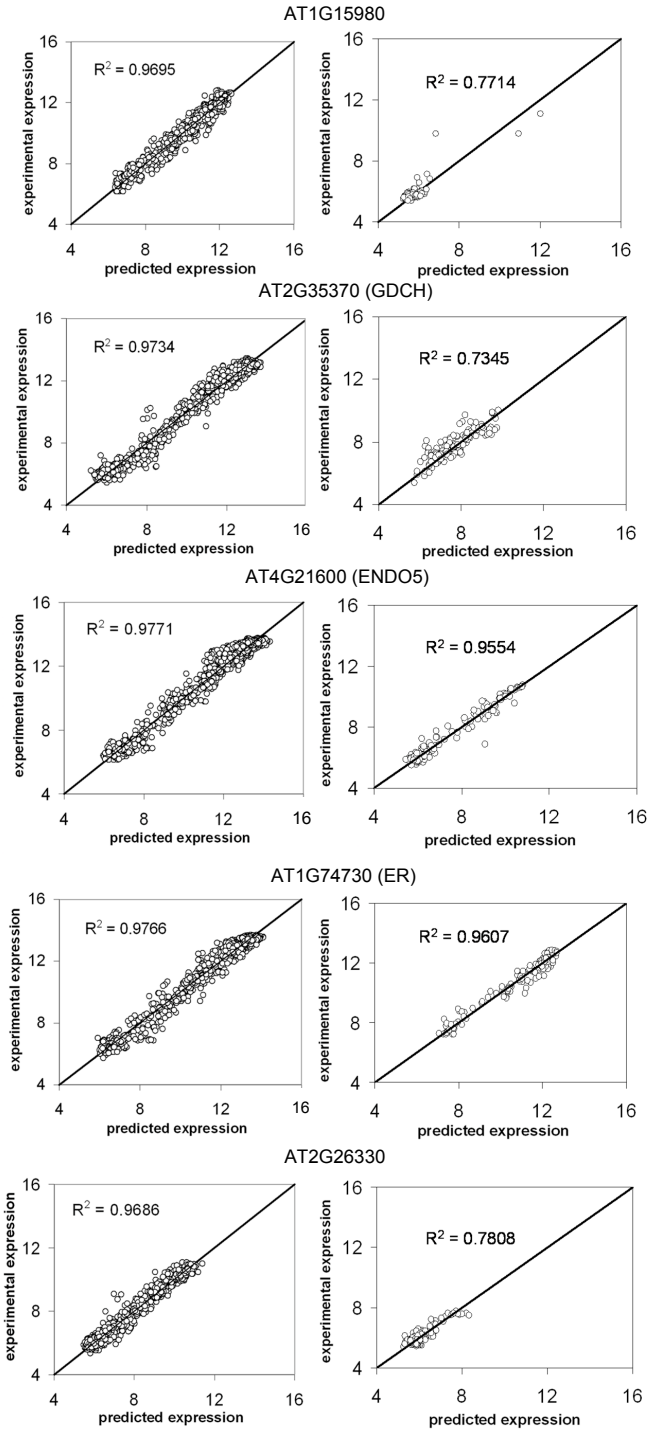

Figure 5.4: Predictive power on gene expression of the transcriptional model of *A. thaliana* inferred from the whole data set (1436 conditions) and the test model from 1292 microarray experiments, used as training set. The left column shows the regression coefficient (R^2) between the model and experimental profiles across the whole data set for the five best predicted genes. The right column shows R^2 between the test model and the 144 experimental profiles used as tester set for the same five genes. In either case, correlation coefficients were highly significant.

left column shows the prediction obtained using the whole dataset (1436 experiments) both as training and as tester sets, whereas the right column shows, for the same five genes, the correlation between the prediction obtained from the test model (inferred from the reduced training set of 1292 experiments) and the observations contained in the tester set (144 experiments). It is remarkable that the quality of the prediction does not change by using a reduced training set, in good agreement with the results reported for *E. coli* [28].

5.3 Selection of Optimality in Changing Environments

Organisms have a high capacity for adjusting their metabolism in response to environmental changes, food availability, and developmental state [34]. On the one hand, we have detected that GO pathways related with response to diverse environmental (*e.g.*, defense against diverse pathogens, response to radiation, temperature, light intensity, or osmotic stress) and internal (development, secondary metabolism, porphyrin biosynthesis, etc) stimuli consists of sets of genes with high incoming connectivity, that is, genes regulated by many different TFs. Therefore, this high degree of interconnection among different stimulus-related pathways allows the cell to rapidly adjust its homeostasis in response to changing environments. On the other hand, functional GO pathways associated to biological functions with expression unaffected by external stresses (*e.g.*, glycerophospholipid and glycerophospholipid metabolic process, sulfur amino acid biosynthetic process, indole and derivative metabolic process, membrane lipid biosynthetic process, sulfured compounds biosynthetic, and Golgi vesicle transport), have low incoming connectivities. Notice that some GO pathways indirectly related with external stresses such as for instance indole derivatives, like camalexin, (involved in response to the bacterium *Pseudomonas syringae*) or lipid biosynthesis pathways (playing a role in defense) were not scored with high levels of connectivity and high number of FFLs involved in the GO pathway. Furthermore, the predicted master regulators of *A. thaliana* belong to biological functions related to transcription and regulation of cellular metabolic processes (containing 812 TFs each) or RNA metabolic processes (536 TFs) that are stimulated by environmental and developmental stresses. After all, the regulatory network of *A. thaliana* governs the intra-cellular processes and modulates and determines the expression of the different programs encoded in the genome.

Networks can be decomposed into subnetworks which can be seen as their building blocks. These building blocks, generally known as motifs, are defined in terms of their frequency and are typically constituted by several promoter regions of genes expressing TFs which regulate each other in a number of well known patterns (*e.g.*, bifans, forward, feedforward, or negative feedback loops) [35]. Certain regulatory network motifs have been described as conferring robustness to perturbations in individual edges, being the coherent feedforward loop (FFL) the prototypical example of such a robustness-conferring motif [36, 39]. Therefore, we sought to characterize our inferred complex network in terms of the presence and abundance of regulatory networks motifs. Some of the overrepresented motifs are shown in Figure 5.5. The third most abundant motif found is, precisely, the FFL (third raw in Figure 5.5A). Indeed, FFL is overrepresented among GO categories involved in stress response compared to non-stress response categories (Fishers exact test, $p < 0.001$).

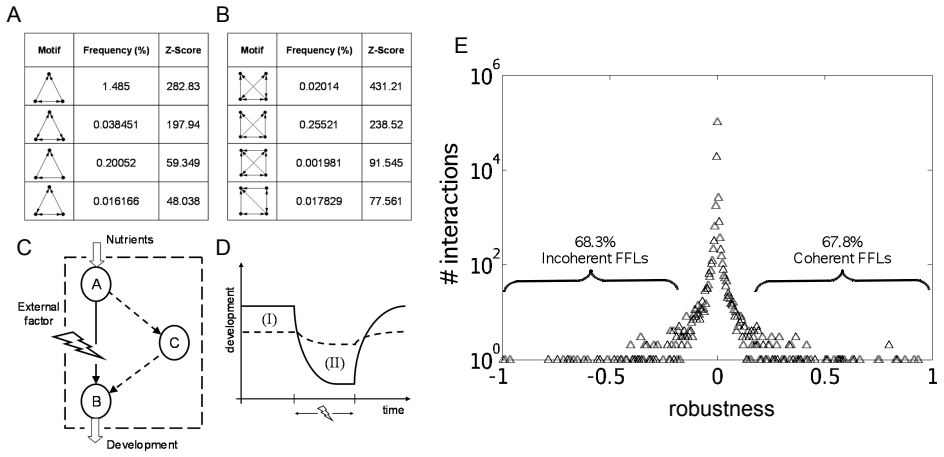

Figure 5.5: Network motifs of three (A) and four (B) genes found in the transcriptional network of *A. thaliana*. Here we plot the most statistically significant ones. We show a motif significantly overrepresented, feed-forward loop (C), where an external factor could inhibit the regulation of the gene A to the gene B, but this structure provides an indirect regulation by means of the gene C. On other hand, we show in (D) the evolution of the qualitative development of a plant with motifs (dashed line) and without motifs (solid line) under changing environments. We note that it exists an evolutionary optimization to include topologic units such as feed-forward loop providing robustness under external factors despite decreasing systems fitness (see area I and II) due to an exceed of gene expression of those genes providing indirect interactions. Panel (E) shows the distribution of normalized robustness coefficients (δ^*) computed for all interactions between TFs and genes.

Next, we sought to test whether the presence of FFL indeed contributed to increment the robustness of the gene expression of the involved genes. To do so, we have computed a score, $\rho*$, quantifying the robustness of gene expression for all predicted TF-gene interactions involving three nodes (Figure 5.5C). Figure 5.5E shows the distribution of the robustness score computed from the inferred regulatory network. Although it may not result apparent after the visual inspection of Figure 5.5E, the distribution is asymmetrical (skewness 1.881 ± 0.007, $p < 0.001$) and strongly leptokurtic (1294.051 ± 0.014, $p < 0.001$), suggesting that there are more data points in the tails than close to the mean. The data points in the upper tail correspond to the more robust interactions and, if coherent FFLs are involved in such type of interactions, they may be over-represented on this tail. This is, indeed, the case. If we look at the upper 1% values, 90.7% of them correspond to coherent FFL. By contrast, if we look at the 1% interactions around the mean value, only 5.7% correspond to FFL. Interestingly, 90.2% of motifs within the 1% lower tail of the distribution correspond to incoherent FFLs.

5.4 Conclusions

We have discussed a reverse-engineered model of the *A. thaliana* cell's gene regulatory network aimed to future research projects focused on distinguishing, *e.g.*, the molecular targets of a plant virus from the hundreds to thousands of additional gene products that may modify levels of gene expression as a side-effect. We have used a recent methodology to infer the global topology of transcription regulation from gene expression data to produce a kinetic model able to predict the alterations in gene expression in plants subjected to different external stimulus. Moreover, we have concluded that the *A. thaliana* inferred transcriptional network presents a hierarchical scale-free architecture where biological functions cluster in modules. We have identified biological functions which are highly controlled by predicted master regulators that could change their operating points in response to dynamic external factors to produce a consistent and robustness response upon different stresses at the expense of decreasing the cellular replication rate. We have successfully applied the inferred model to predict the transcriptomic response of *A. thaliana* under all experimental conditions included in the whole dataset, and also applied the test model to predict the response in the reduced tester set, producing errors of 2 - 10% relative to the experimental value (averaging across all test experiments). Thus, we believe this modeling-validation approach constitutes an important step towards the understanding of the large-scale mode of organisms action to cope with a gen-

erally changing environment. The network model suggests that *A. thaliana* promoters are regulated by multiple TFs, a feature which has been shown to be characteristic of eukaryotes gene regulation [2].

We have discussed a first gene regulatory model based on a transcriptional layer and a second model that embraces the first one by including gene-to-gene interactions that provides an even more accurate prediction of gene expression. Future works would consider just interactions between tissue-specific genes. Next, we have also quantified the presence of network motifs and found that FFL are overwhelmingly common, thus supporting the above notion that robustness against perturbation has been a major driving force during the evolution of plant lineages. Furthermore, we have confirmed that coherent FFL are overwhelmingly over-represented among interactions that are robust against the knock-out of the regulatory TF (Figure 5.5E), while incoherent FFL are so among the most sensitive interactions. Figure 5.5C illustrates a possible mechanism by which FFL would confer robustness. Imagine that the B product is relevant for cell survival. At the one side, deriving regulation flow throughout C is costly because it implies producing a redundant element. However, if perturbations disrupt the direct edge between A and B, the existence of C still allows the cell to obtain the precious B without incurring into a major penalty (Figure 5.5D). Whether a given regulatory network may be selected to contain this sort of regulatory elements depends on the balance between the fitness costs and benefits associated with redundancy [40, 41]. The fact that *A. thaliana* network topology seems to be rich in these transcriptional regulatory elements suggests that it has been evolutionary optimized to allow rapid responses to changes in the external conditions while maintaining cellular homeostasis, and hence maximizing fitness.

The reconstruction of genome-scale regulatory models constitutes a major step towards the understanding of the cellular behavior, but it also is for Synthetic Biology, where predictive models can be applied to engineer synthetic systems for biotechnological applications. Hence, *InferGene* [28] provides a mechanism to predict the changes in the biological processes when perturbing the cell in order to identify the effects of drugs, virus infection and herbicides action in plant interactomes. It may facilitate optimization of cellular processes for biotechnology applications that utilize the complex regulatory properties of genetic networks.

Appendix A Microarray Data

Steady-state mRNA expression profiles derived from transcriptional perturbations collected in the TAIR website [42] have been used in this study. We found 1187 TFs by looking for the motif transcription factor in the functionally annotated A. thaliana genome from TAIR (version 7). The dataset contains pre-processed expression data from 1436 hybridization experiments using the 22,810 probe sets spotted on Affymetrixs GeneChip *Arabidopsis* ATH1 Genome Array [43]. For this study, we consider 22,094 genes. The arrays were obtained from NASCArrays [44] and AtGenExpress [45]. Data were normalized using the robust multi-array average method [27].

Appendix B Inference Procedure

The inference procedure consisted of two nested steps. In the first step, the global network connectivity was inferred using the *InferGene* algorithm [28]. This method uses mutual information (MI) with a local significance (z computation) to obtain the genome regulations [15]. Hence, its potential interaction between a regulator and a gene is z-scored, constituting an estimator of the likelihood of MI. This approach allows eliminating some false correlations and indirect influences [15]. Subsequently, we selected a z threshold for cut-off. In a second step, multiple regressions were obtained to estimate the kinetic parameters of an ODE-based regulatory model. Multilayer model were constructed to account for different types of regulations between genes and TFs. We have constructed two different models, one for transcription regulations and another to account for effective (transcription and non-transcription) regulations. In case of non-transcriptional interactions, LASSO method was used to avoid over-fitting [46] and the effective interactions between genes giving the non-transcriptional layer were unveiled. For that end, we applied a simple and efficient algorithm based on the Gauss-Seidel method [47] that reduces the number of regulators that exceeded the z-score threshold for a given gene. Note that the method enriches in TFs among the predictors of the target for the 33.21% of non-constitutive genes of *A. thaliana* (*i.e.*, the ratio between the number of TFs selected and the total number of predictors of a given gene above a threshold defined as $1187/22{,}094 = 0.0537$).

Appendix C Model Validation

The performance of the inferred model topology was evaluated using a reference network defined by taking those genes with known transcriptional regulation. For that, the AtRegNet platform [48] linking cis-regulatory elements and TFs into regulatory networks was used. Only those interactions among genes included in that reference set were evaluated. The fraction of interactions that were correctly predicted by the model (precision, P) and the fraction of all known interactions that were discovered by the model (sensitivity, S) were used to compute a performance statistic defined as $F = 2PS/(P + S)$ [16]. We have to notice that the number of transcriptional regulations experimentally confirmed and compiled in AtRegNet is quite limited, containing only 448 reported interactions between TFs and genes. Therefore it is difficult to obtain an accurate value for the performance of the model.

To validate the predictive power of the methodology, we constructed two transcription models. The first one was obtained by using the 1436 microarrays for training. For the second model (test model), of all these microarrays, 1292 were used as training set (90%) whereas 144 randomly chosen ones (10%) were retained for validation studies.

Appendix D Motif Detection and Analysis

The FANDOM program [49] has been used to detect motifs of 3 and 4 genes in the predicted *A. thaliana* regulatory model. Those motifs statistically significant have $z > 2$. The robustness of gene expression to perturbations in the underlying motifs was evaluated for each interaction as follows. In a scheme as the one illustrated in Figure 5.5C, TF A operates on gene B but also may act upon a second transcription factor C which, itself, may also interact with the promoter region of B activating its expression. For such a system, we define the robustness score to quantify the impact that removing TF A has in the expression of gene B, $\rho_{AB} = \frac{y_B^+ - y_B^-}{\beta_{AB} y_A}$, where y_B^+ represents the measured expression of gene B when gene A is present and y_B^- after it has been removed. The difference in gene expression is normalized by the expression level of the TF A, y_A, and the strength of its regulation, β_{AB}, on the expression of B. If A is removed ($y_A \to 0$) and no alternative pathway exist, then $\rho_{AB} \to 1$. However, if C exists, as it is the case for the FFL, then $\rho_{AB} \neq 1$, with its sign being determined by $y_B^+ - y_B^-$ and the sign of β_{AB}. This score is unbounded, thus for convenience we further normalized it as $\rho^*_{AB} = (\rho_{AB} - 1)/max_{\forall i,j}(\rho_{ij})$, which is now contained in the interval

$[-1, 1]$. Values of ρ^*_{AB} close to 1 would correspond to maximally robust motifs, whereas values close to zero correspond to motifs not contributing to the robustness of the network. Values close to -1 correspond to incoherent motifs, that is, gene circuits implementing antagonistic regulations [33].

References

[1] Gutirrez-Ros, R.M., Rosenblueth, D.A., Loza, J.A., Huerta, A.M., Glasner, J.D., Blattner, F.R., Collado-Vives, J. (2003). Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* 13, 2435-2443.

[2] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804.

[3] Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., Davidson, G.S. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087-2092.

[4] Ma, S., Gong, Q., Bohnert, H.J. (2007). An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Res.* 17, 1614-1625.

[5] Mentzen, W.I., Wurtele, E.S. (2008). Regulon optimization in *Arabidopsis*. *BMC Plant Biol.* 8, 99.

[6] Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res.* 14, 1085-1094.

[7] Eisen, M., Spellman, P., Brown, P., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863-14868.

[8] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6745-6750.

[9] Ben-Dor, A., Shamir, R., Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.* 6, 281-297.

[10] Dhaeseleer, P., Liang, S., Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707-726.

[11] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat Genetics* 31, 370-377.

[12] Butte, A., Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacic Symp. Biocomput.* 5, 415-426.

[13] Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382-390.

[14] Margollin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., dellaFavera, R., Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7, S7.

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

[16] Meyer, P.E., Kontos, K., Latte, F., Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinf. Syst. Biol.* 2007, 79879.

[17] Husmeier, D. (2003). Sensitivity and specicity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics* 19, 2271-2282.

[18] Yu, J., Smith, V., Wang, P., Hartemink, A., Jarvis, E. (2004). Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 3594-3603.

[19] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C., Ferreira, C.E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Syst. Biol.* 1, 39.

[20] Steinke, F., Seeger, M., Tsuda, K. (2007). Experimental design for efficient identication of gene regulatory networks using sparse bayesian models. *BMC Syst. Biol.* 1, 51.

[21] Ma, S., Bohnert, H.J. (2007). Integration of *Arabidopsis thaliana* stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biol.* 8, R49.

[22] Gardner, T., di Bernardo, D., Lorenz, D., Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression proling. *Science* 301, 102-105.

[23] di Bernardo, D., Thompson, M., Gardner, T., Chobot, S., Eastwood, E., Wojtovich, A., Elliott, S., Schaus, S., Collins, J.J. (2005). Chemogenomic proling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 3, 377-383.

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

[25] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

[26] Bonneau, R. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131, 1354-1365.

[27] Irizarray, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249-264.

[28] Albert, R., Barabasi, A.L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47-97.

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

[30] Barabsi, A.L., Oltvai, Z.N. (2004). Network biology: understanding the cells functional organization. *Nat. Rev. Genet.* 5, 101-113.

[31] Khanin, R., Wit, E. (2006). How scale-free are biological networks. *J. Comp. Biol.* 13, 810-818.

[32] Ravasz, E., Barabsi, A.L. (2003). Hierarchical organization of complex networks. *Phys. Rev. E* 67, 026112.

[33] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabsi, A.L. (2002). Hierarchical organization of modulatiry in metabolic networks. *Science* 297, 1551-1555.

[34] Oltvai, Z.N., Barabsi, A.L. (2002). Lifes complexity pyramid. *Science* 298, 763-764.

[35] Kashtan, N., Itzkovitz, S., Milo, R., Alon, U. (2004). Topological generalizations of network motifs. *Phys. Rev. E* 70, 031909.

[36] Mangan, S., Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11980-11985.

[37] Mangan, S., Zalsaver, A., Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* 334, 197-204.

[38] Hayot, F., Jayaprakash, C. (2005). A feedfoward loop motif in transcriptional regulation: induction and repression. *J. Theor. Biol.* 234, 133-143.

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

[40] Sanjun, R., Elena, S.F. (2006). Epistasis correlates to genomic complexity. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14402-14405.

[41] Dekel, E., Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436, 588-592.

[42] TAIR [www.arabidopsis.org].

[43] ATH1 Genome array [http://www.affymetrix.com/].

[44] NASCArrays [http://affymetrix.arabidopsis.info/narrays/experimentbrowe.pl].

[45] AtGenExpress [http://www.arabiodpsis.org/info/expression/ATGenExpress.jsp].

[46] Tibshirani, R. (1996). Regression shrinkage and selection via de Lasso. *J. R. Statist.* 58, 267-288.

[47] Shevade, S.K., Keerthi, S.S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* 19, 2246-2253.

[48] AtRegNet [http://arabidopsis.med.ohio-state.edu/RGNet].

[49] Wernicke, S., Rasche, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics* 22, 1152-1153.

Chapter 6

Reprogramming Plant Cellular Chassis to Mimic Viral Infection

For over decades, plant molecular virology has been overly focused on the pathogen itself, studying their individual genes and products, and their local effects on certain regulatory pathways related to antiviral responses or to symptoms development. Viral infections typically alter host physiology, notably by diverting almost every cellular resource to the production of virus-specific components, and by actively suppressing host defenses [1, 2]. The recent arrival of genomic tools have allowed high-throughput genetic and metabolic screenings, providing unprecedented views of the plant host-virus interactions from a systemic perspective that would allow us reaching a deeper understanding on how host and virus genotypes interplay in determining the pathological outcome of an infection [3, 4, 5, 6, 7].

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 [2, 8, 9, 10, 11, 12, 13, 14, 15, 16]. 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 virulence of the infection. Imposing the measured transcriptional changes in a biological network context, it was confirmed that host cells undergo a significant reprogramming of their transcriptome during infection [17, 18], which is possibly a central requirement for the mounting of host defenses. Moreover, Rodrigo et al. [19] uncovered a general mode of plant virus action in which

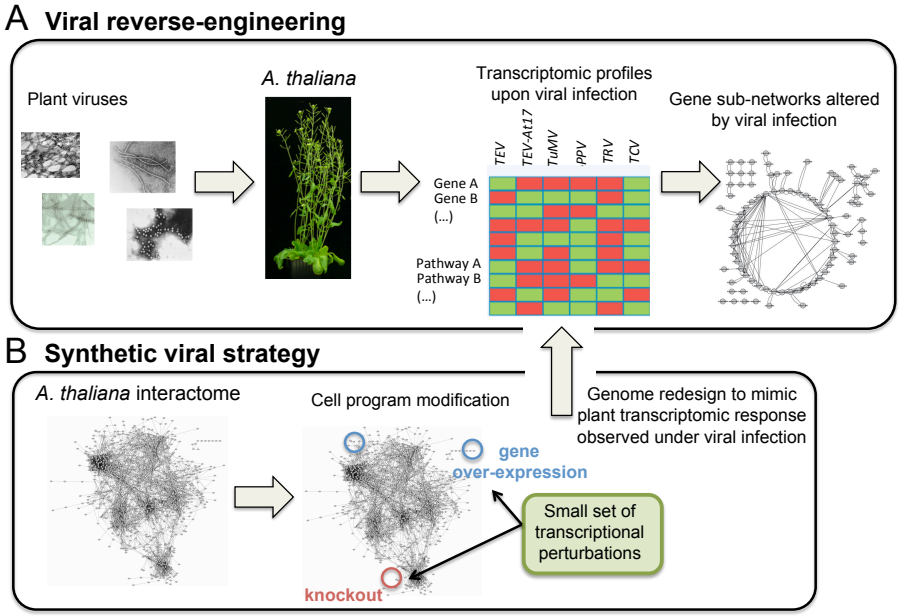

Figure 6.1: (A) Reverse-engineering to reveal gene sub-networks differentially altered by viral infection. (B) Reprogramming cells to mimic the plant transcriptomic responses observed upon viral infection by using computational genome redesign.

perturbations preferentially affect genes that are highly connected, central and organized in modules, a mechanism of action that has been pervasively described for animal viruses [20, 21, 22, 23, 24, 25, 26].

Inspired on an integrated computational-experimental approach for discovering genes and pathways that are targets of specific compounds [27], herein, we aimed to re-design the transcriptional regulatory network (TRN) of *A. thaliana* by altering key transcription factors (TFs) in order to mimic the transcriptional response observed upon infecting the plant with different virus. We will accomplish this goal by re-designing optimal genetic network using as starting point a genome-scale TRN model of the plant [28]. Hence, those re-designs will provide new insights about mechanisms related with virus-target interactions in the plant. Recently, many groups have proposed and implemented different approaches for genome-wide re-design by knocking out and over-expressing genes of prokaryotes and eukaryotes [29, 30, 31, 32] to control global gene expression. Following this synthetic biology strategy, herein we have re-designed *A. thaliana* TRN by exhaustively exploring multiple gene perturbations in the form of gene knockouts or over-expressions. Hence, we have corroborated that several genetic modifications imposed on a critical set of TFs generates a high diversity in the

transcriptome of the plant.

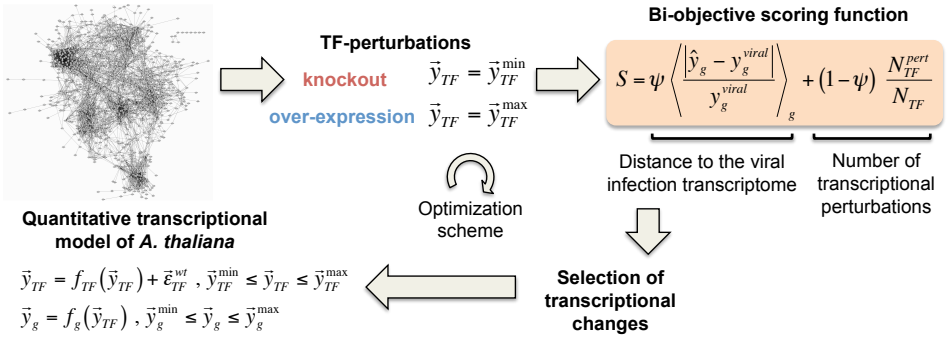

Figure 6.2: Computational methodology to predict optimal redesigns of the *A. thaliana* TRN that mimics the alterations induced by viral infections on the plant transcriptome.

Could a reduced set of perturbed TFs mimic the plants transcriptional response to viral infections? It is of outmost importance to harness the ability of using computational design to predict and optimize *a la carte* synthetic genomes with desired transcriptional responses (Figure 6.1). To address this question, we have developed an algorithm that uses as starting point a wild-type transcription regulation model, inferred from high-throughput microarray data [28]. This TRN is evolved using a heuristic optimization method that at each stage computes the updated gene expression profile and compares it with the one observed during viral infection. With this approach, we explored the space of re-engineered TRNs to find the optimal global network whose expected transcriptional profile minimizes the one characteristic of the viral infection. Consequently, the use of genomic techniques to develop design-guided models, and the application of reverse-engineering methods, open the doors for delineating a high-resolution picture of host-pathogen interactions.

We have developed a methodology to automatically re-design the TRN of *A. thaliana* to mimic the transcriptomic changes induced by perturbations (see Appendix A). In particular, we have focused in the perturbations induced by the infection with a set of eight viruses. For that, we hypothesized that symptoms of viral infections could be recreated in absence of the pathogenic agent by altering a minimal core set of TFs (6.1B). We used a genome-wide model of *A. thaliana* gene transcription based on ordinary differential equations (ODEs) to predict changes in gene expression after modifications such as TF knockout or overexpression [28] (see previous Chapter). This model contains 21929 non-redundant genes, 1187 of

which are putative TFs and consequently, are the potential candidates to be perturbed. Figure 6.2 shows a schematic representation of the evolutionary algorithm implemented, in which single gene mutations were made and then selected according to the fitness functions discussed in the Appendix B. This algorithm explores the landscape of all possible TRN produced by over-expressing or knocking out *A. thaliana* TFs. Operationally, these perturbations were introduced by modifying the corresponding ODEs. At each evolutionary step, a population of such perturbed TRNs was generated and their corresponding expression profiles computed and compared with the target transcriptomic profile observed for each viral infection. Those TRNs showing the better matching with the target were selected for the next round of optimization, as thoroughly discussed in the Appendix B. Fifty independent runs of this evolutionary optimization were generated to evaluate the convergence of the results.

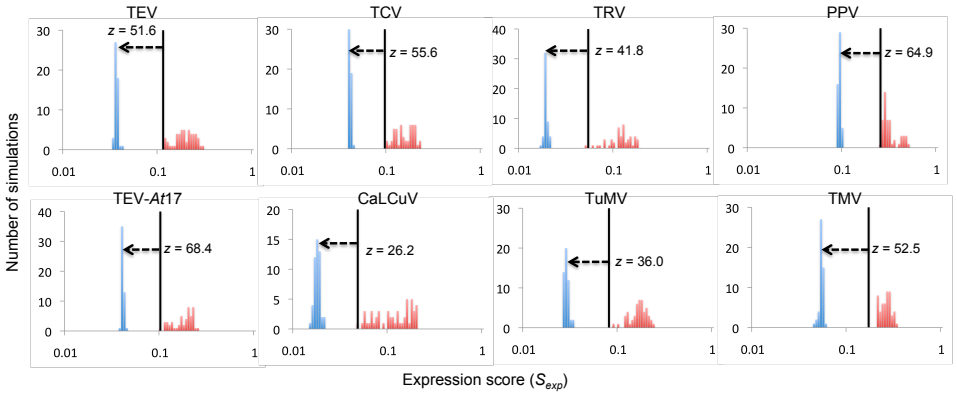

Figure 6.3: Histogram of the scores obtained in 50 optimization processes (blue bars) forced to mimic the plant transcriptome changes observed after infection with eight different viruses. In the optimization process, only TFs were considered in the scoring function (S). Random simulations were computed without imposing any selective pressure (red bars). Black line shows the score obtained using the transcriptional model inferred of *A. thaliana*.

6.1 Re-engineered TRNs that Mimic the Transcriptomic Response Characteristic of Different Viral Infections

Figure 6.3 illustrates that re-engineered TRNs actually show expression profiles that are significantly closer to those observed in infected plants than the profile inferred for non-infected plants (wild-type). Different panels cor-

respond to the optimization done for the different viruses using the scoring method that takes into consideration only perturbations in TFs (Figure 6.2). The ordinates axis shows the computed expression score (S_{exp}). A perfect matching between the observed TRN and that computed for the re-engineered genome would have $S_{exp} = 0$. The larger the departure from zero, the poorest the matching between observed and predicted expression profiles. The black vertical line corresponds to the score computed for the wild-type TRN (*i.e.*, non infected plants). In blue, we show the observed distribution of scores after multiple optimization runs. In all cases, the optimization results in artificial TRNs that are closer to the observed transcriptome than the wild-type. The best fit was obtained for TRV ($S_{exp} < 1.96\%; p < 0.001$) and the worse to PPV ($S_{exp} < 9.51\%; p < 0.001$). As an additional quality control, we also run the optimization algorithm but without the selective constrain imposed by matching the observed transcriptional profiles (red bars in Figure 6.3). In these cases, the distributions of S_{exp} did not show any improvement but, instead, had average values larger than those observed for the wild-type TRN. As we would expect, considering not only TFs in the scoring function but all genes, predictability is not significantly improved, ranging between $S_{exp} < 1.88\%(p < 0.001)$ for CaLCuV and $S_{exp} < 10.36\%$ ($p < 0.001$) for PPV. Indeed, the S_{exp} obtained only with TFs or with all genes are highly correlated ($r = 0.714$, 6 d.f., $p = 0.046$), thus suggesting that the conclusion is robust to the choice of genes to be perturbed during the optimization process. Interestingly, this conclusion does not specifically holds for infection with TCV. For this virus, using only TFs results in a poor optimization.

6.2 A Minimal Set of Transcriptional Factors Guarantees TRN Re-designs that Mimic Viral Infections

Transcriptomic studies have shown that the number of *A. thaliana* TFs altered upon viral infection varies among different viruses, with TRV altering 11 and TEV-$At17$ altering 101 (gray bars in the upper panel of Figure 6.4A) [4, 19]. What would be the number of TFs whose expression has been altered in the re-engineered TRNs? To answer this question we simply counted the number of the transcriptional modification done on each designed TRN. Figure 6.4A shows this information for each virus, both for TRNs evolved with the scoring function that only accounts for changes in TFs (blue bars) and for the scoring function that accounts for alterations in all genes (red

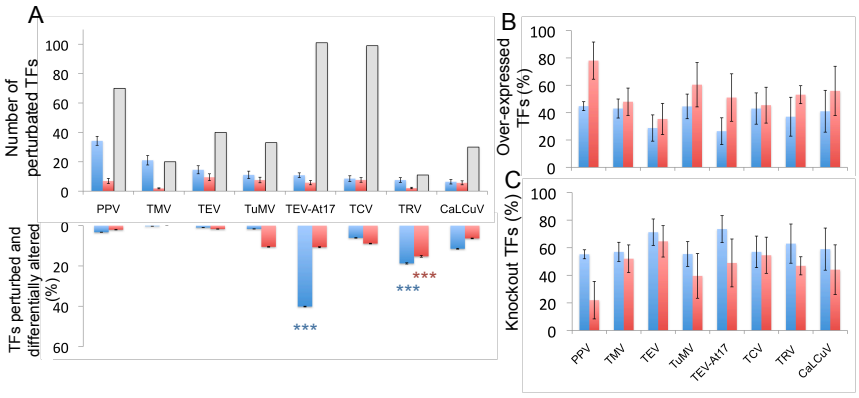

Figure 6.4: (A) Number of TFs to perturb (knockout/over-expression) proposed by the optimization process in which only TFs or all genes are considered in the scoring function (blue and red bars, respectively, in the upper panel). Grey bars show the number of TFs differentially expressed that were identified by gene expression upon infection with eight different viruses. We also show the intersection between the perturbed TFs proposed in the design and those with differentially altered gene expression (bottom panel). Random selections of TFs for designing simulating optimization processes without selective pressure were computed to test statistical significance (*** $p < 0.001$). (B, C) Percentages of over-expressed and knockout TFs proposed to be perturbed in each design for the eight viruses. Error bars show SD from the 50 simulations.

bars). Surprisingly, after optimization, the number of altered TFs necessary to mimic viral infections was quite reduced for all viruses. For example, looking only at TRNs designed using the first scoring function, the minimum number of perturbations necessary was found for CaLCuV infection (6.40) and the maximum of 34.24 TFs perturbations, on average, for the case of PPV (Figure 6.4A). Narrower ranges (from 1.92 to 9.50) were obtained for all viruses when the second scoring function was used instead.

The set of TFs proposed by the design algorithm not necessarily include all those whose expression has been observed in real infections (Figure 6.4A, lower panel). In general, a bootstrap test shows that the intersections between the lists of proposed TFs and observed altered TFs were not significant except in two instances. In the case of TEV-*At*17 40.15% of the proposed TFs were altered in the real infection, being the proposed set significantly enriched in observed TFs (Figure 6.4A lower panel; bootstrap test, $p < 0.001$). This significant enrichment does not exist when all genes were considered in the scoring function. Analogously, 18.73% of the proposed TFs perturbed in the redesigned TRNs were contained in the list of observed altered TFs for TRV (Figure 6.4A, lower panel; bootstrap test, $p < 0.001$), and this result remains significant independently of the gene set

weighted by the scoring function. Next, we observed that the lists of genes proposed by all re-designs were significantly enriched in biological functions related to response to biotic and abiotic stresses, and in developmental processes. This enrichment corroborates that the pathological outcome of viral infection can be reproduced in absence of the viral agent by altering the appropriated plant cellular programs. Hence, the plasticity of *A. thaliana* transcriptome to generate specific expression pattern as a response to multiple genetic perturbations allows verifying that non-infected cells could easily mimic transcriptomic responses to diverse viral infections. What kinds of perturbations contribute the most to this plasticity? Figure 6.4B and Figure 6.4C show, respectively, the percentages of over-expressed and knocked out TFs for each redesigned TRNs (as before, blue and red bars correspond to scoring functions that use only TFs or all genes). Overall, re-engineered TRNs included more gene knockouts than over-expressions for all viruses except for the re-designs mimicking PPV infection, which are more balanced, indicating that gene knockouts generate more plasticity in gene expression than overexpressing genes.

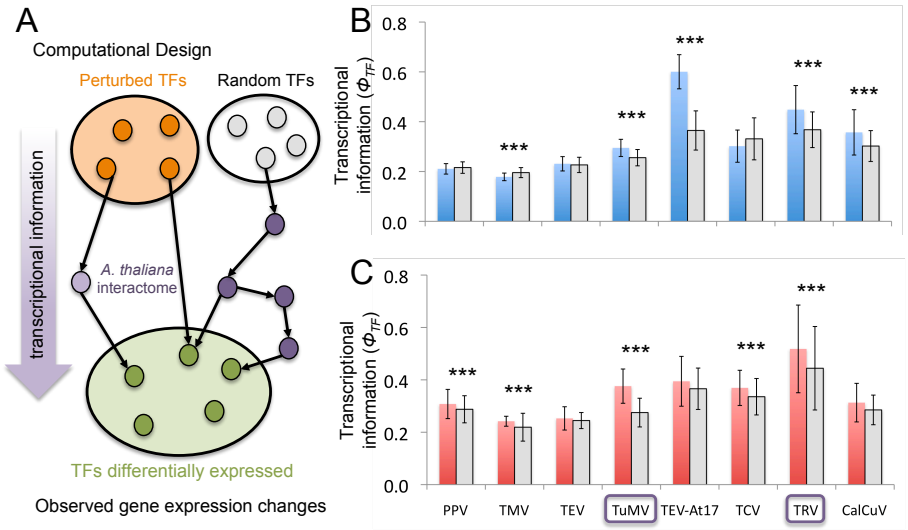

Figure 6.5: (A) Scheme to illustrate the topological properties in a context of the *A. thaliana* TRN between the perturbed TFs proposed by our methodology and those identified differentially altered under viral conditions. (B, C) Transcriptional information of all TFs proposed to be perturbed in the designs. Note that blue and red bars show TFs proposed by using our design methodology evaluating only TFs or all genes in the scoring function, respectively. Random shortest paths were computed to evaluate the statistical significance of the topological distance between the TFs for designing with respect TFs selected randomly in the *A. thaliana* interactome (*** $p < 0.001$). Error bars show SD from the 50 simulations.

We have just shown that the list of TFs altered in the re-engineered TRNs were not necessarily the same set that have been shown to have altered expression in the real infections. Now, we want to explore whether this proposed set of TFs are located in the wild-type TRN close to TFs that are altered during real infections, as described in Figure 6.5A. If this is the case, then it can be argued that the proposed set of TFs does affect exactly the same target genes that may be affected by the TFs altered during viral infection. Otherwise, it can be argued that the proposed set of TFs mimics infection by completely different mechanisms. To evaluate this question we proceeded as follows. For each proposed TF we evaluated a topological parameter, ϕ_{TF}, which takes into account the minimum shortest path between this TF and all the TFs significantly altered by viral infection (see Appendix C). The parameter ϕ_{TF} takes the value one if the proposed TF is included in the list of TFs altered in real infections and tends to zero as the distance to the closest altered TF increases in the network. Figure 6.5B and Figure 6.5C show ϕ_{TF} for each virus (panel B corresponds to scoring functions using only TFs and panel C to all genes). The statistical significance of ϕ_{TF} was evaluated by generating random lists of transcription factors. Only TFs proposed in all re-designs of TuMV and TRV were significantly closer in the wild-type TRN to TFs differentially altered by viral infection (Figures 6.5B and 6.5C), for both scoring functions, than expected by sheer chance. All together, these results confirm that the re-engineered TRNs proposed by our methodology mimic the transcriptomic response observed under real viral infections by altering a smaller and different set of targets than those observed during the transcriptomic characterization of real infections. Indeed, in general, they are not even neighbors in the wild-type TRN.

6.3 The Number of Proposed TFs to be Perturbed Correlates with the Magnitude of the Alterations in Gene Expression Observed upon Viral Infection

The microarray data characterizing the infection of *A. thaliana* with each of the eight viruses shows significant variation in the amount of genes whose expression was altered. Therefore, we hypothesized that in order to match the transcriptomic consequences of infection, our optimization algorithm shall propose more TF alterations for viruses that exert a large impact on the host transcriptome than for viruses that have a mild impact. To test this hypothesis we proceeded as follows. First, to collapse into a single quan-

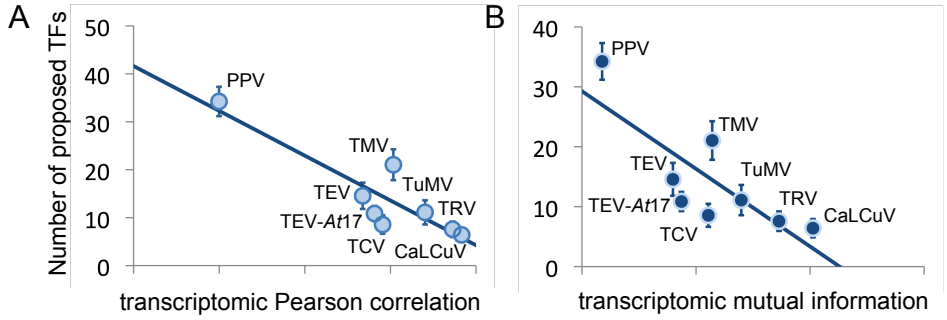

Figure 6.6: Relationship between the number of TFs proposed by the design algorithm to be perturbed and the correlation (evaluated as Pearson correlation (A) or mutual information (B)) between the gene expression profiles of the wildtyype plants and plants infected with each of the eight viruses. Only TFs were considered in the scoring function during the optimization process. Error bars show SD for the 50 simulations.

tity the impact of viral infection in the host transcriptome, we computed two different statistical measures, the Pearsons correlation coefficient and the mutual information (MI) between the transcriptomic profiles of non-infected and infected plants. For viruses having minor impact in the plant transcriptome, the Pearsons coefficient will be close to one and the MI large. By contrast, a small correlation coefficient and low MI will reflect a strong perturbation in the plants transcriptome. Second, we sought for an association between these indexes and the number of predicted altered TFs for each virus. If our hypothesis is correct, we must observe a significant negative correlation between these variables. The largest changes in the transcriptome were found in cells infected by PPV, whereas the action of CaLCuV showed the smallest variation in the host expression profile. Figure 6.6 shows the expected negative and significant correlations between the number of proposed alterations, using the scoring function based only in TFs, and the overall impact of the viral infection measured as the Pearsons correlation (Figure 6.6A: $r = -0.890$, 6 d.f., $p = 0.003$) and as the MI (Figure 6.6B: $r = -0.800$, 6 d.f., $p = 0.017$). Hence, we could conclude that the amount of proposed perturbations needed to mimic a viral infection is directly dependent on the overall effect that the virus exerts on the host transcriptome: the larger the effect, the more perturbations are needed and vice versa.

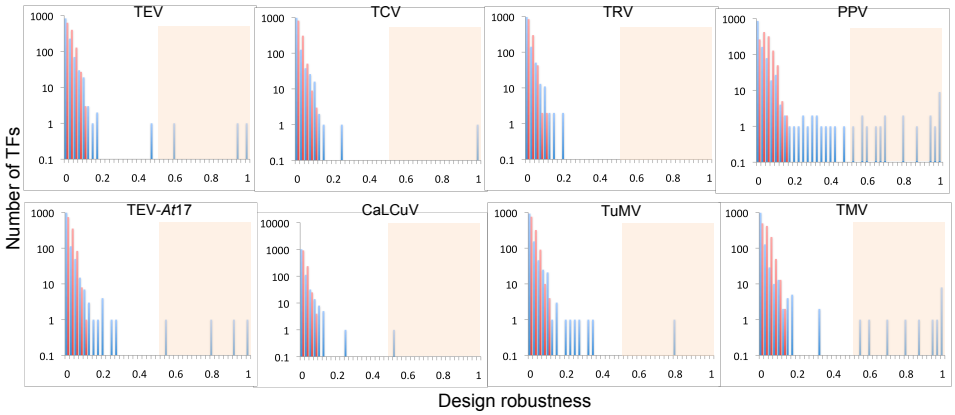

Figure 6.7: Histogram of the robustness of the TFs proposed by the design algorithm for the eight viruses (blue bars). Only TFs were considered in the scoring function for designing. The red bar histograms correspond to the robustness of TFs selected from randomly re-engineered TRNs.

6.4 A Crucial Set of TFs is Pervasively Proposed in the Redesigned TRNs

For each virus, we have run 50 independent optimization processes. Do all of them represent completely different solutions? Or by contrast, do all of them contain a preferred set of TFs? If the second situation is true, then we can suggest that the over-represented TFs represent a critical set of highly relevant TFs that, eventually, may be the focus of future experimental validation. For each virus, we tabulated the 50 lists of proposed TFs and calculated the probability of finding each TFs in the 50 lists. This probability can be taken as a measure of the design robustness of the prediction for each TF. A TF with a very low design robustness (*e.g.*, $< 25/50$) means that it may have very low relevance in mimicking the transcriptomic profile induced by the virus. By contrast, a TF with high design robustness (*e.g.*, $> 25/50$) will be indicative that such TF plays a central role in mimicking the transcriptomic symptoms. Figure 6.7 (blue bars) shows such degree of design robustness for perturbations of all TFs for the eight viruses. In this case the scoring function used in the optimization process only accounted for TFs. The different simulations did not share most of the TFs proposed. However, a certain number of TFs were in common in more than half of the simulation. This number of crucial TFs varies among viruses, ranging from zero (TRV) to 19 (PPV). As a way to evaluate the statistical significance of these results, we generated, for each virus, a new set of 50 simulations but without using the match in expression profile (S_{exp}) in the scoring function.

The distribution of design robustness for these TRNs is shown on Figure 6.7 (red bars). In all cases, the distributions of design robustness did not reach values larger than 0.2, thus confirming that the existence of critical TFs could not be explained by chance. For each virus, the centrality and shape of both distributions were compared, and found significantly different (Mann-Whitney test, $p < 0.001$; Kolgomorov-Smirnov test, $p < 0.001$). Interestingly, among the 47 TFs proposed, 22 are involved in several developmental processes, nine in responses to biotic and abiotic stresses and the rest do not have been assigned to any specific function. This enrichment in TFs involved in development may be justified by the symptoms induced upon viral infection, which in most cases involve dwarfism, leaf malformations and curling and delays in the emergence and development of inflorescences.

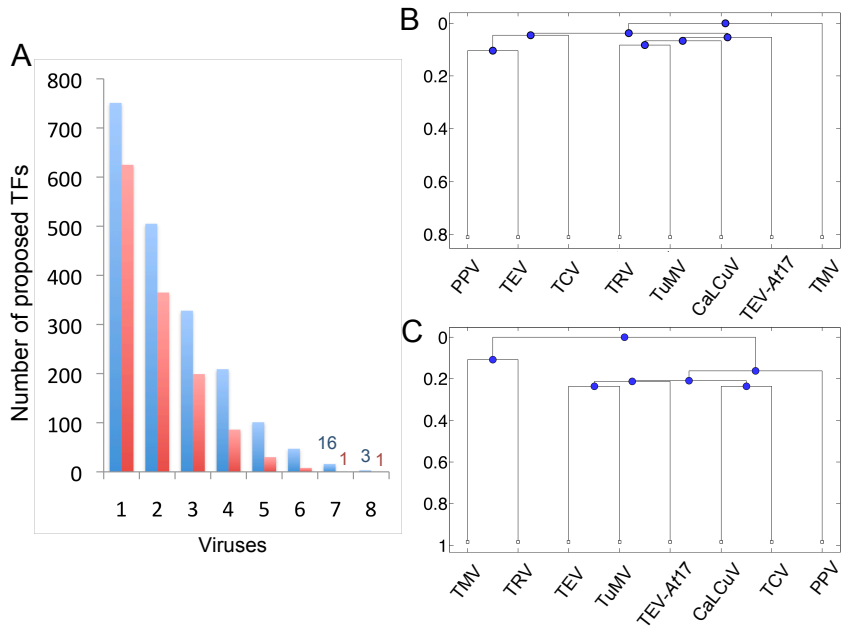

Figure 6.8: (A, B) Neighbor-joining dendograms obtained from the similarity matrix computed from overlapping lists of TFs proposed to be perturbed in the different designs for the eight viruses. (C) Number of common TFs proposed to be perturbed by the model for several viruses. Note that only TFs ((A) and blue bars in (C)) or all genes ((B) and red bars in (C)) were considered in the scoring function for designing.

6.5 Proposed TFs Which Are Common for Different Viruses

Finally, we explored the overlap between the lists of TFs whose expression was altered in the designed TRNs for the different viruses. For each virus, this list includes all the TFs proposed at least in one of the 50 simulations (*i.e.*, not restricting the analyses to the critical set described in the above paragraph). As in previous sections, the lists of TFs were build both using the scoring functions based on TFs and in all genes. Figure 6.8A shows the number of TFs that are found in the lists of one or more viruses (blue and red bars corresponding to whether only TFs or all genes were accounted for in the scoring function). Around 700 TFs were virus-specific, but a large fraction (over 400) was shared by at least two viruses. Among these, ca. 200 TFs were shared by at least three viruses. In the right side of the distribution, we found 16 TFs shared by seven viruses. Finally, all viruses only share three TFs: *At1g50640, At2g35940 and At2g37650. At1g50640* corresponds to the ethylene-responsive transcription factor 3 (ERF3) that negatively regulates the ethylene-mediated signaling pathway and gene transcription. *At2g35940* corresponds to the BEL1-like homeodomain 1 protein (BLH1) that regulates transcription in response to abscisic acid stimulus. *At2g37650* encodes for a TF of the GRAS family involved in root and leaf development and in the negative regulation of flower development.

The overlap between pairs of lists was further quantified using the similarity index $2n_{xy}/(n_x + n_y)$, where n_xy is the number of common entries in the two lists and n_x and n_y the length of each list. A similarity matrix containing all pairwise comparisons was constructed and used to build neighbor-joining dendograms that cluster together viruses according to the similarity of their proposed lists of TFs (Figure 6.8B and Figure 6.8C, using the scoring function based only on TFs or considering all genes, respectively). Three groups result when the first scoring function was used (Figure 6.8B). The first group is formed by PPV, TEV and TCV, the second group by TRV, TuMV, CaLCuV, and TEV-$At17$, while TMV appears as the most dissimilar virus. Given the nature of the data used to build up this dendogram (*i.e.*, similarities among lists of TFs whose alteration mimic the symptoms of infection), not surprisingly, the clustering does not reflect phylogenetic relationships between viruses (PPV, TEV, TEV-At17, and TuMV are all classified within the same taxonomic genus, the Potyvirus) nor whether they share common hosts in nature (TCV, TuMV, CaLCuV, and TEV-At17 infect plants taxonomically related to the experimental host *A. thaliana*, the Brassicaceae). By contrast, when the scoring function used

in the optimization process takes into account all genes, the dendogram obtained clusters together all the brassica-infecting viruses (Figure 6.8C). Thus suggesting that host-driven selection may have determined the set of genes that viruses infecting the same host manipulate to optimize their replication.

6.6 Discussion

Plants have evolved defense mechanisms to recognize pathogens and defeat them, but viruses have developed elements that interfere and suppress these mechanisms. In this article, we have developed a computational methodology to explore the plasticity of the transcriptome of *A. thaliana* in response to the alteration of certain key TFs. Specifically, we have addressed the problem of re-design a plant TRN to mimic the transcriptomic response observed upon viral infection but in absence of any intracellular pathogen. In the case of eight different viruses, our methodology rendered re-engineered TRNs that captured the transcriptional responses of the infected host *A. thaliana*. Surprisingly, this mimicking was obtained by manipulating a reduced number of TFs associated to developmental processes and to responses to biotic and abiotic stresses. As we expected, the complexity of the redesigned TRNs, in terms of number of necessary TF to be perturbed, correlated with the amount of change induced by each viral infection to the transcriptome of infected plants. In addition, we found certain degree of overlap between the TFs selected in each run of the optimization algorithm, providing evidence that a set of essential TFs is able to generate high plasticity in gene expression of the plant.

Our computational work, identifying reduced sets of TFs that result in mimicking the symptoms of infection, opens the doors to future experiments that may use the *A. thaliana* gene knock out collections not only to validate our prediction but also as a way of reach a better understanding of the molecular mechanisms of viral pathology. Even perhaps as potential targets for future therapeutic interventions.

Appendix A Plant Viruses and Transcriptomic Data

Seven positive-sense single-stranded RNA viruses and one virus whose genome is composed by a single-stranded circular ambisense DNA molecule on a common plant host, *A. thaliana*, were selected. The set of RNA viruses comprises three members of the *Potyviridae* family, tobacco etch potyvirus (TEV),

turnip mosaic potyvirus (TuMV) and plum pox potyvirus (PPV), two members of the family *Virgaviridae*, tobacco mosaica tobamovirus (TMV) and tobacco rattle tobravirus (TRV), and one member of the tombusviridae family, turnip crinkle carmovirus (TCV). We also considered a laboratory-strain of TEV (TEV-*At*17) evolved in and adapted to *A. thaliana* [10]. The ssDNA virus included in the study was the member of the Geminiviridae family cabbage leaf curl begomovirus (CaLCuV). TEV and TEV-*At*17 expression data (two-color raw data, NCBI GEO accession GSE11088) were obtained from ecotype *Ler*-0 plants 21 days post-inoculation (dpi) [9], [10]. TuMV data (Affymetrix raw data, ArrayExpress accession e-mexp-509) were obtained 21 dpi from ecotype Col-0 plants [16]. PPV data (Affymetrix preprocessed data, NCBI GEO accession GSE11217) were obtained 17 dpi from Col-0 plants [12]. 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 [14]. TRV data (two-color raw data), NCBI GEO accession GSE15557, GSE155562 and GSE15558) were measured 21 dpi from Col-0 leafs. TCV data (two-color raw data, NCBI GEO accession GSE29387) were quantified 10 dpi in Col-0 plants. Finally, CaLCuV data (Affymetrix raw data, ArrayExpress accession E-ATMX-34) were collected from Col-0 plants 12 dpi [11]. The list of differentially expressed genes was obtained by performing a fold-change criterion of $z > 1.96$ over all genes (averaging replicates).

Appendix B Genome-Wide Multiple-Optimization

Our algorithm searches possible reconfigurations of the global transcription regulatory network such as that the expression profile of the re-engineered genome mimics the transcriptional response of the host infected by different viruses. We address this problem by using a high efficient heuristic optimization. We suggest genome reconfigurations that include multiple TF knockouts or over-expression by imposing in the model (see previous Chapter), $y_{TF} = y_{TF}^{min}$ or $y_{TF} = y_{TF}^{max}$, respectively, or both types of perturbations in order to enlarge the combinatorial space of perturbed genomes targeting the transcriptional response given under viral infection. We started from the inferred model and at each step in the optimization process, we randomly selected a new TF to evaluate its three states (knockout, over-expression or wild-type) and update the model with the best-scored scenario until all TFs have been manipulated. Hereafter, we looped back and introduced new tran-

scriptional modifications restricted to a maximum number of perturbations in order to be able to implemented experimentally re-designs.

We use a bi-objective scoring function, ϕ, to force a minimum average quadratic deviation to the viral infection expression profile in steady state (S_{exp}) with the minimum number of genetic perturbations (S_{pert}), $\phi = \theta S_{exp} + (1-\theta)S_{pert}$, with $S_{exp} = \sum_g \lambda_g \frac{|y_g^v - y_g|}{y_g^v}$ and $S_{pert} = \frac{N_{TF}^{pert}}{N_{TF}}$, where θ is the weighting factor of each objective function, y_g^v is the expression profile under viral infection of all genes (N_g) of *A. thaliana* and $\lambda_g \in (0,1)$ is a parameter defined for each gene that differences those genes differentially expressed in the microarray data measured under viral conditions. In fact, we can divide the objective function representing the expression score into two terms: a first that quantifies expression deviations from genes that have been identified as altered in the transcription and a second to compute the rest of genes. N_{TF} is the total number of TFs annotated in *A. thaliana* and N_{TF}^{pert} is the number of TFs that the model suggests to be perturbed.

To measure the degree of contribution that a TF has to mimic a target gene expression of the TRN of *A. thaliana*, we defined a parameter $\phi_{TF} = 1/(1 + min(\delta_{TF-g}))$ that evaluates the minimal number of links, δ_{TF-g}, between a TF (gene perturbations proposed in the a given design) and a set of TFs potentially affected by global changes in gene expression, g (genes differentially altered by viral infection).

References

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

[2] Jenner, R.G., Young, R.A. (2005). Insights into host responses against pathoghens from transcriptional profiling. *Nat. Rev. Microbiol.* 3, 281-294.

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

[4] Elena, S.F., Carrera, J., Rodrigo, G. (2011). A systems biology approach to the evolution of plant-virus interactions. *Curr. Op. Plant. Biol.* 14, 372-377.

[5] Friedel, C.C., Haas, J. (2011). Virus-host interactomes and global models of virus-infected cells. *Trends Microbiol.* 19, 501-508.

[6] Peng, X., Chan, E.Y., Li, Y., Diamond, D.L., Korth, M.J., Katze, M.G. (2009). Virus-host interactions: from systems biology to translational research. *Curr. Opin. Microbiol.* 12, 432-438.

[7] Tan, S.L., Ganji, G., Paeper, B., Proll, S., Katze, M.G. (2007). Systems biology and the host response to viral infection. *Nat. Biotech.* 25, 1383-1389.

[8] Wise, R.P., Moscou, M.J., Bogdanove, A.J., Whitham, S.A. (2007). Transcript profiling in host-pathogen interactions. *Annu. Rev. Phytopathol.* 43, 329-369.

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

[10] Agudelo-Romero, P., Carbonell, P., Perez-Amador, M.A., Elena, S.F. (2008). Virus adaptation by manipulation of host's gene expression. *PLoS ONE* 3, e2397.

[11] Ascencio-Ibez, J., Sozzani, R., Lee, T.J., Chu, T.M., Wolfinger, R.D., 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.

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

[13] 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.

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

[15] Ishihara, T., Sakurai, N., Sekine, K.T., 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.

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

[17] Whitham, S.A., Yang, C., Goodin, M.M. (2006). Global impact: elucidating plant responses to viral infection. *Mol. Plant Microb. Interact.* 11, 1207-1215.

[18] Whitham, S.A., Wang, Y. (2004). Roles for host factors in plant viral pathogenicity. *Curr. Op. Plant Biol.* 7, 365-371.

[19] Rodrigo, G., Carrera, J., Ruiz-Ferrer, V., del Toro, F.J., Llave, C., Voinnet, O., Elena, S.F. (2011). Characterization of the *Arabidopsis thaliana* interactome targeted by viruses. *SFI Working Papers* 11-10-049 (www.santafe.edu/media/workingpapers/11-10-049.pdf).

[20] Bushman, F.D., Malani, N., Fernades, J., DOrso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., Knig, R., Bandyopadhyay, S., Ideker, T., Goff, S.P., Krogan, N.J., Frankel, A.D., Young, J.A.T., Chanda, S.K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 5, e1000437.

[21] Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., Ewence, A.E., Li, N., Hirozane-Kishikawa, T., Hill, D.E., Vidal, M., Kieff, E., Johannsen, E. (2007). Epstein-Barr virus and virus human protein interaction maps. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7606-7611.

[22] De Chassey, B., Navratil, V., Tafforeau, L., Hiet, M.S., Aublin-Gex, A., Agaugu, S., Meiffren, G., Pradezynski, F., Faria, B.F., Chantier, T., Le Breton, M., Pellet, J., Davoust, N., Mangeot, P.E., Chaboud, A., Penin, F., Jacob, Y., Vidalain, P.O., Vidal, M., Andr, P., Rabourdin-Combe, C., Lotteau, V. (2008). Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230.

[23] MacPherson, J.I., Dickerson, J.E., Pinney, J.W., Robertson, D.L. (2010). Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comp. Biol.* 6, e1000863.

[24] Uetz, P., Dong, Y.A., Zertzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S.V., Roupelieva, M., Rose, D., Fossum, E., Haas, J (2006) Herpesviral protein networks and their interaction with the human proteome. *Science* 311, 239-242.

[25] Watanabe, T., Watanabe, S., Kawaoka, Y. (2010). Cellular networks involved in the influenza virus life cycle. *Cell Host Microbe* 7, 427-439.

[26] Wuchty, S., Siwo, G., Ferdig, M.T. (2010). Viral organization of human proteins. *PLoS ONE* 5, e11796.

[27] Di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., Collins, J.J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotech.* 23, 377-383.

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

[29] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

[30] Segre, D., Vitkup, D., Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112-15117.

[31] Burgard, A.P., Pharkya, P., Maranas, C.D. (2003). OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647-57.

[32] Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M., Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840-845.

Part III

Fine-Tuning of the Tomato Fruit Agronomic Properties by Computationa Design

Chapter 7

Computational Optimization of the Tomato Fruit Agronomic Traits

Introduction

Considering a cell as a DNA-based molecular factory [1] and applying the principles drawn from industrial engineering provides new approaches to optimize cellular performance. This approach adopts the new philosophy implemented nowadays by large industries that is known as Lean Manufacturing (LM). LM consists in the implementation of standards based on elimination of bottlenecks and processes without mark-up and minimization of pathways and excessive costs. This approach can be applied to the emerging fields of systems and synthetic biology, and allows translating engineering concepts into biotechnology [2, 3, 4].

Our main goal is to optimize the phenotypic response of a natural plant biofactory, exemplified here by the edible tomato fruit, by using a combined experimental and computational synthetic biology approach. The approach involves re-designing the fruit factory from within; *i.e.*, by modeling and identifying the important genes and intermediates for a given trait of agronomical interest (see Figure 7.1).

Previous works have considered modeling the global metabolism [5], transcription [6, 7, 8, 9, 10, 11] or the integration of both in microbial organisms [12, 13, 14] from the point of view of systems biology. Many groups, using a re-designing strategy which is characteristic of synthetic biology, have implemented genome-scale re-designs and explorations of the gene knockout landscape both in prokaryotes [15, 16, 17] and eukaryotes [19]. More recent reports have tackled the prediction of phenotypes from metabolic

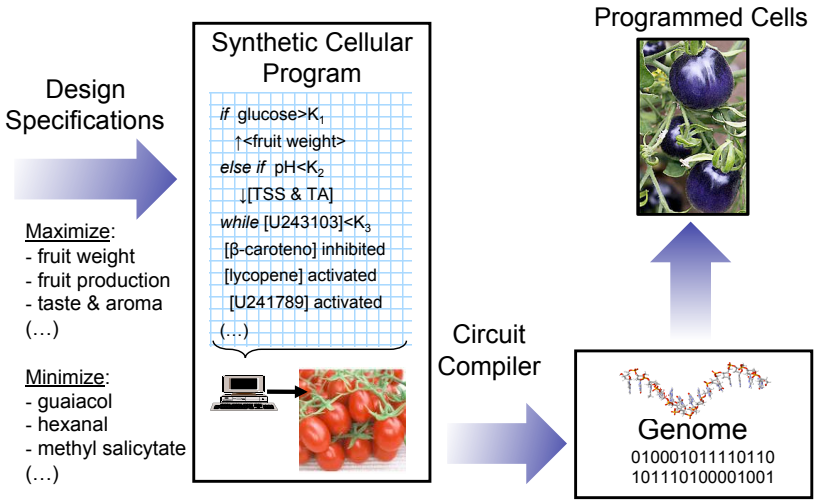

Figure 7.1: Synthetic biology of tomato fruit *vs* computer science.

data based on statistical models for microbes [12], plants [19, 20, 21]. The next logical and desirable development should consist in modeling phenotypes of interest in a complex organism from metabolic and gene expression data. For that purpose we have chosen tomato: a model plant for fleshy fruit -this being a natural biofactory of nutrients and healthy compounds, and a plant of agronomic interest with well-developed genetics and genomics (http://solgenomics.net) and with extensive work on metadata analysis [38, 23, 24]. We have assumed that at least in part the genetic program of the fruit at the ripe stage should have an impact on the metabolite content and also in other high order fruit traits. In this study, we have used omic data that have been experimentally obtained by means of transcriptomics, metabolomics and phenomics for a large number of recombinant inbred lines (RILs) derived from a cross of *Solanum lycopersicum* x *S. pimpinellifolium*. Following the LM approach, we have developed here a novel *in silico* optimization method that extensively explores single and multiple genetic perturbations to render a series of desired tomato phenotypes; *i.e.*, show agronomical properties of biotechnological interest. Techniques of reverse engineering were applied to this large set of experimental omics data to obtain a kinetic model based on ODEs. This model describes the fruit metabolic profile from gene expression data for an autonomous subset of genes with potential effect on transcription regulation [25]. By capturing relationships between metabolic profiles and high-throughput phenomic data, our model was extended to predict changes in agronomic properties that would be produced by specific changes in genetic expression.

Finally, in order to close the design cycle imposed by LM, the genetic modifications suggested by our computational approach were experimentally verified. This was done by demonstrating the predicted ability of the *in silico* modified fruit genomes to reconstruct the correlations found between the metabolites actually measured in the fruit. We propose that the principles and practices learned from these engineering success cases can help to formulate a model to guide the design of new organisms with biotechnological applications.

7.1 Dissecting Tomato Genome: Kinectics-Based Models of Transcription, Metabolism and Phenotype

7.1.1 A Genome-Wide Transcriptional Model Allows the Integration of Tomato Fruit Metabolism

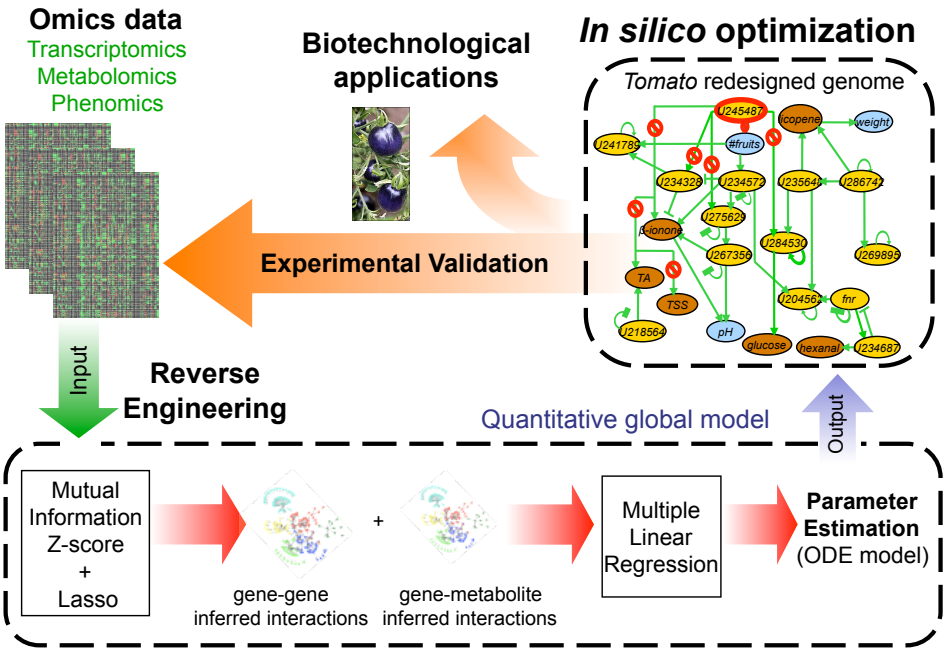

Figure 7.2: Lean Manufacturing as a model applied in systems and synthetic biology. From omic data (transcriptomics, metabolomics and phenomics), a quantitative global model was constructed using reverse engineering methods. The predictive model was used to propose genome perturbations, to improve desired phenotypes with relevant biotechnological applications. The genome perturbations were guided by an *in silico* optimization that imposed the desired selective pressure.

We have extended our developed inference methodology, *InferGene* [7],

to obtain a gene regulatory model coupled to metabolism that allows us analyzing optimality in terms of specified agronomic and organoleptic properties of the tomato fruit (Figure 7.2). For this, we have taken advantage of an experimentally characterized subset of the metabolome of 169 tomato RILs, which includes the accumulation levels in 67 metabolites of the fruit that contribute to the flavor (sugars, acids and some volatiles), aroma (volatiles) as well as other quality traits (such as color and healthy carotenoids and vitamins). Moreover, we have also used the information on transcript levels from fruits for a subset of the 50 RILs analyzed at the metabolic level, to select 5592 non-redundant genes that were consistently expressed in those fruit samples (see Appendix A).

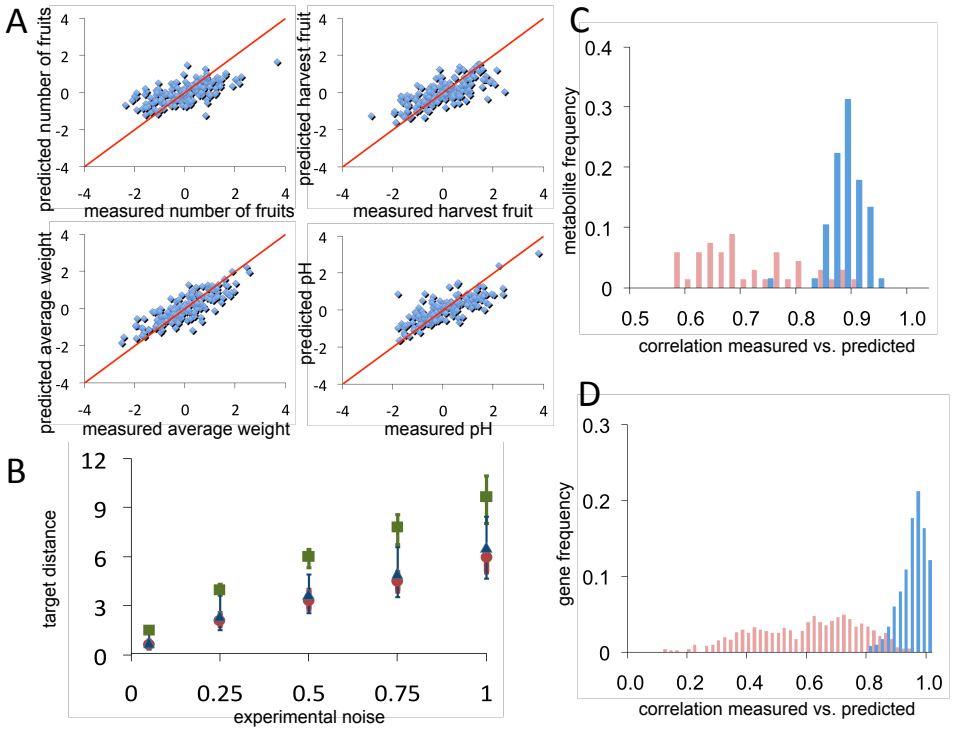

Figure 7.3: Lean Manufacturing as a model applied in systems and synthetic biology. From omic data (transcriptomics, metabolomics and phenomics), a quantitative global model was constructed using reverse engineering methods. The predictive model was used to propose genome perturbations, to improve desired phenotypes with relevant biotechnological applications. The genome perturbations were guided by an *in silico* optimization that imposed the desired selective pressure.

Transcriptomic and metabolomic data from these 50 RILs were normalized by the LOWESS method [26] and used to construct a model that predicts components of the fruit quality metabolome from transcriptome data;

i.e., level of a given metabolite is effectively determined by the expression of a minimal set of genes. The size of the space of possible gene-predictors was reduced in one order of magnitude by using a CLR method. After that, LASSO method was used to find a minimal set of potential predictor genes for each metabolite; subsequently, multiple regressions were obtained to estimate the effective kinetic parameters of a linear model based on ODEs that integrates transcription and metabolism processes (Figure 7.3) [7]. Values $z > 3$ were used as optimal threshold in order to limit the number of possible gene-metabolite interactions and minimize the distance between the predicted and measured metabolic profiles over the training set in terms of average Pearson correlations (blue bars in Figure 7.3C; $r = 0.85$, $p < 0.001$). Hence, on average, each metabolite required 18 genes for explaining its behavior, thus a total of 959 genes was required to describe our tomato fruit metabolome. This subset of genes constitutes the effective transcription network. We performed a 5-fold cross-validation test to rule out dependence of the testing set, this reducing the metabolite average prediction (red bars in Figure 7.3C; $r = 0.42$, $p < 0.1$) with a mean false positive rate (FPR) of 14% and a 56%mean positive predictive value (PPV) of predictors.

The next step was to construct an effective gene regulatory model able to predict autonomously the transcriptional processes that, by means of the model previously described, would generate a quantitative metabolic response. In this way changes at the transcriptional level resulting from the proposed genetic perturbations could be translated and predicted effectively into metabolic changes. For doing that, we used the microarray data obtained from fruits of 50 of the RILs to infer a network of gene-gene interactions. The CLR method provided the first sets ($z > 2$) of predictor genes for each gene considered. Afterwards, LASSO method reduced the number of regulations per gene to a scale-free space following a power-law with exponent $\gamma = 5.47$ ($R^2 = 0.91$) and an average of 26 interactions per gene. High values of similarity between the predicted and measured gene expression (blue bars in Figure 2D) were computed for the whole training set ($\langle r \rangle = 0.793$, $p < 0.001$) while for a 5-fold cross validation the average similarity (red bars in Figure 7.3D) was $r = 0.59$ ($p < 0.1$) with a mean FPR of the 25% and a 63%mean PPV of predictors.

7.1.2 Global Transcriptional Model Integrating Metabolism

We wonder whether the agronomic/phenotypic properties of the tomato fruit could be controlled by/or be the consequence of their metabolite composition. To provide some insight into this question, we studied the relationship

between agronomic properties and metabolic composition across 169 tomato RILs. We applied LASSOs method to select a set of metabolites that may act as predictors for each agronomic property (Supplementary Data 1). Our model included 47 metabolites observing considerably high Pearson correlations between the measured and predicted phenotypic responses over the 169 RILs for number of fruits per plant and fruit harvested across two different seasons, (Figure 7.3A; $r = 0.62$ and $r = 0.73$ respectively, $p < 0.001$ in both cases). A reduction to $r = 0.46$ ($p < 0.1$) and $r = 0.62$ ($p < 0.05$) in the median correlation was computed in a 10-fold cross validation, with 84%mean PPV and mean FPR of 33% and 35%, respectively. Average fruit weight and pH required as many as 44 metabolites as potential predictors with high reliability levels. Reliability was assessed by comparing the corresponding predicted and measured values for the 169 RILs (Figure 7.3A; $r = 0.85$ and $r = 0.80$, respectively, $p < 0.001$ in both cases). A 10-fold validation only reduced those similarities to $r = 0.73$ and $r = 0.63$ ($p < 0.05$ in both cases), with mean FPRs of 37% and 22% and mean PPVs of 81% and 88%, respectively. To test the specificity of the inferred model parameters, we perturbed the target phenotypic profile for each RIL adding different levels of noise. Figure 7.3B shows the distance between predicted and measured values (green points) and mean correlations for different noise levels. A similar approach was performed by using the metabolic and gene expression profiles (red and blue points, respectively). Correlations with significance levels higher than the indicated above were not considered in the cross-validations. In addition, we estimated a very low mean error in predicting the agronomic properties across the training set ($0.45 \left\langle \sigma^{AV} \right\rangle_{RIL}$, see Appendix).

7.2 Computational Optimization of the Agronomic Properties

7.2.1 Genome Redesign by Using Single and Multiple Genetic Perturbations

Here, our main goal is to redesign the genome of tomato to generate an engineered surrogate that, if viable, would be easier to study and of greater potential biotechnological interest. Our design approach was inspired by the practice of *in silico* optimization over a predictive global model. Our next step was to test the possibility of improving agronomical properties of interest. We tested several scoring functions that fall into two global types: on the one hand, agronomical variables measured experimentally such as

the number of fruits harvested per plant, the average fruit weight or its pH; and on the other hand, more complex fruit attributes that could be defined according to some of the components of the metabolic profile and are related to organoleptic properties of the fruit. In this later case, we first evaluated as proof of concept: fruit acceptability according to criteria based on acidity and sugars [27], quality as defined by the contribution of specific volatiles to aroma and by a reported [28] panel assessments of the tomato fruit and consequently on organoleptic acceptance. For this latter case we assumed a strong influence of a set of metabolites to be either maximized (β-ionone, β-damascenone, 2-phenylethanol and benzaldehyde) or minimized (methyl salicylate, guaiacol, hexanal, 1-penten-3-one and (E)-2-hexenal) using balanced weighting factors to account for their positive or negative contribution to quality. Moreover, all single metabolites were also optimized in single target analyses. Finally, a bi-objective function that included a high trade-off was proposed to optimize fruit quality and its production. As a first approach, we re-engineered tomato genome by perturbing independently the 959 genes included in the model, then we re-computed the scoring functions for all RILs enumerating all single knockouts and finally, all gene over-expression models were obtained.

Hence, mimicking the optimization patterns typical from LM, the landscape of desired agronomic properties of tomato fruit was exhaustively explored perturbing its effective transcriptional regulatory network (TRN) with single-gene alterations. Figure 7.4A shows the improvement of two of the agronomic properties mentioned above (fruit acceptability and quality vs production) as result of single gene perturbations according to our model. The success of the approach is shown by the efficiency function obtained for each transcriptional perturbation computed and which is defined by the normalized ratio between the agronomic property obtained for the re-engineered TRN and that for the wild-type TRN. Both agronomic properties and efficiencies in the case of single-perturbations were computed for each of the 169 RILs, resulting in a high variability between the lineages for all knockouts and over-expressed gene re-engineered TRN cases. We corroborated that there is a highly significant linear correlation ($R^2 > 0.99$; $p < 0.001$) between the average value of the improved agronomic properties and the efficiencies reached across the set of RILs for all transcriptional perturbations. Both gene knockout and over-expression models resulted in similar linear regression slopes when considering acceptability and quality vs. production together (0.05 and 0.24, respectively, Figure 7.4A). In addition, we also explored the possibility of tuning a given agronomic property

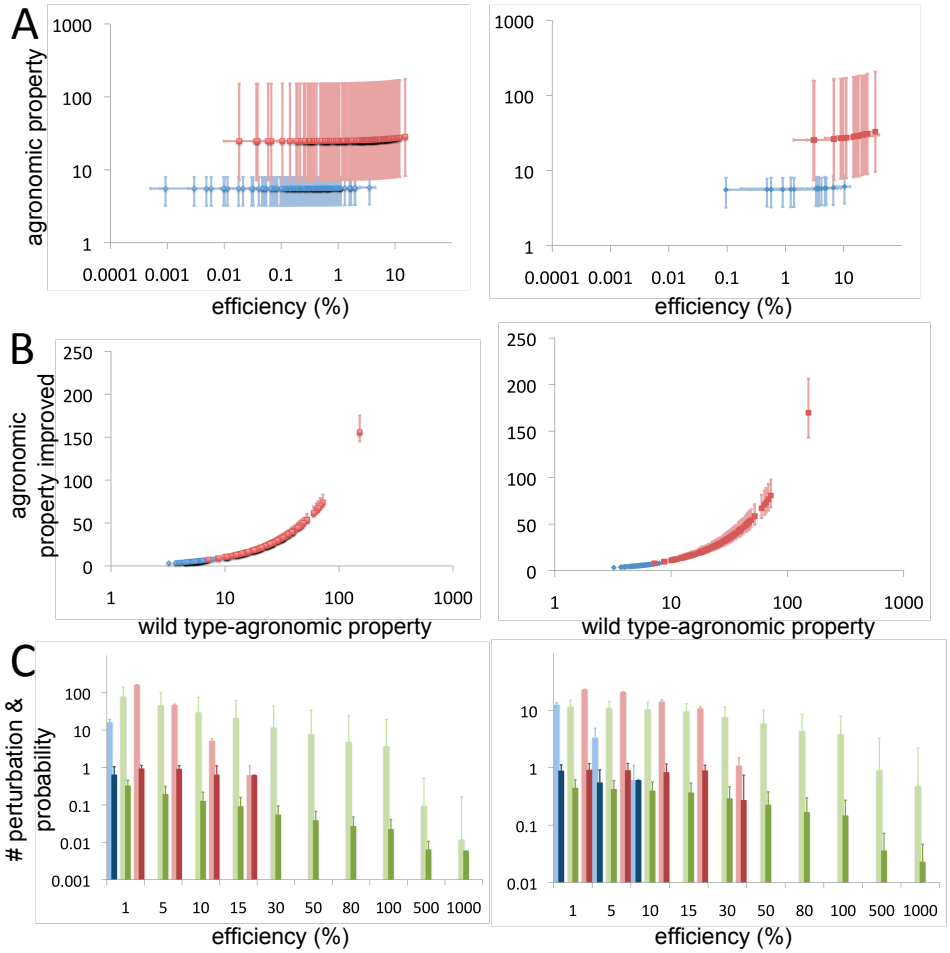

Figure 7.4: Exploration and statistical significance of the landscape of multiple agronomic properties of interest for tomato fruit applying local perturbations in its effective TRN. (A) Agronomic properties improved by perturbing a single gene as function of efficiency reached by that transcriptional perturbation with respect to the wild-type scenario; only perturbations causing positive mean efficiencies are plotted. Both agronomic properties and efficiencies of a single perturbation are tested on the 169 RILs and error bars represent their minimum and maximum values in both axis. (B) Relationship between agronomic properties in the wild-type genome and the average of the agronomic properties resulting of all single perturbations in the wild-type TRN for each RIL; vertical error bars represent the best and worst optimized re-engineered TRN for a given RIL. (C) Average number of single gene perturbations that overcome a given efficiency threshold in the 169 RILs (light bars; error bars represent standard deviation for the 169 RILs) and average probability of selecting the same gene-perturbation in a set of RILs (dark bars; error bars show standard deviation for all genes of the TRN). Left and right columns represent perturbations of single gene in case of knockout or over-expression, respectively. (A, B) show fitness as related to the acceptability of tomato fruit (blue) and production vs. quality (red); (C) and fitness values associated to maximize only fruit quality (green). Agronomic properties are plotted in arbitrary units.

towards a defined value, as it is desired for some biotechnological applications, achieving also in this case high efficiency values.

After this, we ranked the list of knockout/over-expressed genes of the TRN according to two criteria directed to maximize: (i) the mean efficiency across all lineages in the case of goals such as acceptability and quality vs production; and (ii) the average of the maximum agronomic property reached by all possible TRN reconfigurations in the case of fruit quality. Fruit acceptability could be improved to 2.91% or 8.84% using gene knockout (*i.e.*, *LE24K20*) or over-expression (*i.e.*, *LE13M10*) in all lineages, respectively. By contrast, quality was highly increased achieving improvement ratios of 43.34% by gene knockout (*i.e.*, *LE24K20*) and 227.31% by over-expression of *LE15D07*. Finally, taking into account not only the quality but also fruit production, ratios decreased to 15.32% (*i.e.*, *LE13F23*) and 35.94% (*i.e.*, *LE14B20*) using the two types of perturbations, respectively. Notice that all these rates of improvement were achieved in the lineages that provided maximum fitness in the wild-type TRN.

Lineages exhibited variability in their resistance to be optimized and this resistance changed with each target agronomic property. Figure 7.4B shows a strong linear dependence between the level of the agronomic property in the wild-type TRN and the average level of the agronomic properties resulting from all single perturbations in the TRN for each RIL (linear regression slope in the range 0.99 - 1.12 and $R^2 > 0.99$; $p < 0.001$). Interestingly, we observed that the effect of predicting agronomic properties under genetic perturbations was not dependent on the lineage selected. This provided a high level of robustness when we selected the lineages to implement experimentally re-designed TRN.

We computed the average number of single-gene perturbations to overcome an efficiency threshold given in the 169 RILS and the average probability of selecting the same gene-perturbation commonly for the whole set of RILs. The right panel in Figure 7.4C shows that only a few gene knockouts were able to improve fruit acceptability with a high probability in all lineages whereas, on the other hand, tens of gene knockouts could be proposed for increasing fruit quality and for the quality and production. On the other hand, the left panel in Figure 7.4C allowed re-asserting that re-engineering the TRN by gene over-expression could result in higher increments in the agronomic properties and with a higher density of suggested perturbations across the RILs.

The next step in our study was to propose new genome re-designs including multiple perturbations. To do this, we sampled widely the landscape of

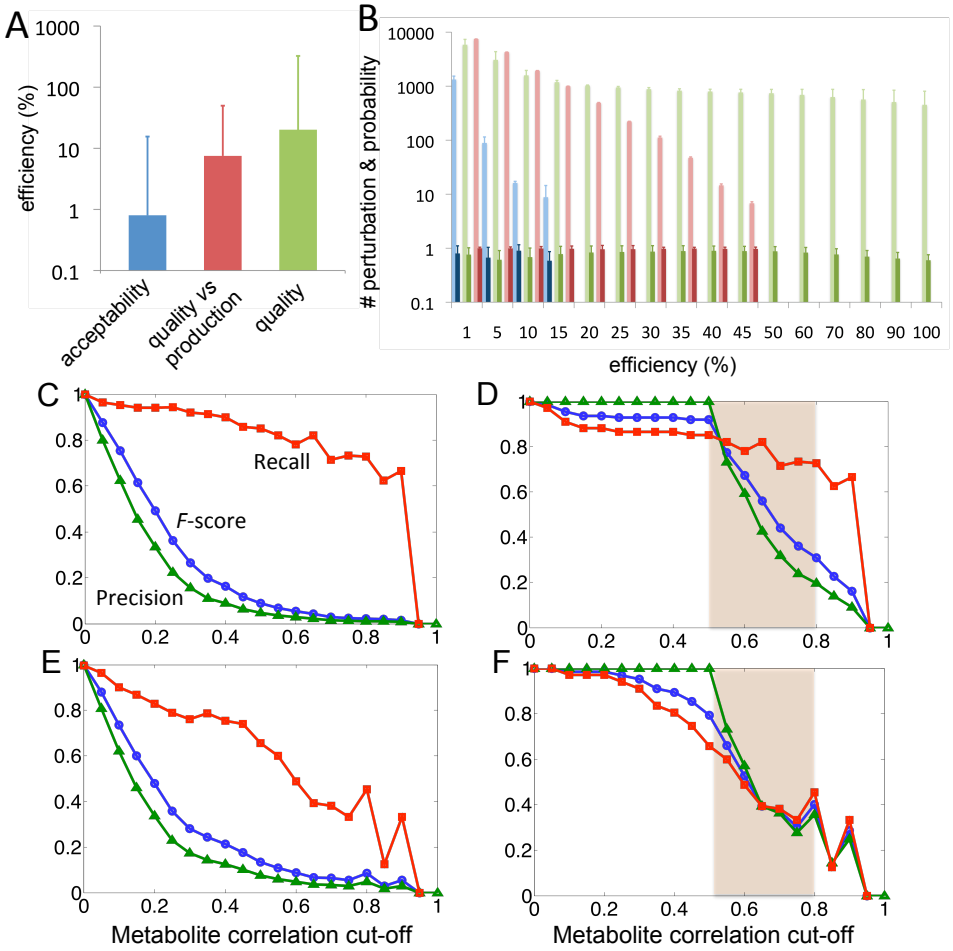

Figure 7.5: Heuristic exploration (A) and statistical significance (B) of the landscape of multiple desired agronomic properties of tomato fruit perturbing its effective TRN adding multiple genetic changes and, predictive power (C-F) for optimizing the levels of volatile compounds and identifying compounds in closed metabolic pathways. (A) Median efficiencies reached by transcriptional perturbation based in gene knockouts or over-expression to improve agronomic properties. (B) Average number of single gene perturbations that overcome an efficiency threshold in the top 5 RILs scored by single perturbation (light bars; error bars represent standard deviation for the selected RILs) and average probability of selecting the same multiple-perturbation commonly in a set of RILs (dark bars; error bars show standard deviation for all genes of the TRN). Precision, recall and F-score (green, red and blue lines, respectively) compare observed experimentally volatile compound correlations vs inferred set of potential genetic perturbations (gene knockout (C, D) or over-expression (E, F)) shared to optimize each compound independently. Note that experimental metabolite correlations < 0.5 were not considered in (D, F).

the acceptability, quality and quality vs production of tomato fruits by introducing two-gene perturbations either by knockouts and over-expressions. Figure 7.5A shows the median efficiencies reached by two-gene transcriptional perturbations based on knockouts and over-expression in order to improve the agronomic properties defined as multiple-objective. As expected, we corroborated that multiple perturbations, located in different pathways, could improve the agronomic properties significantly better than single perturbations. Figure 7.5B shows the average number of single gene perturbations that are able to overcome a given efficiency threshold for the top 5 RILs when ranked for single perturbations as well as the average probability of selecting the same multiple-perturbation commonly in a set of RILs.

7.3 Experimental Validation Via Prections of Volatile Compounds Correlations

After generating our predictive model for the TRN and metabolism of tomato fruit, we use it to automatically design tomato genomes with extreme alterations for each of the 56 volatile compounds by introducing a set of genetic perturbations. We compared sets of genetic perturbations for all pairs of volatile compounds and then inferred their levels of correlations (see Appendix E). Hence, these predicted correlations were compared to the levels of correlations obtained from the experimental values for each volatile pair that often reflects their belonging or not to the same metabolic/ regulatory pathway or to be or not structurally related. Figure 7.5C-F shows the predictive power of our model to determine correlations between all the volatile compounds. Interestingly, selecting a correlation cut-off between 0.5 and 0.8 we obtained high performance F-scores (see Appendix E) ranging between 0.32 and 0.91 (Figure 7.5D) for gene knockouts and between 0.31 and 0.80 when model selected genes by over-expression (Figure 7.5F). Notice that only pairs of experimental volatile compounds with $r > 0.5$ were considered. Predictions decreased when we incorporated all pairs of compounds (Figure 7.5E-F) indicating that our model captured high correlations observed experimentally with more precision. We computed the dendograms of the volatile compound obtained from the correlation of experimentally obtained volatiles levels and the dendograms obtained using as distance between volatile compounds the number of common genetic perturbations proposed by the model. We observed that perturbations proposed by gene over-expression were pivotal to predict computationally significant distances between volatile compounds (Mantel test, $r = 0.54$; $p < 10^{-5}$)

thus providing high support to our model. By contrast, predicted perturbations based on gene knockout could only identify a small fraction of the entire dendogram (Mantel test, $r = 0.38$; $p < 10^{-5}$).

To give further support to our model we designed experimentally two inbred lines (ILs) derived from another interspecific cross whose transcriptome and metabolome were also experimentally measured. We corroborated that a significant set of genetic perturbations suggested by computational design to optimize the phenotype observed were identified as genes differentially altered in the target phenotype.

7.4 Conclusions

LM is a methodology that is being implemented by large industries to optimize their production. In the process of decision making applied to the redesign of production systems, firstly, engineers evaluate systematically the addition or elimination of resources in each of the participating single processes; afterwards, multiple changes are considered trying to achieve maximum quality and production [29]. Translating this engineering approach to a cellular molecular factory and identifying the basic functional elements has allowed us to develop a design methodology which optimizes the genome as it should result in a more desirable phenotype [25]. In addition, by mimicking the methodology from LM we have provided a first robust optimization to redesign an optimal genetic network based on the systemic exploration of the effects of a large number of single gene knockout and over-expression genotypes; then, a second multiple-optimization of random paths allowed improving substantially the desired agronomical properties. The success of this approach indicates that despite the existence of molecular interactions, the model is able to overcome this limitation and results in a good predictor.

We have proposed several re-engineered genomes that improve desired agronomic properties of the fruit by targeting single or multiple genetic modifications. It has been previously reported that single under-/over-expressed of certain genes may affect fruit quality traits, being these key genes involved in the biosynthesis of a product of fruit metabolism or to a general ripening regulators (*i.e.*, carotenoids [30]). We have explored single perturbations by gene knockout or over-expression and our results indicated that a significantly better fine-tuning could be obtained by using over-expression approaches. We observed that improvement ratios could reach even more than 4-fold the wild-type value of most of phenotypes desired by designing genomes with only two genetic perturbations (Figure 7.5A). The magnitude of the predicted change sometimes may appear low but an improvement in a

quantitative trait, if consistent and predictable, maybe economically important. Indeed, a good combination of high yield with even slightly increased solid solids content is a major breeding goal for processing tomatoes that it is difficult to be achieved [31] because of polygenic nature and pleiotropic relationships of both traits [32].

Although it is not the objective of this paper, it does not escape our attention that some of the perturbations proposed are consistent with the biological processes associated to the trait and therefore the model could be used to reveal the molecular underpinnings of quality traits. For instance the role of *YABBY* (a gene proposed by our model to affect quality) in controlling fruit size probably through the auxin pathway and the effect of auxin in altering fruit growth and ripening has been previously reported [33, 34]. Similarly the importance of phytoene desaturase to affect carotenoids and carotenoid derived volatiles has been reported [35]. Most of the genes proposed by the models however are new, therefore opening new avenues of research either by targeting in transgenic plants, identification of mutants in those genes by TILLING [36] or by TAL engineering [37], as well as to be used as an additional guide during plant breeding. In principle these modifications are to be implemented in red fruit or around red fruit stage either genetically or by the use of external elicitors (physical or chemical) and our model provides roadmap for those approaches. Our methodology takes advantage of our ability to predict variations in fruit cell phenotype based on changes in the transcriptome. The linear relationships shown in Figure 7.4 (A, C, and D) guarantee that by optimizing our effective transcriptomic, metabolic or phenotypic fitness we are also optimizing the phenotype measured experimentally of the tomato fruits. While it is true that complex multi-organ organism such as tomato rely on the coordination and transport of multiple signals and nutrients from different parts of the plants to achieve the final phenotype, and this is especially true for the fruit [20, 38], it not less true that the most important part of the fruit characteristics at ripening depends basically on the fruit program before around the ripening stage [39, 40].

The ability to target redesign crops for enhanced content of metabolites of interest has been experimentally achieved in a number of cases (for instance vitamin C [41]; vitamin E [42]) using transgenic approaches and the information of bottlenecks or limiting steps for the biochemical pathways of the compounds of interest. The most dramatic examples of this have been introducing the new trait in a background with very low value for it (*i.e.*, golden rice [43]) using ectopic expression of one or several foreign genes.

The use of natural genetic variability in combination with our nonbiased (hypothesis-free) modeling approach allows us to identify new candidate genes as potential targets to engineer the plant (although the biotechnological use of more active orthologs from other organisms is not discarded in our approach). The existence of regulatory networks connecting primary and secondary metabolism in plants should also be taken into consideration in future attempts to metabolically engineer the various classes of plant secondary metabolites [44]. It is interesting that known genes in the biosynthesis path often do not co-localize with quantitative trait locus for the metabolites in the path [35] indicating that there is ample of opportunities to be explored for metabolite and quality improvement, and our model fits nicely in this gap.

Appendix A Plant Material, Transcriptomic, Metabolomic and Phenomic Data

The construction of the tomato RILs used in this study has been described elsewhere [45]. Triplicate samples of red ripe fruits (each representing at least 5 fruit) from each of 169 RILs were harvested and analyzed for volatile compounds as described in [46]. For method validation, red ripe fruits from five ILs with a different genetic background [47] were used. Transcript profile datasets (11,876 x 3 x 50 data points) were obtained from triplicate fruit samples of 50 selected RILs using TOM2 microarray, as previously reported [48]. Data sets corresponding to the rest of metabolites and phenomic data were obtained as in [46] from triplicate samples of the 169 RILs. To decrease experimental variability, the same fruits representing each RIL were homogenized and divided in different aliquot samples for the different metabolite or transcript profiling techniques. Before use all transcriptomic, metabolomic and phenomic data were normalized and transformed to log-scale. The ILs used for model validation have been described previously [38].

Appendix B Mathematical Model

An effective linear model based on ODEs each providing the steady states of tomato fruit mRNA was used to describe transcriptional gene regulations [7]. Thus, the mRNA steady state from the i^{th} gene, g_i, is given by $\frac{dg_i}{dt} = \sum_j \theta_{ij} g_j - \delta^g g_i + \Delta_i$, where θ_{ij} represents the regulatory effect that gene j has on gene i. Each gene expression value is contained ($\xi g_i^{min} \leq g_i \leq \xi^{-1} g_i^{max}$) in a range interval defined by the minimum (g_i^{min}) and maximum

(g_i^{max}) value of all its experimental measurements obtained from the subset of 50 RILs used for transcript profiling. $\xi \leq 1$ is a tunable parameter that decreases the gene expression range to improve the predictive capacity of the presented model under genetic predictions. The dynamics of metabolic profile was computed by $\frac{dm_i}{dt} = \sum_j \gamma_{ij} g_j - \delta^m m_i + \Gamma_i$, where m_i is the steady-state concentration from the i_{th} metabolite, γ_{ij} is the regulatory strength that gene j has on metabolite i. Hence, agronomic variables (AV) were predicted by means of a linear combination of the metabolic profile, $AV_i = \sum_j \beta_{ij} m_j + \Omega_i$, where β_{ij} is the regulatory effect that metabolite j has on agronomic variable i. Δ, Γ and Ω are the perturbation terms that allow to calibrate gene expression, metabolic profiles and predicted agronomic properties, respectively, for all RILs. Notice that degradation coefficients of genes and metabolites ($\delta^g = \delta^m = 1$, respectively) scaled time conveniently and that we assumed the model in steady state ($g_i = \sum_j \theta_{ij} g_j + \Delta_i$ and $m_i = \sum_j \gamma_{ij} g_j + \Gamma_i$).

Appendix C Construction of an Effective Transcriptional Regulatory Network Connected with Metabolism to Explain Agronomic Properties

Our global model consists of three blocks of algebraic equations covering respectively from gene expression, through metabolic profile until agronomic properties, and in all three cases the same methodology was applied. The inference procedure consisted of two nested steps. Firstly, the network connectivity was inferred by using the *InferGene* algorithm [7]. This method uses mutual information with a local significance value (z computation) to obtain the effective regulations. Hence, the potential interaction between a predictor and a target is z-scored, constituting an estimator of the likelihood of mutual information. Subsequently, we selected a z threshold for a predictor cutoff. In a second step, LASSO method was used to avoid over-fitting and to estimate the kinetic parameters of each effective model. Notice that the 8.7% of the selected genes in the TRN were annotated as TFs and 16.2% as encoding enzymatic activities and, in neither case, they were over-represented since both the tomato genome and the whole array contain similar fractions of TFs (8.8%) and enzymes (17.1%).

For the construction of the effective TRN model and its later integration with the metabolism, we used steady-state mRNA expression profiles derived from RILs transcriptionally and metabolically characterized. The dataset contains pre-processed expression data from 50 x 3 hybridization

experiments using an array with 11,876 probe sets spotted, and data for levels of 67 metabolites that were quantified over the same sample set. For this study, we only considered the 5592 genes whose expression values could be consistently found in more than 80% of the microarrays. We found 1057 TFs and 1962 genes with enzymatic activity after searching for the motifs transcription regulator and enzyme activity respectively in the functionally annotated tomato genome (*Tom*2). Moreover, all 169 RILs (including the previous 50 ones) for which we had metabolite and phenotype data were used to train a linear model able to predict agronomic properties of the fruit from potentially predictor metabolites. In all cases transcriptomic and metabolomic data were first normalized using the LOWESS procedure [26] and subsequently converted into z across the RILs. In order to calibrate gene expression and metabolite concentration, both models included a perturbation term (Δ_i^{RIL} and (Γ_i^{RIL}, respectively) to fit all their i-genes and j-metabolites for a given RIL. We assumed a constant perturbation in the gene expression prediction because of its low variation across the training set (standard deviation of $\langle \Delta/g \rangle$ for all RILs is 0.072-fold the standard deviation of gene expression, $\langle \sigma^g \rangle_{RIL}$) with respect to the mean value, $0.22 \langle \sigma^g \rangle_{RIL}$. Similarly, the average error to predict the metabolic profile across the training set was increased to $0.99 \langle \sigma^g \rangle_{RIL}$.

Appendix D Genome-Wide Multiple-Optimization

Our algorithm searches possible reconfigurations of the global effective transcription regulatory network of tomato such as that the specified agronomic properties are improved (maximized or minimized) with respect to the properties of interest obtained in a given RIL. Different properties of interest have been optimized, ranging from single metabolites defining the sweetness or sourness of the fruit, to linear combinations of a set of metabolites determining the quality in terms of flavor and taste and even further to include objective functions that try to integrate two of those goals with a trade-off and balanced weighting factors such as fruit quality and yield.

We have addressed this optimization problem using two approaches. Firstly, we exhaustively enumerated all possible single gene knockouts and over-expression for each case to be optimized under a given selective pressure of interest. Second, we ranked all possible perturbations according to the new agronomic properties they would generate. The third step was to suggest genome reconfigurations that include multiple actions: gene knockouts,

over-expressed genes, or both, in order to enlarge the combinatorial space of perturbed genomes. To do that, we have used an exhaustive method aimed at finding the global optimum in the space of all possible synthetic TRN. We started from the inferred model (see Appendix A) and applied an optimization scheme. At each step of the optimization process, we selected each gene among the ones involved in the transcriptomic model to evaluate the effect of three possible approaches (knockout, over-expression or wild-type scenario); we updated the model with the genetic perturbation that provided the best score. Note that to simulate knockout or over-expression in the gene i, we substituted its ODE by the minimum (ξg_i^{min}) or maximum ($\xi^{-1} g_i^{max}$) values respectively observed in the range of diversity of the 50 RILs.

Appendix E Experimental and Computational Metabolite Correlation

We computed the sets of single-gene perturbations, Δ, by gene knockout or over-expression that alter significantly the levels of the 56 volatile metabolites representing the volatile compounds taking into account the global model. For the sake of the model we considered only those gene perturbations that would cause significant changes in the metabolite concentration higher than 1% ($p < 0.01$). Δ can be divided into genetic modifications that increase (Θ) or decrease (Ξ) the metabolite concentrations, respectively. Hence, correlations between metabolite pairs i and j (C_{ij}) were calculated as the difference between C_{ij}^{+} and C_{ij}^{-} by using the set of single-gene perturbations proposed by the model $C_{ij}^{+} = max\left(\frac{\Theta_i \cap \Theta_j}{\Theta_i \cup \Theta_j}, \frac{\Xi_i \cap \Xi_j}{\Xi_i \cup \Xi_j}\right)$ and $C_{ij}^{-} = max\left(\frac{\Theta_i \cap \Xi_j}{\Theta_i \cup \Xi_j}, \frac{\Xi_i \cap \Theta_j}{\Xi_i \cup \Theta_j}\right)$.

where C_{ij}^{+} and C_{ij}^{-} is the maximum normalized intersection predicted between the set of gene perturbations proposed by altering positively or/and negatively, respectively. We used these correlations to compute dendograms of all volatile compounds by using the distance inferred by the model $(1-C_{ij})$ depending on the Δ selected by gene knockout or over-expression. The performance of the inferred metabolite correlations was evaluated using as a reference a set of empirical correlations previously obtained among these metabolites. We used different cut-offs, k, to identify metabolite correlations ($C_{ij} > k$). The fraction of metabolite pairs that were correctly predicted by the model (precision, P) and the fraction of all known correlations that were discovered by the model (sensitivity, S) were used to compute a performance statistic defined as $F = 2PS/(P + S)$.

References

[1] Baker, D., Church, G., Collins, J. J., Endy, D., Jacobson, J., Keasling, J., Modrich, P., Smolke, C., Weiss, R. (2006). Engineering life: building a fab for biology. *Sci. Am.* 296, 44-51.

[2] Endy, D. (2005). Foundations for engineering biology. *Nature* 438, 449-453.

[3] Knight, T.F. (2005). Engineering novel life. *Mol. Syst. Biol.* doi:10:1038/msb4100028.

[4] Andrianantoandro, E., Basu, S., Karig, D.K., Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.* doi:10.1038/msb4100073.

[5] Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L., Palsson, B. O. (2008). Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129-143.

[6] di Bernardo, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., Wojtovich, A. P., Elliott, S. J., Schaus, S. E., Collins, J. J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotech.* 23, 377-383.

[7] Carrera, J., Rodrigo, G., Jaramillo, A. (2009). Model-based redesign of global transcription regulation. *Nucleic Acids Res.* 37, e38.

[8] Carrera, J., Rodrigo, G., Jaramillo, A., Elena, S.F. (2009). Reverse-engineering the *Arabidopsis thaliana* transcriptional network under changing environmental conditions. *Gen. Biol.* 10, R96.

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

[10] Bonneau, R. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131, 1354-1365.

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

[12] Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92-96.

[13] Endy, D., Brent, R. (2001). Modelling cellular behaviour. *Nature* 409, 391-395.

[14] Joyce, A. R., Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* 7, 198-210.

[15] Burgard, A. P., Pharkya, P., Maranas, C. D. (2003). OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647-57.

[16] Segre, D., Vitkup, D., Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15112-15117.

[17] Rocha, M., Maia, P., Mendes, R., Pinto, J. P., Ferreira, E. C., Nielsen, J., Patil, K. R., Rocha, I. (2008). Natural computation meta-heuristics for the *in silico* optimization of microbial strains. *BMC Bioinfo.* 9, 499.

[18] Fraser, P. D., Enfissi, E. M. A., Bramley, P. M. (2009). Genetic engineering of carotenoid formation in tomato fruit and the potential application of systems and synthetic biology approaches. *Arch. Biochem. Biophys.* 483, 196-204.

[19] Meyer, R. C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Torjek, O., Fiehn, O., Eckardt, A., Willmitzer, L., Selbig, J., Altmann, T. (2007) The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 4759-4764.

[20] Mounet, F., Moing, A., Garcia, V., Petit, J., Maucourt, M., Deborde, C., Bernillon, S., Le Gall, G., Colquhoun, I., Defernez, M., Giraudel, J. L., Rolin, D., Rothan, C., Lemaire-Chamley, M. (2009). Gene and metabolite regulatory network analysis of early developing fruit tissues highlights new candidate genes for the control of tomato fruit composition and development. *Plant Physiol.* 149, 1505-28.

[21] Garcia, V., Stevens, R., Gil, L., Gilbert, L., Gest, N., Petit, J., Faurobert, M., Maucourt, M., Deborde, C., Moing, A., Poessel, J. L., Jacob, D., Bouchet, J. P., Giraudel, J. L., Gouble, B., Page, D., Alhagdow, M., Massot, C., Gautier, H., Lemaire-Chamley, M., de Daruvar,

A., Rolin, D., Usadel, B., Lahaye, M., Causse, M., Baldet, P., Rothan, C. (2009). An integrative genomics approach for deciphering the complex interactions between ascorbate metabolism and fruit growth and composition in tomato. *C. R. Biol.* 32, 1007-21.

[22] Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotech.* 24, 447-54.

[23] Osorio, S., Alba, R., Damasceno, C. M., Lopez-Casado, G., Lohse, M., Zanor, M. I., Tohge, T., Usadel, B., Rose, J. K., Fei, Z., Giovannoni, J. J., Fernie, A. R. (2011). Systems biology of tomato fruit development: combined transcript, protein, and metabolite analysis of tomato transcription factor (nor, rin) and ethylene receptor (Nr) mutants reveals novel regulatory interactions. *Plant Physiol.* 157, 405-25.

[24] Rohrmann, J., Tohge, T., Alba, R., Osorio, S., Caldana, C., McQuinn, R., Arvidsson, S., van der Merwe, M. J., Riao-Pachn, D. M., Mueller-Roeber, B., Fei, Z., Nesi, A. N., Giovannoni, J. J., Fernie, A. R. (2011). Combined transcription factor profiling, microarray analysis and metabolite profiling reveals the transcriptional control of metabolic shifts occurring during tomato fruit development. *Plant J.* 68, 999-1013.

[25] Carrera, J., Elena, S. F., Jaramillo, A. *Escherichia coli* genome can be refactored into fewer operons while still maintaining its environmental responsiveness (submitted).

[26] Magniette, F., Renou, J. P., Daudin, J. J. (2008). Normalization for triple-target microarray experiments. *BMC Bioinf.* 9, 216.

[27] Malundo, T. M. M., Shewfelt, R. L., Scott, J. W. (1995). Flavor quality of fresh tomato (*Lycopersicon esculentum* Mill.) as affected by sugar and acid levels. *Postharvest Biol. Tech.* 6, 103-110.

[28] Buttery, R. G., Teranishi, R., Flath., R. A., Ling, L. C. (1989). Fresh tomato volatiles: Composition and sensory studies, p. 213-222. In: R. Teranishi, R.G. Buttery, and F. Shahidi (eds.). *Flavor chemistry: Trends and developments.* Amer. Chem. Soc., Washington, D.C.

[29] Shaha, R., Ward, P. T. (2003). Lean manufacturing: context, practice bundles, and performance. *J. Operations Management* 21, 129-149.

[30] Rosati, C., Diretto, G., Giuliano, G (2010). Biosynthesis and engineering of carotenoids and apocarotenoids in plants: state of the art and future prospects. *Biotechnol. Genet. Eng. Rev.* 26, 139-62.

[31] Fridman, E. F., Liu, Y. L., Carmel-Goren, L., Gur, A., Shoresh, M., Pleban, T., Eshed, Y., Zamir, D. (2002) Two tightly linked QTLs modify tomato sugar content via different physiological pathways. *Mol. Genet. Genom.* 266, 821-826.

[32] Emery, G. C., Munger, H. M. (1970). Effects of inherited differences in growth habit on fruit size and soluble solids in tomato. *J. Amer. Soc Hort. Sci.* 95, 410-412.

[33] Cong, B., Barrero, L. S., Tanksley, S. D. (2008). Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat. Genet.* 40, 800-4.

[34] Wang, H., Schauer, N., Usadel, B., Frasse, P., Zouine, M., Hernould, M., Latch, A., Pech, J. C., Fernie, A. R., Bouzayen, M. (2009). Regulatory features underlying pollination-dependent and -independent tomato fruit set revealed by transcript and primary metabolite profiling. *Plant Cell* 21, 428-52.

[35] Klee HJ (2010). Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. *New. Phytol.* 187, 44-56.

[36] Minoia, S., Petrozza, A., DOnofrio, O., Piron, F., Mosca, G., Sozio, G., Cellini, F., Bendahmane, A., Carriero, F. (2010). A new mutant genetic resource for tomato crop improvement by TILLING techonology. *BMC Res. Notes* 12, 3-69.

[37] Bogdanove, A. J., Voytas, D. F. (2011). TAL effectors: customizable proteins for DNA targeting. *Science* 333, 1843-6.

[38] Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotech.* 24, 447-54.

[39] Hetherington, S., Smillie, R., Davies, W. (1998). Photosynthetic activities of vegetative and fruiting tissues of tomato. *J. Exp. Bot.* 49, 1173.

[40] Fridman, E., Carrari, F., Liu, Y. S., Fernie, A. R., Zamir, D. (2004). Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* 305, 1786-9.

[41] Agius, F., Gonzlez-Lamothe, R., Caballero, J. L., Muoz-Blanco, J., Botella, M. A., Valpuesta, V. (2003). Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. *Nat. Biotech.* 21, 177-81.

[42] Cahoon, E. B., Hall, S. E., Ripp, K. G., Ganzke, T. S., Hitz, W. D., Coughlan, S. J. (2003). Metabolic redesign of vitamin E biosynthesis in plants for tocotrienol production and increased antioxidant content. *Nat. Biotech.* 21, 1082-7.

[43] Ye, X., Al-Babili, S., Klti, A., Zhang, J., Lucca, P., Beyer, P., Potrykus, I. (2000). Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* 87, 303-5.

[44] Aharoni, A., Galili, G. (2011) Metabolic engineering of the plant primary-secondary metabolism interface. *Curr. Opin. Biotechnol.* 22, 239-44.

[45] Alba, J. M., Montserrat, M., Fernandez-Munoz, R. (2009). Resistance to the two-spotted spider mite (Tetranychus urticae) by acylsucroses of wild tomato (*Solanum pimpinellifolium*) trichomes studied in a recombinant inbred line population. *Exp. App. Acar.* 47, 35-47.

[46] Zanor, M. I., Rambla, J. L., Chaib, J., Steppa, A., Medina, A., Granell, A., Fernie, A., Causse M. (2009). Metabolic characterization of loci affecting sensory attributes in tomato allows an assessment of the influence of the levels of primary metabolites and volatile organic contents. *J. Exp. Bot.* 60, 2139-2154.

[47] Eshed, Y., Zamir, D. (1995). An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* 141, 1147-1162.

[48] Lytovchenko, A., Eickmeier, I., Pons, C., Szecowka, M., Lehmberg, K., Arrivault, S., Tohge, T., Pineda, B., Anton, M. T., Hedtke, B., Lu, Y., Gupta, K. J., Fisahn, J., Bock, R., Stitt, M., Grimm, B., Granell, A., Fernie, A. R. (2011). Tomato fruit photosynthesis is seemingly unimportant in primary metabolism and ripening but plays a considerable role in seed development. *Plant Physiol.* 157, 1650-63.

Chapter 8

General Discussion

A milestone in the computational genome design will be the integration of the transcriptomic and signaling networks in the metabolic model of the cell. The coordinated expression of hundreds, even thousands, of genes in bacteria could be seen as the result of a program encoded in the DNA, at a given time point of the cell cycle, and in response to external stimuli. The cellular machinery, especially polymerases and ribosomes, would play the role of compiling the program into proteins that perform the biological functions.

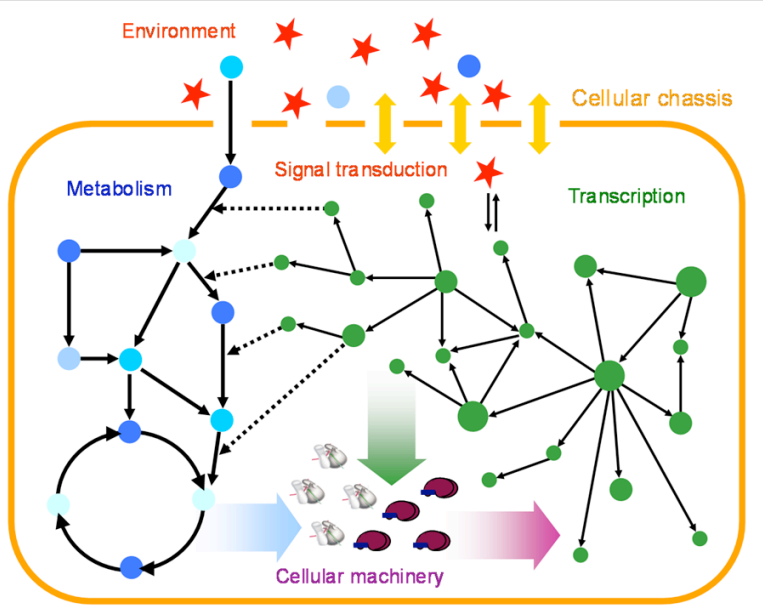

Figure 8.1: Scheme of the cellular chassis involving transcription, metabolism, machinery and signal transduction. The modules are interconnected and are context-dependent.

The first step to construct a global multi-scale model (Figure 8.1) consists of modeling the steady state of the cell. Since transcription is context-dependent, signaling mechanisms (*e.g.*, two-component or transportation systems) allow the cell to sense the environment and act accordingly. Basically, varying compounds (*e.g.*, glucose, oxygen or metals) and external factors (*e.g.*, temperature, pH or light) regulate genes, especially those with regulatory functions [1]. For the purposes of this review we could assume for simplicity that the sensory machinery is only constituted by TFs and the signal transduction machinery. Bottom-up signaling networks [1, 2, 3] are combined with optimization-based methods,[4] which can overcome the lack of knowledge on the structure of signal transduction pathways, to elucidate the global sensory machinery. Furthermore, by using -omics data from environmentally perturbed cells, regulatory models integrating external influences can be inferred [5]. In steady state, the gene expression profile, in particular those enzymes catalyzing biochemical reactions, specifies the cell metabolism in a given environment. In that way, regulatory FBA models have been developed to predict phenomic data in bacteria [6]. This data consists on a high-throughput analysis of cell growth using multiple plate readers with cells in wells with defined conditions. Thus, integrated metabolic, transcription and signaling networks [6, 7, 8, 9] can better explain physiological and intracellular changes and constitute the next step towards the project of modeling a cell. There are software initiatives such as the *E*-cell project [10, 11] that could be used to implement in a user-friendly software package the methodologies presented here, once they become accurate enough.

The continuous developments in the sequencing and synthesis of DNA [12] have extended our understanding of the workings of living organisms. Synthetic genomics is focused on the powerful step of synthesizing and programming genetic material, DNA or RNA. Several works have been focus on the synthesis, assembly, and cloning of a viral genome [13, 14, 15, 16, 17]. In that direction, the J. Craig Venter Institute has proposed an approach to produce reduced genome of a mycoplasma by complete chemical synthesis [18]. Moreover, they verify the capacity of the cells with the new genome to provide the essential genetic functions for life.

The main ingredients to obtain a methodology for automatic genome design are the same that are needed in any evolutionary procedure: (i) A given genome to be used as starting point, (ii) evolutionary steps and (iii) fitness function. We could use well-characterized plasmids and minimal genomes for (i). Step (ii) would require using a modular approach where the genome

is decomposed in elementary modules (such as the biological part models), then the evolutionary procedures would add/remove/modify such modules. In this step is very important that the genomes used in (i) be also completely understood in terms of such modules. This would be an impossible task in general, as modularity is not enforced in natural genomes. This will require the future refactoring of natural genomes making them modular enough. The bottleneck here is in the appropriate understanding of the different genetic elements in different genomic contexts, which requires a large community effort. The enforcing of standardization and characterization of genetic parts worldwide will be key in this endeavor.

The step (iii) is the most complicated of all, as it requires a quantitative model of the whole cell able to predict cell growth for a given genome. In this respect, the integration of the transcription with metabolism is poorly understood. But we also need an appropriate incorporation of signal transduction or cell machinery. Also it is unclear whether we really need a spatial model to predict cell growth. The state-of-the-art here is in the use of high-throughput data to infer global models. As we have shown in the previous section, this has been done with relative success for the metabolism, but a proper coupling with the transcription regulation is still missing. When operons are activated or repressed by transcription factors they change the expression level of enzymes that in turn affect the metabolic fluxes. In addition, there are the post-transcriptional and post-translational regulations, which have a prevalent role in signal transduction.

The construction of large-scale models by means of reverse-engineering methods has allowed to predict quantitatively the cellular response under global redesign. However, the missing significant elements in the initial models could lead to inaccurate designs. Experimental noise on the gene expression and missing/inaccurate reactions in the metabolism are examples of possible deficiencies in the predictive value of the models. These inaccuracies lead to some limitations to the feasibility of designing synthetic genomes. One way to extrapolate on the limitations of the genome design is by estimating the errors of predicting cell growth. As the predictions are done using a model trained with experimental data for a given genome, as soon as we start modifying the genome, the predictions will start to degrade. This could be seen as the propagation of an error: every time we modify the genome (by making a knockout or modifying a promoter) the error in prediction will add up to the previous one. Towards this end, Figure 8.2 shows how the propagation of error between the predicted and experimental measurement increases as the genetic modifications imposed by a redesign

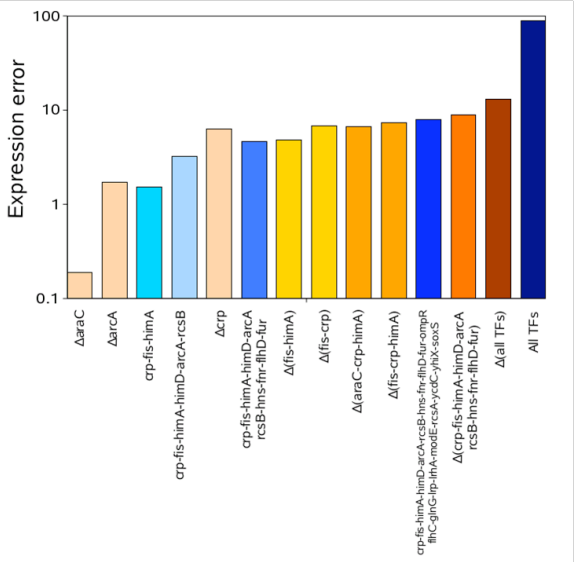

Figure 8.2: Normalized deviation with respect to the wild-type of gene expression for different genome transcriptional perturbations involving knockouts and changing the regulation of TFs. We have considered single and multiple knockouts (prefixed with a delta symbol), which are shown in different grades of red and yellow. On the other hand, we have also considered changing the promoter controlling TFs, to simplify the notation we have named the promoters after their downstream TF.

at transcriptional level.

The development of a computational platform to design genomes requires large-scale models able to predict the behavior of a set of genes. We have discussed various genome-scale techniques that will allow the development of predictive tools for genome design. The main difficulty resides in the quantitative prediction of cell growth from a genome in such a way that our predictions remain valid when we reshuffle and evolve the genome. We started with the design of metabolic pathways, where current tools are able to predict the growth rate given a genome-scale metabolic model. For the design of synthetic genomes, it is important to have tools that are accurate enough to predict large modifications of a natural metabolism. As a first step towards the *de novo* design of genomes, we should be able to redesign existing genomes by modifying large parts of its metabolic network. The modification of such metabolic pathways can be performed by using enzymes from various organisms, which would have to be later codon-optimized for improved expression in the targeted host. Even after circumventing possible problems with protein expression (we could get protein aggregates), we should also face another problem: the matching of enzyme expressions. Most often the level of expression of each enzyme from a heterologous metabolic

pathway is different from its original one. It is reasonable to assume that in the wild-type organism the expression of enzymes is optimized to maximize cell growth, but after grafting the enzymes from various organisms we will have to re-optimize their expression. This is more important for the case where the enzymes are placed polycistronically. Recently, the group of Keasling proposed a method [24] to optimize the relative expression of enzymes, which could be in the same operon, by modifying the nucleotide sequence around the ribosome-binding site. For the time being, this optimization can only be performed experimentally. It will be very useful if this post-transcriptional optimization of relative enzyme expression within an operon could be done in such a way that the flux of the operons pathway is maximized, so the operon could be later introduced in alternative organisms after an optimization through transcriptional regulation. This modularity is one of the essential elements in synthetic biology. We should also take care of including the appropriate essential genes.

After having designed and optimized the metabolic pathways, placed into suitable synthetic operons, we have to choose the appropriate promoters and TFs. This could be viewed as a second optimization of the expression of our enzymes, but this time we could change the regulation dynamically. Here it is important to have a model of signal transduction in order to know the effects of external signals on the TFs. Once we have a model for the concentration of transcription factors in a given environmental condition we can use recent quantitative models of global transcription to predict the global expression profile of all operons. As a challenge for the future will be the integration of this global transcription model with the metabolic model. This is an important step after which we would be able to use computational design to evolve new genomes that could better adapt to a given temporal pattern of variability.

As already happens with the computational design of proteins [25] the computational genome design could be improved by using directed evolution. It is also expected that new computational methodologies will arise that will suggest the appropriate experiments, such as the most appropriate nucleotide sequences to be randomized.

Finally knowing the operons, promoters, transcription factors and protein transducers, is not enough to obtain a synthetic genome sequence. We need to place the operons in the genome in suitable position and orientations. The genome should also contain the necessary elements for replication. It has been argued [23] that transcription factors should be placed close to their regulating operons. There are also new attempt to use physical mod-

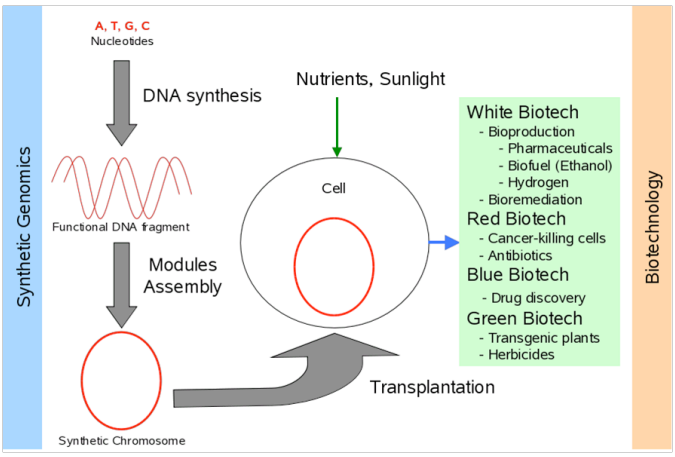

Figure 8.3: Schematic for synthetic genomics and biotechnology. A synthetic chromosome is designed, synthesized and transplanted into the host. The cell has a new code but the same compiler. Those cells perform tasks for many biotechnological applications, such as the bio-production of chemical compounds of interest.

els to represent the genome in 3D, [20] where a local DNA model based on the nucleotide sequence is used to predict the global structure of the chromosome. Once this methodology incorporates longer-range physical models it will be able to be used to engineer a given 3D genome topology.

New microfluidic techniques are allowing the production of more reproducible measures of cell function [26]. In particular, the availability of microchemostats is allowing cell growth in controlled environments over long periods. Further, now that it is also possible to measure gene expression at spatial resolutions, new quantitative models that incorporate stochasticity will be required.

In this thesis, we have discussed the methodologies that could be used for the automatic design of genomes. This will help to stimulate the development of an integrated software platform able to design genomes using unsupervised methods. The engineering of genomes will open new avenues in science and biotechnology (see Figure 8.3) and will require automatic methodologies able to design genomes with targeted functions.

References

[1] Wall, M. E., Hlavacek, W. S., Savageau, M. A. (2004). Design of gene circuits: lessons from bacteria, *Nat. Rev. Genet.* 5, 34-42.

[2] 101 Martinez-Antonio, A., Janga, S. C., Salgado H., Collado-Vives, J. (2006). Internal-sensing machinery directs the activity of the regulatory

network in *Escherichia coli. Trends Microbiol.* 14, 22-27.

[3] Hlavacek, W. S., Faeder, J. R., Blinov, M. L., Posner, R. G., Hucka M., Fontana, W. (2006). Rules for modeling signal-transduction systems. *Sci. STKE* 344, r6.

[4] Dasika, M. S., Burgard, A., Maranas, C. D. (2006). A computational framework for the topological analysis and targeted disruption of signal transduction networks. *Biophys. J.* 91, 382-398.

[5] Bonneau, R., Facciotti, M. T., Reiss, D. J., Schmid, A. K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M. H., Bare, C., Longabaugh, W., Vuthoori, M., Whitehead, K., Madar, A., Suzuki, L., Mori, T., Chang, D.E., DiRuggiero, J., Johnson, C.H., Hood L., Baliga, N. S. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131, 1354-1365.

[6] Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429, 92-96.

[7] Shloni, T., Eisenberg, Y., Sharan, R., Ruppin, E. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.* 3, 10.

[8] Covert, M. W., Xiao, N., Chen, T. J., Karr, J. R. (2008). Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli. Bioinformatics*, 24, 2044-2050.

[9] Lee, J. M., Gianchandani, E. P., Eddy, J. A., Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic and regulatory networks. *PLoS Comput. Biol.*, 4, e1000086.

[10] Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C., Hutchison, C. (1999). E-CELL: Software environment for whole cell simulation. *Bioinformatics* 15, 72-84.

[11] K. Takahashi, N. Ishikawa, Y. Sadamoto, S. Ohta, A. Shiozawa, F. Miyoshi, Y. Naito, Y. Nakayama and M. Tomita, E-Cell 2: Multi-platform E-Cell Simulation System, Bioinformatics, 2003, 19, 17271729.

[12] Endy, D. (2008). Reconstruction of the genomes. *Science* 319, 1196-1197.

[13] Blight, K. J., Kolykhalov, A. A., Rice, C. M. (2000). Efficient Initiation of HCV RNA replication in cell culture. *Science* 290, 1972.

[14] Cello, J., Paul, A. V., Wimmer, E. (2002). Chemical synthesis of Poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297, 1016.

[15] Smith, H. O., Hutchison III, C. A., Pfannkoch, C., Venter, J. C. (2003). Generating a synthetic genome by whole genome assembly: fX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 100.

[16] Kodumal, S. J., Patel, K. G., Reid, R., Menzella, H. G., Welch, H. G., Santi, D. V. (2004). Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15573.

[17] Lartigue, C., Glass, J. I., Alperovich, N., Pieper, R., Parmar, P. P., Hutchison III, C. A., Smith, H. O., Venter, J. C. (2007). Genome transplantation in bacteria: changing one species to another. *Science* 317, 632-638.

[18] Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., Stockwell, T. B., Brownley, A., Thomas, D. W., Algire, M. A., Merryman, C., Young, L., Noskov, V. N., Glass, J. I., Venter, J. C., Hutchison III, C. A., Smith, H. O. (2008). Complete chemical synthesis, assembly and cloning of a *Mycoplasma genitalium* genome. *Science* 319, 1215-1220.

[19] Carlson, R. (2003). The pace and proliferation of biological technologies. *Biosecur. Bioterror.* 1, 203-214.

[20] Herisson, J., Ferey, J., Gros P.E., Gherbi, R. (2007). ADN-Viewer: a 3D approach for bioinformatic analyses of large DNA sequences. *Cell. Mol. Biol.* 52, 24-31.

[21] Bindewald, E., Grunewald, C., Boyle, B., OConnor, M., Shapiro, B.A. (2008). Computational strategies for the automated design of RNA nanoscale structures from building blocks using NanoTiler. *J. Mol. Graphics Modell.* 27, 299-308.

[22] Mehta, P., Goyal, S., Wingreen, N.S. (2008). A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol. Syst. Biol.* 4, 221.

[23] Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand M.S., Mirny, L. A. (2007). How gene order is influenced by the biophysics of transcription regulation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1394813953.

[24] Pfleger, B. F., Pitera, D. J., Smolke, C. D., Keasling, J. D. (2006). Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.* 24, 1027-1032.

[25] Suarez, M., Jaramillo, A. (2009). Challenges in the computational design of proteins. *J. R. Soc. Interface* DOI: 10.1098/ rsif.2008.0508.focus.

[26] Bennett, M. R., Pang, W. L., Ostroff, N. A., Baumgartner, B. L., Nayak, S., Tsimring, L. S., Hasty, J. (2008). Metabolic gene regulation in a dynamically changing environment. *Nature* 454, 1119-1122.

Acknowledgements

I am forever indebted to my advisors Alfonso Jaramillo and Santiago Elena for bringing me to synthetic biology. Their willingness to tackle hard problems and to do so when everyone thinks youre crazy has given me a more valuable experience than I could have possibly imagined. They have inspired me by their ability to see the world for what it should be rather than what it is and their tireless work to make it so.

To Guillermo, thank you specially for discussing every failed computational experiment and wild idea any time, day or night, in Boston or California. He has been my office mate and friend since I entered the world of engineering. Also thanks to the past and current members from the IBMCP, Guillaume, Paqui, Julia, Regi, Lesia, Miguel, Maria, Ricardo, Laura, Jose Luis, Susana, Patricia, Clara, Puri, Jasna, Angels, Stephanie, Josep and Mark. Specially, I would thank a lot to Nicolas and Jorge for providing their daily advice throughout my thesis and a great friendship. To the Jaramillo group, Pablo, Maria, Josselin, Filipe, Daniel, Boris, Vijai, Satya, and Thomas. I would like to mention the interesting discussions about Molecular Biology with the 2006 Valencia iGEM team, Miguel Blazquez, Mario Fares and, specially, Toni Granell.

To the Knight, Endy and Prather labs from MIT where I honor to work with such a talented and passionate group of people. It has been a fantastic ride. I want to recall the friends I met there, Barry, Jason, John and Francois.

I would like to highlight the role played by all members of swimming clubs I've enjoyed during my last years (Club Natacion Liria, C.N.O. Dom Bosco, Stanford Masters Swimming and U.P.V. Natacion Master). They have given me a boost daily from 7 a.m. in the morning. I especially want to thank Luis Thorpe, a great swimmer without any knowledge of biological science who has greatly influenced this thesis.

Over the last months, Raissa has become the most special colleague, friend, and girlfriend I could have hoped for. At this moment, I would not

have taken certain decisions without their support and her special stabilizing influence, and I would be a lot less happy. To my sister Patri, my aunt Mari and my grandparents, thank you for your love, wrestling fights, and recent fort-building. And to the rest of my family spread out all over the world, your love and support means a lot. Finally, I dedicate this thesis to my mom and dad for their belief in me, their unconditional love, and their impressive sacrifices. Undoubtedly, everything I am I owe to them.